Behavioral Dynamics of Public Transit Ridership in Chicago and Impacts of COVID-19

by

Mary Rose Fissinger

B.S., Boston College (2015)
M.S., University of California, Berkeley (2016)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Department of Civil and Environmental Engineering
August 17, 2020

Certified by .....
Jinhua Zhao
Associate Professor
Thesis Supervisor

Certified by .....
John Attanucci
Research Associate
Thesis Supervisor

Accepted by .....
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Behavioral Dynamics of Public Transit Ridership in Chicago and Impacts of COVID-19

by

Mary Rose Fissinger

Submitted to the Department of Civil and Environmental Engineering
on August 17, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

Abstract

Public transportation ridership analysis in the United States has traditionally centered around the tracking and reporting of the count of trips taken on the system. Such analysis is valuable but incomplete. This work presents a ridership analysis framework that keeps the rider, rather than the trip, as the fundamental unit of analysis, aiming to demonstrate to transit agencies how to leverage data sources already available to them in order to capture the various behavior patterns existing on their transit network and the relative prevalence of each at any given moment and over time. In examining year over year changes as well as the impacts of the COVID-19 pandemic on ridership, this analysis highlights the complex landscape of behaviors underlying trip counts. It keeps riders' mobility patterns and needs as the focal point and, in doing so, creates a more direct line between results of analysis and policies geared toward making the system better for its riders.

This work makes use of two primary methodological tools: the k-means clustering algorithm to identify behavioral patterns, and linear and spatial regression to model metrics of urban mobility across the city. The former is chosen because of its established history in the literature as a technique for classifying smart cards, and because its simplicity and efficiency in clustering high numbers of cards made it an attractive option for a framework that could be adopted and customized by various transit agencies. Spatial regression is employed in conjunction with classic linear regression to capture spatial dependencies inherent in but often ignored in the modeling of urban mobility data.

Chapter 3 of this work identifies the behavioral dynamics underlying top-level ridership decreases between 2017 and 2018 on the Chicago Transit Authority (CTA) and finds that riders decreasing the frequency with which they ride, rather than leaving the system, is the primary driver behind the loss of trips on the system, despite growth in the number of frequent riders using the system for commuting travel. The following chapter applies a similar framework to understand the precipitous ridership drop due to COVID-19 and discovers distinct responses on the part of two frequent rider groups, with peak rail riders abandoning the system at rates of 93% while

half of off-peak bus riders continued to ride during the pandemic. Chapter 5 uses linear and spatial regression to model the percent change in trips due to COVID by census tract and finds that even when controlling for demographics, pre-pandemic behavior is predictive of the percent loss in trips. Specifically, high rates of bus usage and transfers, along with pass usage, are associated with smaller drops in trips, while riding during the peak is predictive of larger decreases in trips. Chapter 6 presents preliminary thoughts on employing a spatial regression framework on high-dimensional data to learn urban mobility patterns.

This work highlights the insights to be gained from an analysis framework that reveals the complex behavioral dynamics present on a transit network at any given time. It further connects these behaviors to other rider characteristics such as home location and response to the COVID-19 pandemic, painting a rich picture of an agency's riders with their existing data and allowing for informed, targeted policy creation. A key finding was that frequent, off-peak bus riders who frequently have to transfer are one of the largest groups of riders and the group most associated with continued ridership during the pandemic. Future policies should recognize that this group uses the system when and where overall ridership is low, and direction of resources away from these parts of the system will disproportionately hurt riders who are most reliant on public transit and therefore have the most to gain from increased investment. The CTA should work in conjunction with other stakeholders to ensure that as public transit ridership recovers from the pandemic, attention is paid not only to those riders who need to be brought back onto the system, but also those who never left it.

Thesis Supervisor: Jinhua Zhao
Title: Associate Professor

Thesis Supervisor: John Attanucci
Title: Research Associate

Acknowledgments

This work is indebted to the Chicago Transit Authority. I would like to especially thank President Dorval R. Carter for his continued support and enthusiasm for the partnership between the CTA and MIT. His active engagement with my work and that of my classmates served as inspiration and fuel throughout this process.

This work would also not be possible without Maulik Vaishnav, who answered my endless questions about the Ventra data and provided invaluable insight into the workings of the CTA and the city of Chicago that informed much of this thesis. His genius policy analysis contributed immensely to Chapter 3 and taught me lessons that I will take with me into my career.

Tom McKone and Scott Wainwright additionally provided thoughtful guidance along the way, helpful context in which to situate my work, and willing direction to answers if they themselves could not provide them. Molly Poppe proved to be a reliably active listener and advocate for this work, and I appreciate her immediate and enthusiastic investment in the MIT partnership. Laura De Castro is the glue that holds everything together, and I and everyone at MIT who works with the CTA are deeply grateful for all that she does. Additionally, I would like to thank Jeremy Fine, Paris Bailey, Ray Chan, Bryan Post, Elsa Gutierrez, and Emily Drexler for their conversations and solutions along the way. I want also to express gratitude for every employee at the CTA who works each day to keep the system running, power an extraordinary city, and offer such a successful example of American public transit to the people like me who sit at a computer and crunch the numbers.

To Daiva Siliunas, thank you for sheltering me whenever I came to Chicago. You are a phenomenal host and an even better friend.

At MIT, I would like to express gratitude for the guidance of Jinhua Zhao, John Attanucci, and Fred Salvucci for providing wisdom that improved this work and me as a thinker. I would also like to thank my talented classmates and colleagues for filling the environment with rich and varied public transit knowledge. In particular, Shenhao Wang brought structure and rigor to Chapter 6, Hui Kong offered feedback

that vastly improved Chapters 3, 4, and 5, and Joanna Moody's edits to Chapter 4 transformed it into something much better. Lastly, thanks especially to Annie Hudson for the 5PM Tuesday beers at the Muddy that got me through it all.

To my parents, thank you for the support and love that has been the most powerful and important constant in my life. I am incredibly blessed and owe it all to you.

Lastly, thank you to David. You do more than you know. I love you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

As recently as five years ago, the story looked promising for public transit ridership in America. Except for a couple relatively minor dips in trip numbers following the economic recessions in 2001 and 2008, from which public transit ridership recovered in about three years' time, yearly counts of unlinked passenger trips had enjoyed two decades of steady of growth [American Public Transportation Association, 2020, Mallett, 2018]. In 2010, the United States Government Accountability Office delivered a report to the U.S. Senate Committee on Banking, Housing, and Urban Affairs entitled "Transit Agencies' Actions to Address Increased Ridership Demand and Options to Help Meet Future Demand" [Wise, 2010]. The report notes that ridership growth between 1998 and 2008 outpaced the growth in service provision for all modes — light rail, heavy rail, and bus — and attributes the growth in ridership to population increases, employment growth, higher prices for gasoline and parking, as well as additional measures taken by individual agencies, such as the creation of partnerships with local businesses to encourage commuting by public transit. As the title suggests, the report is primarily concerned with recommending changes to public transportation funding that could best help transit agencies meet a continued growth in demand.

Just one decade later, it is hard to imagine such a time in the public discourse surrounding mass transit in America. A few years ago, as it became clear that the dip

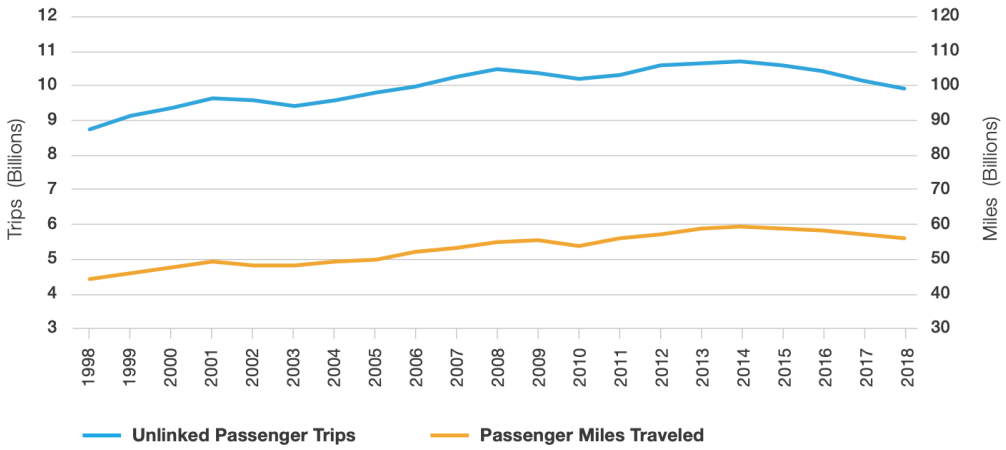


Figure 1-1: Yearly Public Transportation Ridership in the United States, 1998 - 2018

Source: 2020 APTA Fact Book

in ridership from 2014 to 2015 was not a blip but rather the beginning of a sustained downward trend, the conversation around mass transit in the U.S. changed markedly. News publications and transit blogs around the county adopted language that ranged from colorful to apocalyptic to describe the current state of affairs. *The Washington Post* used a headline quoting experts describing the situation as an "emergency" [Siddiqui, 2018] while *The Los Angeles Times* ran a story describing the city's bus system as "hemorrhaging" riders [Laura J. Nelson, 2019].

These stories in turn quickly came to feel like historical documents after March of 2020 and the spread of the COVID-19 virus in America. As schools shut, businesses closed down or turned to remote work, and state and local governments urged people to remain home as much as possible and avoid crowded indoor areas, public transit ridership plummeted across the world [Transit, 2020]. According to the mobile app Transit, which allows users to track locations of trains and buses in their city, demand for public transit as measured by use of their app was down 75% in the month of April. Headlines on articles addressing transit ridership in America spoke about the end times for mass transit: Time Magazine published an article in July whose title asserted that "COVID-19 Has Been 'Apocalyptic' for Public Transit" [De la Garza, 2020] and Forbes published a piece posing the question "Will COVID-19 Sound The

Permanent Death Knell For Public Transit?" [Templeton, 2020].

The sudden and precipitous drop in ridership that occurred in US cities made the fluctuations in trip numbers over the previous few decades seem incredibly stable. The temporal and spatial distribution of trips within cities changed shape almost overnight, and many things that had once felt like accepted facts about transit usage in a city had to be re-investigated and re-learned.

As cities and transit agencies across the country are re-grouping and attempting to learn all they can about what is likely to be a new normal of significantly reduced transit ridership for some time, they have an opportunity to rethink and enrich how they measure and track ridership on their systems. This work offers one potential avenue for doing so. Specifically, it puts forth a ridership analysis framework that centers around the *rider* instead of around trip counts. It demonstrates the usefulness of such a framework for, first, understanding the mobility needs of riders at any given time, for example during a global pandemic, and, second, crafting policy based on these needs.

1.2 Motivation

Public transportation ridership analysis in the United States has traditionally centered around the tracking and reporting of the count of trips taken on the system. These numbers can be disaggregated spatially to the zone, census tract, line, or stop, or disaggregated temporally to weekends, weekdays, peak periods, and off-peak periods. They can be normalized by capita or by revenue vehicle mile or available capacity, and they can be tracked across months and years. Such analysis provides valuable information about the health of public transit systems within and across US cities and how this is trending over time. The value of analysis based on counts of trips will never be supplanted, but it is incomplete. By using the passenger trip as the fundamental unit of analysis, it obscures the reality of a city as a place full of people who are living, working, visiting, and, to various extents, making use of the public transit system to get them where they need to go. A ridership analysis framework

that keeps the rider, rather than the trip, as the fundamental unit of analysis, on the other hand, seeks to understand the patterns in which trips across hours, days, or months and across neighborhoods, lines, and modes are tied to an individual person. Its aim should be to capture the patterns of behavior that exemplify how people typically use their transit network and the relative prevalence of these behaviors. This type of analysis re-centers the question on the people rather than the trips, and leads to answers that are focused around "who?" instead of "how many?" It keeps riders' mobility patterns and needs as the focal point and, in doing so, creates a more direct line between results of analysis and policies geared toward making the system better for its riders.

1.3 Research Aims

The primary aim of this work is to demonstrate to transit agencies how to leverage data sources already available to them in order to better understand who their riders are and how they use the system. I furthermore seek to show how this knowledge can form the baseline of deeper analysis that addresses a few crucial questions facing American transit agencies today, and how, by keeping the rider as the fundamental unit of analysis, the results can directly inform policies aimed at riders. This work was begun in pre-pandemic times and as such, I first demonstrate how to track changing behaviors across years to uncover the behavioral dynamics underlying relatively minor top-level changes in trip counts. Next, I leverage this framework to explore the distinct ridership responses to the COVID-19 pandemic by behavior group and use this rider-centric knowledge to craft policy recommendations for ridership recovery. Then I employ linear and spatial regression to identify the behavioral and demographic traits most predictive of COVID-related ridership loss. Lastly, I offer some initial thoughts on how to better understand the urban mobility landscape of a city by leveraging high dimensional data to capture the dynamics among multiple modes' usage patterns.

1.4 Data and Methods

This work was sponsored by and done in conjunction with the Chicago Transit Authority (CTA) and thus uses Chicago as the case study for all analyses. All data on public transit usage comes from the CTA's account-based fare payment system Ventra, but the framework could be applied by any agency with a widely used smart card fare payment system in place.

The primary methodology employed for the identification of key behavioral patterns on the CTA's network was the k-means algorithm applied to a dataset in which each point to be clustered was a single Ventra card whose behavior was captured by a vector of unit-standardized attributes summarizing key aspects of the card's usage, such as the percent of trips taken during peak hours or the average number of weekly trips taken on that card. The k-means algorithm was chosen because of its established history in the literature as a technique for classifying smart cards, and because its simplicity and efficiency in clustering high numbers of cards made it an attractive option for a framework that will ideally be adopted and customized by various transit agencies. This work's contribution lies not within the realm of rider segmentation methodologies, but in establishment of a framework for segmentation that is accessible to transit agencies and can easily serve as the foundation for deeper analysis and informed policy creation.

The other methodology employed in this work is that of linear and spatial regression. Spatial regression models, and particularly spatial lag models, are employed as alternatives to linear regression that should be explored in the modeling of mobility data that is located in space. The use of such models in ridership analysis within a single city has been limited, and this work does not seek to definitively establish its usefulness in the urban mobility context, as much more research is needed on this front, but it does offer it as a model worth considering, at least in the situations in which I employ it. The first such situation is a model in which aggregate traits of individuals assigned to a census tract based on inferred home location are used to explain ridership changes in that tract. Here, the likely tract spillover of transit use

due to people using multiple stops near their home, as well as the interconnectedness and resulting lack of independence of the network in general, motivated an exploration of a spatial lag model that used nearby ridership loss to explain tract-specific transit ridership loss. One can easily imagine other models of trip volume or ridership changes by geographic area that could benefit from the inclusion of spatially lagged dependent or independent variables. One such example is from the final analytical chapter of this work, which is motivated by the hypothesis that TNCs are in competition with public transit systems, and seeks to provide a set of models that can be used to explore the relationship between usage of the two modes across spatial and temporal dimensions. A straightforward way to explore this question is to model TNC trips as a linear combination of public transit trips in and around the TNC trips' origin locations. Such a model involves a spatial lag of public transit usage. Capturing the dynamics between modes is a situation that could likely benefit from exploration of spatial regression models, definitions of "neighborhoods," and other related topics. This work begins to explore these questions.

1.5 Organization of Thesis

Chapter 2 provides background on relevant topics, specifically public transit ridership behaviors in America, the city of Chicago and the CTA system, and the COVID-19 pandemic and the response of transit agencies and riders across the United States.

Chapter 3 offers a framework for transit agencies with account-based fare systems to capture changing behavior dynamics on their systems using smart card clustering on data from multiple years. The CTA is used as a case study, with the data coming from their account-based Ventra system. The behavior changes on the system from 2017 to 2018 are identified and summarized, and then used to pinpoint particular rider groups of interest, who are then investigated in greater depth, using the fact that the Ventra system is rich with data that can be layered onto each card at any point in the analysis. Specifically, the additional analysis at the end of Chapter 3 is performed in order to more deeply understand the impacts of the January 2018 fare

increase on the CTA system.

In Chapter 4, the same clustering technique is applied (in a slightly modified fashion) to establish the baseline behaviors present on the CTA system in the months just prior to the stay-at-home order issued in Chicago in response to the COVID-19 outbreak. Ridership data from the beginning of the stay-at-home order and from two months after is then analyzed to paint a picture of how the pandemic has affected transit ridership in America's third largest city. This section concludes with policy recommendations for the CTA in light of the findings.

Motivated by the findings from Chapter 4, Chapter 5 employs a different methodology to understand the factors associated with the steepest declines in transit ridership during the COVID period when compared with the baseline. This section employs classic linear regression as well as spatial regression models to quantify the relationship between demographics and baseline ridership behavior as the explanatory variables and ridership decline as the dependent variable at the census tract level.

Chapter 6 presents preliminary work applying spatial regression concepts to higher dimensional data and begins to explore models that incorporate both the space and time dimension to capture the dynamics of Transportation Network Company (TNC) trips in Chicago. This chapter also lays out ways that this structure could be used to more deeply understand the extent to which TNC ridership is related to transit ridership.

Chapter 7 summarizes the findings and offers concluding thoughts regarding recommendations for the CTA and directions for future work.

Chapter 2

Background

This work is motivated by the idea that a deep and continually evolving understanding of how public transit riders make use of their cities' transit systems is a crucial part of providing good service as a transit agency. Analyses of trip counts tell an incomplete story about the state of transit ridership. Beneath these trip numbers are thousands or millions of people moving about their city, living their lives. Some of them have no other option but to use mass transit. Some use it only to commute, opting for other modes on the weekends or in the evenings. Some use it every day, others once a month. From only aggregate trip counts, one cannot deduce the set of behaviors existing on a system. Yet, knowing these behaviors can inform policies in very valuable ways. Policies aimed at increases in ridership or revenue will be more effective if the target is not merely "more trips" but a person whose mobility needs and challenges are well-understood. Furthermore, if several dominant behaviors can be uncovered, more targeted policies can be directed to each in turn. Transit agencies can meet riders where they are, and then get them where they need to go.

This is not novel thinking — transit agencies have understood this for a long time. But up until recently, their primary method of learning about their riders was surveys, which capture ridership behaviors at a (often very brief) snapshot in time. Longitudinal tracking of riders via surveys is expensive, and the validity of the conclusions that can be drawn is sensitive to the sample that is reached and the accurate reporting on the part of the survey takers. While these methods undoubtedly provide valuable

insights, in part because they have the benefit of capturing rich demographic data along with ridership behavior, the picture they capture of an individual's ridership behavior is limited, either in terms of detail or duration.

With the emergence of Automated Fare Collection (AFC) smart card technology, however, transit agencies can now connect each trip to a fare card, and observe behaviors on cards that are used for an extended duration. In cities such as Chicago, where the fare payment system is account-based, meaning that even replacement cards can be tied to the same person, the implications are especially powerful. Multi-year ridership trends can be analyzed in terms of underlying changing behaviors, weighing these against volumes of churned riders versus new riders. The impacts of service changes or disruptions can be looked at through the lens of the people affected. Changes to fare policies can be evaluated based on which groups prove most or least elastic. In short, transit agencies now have the ability to more fully understand and meet the needs of the riders they serve.

In the next section, I will offer a review of work that has looked at travel behaviors among public transit riders and work that has studied changes in these behaviors over time. Next I will offer some context on the city of Chicago and CTA system, as that will be the subject of the case studies in this work. Lastly, I will give background on the COVID-19 pandemic, which motivates the analysis in chapters 4 and 5, and explain the ways in which transit agencies, riders, and analysts had responded to the outbreak at the time of this writing.

2.1 How Americans Use Public Transit

Within the realm of travel behavior research, the primary question for the past several decades has been that of mode choice: what makes someone choose one mode over the other? The methodology employed to answer this question is typically a logit model that takes in the riders' demographics and the trip's attributes for each mode and outputs the mode that such a traveler would most likely choose. Implicit in this is the idea that individuals make several trips throughout the course of the day, some of

which may be on public transit, some via private automobile, some using a rideshare service, etc. These models try to capture the context in which someone lives their life and makes travel decisions. The insights such studies can provide are incredibly valuable, as they shed light on the factors riders weigh before setting off on their mode of choice, but the downside is that these models are incredibly data-intensive, requiring information on the decision-maker and each of the travel modes available to her. Studying travel behavior as observed on only a single mode removes much of this rich context but is, with smart card data, readily possible for public transportation agencies. Looking at only public transit ridership behavior can by itself communicate a lot about a person and, coupled with knowledge about how much daily travel the average person engages in, give us a good idea of the extent to which people are using mass transit for all or most of their travel needs.

The mode choice literature helps inform decisions about how to entice more people and trips onto public transit— an extremely important goal, but not the sole objective of a public transit agency. How people currently use the system contains a wealth of information regarding the mobility needs of riders, and understanding these needs so that they can be best met should be an equally important, if not more important, goal of public transit agencies. Thus, an understanding of existing public transit rider behavior is crucial. Furthermore, insight into how these behaviors change over time can shed light on where transit systems need to become more competitive, and where resources should be invested.

Capturing predominant public transit ridership behaviors has received less attention than understanding factors driving mode choice, or determining the demographic profile of transit riders, but with the advent of AFC technology, the question has gained more attention. In Chapter 3 I provide a literature review specifically on how smart card data has been mined to uncover transit ridership behaviors, but none of the studies use data from an American city, so here I will present details on how we currently understand Americans to use public transit and how that is changing over time.

A January 2017 report from the American Public Transportation Association

entitled "Who Rides Public Transportation: The Backbone of a Multimodal Lifestyle" drew on ridership reports from 163 transit systems in the U.S. that surveyed over 650 thousand riders in total [Clark, 2017]. This report found that fully half of all respondents used public transit five times per week. It also found that half of the survey respondents' most typical transit trips involved a transfer. Among all riders, 29% had been using transit for under two years, and 53% had been using transit for more than five years. Rail riders were more likely to be long-term users of mass transit than bus riders.

TransitCenter's 2016 report "Who's On Board" summarized findings from focus groups and online surveys of public transit riders in 17 large and medium-sized cities [Higashide, 2016]. The report found three general behaviors to be predominant on public transit systems: occasional riders, commuters, and all purpose riders. The latter group was most prevalent in cities with strong transit networks offering frequent service to many destinations. This report stressed that all riders were sensitive to transit quality and that the traditional distinction of "choice" and "captive" riders was detrimental.

Using a similar methodology, TransitCenter followed this report up three years later with "Who's On Board 2019: How To Win Back America's Transit Riders" [Higashide and Buchanan, 2019]. In it, they determine that declining public transit ridership is driven by people scaling back their use of public transit systems and largely replacing trips with private vehicles, rather than abandoning mass transit altogether. This was reflected in the fact that more riders fell into the "occasional" category than had in 2016.

Despite the outcry over declining public transportation ridership over the past few years, few studies besides the TransitCenter reports mentioned above have examined the behavioral trends underlying these declines. Chapter 3 of this thesis outlines a framework for how transit agencies with access to smart card data can leverage it to answer the same questions answered by the 2019 TransitCenter report: Who *is* on board their buses and trains, and how are those people shifting their behavior over time? Our analysis demonstrates that many of the nationwide findings from

TransitCenter's report hold in Chicago as well, with overall ridership declines being driven by people using the system less, rather than fewer people using the system at all.

2.2 Chicago and the CTA

The case studies in this work all focus on the city of Chicago, as all of the public transit ridership data comes from the CTA's account-based fare payment system, Ventra.

Chicago is located on the coast of Lake Michigan in Illinois and is the third largest city in the United States with a population of 2.6 million, surpassed only by New York City (8.3 million) and Los Angeles (4 million). The CTA is the second largest transit agency in the country behind the Metropolitan Transportation Authority (MTA) in New York in terms of unlinked passenger trips in 2018. Broken out by mode, the CTA provided more trips on heavy rail than any agency other than the MTA and the Washington Metropolitan Area Transportation Authority (WMATA) in 2018, and provided more bus trips than all agencies other than the MTA and the Los Angeles County Metropolitan Transportation Authority (LACMTA) [American Public Transportation Association, 2020].

Outside of the CTA, the Chicago metropolitan area is also home to Metra, the largest commuter rail network outside of the New York City area as measured by unlinked passenger trips, and Pace, a large suburban bus and regional paratransit network. These services exist to complement rather than compete with the CTA, however, and largely bring people from the suburbs into the city or to other areas outside the city. Within the city boundaries, the CTA is the primary provider of mass transit services, and thus analysis using CTA data provides a nearly complete picture of public transit usage in the urban area.

Public transit has been an integral part of Chicago for a long time, with the first rapid transit line opening in 1888. The unique structure of the rail network gave the downtown core of the city its now official name — "The Loop." The CTA's eight rail

lines are arranged in a spoke-hub formation, all feeding into the dense downtown area (except the Yellow line, which acts as a feeder branch to the Red and Purple lines in the north). Five of the seven lines that reach the downtown area traverse at least some portion of the 1.79 mile long elevated loop of rail tracks running above Lake Street in the north, Wabash Avenue in the east, Van Buren Street in the south and Wells Street in the west. The other two lines— the Blue line and the Red line — are underground at this point. Figure 2-1 shows the layout of the CTA rail network.

This network structure has contributed to the continued concentration of jobs in and around the Loop. More than half of Chicago's jobs are located in the downtown area, and the total number of downtown jobs as well as the proportion of jobs located in downtown has been growing since 2010 (Figure 2-2). Patterns of usage on the CTA's rail system reflect its role as a connection to jobs: in 2019, just over half of all Ventra card taps on CTA's rail system occurred between the hours of 6AM and 10AM or between 4PM and 8PM on weekdays.

The bus network structure, on the other hand, largely reflects the city's grid layout, with most routes running north-south or east-west, many along a single street. Bus ridership on the system also exhibits peak patterns, but to a lesser extent than rail. About 45% of Ventra card taps of bus occurred during weekday peak hours during 2019.

Ridership by year for each mode between 2006 and 2018 is shown in Figure 2-3. While bus ridership has declined each year since 2012, rail mostly continued to grow ridership until 2015. Since then, until the COVID pandemic, total ridership on the system had declined by about 3% yearly, with slightly steeper losses coming from bus rather than rail [Chicago Transit Authority, 2020b]. This is largely in line with nationwide trends in transit ridership, which have been declining since a peak in 2014 [American Public Transportation Association, 2020].

This work began as an effort to explain the CTA's ridership decreases through the lens of changing individual behaviors. This topic is the focus of Chapter 3 and explores the system's changing behavioral dynamics between the fall of 2017 and the fall of 2018. Several months after that analysis was completed, however, the COVID-

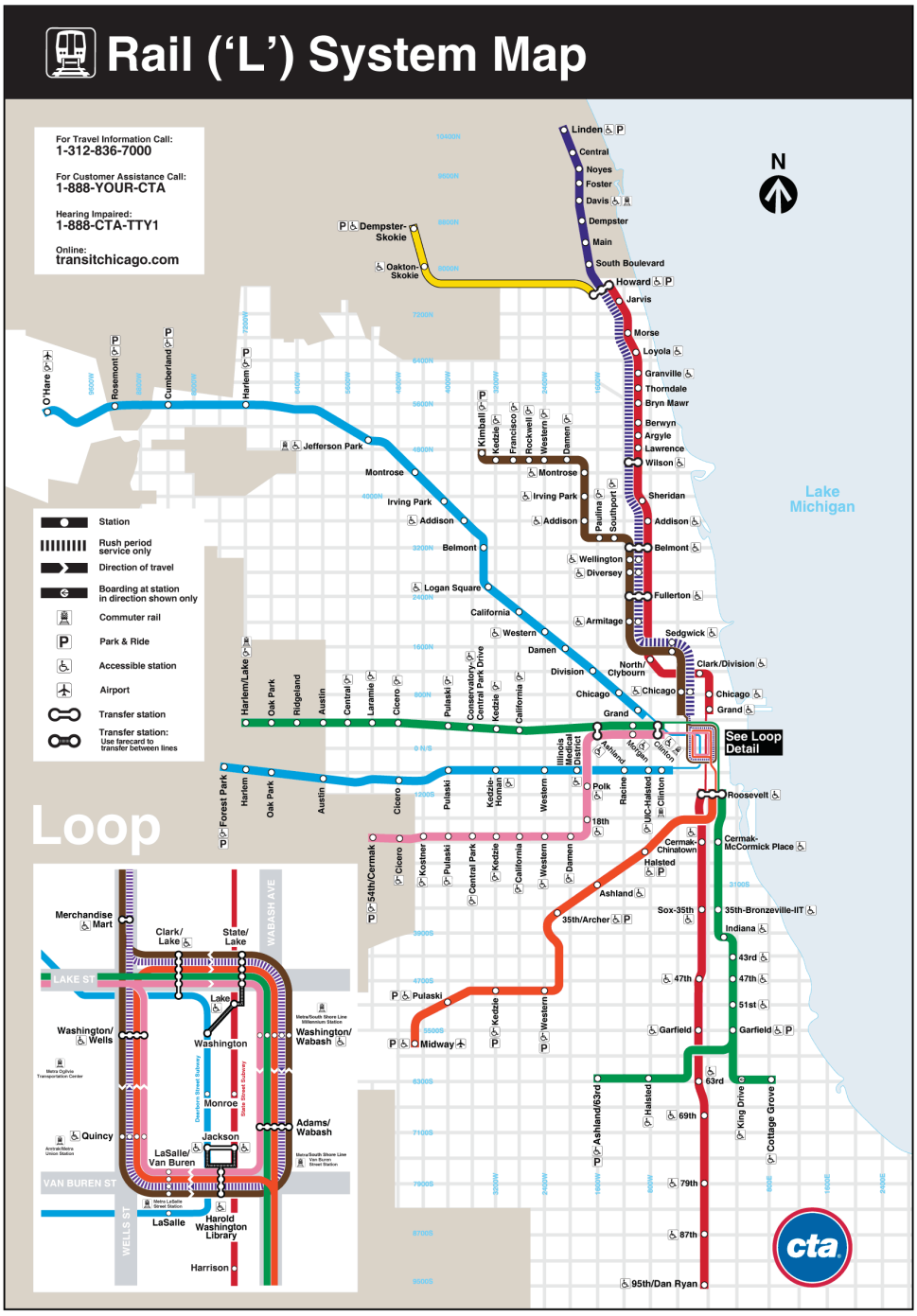


Figure 2-1: CTA Rail Map Network

19 pandemic abruptly and dramatically changed daily life for nearly everyone in the world. One element of these changes was people's travel needs and behaviors. Chap-

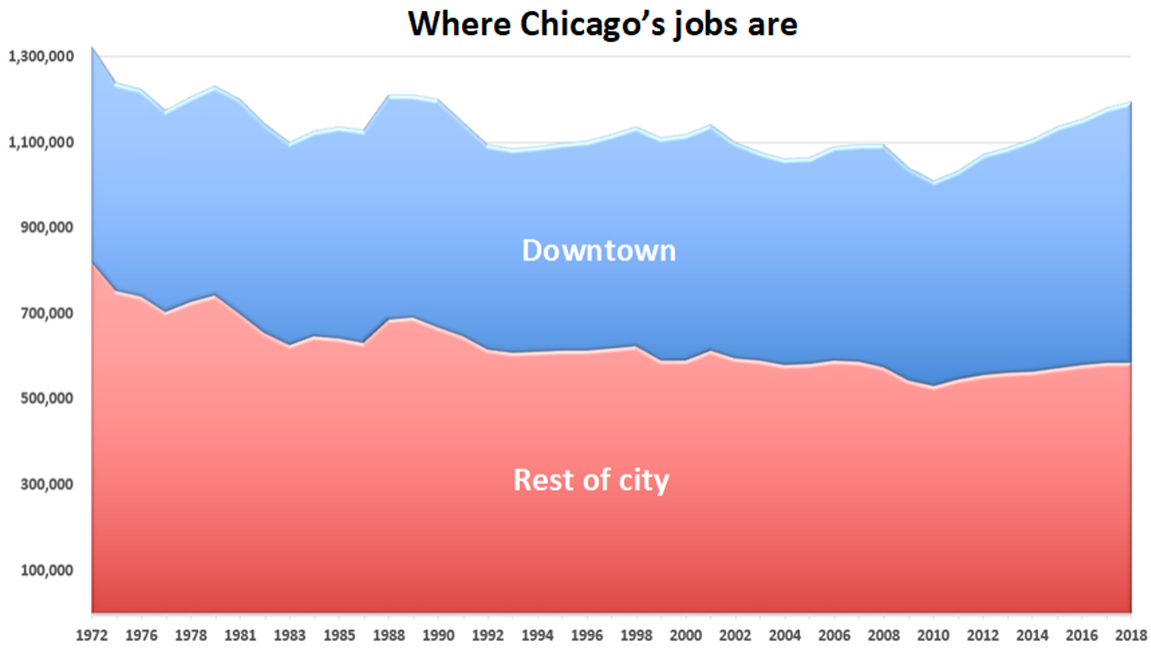


Figure 2-2: Number of Jobs in Chicago by Location (Downtown or Elsewhere)

Source: Chicago Sun Times

ters 4 and 5 of this work build off the foundations set forth in Chapter 3 to capture the heterogeneous individual behaviors underlying the massive drop in overall transit trips occurring in Chicago due to the pandemic. Before concluding this chapter, I provide some context on the timeline of the COVID-19 pandemic and what we know about reactions of transit agencies and riders as of July 2020.

2.3 The COVID-19 Pandemic

At the time of this writing, the United States is four months removed from the initial escalation in COVID-19 cases that occurred in the second half of March. We now know that the virus is spread primarily via respiratory droplets, such as those produced when someone coughs, sneezes, or talks, and that people exhibiting no symptoms can still spread the virus [Center for Disease Control and Prevention, 2020a]. There is still much that is unknown about the virus, however, including how unsafe an activity like riding public transit really is for the general population. The body of research on that is growing, and a brief summary of it will be provided here, along with accounts

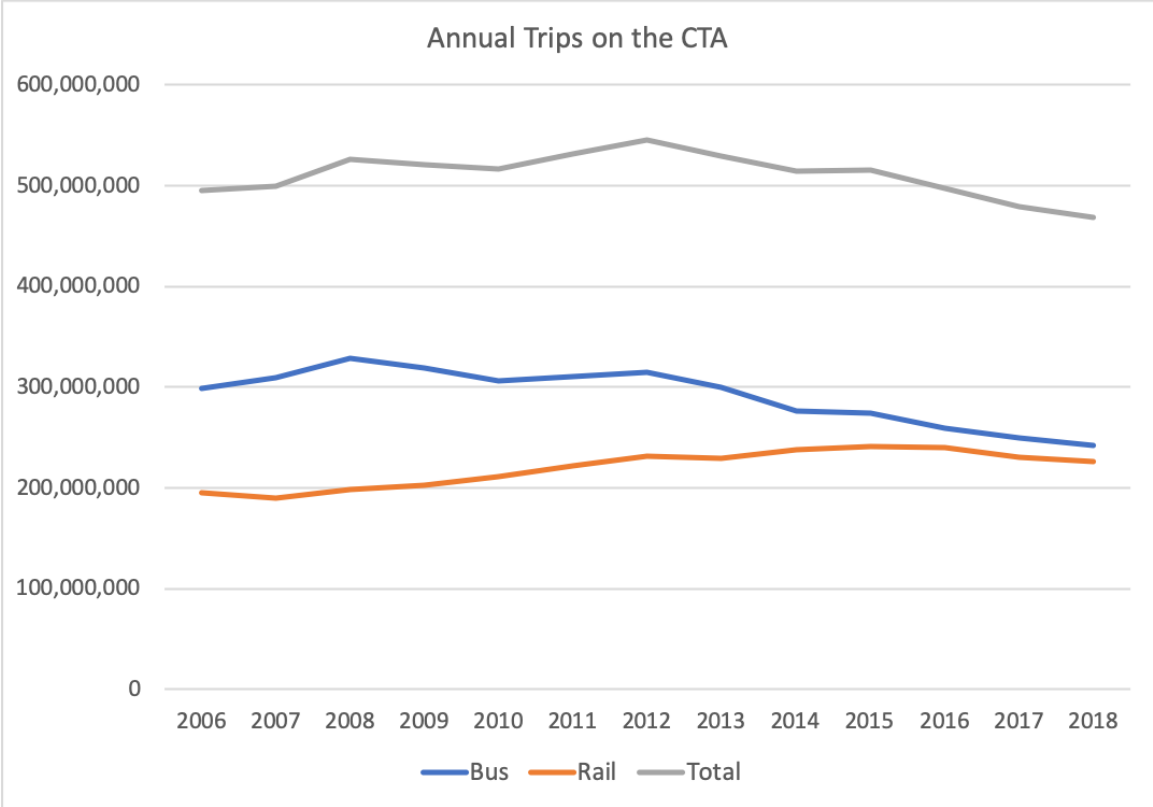


Figure 2-3: Count of Yearly Trips on CTA

Source: Chicago Open Data Portal

of transit ridership responses across the country.

In April, an MIT economics professor published a study claiming that the New York City subway was "a major disseminator—if not the principal transmission vehicle" of the disease in the city and the reason the outbreak was so much worse there than elsewhere in the country [Harris, 2020]. His methodology, which involved overlaying declines in subway ridership with infection rates by zip code, quickly drew many critics, who noted his failure to account for any of the obvious cofounders, such as the decline in activities that was driving the decline in transit use [Bliss, 2020]. They also noted that many zip codes with the highest density of transit stations had some of the lowest infection rates [Levy, 2020]. Transit analysts, epidemiologists, and mathematicians alike have concluded that the paper provides no concrete evidence that the subway explains why the outbreak was so much worse in New York than elsewhere in the country in the early days of the pandemic [Sadik-Khan and Solomonow, 2020].

Since this paper, several other studies have provided evidence that subway systems are not, in general, responsible for a significant portion of disease spread. Epidemiologists found that in Paris, none of the 150 identified coronavirus infection clusters between early May and early June were traced to public transit usage or transmission [Berrod, 2020]. Likewise, researchers investigating the outbreak in Austria in April and May found that none of the 355 clusters could be connected to transit [Austrian Agency for Health and Food Safety, 2020]. Furthermore, several cities, particularly in Asia, with transit use on par with or higher than New York's and smaller declines in that usage saw outbreaks that were much more successfully contained [Mahtani et al., 2020]. Hong Kong, for example, had only 1,655 confirmed cases as of this writing or about as many as Duplin County, North Carolina, whose population is just over 59,000, compared with Hong Kong's 7.5 million [Johns Hopkins University and Medicine, 2020]. Japan, home to the world's busiest rail network in Tokyo, along with several other major transit systems, has had just over 25,000 confirmed cases compared to the U.S.'s 3.8 million.

While there is growing evidence that public transit systems are not a unique evil in terms of risk of transmission, there is no question that Americans with the option to stay home or use another mode have abandoned it in droves. According to the mobile app Transit, public transit usage was down 77% across the country [Transit, 2020]. The app surveyed the remaining riders and found that they were overwhelmingly (92%) using transit to get to work, and that they were predominantly women of color and 70% of them made under $50,000 a year. Meanwhile, Transit app users in higher paying jobs had been able to shift to working from home. The Eno Center for Transportation analyzed news reports from transit agencies across the country and showed that the drop in transit ridership due to COVID differed significantly by mode, with commuter rail lines seeing the largest drops, followed by urban heavy rail, and then bus [Puentes, 2020]. The difference between commuter rail and bus is stark, with commuter rail ridership down more than 90% in many places, while major bus systems were maintaining up to two-thirds of their baseline ridership. These findings are consistent with those from the survey of Transit app users, as it is well-established

in the literature that bus riders more likely to be lower income than rail riders [Maciag, 2014].

Transit agencies have responded to the disease and resulting drops in ridership in various ways. Many have significantly reduced their service in order to save money and accommodate staff shortages. New York City's MTA cut subway service by 25%. WMATA in Washington, D.C. shut down 19 rail stations in response to the pandemic in March, reopening them at the end of June [Washington Metropolitan Area Transit Authority, 2020]. The San Francisco Municipal Transportation Agency (SFMTA) closed all subway stations and replaced all Muni Metro and light rail routes with buses in order to "redirect custodial resources to other, higher-use facilities," specifically those on routes connecting people to essential jobs and services [Fowler, 2020]. The CTA, on the other hand, made no permanent service cuts despite drops in ridership around 80%, canceling trips only as a result of staff shortages. The CTA also replaced 40-foot buses with 60-foot buses on certain routes that maintained particularly high ridership [Chicago Transit Authority, 2020a]. Aside from changes to service volumes, many transit systems have implemented rear-door boarding for buses to limit passenger contact with operators, rendering bus travel essentially free in these cities. Several have also authorized their bus drivers to maintain capacity caps on buses. In addition, nearly all have increased communication about how to ride safely during COVID, suggesting or requiring masks, advising maintaining a safe distance between passengers, and recommending frequent hand sanitation, among other guidelines.

Figure 2-4 shows the daily count of Ventra taps by mode on the CTA from the start of 2020 through July 19, along with key dates in Chicago's management of the disease spread. Although a Chicago woman on January 24 became the second confirmed case of COVID-19 in the United States, the city, like the rest of the country, maintained business as usual until the early part of March. On March 9, the Governor of Illinois, J. B. Pritzker, issued a disaster proclamation, allowing the state to take advantage of additional state and federal resources to fight the disease. Over the next two weeks, in quick succession, the governor banned gatherings of over 1,000 people,

ordered all bars and restaurants closed, shut down public and private schools, and on March 21, issued a stay-at-home order [Tribune staff, 2020]. Over the same time frame, the number of rides occurring on the CTA dropped by more than 80%. On April 9, the CTA implemented rear-door boarding on buses, leading to effectively free bus service in Chicago. On June 3, after two and a half months, the stay-at-home order was lifted in Chicago as part of "Phase III" of reopening, which allowed some non-essential businesses to resume operations with capacity limitations [Munks and Anderson, 2020]. Restaurants and coffee shops were permitted to allow outdoor dining, and personal services such as hair salons reopened [NBC Chicago, 2020a]. On June 26, Chicago moved on to Phase IV, which allowed indoor dining at restaurants as long as tables were more than six feet apart, museums were permitted to operate at 25% capacity, and gatherings could occur of up to 50 people, up from 10 in Phase III [NBC Chicago, 2020b].

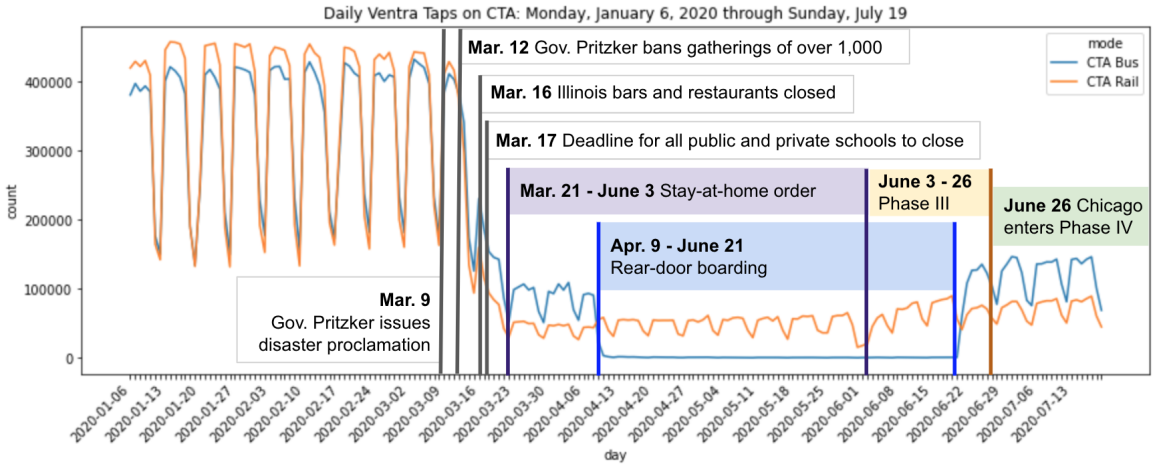


Figure 2-4: Daily Ventra Taps in 2020 with Key Dates from COVID-19 Management in Chicago

Source: Chicago Tribune, City of Chicago

At the time of this writing, despite being deeply uneasy about the massive revenue drops they are are sure to see for some time, transit agencies are still largely following the suggestions of the CDC and urging riders not to travel unless necessary, so that public transit is as safe as possible for those who need it [Center for Disease Control and Prevention, 2020b]. Eventually, however, transit agencies will need to recover a

significant portion of their baseline ridership; otherwise, U.S. cities will be gridlocked with private vehicles as people begin moving again. Charting a path forward will require an understanding of who riders were before the pandemic, how their travel behavior shifted in response, and what this information tells us about their mobility needs and the challenges to bringing them back on board. Chapters 4 and 5 of this work use Chicago as a case study to examine the differential impacts of the pandemic on the ridership of a few key groups, and Chapter 4 uses this analysis to craft a multi-pronged policy approach to meet the mobility needs of riders using the service during the pandemic, to get riders back on buses and trains, and to help the CTA be more reflective of the needs of its riders going forward.

Chapter 3

Customer Segmentation Framework

The introduction of account-based automated fare collection technology has given transit agencies a new level of depth in their data, allowing for deeper analysis and understanding of the underlying ridership trends present on their system. Whereas before, most transit agencies could only quantify the number of trips made by time of day, day of week, mode, route, or stop, now they can quantify the number of trips made by person, as well as the spatio-temporal distribution of those trips. This allows and, I would argue, demands full recognition of the fact that a transit system is built for people, and that every trip occurs because of the person who decided to make it. This section will demonstrate how to leverage AFC data to uncover how behavior trends are driving top-level ridership changes. The framework presented here will enable transit agencies to answer questions such as whether a ridership decline was due more to riders churning from the system altogether or decreasing the number of trips they took. Which ridership behaviors are most stable? Most unstable? Having identified some key behavior groups of interest, what else can we learn about these riders? And finally, how can we use these insights to inform policy analysis?

3.1 Background

Since the emergence of AFC data, a robust literature on methodologies for mining this data for ridership behavior patterns has emerged. This particular work draws heavily

from that of Basu, who used k-means to cluster cards from Hong Kong's Mass Transit Railway (MTR) system [Basu, 2018]. He used one month of data and characterized ridership using both temporal and spatial features with the goal of allowing MTR to target system information to only the riders for whom that information was relevant. The literature on this topic has grown to cover not only a large range of clustering methodologies but also types of systems analyzed and sets of input features. I will touch on only some of them here.

Many of the studies in this body of literature leverage unsupervised learning algorithms to uncover patterns in the data. The two most popular within this body of work are k-means and Density-Based Spatial Clustering of Application with Noise (DBSCAN), and they are applied to a variety of different types of input data at various steps in the customer segmentation process. Morency et al., for example, apply k-means in a comparison of just two individuals, and use the algorithm to uncover days of travel with similar patterns [Morency et al., 2006]. Agard et al. use k-means along with Hierarchical Agglomerative Clustering (HAC) on binary features indicating day of week and time of day to explore the relationship between temporal ridership patterns and fare type [Agard et al., 2006]. To separate infrequent from frequent passengers, Kieu et al. apply k-means to the number of trip chains evident from card usage, and then apply DBSCAN to the frequent group to refine the differentiation by spatial and regularity metrics [Kieu et al., 2013]. In a later paper, Kieu et al. employ DBSCAN to separate transit riders into 4 groups using data on typical times of travel and origin and destination locations [Kieu et al., 2015]. DBSCAN is also used by Ma et al. on identified trip chains made by riders in Beijing [Ma et al., 2013] to classify behaviors there.

Others have explored different methods of classifications to attempt to capture even more nuance in the data. El Mahrsi et al. apply two clustering approaches to two problems: they use Poisson mixture models to cluster transit stations by their usage problems, modeled after similar work on bike share stations by Come et al., and they cluster passengers by estimating a mixture of unigram models, based on work by Nigam et al. on document classification [El Mahrsi et al., 2017, Côme and Oukhellou,

2014, Nigam et al., 2000]. Ghaemi et al. propose a technique for projecting high dimensional binary data onto the three-dimensional plane and then applying HAC to cluster the vectors [Ghaemi et al., 2017]. Gaussian Mixture Models were used by Briand et al. in order to group riders based on temporal features while maintaining the continuous nature of temporal data [Briand et al., 2016]. More recently, He et al. have explored the tradeoffs between cross-correlation distance (CCD) and Dynamic Time Warping (DTW) for assessing the difference in travel patterns represented by time-series data, and found that CCD outperforms DTW [He et al., 2020].

Most of the work mentioned above relies upon around one month's worth of data for a transit agency, thus providing an informative glimpse into transit behavior at a point in time. Less work has been focused on applying these clustering techniques longitudinally as a way of understanding how behavior is changing. Briand et al. have followed up their work with Gaussian Mixture Models with a paper that analyzes behavior changes by investigating year-to-year cluster membership changes over five years using data from a medium sized transit agency serving Gatineau, Canada. They then used HAC on the clusters, and found that there was higher switching from year to year among clusters that were more similar in temporal patterns, as judged by the HAC output [Briand et al., 2017]. Additionally, Viallard et al. studied the same transit system to understand behavioral evolution on a week-to-week basis, using k-means on 7 features summarizing behavior for each day of the week [Viallard et al., 2019].

As researchers probe the frontier of classification algorithms, there is much that we can learn from what they uncover to be differentiating factors among riders in their data sets, and it is helpful to see how they have used knowledge on how transit systems work and what the key features of urban mobility are to inform their work. The framework presented here, however, is not intended to push forward that frontier but rather to bring the fruits of these labors within the grasps of American transit agencies. It aims to employ a tried and tested method— k-means— on input features deemed important by the CTA and validated as informative based on the literature above, in order to uncover the dominant behaviors among cards in the Ventra system.

This framework is flexible regarding the duration of time on which to calculate the input features as well as the set of cards that are clustered. In this chapter, we use four months of data for each clustering period and apply the algorithm to all cards in the system, thus demonstrating the usefulness of such a practice for the scale of data available to a large American city's transit agency. It aims to be easy and quick to reproduce as well as straightforward to interpret. In this way, we hope to offer similar agencies a process for identifying behavioral archetypes within the their ridership database and uncovering the types of behavioral shifts that are leading to overall changes in the number of trips or riders in their system.

3.2 Data

The data used in this analysis is from the CTA's Ventra account-based automated fare collection database, which houses the sale and use history of all Ventra cards. As of 2017, the first year considered in this analysis, the Ventra system captured 95% of all rides taken on the CTA [Vaishnav, 2019]. The database houses information on each card transaction, which encompasses each trip taken. The tap-in station and time are recorded for each trip, along with other information such as the cost of the trip, the fare product used to pay for it, and whether it was considered a transfer.

In addition to information on trips, the database also contains information on the purchase of fare products, including the time of purchase, the payment method, and whether it occurred via the Ventra mobile application, at a vendor located in the city, or via some other method, such as through an employer. While one could feasibly leverage all this information for input features and allow characteristics such as typical payment to contribute to the definition of the various clusters, this analysis opts to limit the input features to a small set of values that describe the temporal dimensions of each card's transit ridership. The additional information available from the Ventra database can later be layered on top of cluster assignments in order to observe how other rider characteristics, such as inferred home location or payment method, break down along behavioral lines.

We consider a transit account ID to be equivalent to one person. This is not a perfectly accurate assumption, as people who do not register their Ventra card are given a new transit account ID if they replace it. Thus, this analysis will count as separate people those who are issued a new transit account ID. While further work should address this issue, we believe it is not substantial enough to change the overall picture of behavior trends in the city. Cards that were completely free of cost were removed from consideration because these cards are frequently passed around among many users and inflate the number of riders who appear to be using the system with an extremely high frequency.

Notably, the Chicago system has only tap-in data, and thus the data points used to capture and distill riders' behaviors are limited to information that can be obtained from a tap-in system. As will be discussed in more depth in the next section, selection of the input features which will determine the dimensions along which the clusters are defined is a crucial step in this process, but one for which there is no clear correct answer. Each transit agency must decide on input features based on the data they have available and the goal of their analysis.

3.3 Methods

3.3.1 K-Means Clustering

The k-means clustering algorithm is a well-known and widely used machine learning algorithm for uncovering structure in large data sets. Within the realm of machine learning, it falls under the umbrella of "unsupervised learning" because it does not require a set of observation inputs and labels to learn the structure it is trying to uncover. Rather, the data, with each observation summarized by its values for the chosen input features, is taken in by the algorithm, which then outputs the labels for us.

The k-means algorithm works as follows [Lloyd, 1957]:

1. A pre-specified number (k) of cluster centroids are each assigned random values

for each input feature, locating them at random in m-dimensional space, where m is the number of input features.

2. Each data point is assigned to the nearest centroid, resulting in k clusters.

3. New centroids are calculated by taking the mean value of the data points within each cluster.

4. Steps 2-3 are repeated until the iteration in which no data point changes cluster assignment.

K-means works best when data across input features have similar scales. This is typically achieved by standardizing the data for each feature so that the values approximate a normal distribution, or by scaling the data for each feature so that all data falls along the unit scale. In this analysis, we choose the latter approach. We then match cluster assignment by transit account ID to the original data in order to investigate the results using the true values of the input features.

The k-means algorithm has some shortcomings that should be noted. First, it tends to bias results towards clusters that are roughly similar in size. Secondly, it assigns each data point to a cluster, regardless of how significant of an outlier that data point is. In this work, where all revenue-generating riders are included, this could lead to some non-intuitive cluster results for very infrequent riders. Further work on this topic should experiment with other clustering algorithms, including those that either do not assign every data point to a cluster, or those that allow for "fuzzy" cluster membership, where each data point can be associated with more than one cluster. For our goal of capturing the predominant behavioral archetypes present in a large transit system and investigating the stability of these behaviors over time, in aggregate and individually, k-means offers a quick and interpretable method of doing so.

Feature	Description
Weeks Rode	Number of weeks in which the rider used the system at least once
Percent Peak	Percent of all rides taken between 6AM and 10AM or between 3PM and 7PM on weekdays
Percent AM Peak	Percent of all rides taken between 6AM and 10AM on a weekday
Percent Weekend	Percent of all trips taken on a weekend
Range	Number of days between the riders' first and last trip during the study period
Average Weekly Rides	The average number of trips taken in weeks where at least one trip was taken

Note: Journeys involving a transfer are counted as one trip

Table 3.1: Description of Input Features for Longitudinal Cluster Analysis

3.3.2 Input Feature Selection

The selected features are outlined in Table 3.1. These six features were settled on by drawing upon the literature, specifically Basu's work on clustering groups of cards that included infrequent riders [Basu, 2018], as well as in consultation with the CTA. Additional temporal features were investigated, such as average daily rides, but ultimately excluded due to the low variability in this value across riders and the subsequently small role they played in dictating cluster assignment.

The set of features used in this section is rather limited, and excludes some, such as mode share and transfer rate, that will be used in the next chapter, which applies customer segmentation analysis to understand the transit ridership impacts of the COVID-19 pandemic. For the purposes of establishing the framework, we stick with including only temporal features in our clustering algorithm, but stress that this same procedure could be followed with a wide variety of feature sets.

3.3.3 Segmentation

To perform this analysis, we first clustered cards that were present in the Ventra system in the 17 complete weeks (Monday-Sunday) preceding December 31, 2017. The last day of 2017 happened to be a Sunday, so the study period for 2017 ran from

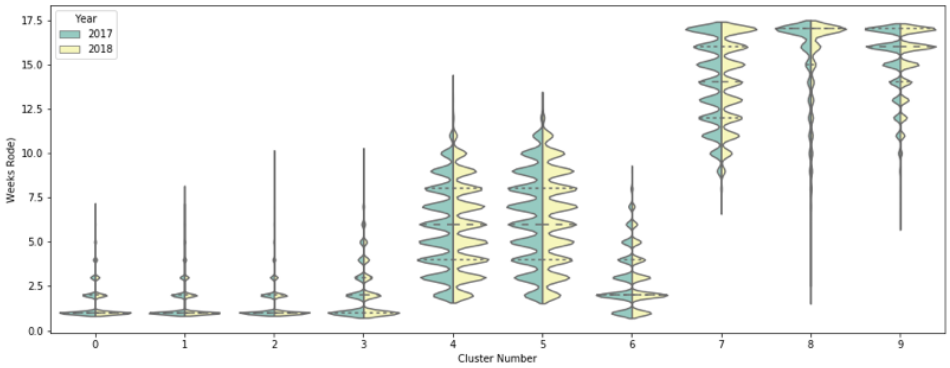
Monday, September 4 until Sunday, December 31. Each observation corresponded to a single transit account ID and consisted of a vector of six values, one corresponding to each of the six input features.

To determine the optimal number of clusters, the Elbow Method was used. This practice involves running the algorithm using multiple different values of k, plotting the intra-cluster variation as a function of the number of clusters, and selecting the number at which this variation begins to flatten out. This, combined with investigation of the outputs for various numbers of clusters and a desire for the clusters to be easy to digest and interpret, led us to settle on using 10 clusters.

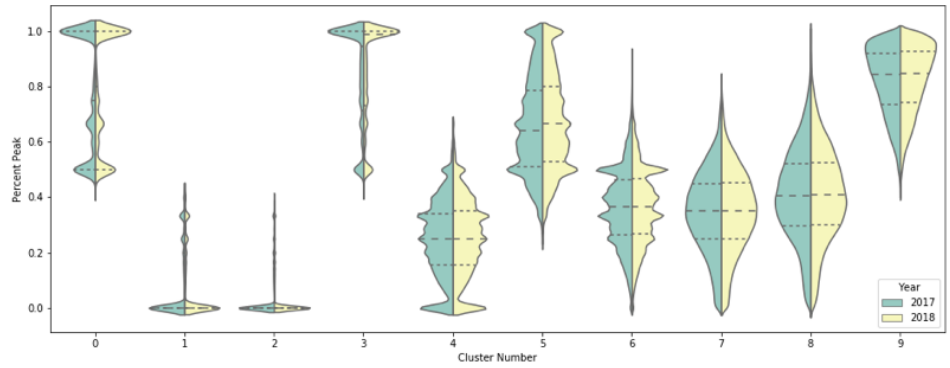
3.3.4 Establishing Stability

Next, the cards from the analogous time period in 2018 were clustered. Again we used the 17 complete weeks of data preceding December 31. For 2018, this led to a study period beginning on Monday, September 3, 2018, and ending on Sunday, December 30, 2018.

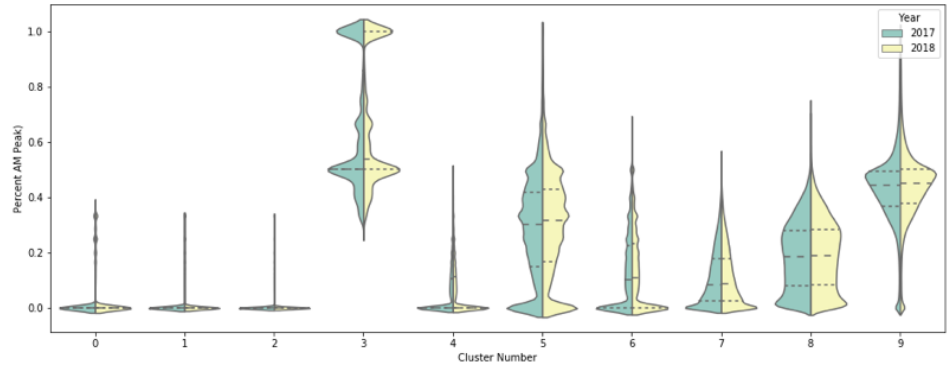
We then explored the stability of the clusters. We matched each 2017 cluster to the closest 2018 cluster, as measured by the Euclidean distance between the centers. Next, we quantified the percent change in the shifts of the centers for each cluster and observed that they were uniformly very small ($<1\%$). We further plotted the distribution of the true values of the features in each cluster and compared these across the two years. Figures 3-1 and 3-2 show the comparison between the 2017 and 2018 distributions for each feature and illustrates the nearly identical shapes and quantiles of the two years' data. These comparisons convinced us of the clusters' stability across these two years, and justified the following step, in which we fix cluster centers to be the same for both years so that a "cluster" has a single definition and we are able to perform longitudinal analysis.



(a) Weeks Rode



(b) Percent Peak

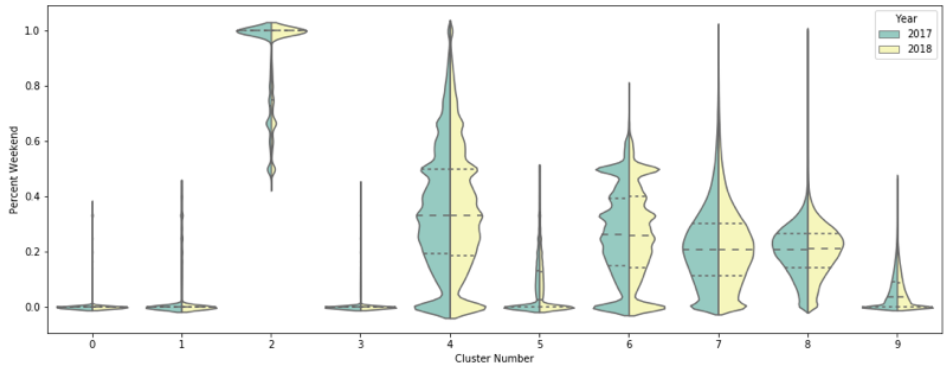


(c) Percent AM Peak

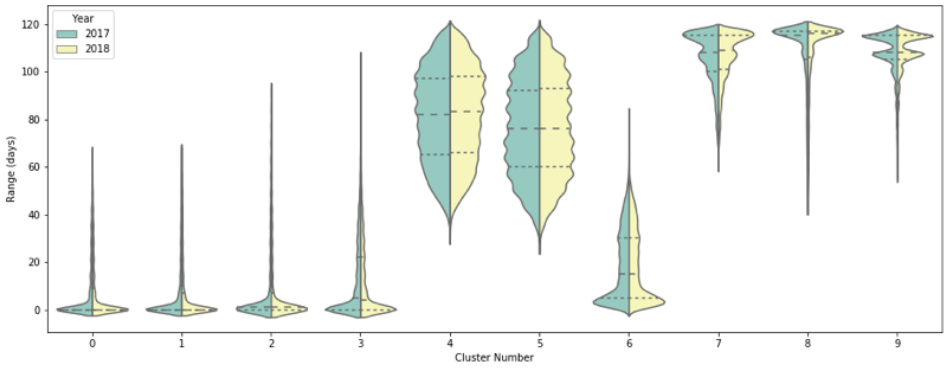
Figure 3-1: Distribution of Cluster Values for 2017 and 2018 (Part 1)

3.3.5 Longitudinal Comparison

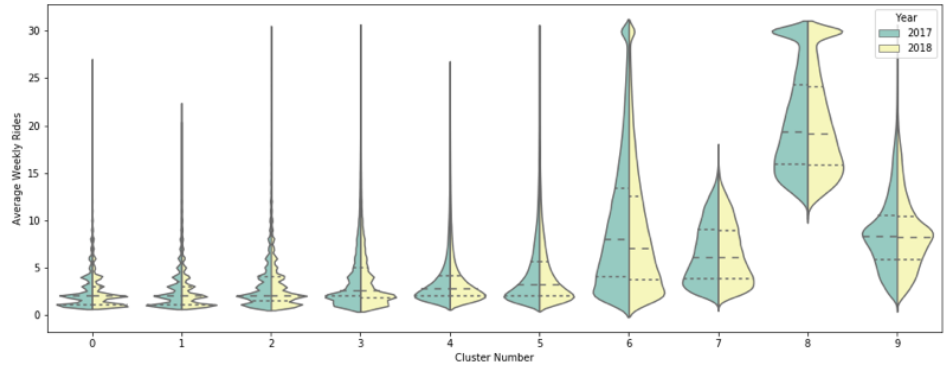
To allow for straightforward comparison across years, we fixed the center of each of the behavioral clusters to be the mean of the 2017 and 2018 centroids for that



(a) Percent Weekend



(b) Range



(c) Average Weekly Rides

Figure 3-2: Distribution of Cluster Values for 2017 and 2018 (Part 2)

cluster. All cards from both years were then reassigned based on these new, fixed centroids. Only 0.6% of all cards changed cluster assignment as a result, providing further evidence of the stability of the clusters.

Establishing a fixed definition for these behavioral clusters across years allowed us to use them as a basis for comparing the distribution of behaviors between the two time frames. Investigating which clusters grew in number and which clusters decreased in size gave us insights into the behavioral dynamics behind the overall drop in trips that occurred on the system from the fall of 2017 to the fall of 2018. It also allowed us to identify the cards that churned from the system after 2017, the cards that entered the system in 2018, the cards that were present in both years but exhibited changing behavior, and the cards that exhibited consistent behavior in both years.

In this chapter, we performed the clustering algorithm on all of the cards in the system (except the free cards mentioned earlier). Because of the relatively small number of features and output clusters, this process was not time-intensive (taking fewer than 10 minutes). An alternative method, however, is to cluster several smaller random samples from each year, determine the inter- and intra-year stability, assign fixed centroids based on some combination of the random samples, and then assign each card from the entire set to a cluster determined by the centroid to which it is nearest. Depending on the extent to which stability can be assumed or proven, this method would likely be the most expedient for transit agencies looking to implement this analysis, as assigning each card to the nearest centroid can be accomplished in seconds.

3.4 Results

In the end, we clustered 1,698,851 accounts that were active in 2017 and 1,692,086 accounts that were active in 2018 (including those that were also present in 2017). This section begins with a description of the 2017 clusters themselves and moves on to discuss findings from the longitudinal comparison.

Cluster Group	Cluster No.	Cluster Name	% of All Riders	% of All Trips
Infrequent	**0**	**Infrequent PM Peak**	7.9%	0.5%
	1	**Infrequent Weekday Off Peak**	8.3%	0.6%
	2	**Infrequent Weekend**	8.7%	0.7%
	3	**Infrequent AM Peak**	7.2%	1.1%
Occasional	**4**	**Occasional Off-Peak**	10.4%	3.7%
	5	**Occasional Peak**	9.6%	4.6%
	6	**Short Term High Frequency**	10.3%	4.7%
Regular	**7**	**Regular Off-Peak**	10.7%	15.8%
	8	**All Day**	5%	25.2%
	9	**Regular Peak**	13.5%	27.6%

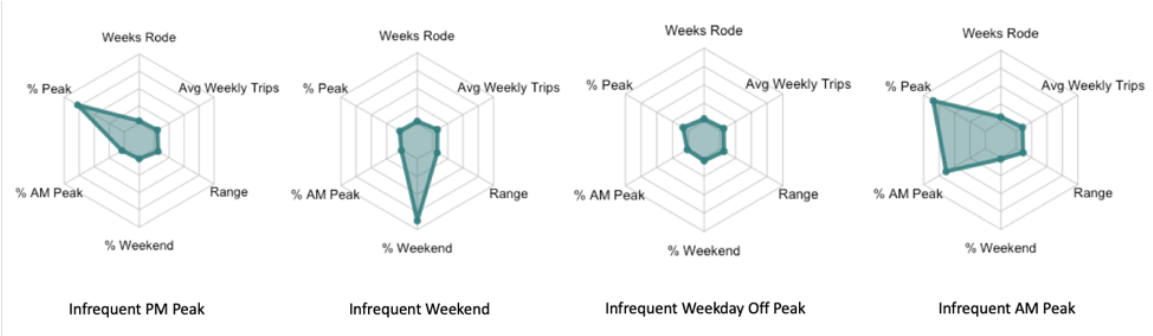
Table 3.2: Percent of Riders and Trips Belonging to Each Cluster - 2018

3.4.1 2017 Clusters

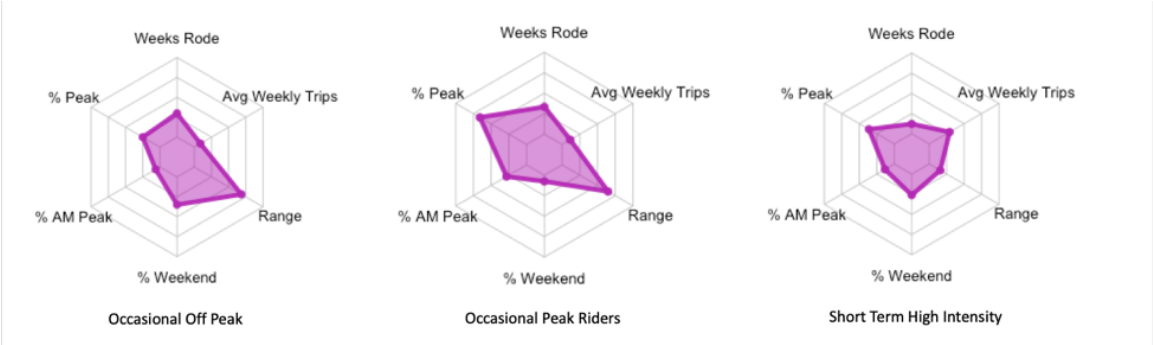
The ten behavioral clusters can be categorized into three groups based on the percentage of all trips that they represent. Aggregating up to this higher level allows for initial analysis that approximates work done by the CTA in the past, in which cards are classified exclusively by the frequency of their usage. Having the underlying clusters enables us to deepen the understanding of behavioral dynamics within each of these groups. Figure 3-3 provides a depiction of the relative centroid location for the features within each cluster, and Table 3.2 lists the clusters, identified by cluster group, as well as the percent of all trips and riders represented by each.

The Infrequent group encapsulates four clusters of infrequent or short term riders who together represent about 3% of all trips taken on paid cards in the fall of 2018. They are differentiated by the time at which these trips are typically taken. These four clusters account for 32.1% of all riders considered in the analysis.

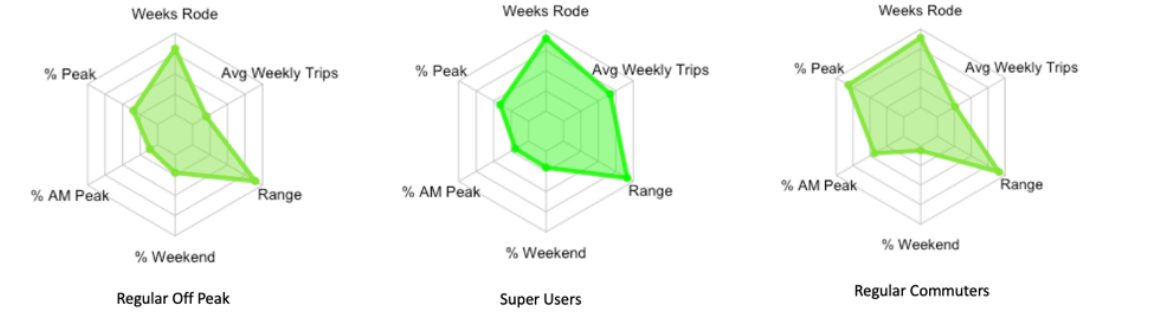
The next group contains three clusters, each of which accounts for between 3.5% and 5% of all trips taken by riders in this analysis. Two of these clusters represent occasional riders—— their first and last rides in the study period are nearly as far apart as the first and last rides of a regular rider, but they typically average only about four rides per week. The last of the three clusters in this group contains riders



(a) Infrequent Clusters



(b) Occasional Clusters



(c) Frequent Clusters

Figure 3-3: Relative Centroid Values by Cluster

whose first and last rides are much closer together in time (the average is about 19 days – for reference, the study period is 17 weeks), but who average slightly more rides per week than a regular commuter. We call these "Short Term High Frequency" riders, but for the sake of simplicity refer to the group as a whole as "Occasional" riders. Together this group accounts for 13% of trips and 30.3% of riders.

The final group contains the high frequency or regular riders: Regular Commuters ride often and at peak hours, Regular Off-Peak riders ride similarly frequently but

at off-peak hours, and All Day riders ride very frequently both on and off peak. The latter of these three, which we also call Super Users, is the smallest of the three. This group accounts for 5% of all riders but 25% of all trips. The Regular Commuters, on the other hand, account for 13.5% of all riders and 27.6% of all trips. The Regular Off-Peak riders are less numerous and account for fewer trips (15.8%).

As mentioned, there are limitations to this straightforward application of k-means. In order to get as holistic a picture of paying customers as possible, we did not filter out any riders based on usage criteria. Thus, someone who moved to the city at the end of the study period and began using the system regularly might be classified as a Short Term High Frequency rider when in reality their longer term behavior fits more appropriately with the Regular Commuter group. Despite these drawbacks, our clustering output is largely consistent with previous customer segmentation work done by the CTA using different methods and time frames. In addition, by incorporating the maximum number of cards, we have made it easier to translate the outputs of the analysis into intuition about what is going on system wide at the CTA.

3.4.2 Change in Cluster Groups Over Time

The goal of this exercise is to provide deeper insight into the often-reported top-level trends in the number of trips and riders on a given system. For the CTA during this time frame, the analyzed cards revealed a 0.4% drop in riders on the system and a 1.3% drop in trips taken. This section explores how an agency can use the methodology outlined above to uncover the underlying behavioral trends driving these numbers.

Our first step is to investigate change at the highest level of aggregation – the three cluster groups that correspond to volume of trips. Overall, there was a drop of about 0.4% in the number of cards in the analysis in 2018 compared with 2017. This drop was not uniform across the three cluster groups, however. Both the Infrequent and the Regular groups grew in size, the former by 1.7% and the latter by 0.7%. The Occasional group, however, dropped in size by 3.6%.

The set of factors contributing to the changing size of each of these groups is twofold. First, how does the number of new riders to the system entering this group

compare with the number of riders churning? And secondly, how does the number of riders shifting their behavior away from this group compare with the number of riders shifting their behavior to this group?

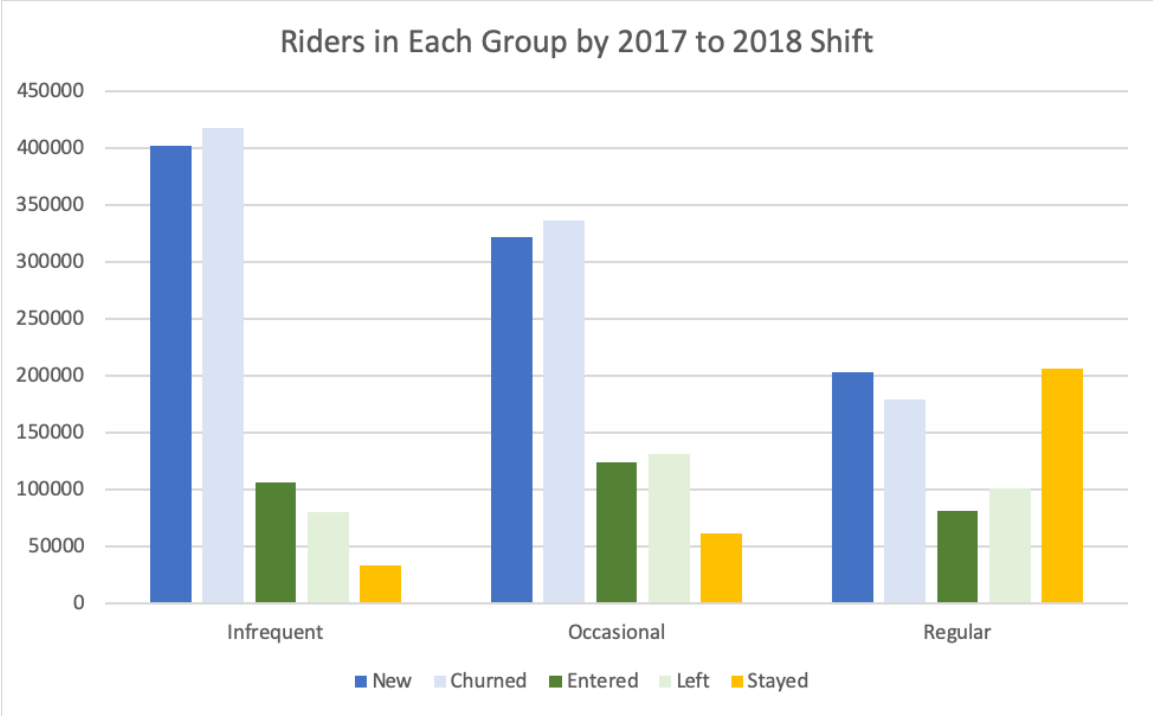


Figure 3-4: Number of Riders By Cluster Group and Observed Behavior Shift

Figure 3-4 illustrates that different trends are at work within each of the groups to yield the top-level losses and gains that we observe. "New" riders refer to those who were in the system in 2018 but not in 2017, "Churned" riders refer to those who were in the system in 2017 but not 2018, "Entered" and "Left" both refer to riders who were in the system both years but different cluster groups. The dark green bars indicate the riders who entered that cluster group in 2018 from a different cluster group, and the light green bars indicate the riders who were in that cluster group in 2017 but a different one in 2018. "Stayed" refers to riders who were in that particular cluster group in both years. Taken together, the light blue, light green, and yellow represent all the riders in each cluster group in 2017, while the dark blue, dark green, and yellow represent all the riders in each cluster group in 2018.

In the Infrequent group, more riders are churning from the system entirely than

new riders are entering that group. This is offset, however, by the fact that more riders are shifting to this behavior from other behaviors than are shifting away from this behavior. This leads to a slight gain in the number of Infrequent riders in 2018 compared with 2017. The Occasional group, however, is losing on both fronts, which drives the overall decline we see in this group in 2018. Regular riders, like Infrequent riders, are growing in number, but for the opposite reason. While more people are shifting away from this behavior than towards it, the number of new riders entering these groups not only is larger than the number churning, but makes up for the loss from the first type of shift.

These numbers hint at the complicated dynamics behind the relatively small overall changes in trip and ridership numbers presented at the top of this section. On the brighter side, the growth of regular riders is a positive sign for the agency, as this group accounts for about 79% of all revenue. The ability to attract new riders to this group at rates higher than those churning from the system bodes well for future revenue streams.

At the same time, however, there is a clear trend of riders who remain in the system decreasing their usage of it. While most riders who were present in both 2017 and 2018 were in the same behavior group both years (51%), of those who changed behavior, 56% moved to a less frequent cluster group compared with 44% who moved to one characterized by more frequent ridership. These shifts are the reason behind the absolute growth in infrequent riders —- their gain comes from riders decreasing their frequency and falling into the bottom tier of riders.

These trends hold both in terms of the absolute numbers of riders switching behaviors and the fraction of riders from each group that are switching. For example, 12.5% of Occasional riders (72,590) moved to the Infrequent group, while only 9.8% of Infrequent riders (56,817) made the opposite switch. In fact, we see very similar rates of exchange between Occasional and Regular riders: 12.5% of Regular riders (67,328) became Occasional riders while 10% of Occasional riders became Regular (58,173). The volumes are smaller for the exchange between Regular and Infrequent, but the imbalance holds: 6.2% of Regular riders (33,349) dropped all the way down

to infrequent riders, while only 4% (23,411) of infrequent riders climbed to Regular. A depiction of the volume of riders switching between groups is given in Figure 3-2b.

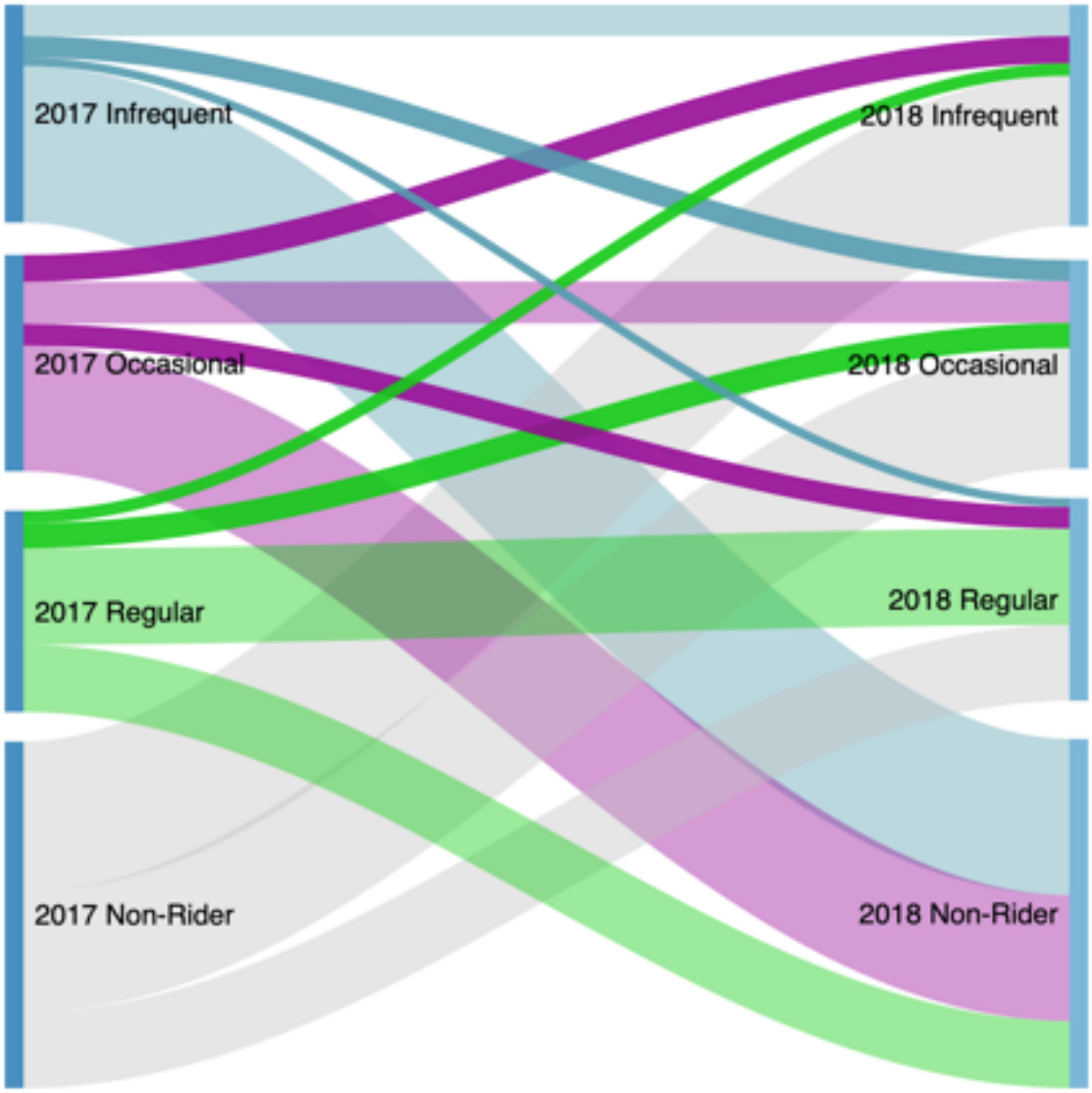


Figure 3-5: Size of Rider Behavior Shifts from 2017 to 2018

Taken together, we see two clear phenomena emerging. First, the CTA system is largely replacing churning riders with new ones. While the number of new riders is slightly below that of churned riders (0.7%), the distribution of new riders is slightly more in favor of regular riders than was the distribution of the churned riders, leading to an overall increase in the number of trips from this group, as well as revenue.

Secondly, and oppositely, the CTA is seeing a general trend toward decreased

usage among those who stay in the system. Those cards which were present in both 2017 and 2018 collectively took 6.7% fewer trips during the fall of 2018 as compared with 2017.

3.4.3 Change in Clusters Over Time

We now turn to investigate the individual clusters in order to gain deeper insight. Table 3 shows the percent change in size of each of the individual clusters from 2017 to 2018. This reveals that while all the clusters in the Infrequent group are growing in size and all clusters in the Occasional group are shrinking in size, the overall growth in the Regular group is driven entirely by growth in the Regular Commuters group, which is compensating for decreases among Super Users and Regular Off-Peak riders.

We also note that in general, clusters characterized by ridership in the peak hours are faring better than those characterized by off peak ridership. Aside from the growth in Regular Commuters, we also see relatively small losses in Occasional Peak riders compared with the other Occasional clusters, and we see noticeably more growth in the Infrequent AM Peak cluster than in the other Infrequent clusters. This is validated by the fact that the 1.3% drop in overall trips noted above is not uniform. Rather, weekend trips dropped by 1.9%, weekday off peak trips by 3%, and peak trips by only 0.1%. By looking more closely at what is happening within each of these clusters, we can understand how rider behavior is behind these numbers.

Figure 3-6 shows the count of new and churned riders by cluster, and Figure 3-7 shows the counts of people shifting to and from each of the clusters. The darker blues and greens represent the people that were in that cluster in 2018, while the lighter colors represent the people that were in that cluster in 2017.

We see that within each of the three larger cluster groups, the tradeoffs between new and churned or between shifts to and away from clusters are directionally consistent among the member clusters. That is, all the Infrequent clusters see more churned than new riders and more shifts to them than away from them; all the Occasional clusters see more churned than new riders and more shifts away than to; and all the Regular clusters see more new riders than churned and more shifts away than to

Cluster Group	Cluster No.	Cluster Name	Change in Number of Riders
Infrequent	**0**	**Infrequent PM Peak**	+0.5%
	1	**Infrequent Weekday Off Peak**	+0.5%
	2	**Infrequent Weekend**	+2.8%
	3	**Infrequent AM Peak**	+5.2%
Occasional	**4**	**Occasional Off-Peak**	-4.2%
	5	**Occasional Peak**	-1.7%
	6	**Short Term High Frequency**	-4.6%
Regular	**7**	**Regular Off-Peak**	-2.6%
	8	**All Day**	-1.5%
	9	**Regular Peak**	+3.1%

Table 3.3: Change in Cluster Membership Size from 2017 to 2018

them. The differences in cluster size changes among these groups then comes down to the relative size of each of the gaps – between new and churned riders and between behavior change and behavior adoption.

The Regular Commuter group is growing while the Super Users and Regular Off-Peak group shrink, for example, because the gap between new and churned riders is larger than the gap between those shifting away from being Regular Commuters and those adopting that behavior. In the other Regular clusters, the gain of new over churned riders is not enough to make up the loss from people changing behaviors.

We can also investigate in detail from which cluster those who change behavior are coming. Figure 3-8 has 2018 cluster membership on the x-axis and shows the breakdown of 2017 behaviors among riders that entered that cluster in 2018.

Each cluster pulls from all of the other clusters to some extent, though we note that in almost every case, the largest share came from a cluster characterized by the same peak level but a different frequency level. Exceptions are Infrequent PM Peak, which drew about equally from Occasional Peak and Occasional Off-Peak, and Regular Commuters, which adopted Regular Off-Peak riders at about the same rate that it adopted Occasional Peak riders.

Figure 3-6: Count of Churned and New Riders by Cluster

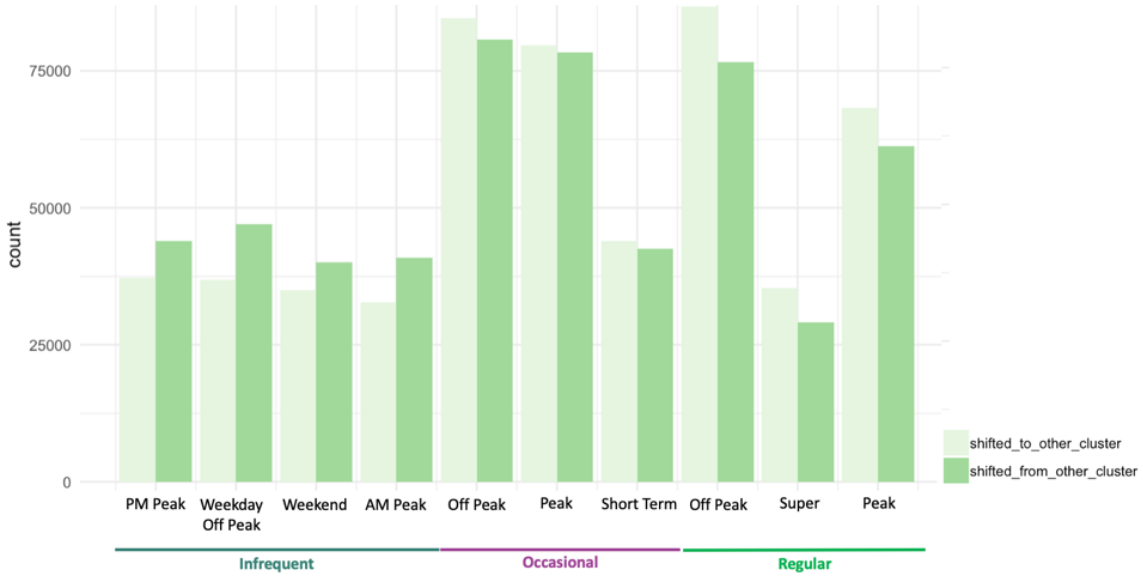


Figure 3-7: Count of Riders Shifting To and Away from Each Cluster

3.5 Case Study: January 2018 Fare Increase

Beyond helping transit agencies uncover the behavioral trends driving overall ridership loss (or gain) on their systems, this analysis can also aid in the diagnosis of policy interventions by breaking down the responses by cluster to see how various groups reacted. Here we take the January 2018 fare increase on the CTA as a case study and offer an example of how this framework can be used to explain the better-than-

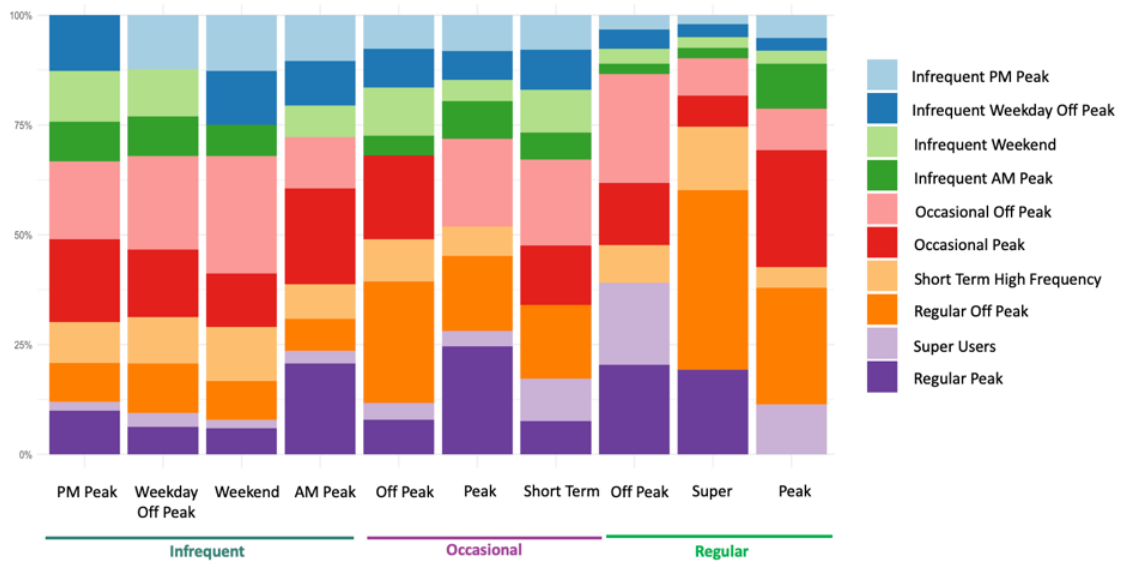


Figure 3-8: Percent of Non-New Riders in Each 2018 Cluster by 2017 Cluster

predicted results of the fare increase, enable deeper analysis into specific rider groups of interest, and inform future policies. This analysis was done in collaboration with Maulik Vaishnav at the CTA.

3.5.1 Fare Increase Outcome and Diagnosis

CTA ridership in 2018 declined for a third consecutive year. In January 2018, the agency increased the base fare by $0.25 and 30-Day Pass price by $5. The agency budgeted annual ridership of 462 million, down from 479 million in 2017 and anticipated revenue to grow by $23 million. At year end, ridership reached 468 million and CTA generated $27 million in additional revenue. Our clustering segmentation analysis helps shed light on these better-than-anticipated results: growth in Regular Commuters due to a robust downtown economy helped increase revenue and ridership. Their growth offset losses seen in other larger segments, such as super users and off-peak users.

3.5.2 Deeper Investigation of Regular Commuters

Because this segmentation methodology assigns a cluster label to each account in the system, we were able to delve into more detail about Regular Commuters to understand that group. As anticipated, many Regular Commuters begin their trips (their inferred home location) on the north-side of the Chicago. These relatively price-inelastic 80,000 riders accounted for 15% of fare revenue in 2018. When compared with other 2017 Regular Commuters who were geographically stable (did not change inferred home location) between 2017 and 2018, north-siders were significantly more behaviorally stable (did not change cluster assignment). Fully half of north-side Regular Commuters remained Regular Commuters. In other regions, the share of Regular Commuters maintaining their behavior only reached as high as 42% but was as low as 25% in some places. This may speak to the many transit-rich neighborhoods in the north, as well as the fact that these neighborhoods are typically wealthier and home to individuals commuting to and from downtown. Furthermore, the share of riders using a 30-Day Pass increased slightly even as the price increased.

3.5.3 Policy Implications

Many large cities in the US have seen similar growth in population and employment along transit-rich corridors. Our analysis indicates that this market is using transit mainly during the morning and evening peaks on weekdays for commuting, and they are relatively price-inelastic. However, as in Chicago's case, most other major groups decreased their membership numbers or use over the year. While many reasons may have contributed to these declines, it is important to design future policies that target growing vs. declining segments differently. For example, a future fare increase may be more successful if peak rail fare was introduced that mostly targets this inelastic market more than other segments.

We also examined the change of behavior in people who switched their pass type in response to the fare increase. We found that regular commuters who switched from pay-per-use fares to a 30-Day Pass increased their ridership by an average of 30

percent, while the group doing the reverse decreased their use by 16 percent. Their use increased on weekends as well as both peak and off-peak periods on weekdays, but increased by a higher proportion in off-peak and on weekends. Notably, 7.8% of this cohort with no pass use in 2017 moved up to become super-users with a 30-Day Pass in 2018. Fare policies that prioritize pass use and keep their prices affordable relative to base fare can therefore anticipate an increase in ridership, not only in peak times, but also in off-peak and weekends when transit travel times are slower.

3.6 Conclusion

This chapter develops a framework for using AFC data to identify the behavioral shifts and trends that are underlying the change of top-level ridership and trip numbers frequently reported by transit agencies. The analysis focuses on a comparison between fall of 2017 and 2018 data in Chicago to illustrate the amount of insight that can be gained from data that is even just a single year apart.

In this chapter, we start with the fact that the number of cards in the system has declined by 0.4% and the number of trips has declined by 1.3% We then examine the three cluster groups to determine that these numbers can be explained more by remaining riders decreasing their usage than by new riders using the system less than churning riders. By diving deeper into the individual clusters, we learn that new riders entering the system as Regular Commuters are largely responsible for limiting the drop in trips on the system. We also noted a slight shift toward peak travel and a tendency for people to change the frequency with which they ride at higher rates than they change the time of their typical travel during a given week (peak/off-peak).

Evaluating this information in the context of the January 2018 fare increase reveals that continued growth of the Regular Commuter group helps explain why the CTA outperformed revenue and ridership predictions for this year. Delving deeper into this slice of the ridership offered insights that can help inform future fare policies at the agency.

The framework provided in this chapter offers several advantages for transit agen-

cies hoping to make ridership behavior a fundamental part of their regular analysis. First, it uses a well-established and computationally efficient algorithm to create behavioral profiles that contain multiple relevant dimensions and are easily digestible. In other words, it is straightforward both to implement and to interpret. Next, it can easily be replicated in the future to investigate how these trends progress. Once fixed cluster centroids have been determined, cards can easily be assigned to a behavior group for discretized time frames. Periodic re-clustering is advised to ensure that the fixed clusters remain close to independent clustering on a newer set of data. Third, the output of this methodology can be easily layered with other analyses. We have captured temporal behavior in a single variable, which can now be interacted with a host of other aspects of the ridership experience, such as mode choice, location choice, or pass purchase behavior. Lastly, it enables analysis that is rider-centric. Issues of decreasing ridership, whether they be across the system, on certain modes, on certain lines, or in certain regions, are ultimately the result of individuals choosing to alter their ridership behavior. This method puts the question of "who?" at the forefront of investigating such issues.

In the following chapter, we employ a similar framework to understand the implications of a shock to the system much larger than a fare increase— the COVID-19 pandemic. Because of the magnitude of the change in ridership behavior, we do not attempt to establish stable clusters, but rather only segment clusters based on pre-pandemic behavior and examine behavior changes by group. Such an extreme alteration to typical transit behavior patterns suggests an extension of this work that does not seek to establish the same behavioral clusters over time, but rather identifies the behavioral segments most indicative of riders in each specific time frame. While this approach will be more complex to interpret and analyze, as the number of behavioral profiles will be much larger if not consistent over time, it will likely be necessary for the time being as urban areas deal with the repercussions of the pandemic.

Chapter 4

Customer Segmentation Case Study: Ridership Impacts of COVID-19

On January 20, 2020, the Center for Disease Control and Prevention confirmed the first positive test for COVID-19 in the United States, a 35-year-old man in Snohomish County, Washington [Holshue et al., 2020]. Over the course of the next two months, the number of confirmed cases increased slowly but steadily, reaching 100 on March 2 [Johns Hopkins University and Medicine, 2020]. In early March, as the United States began to greatly increase its testing capacity, the number of confirmed cases grew more rapidly, jumping from 100 on March 2 to 4,604 two weeks later. On March 11, the World Health Organization officially declared the outbreak to be a pandemic, and US state and local governments that had not already done so began to enact sweeping restrictions regarding which establishments could remain open, how large gatherings could be, and to what extent citizens should spend time outside their residences.

Along with these restrictions came a dramatic drop in the number of trips taken on public transport as people's workplaces closed, nearly all events were canceled, and many large cities issued "shelter in place" orders. The latter half of March, along with April and May saw public transit trips at 10-30% of their typical levels, though there was heterogeneity in the size of the drop by city, mode, and demographics. At the time of this writing, the future of public transit in American cities is still very much unknown as people grapple with changing employment circumstances and the

public health implications of riding mass transit. At the end of April, 30 million Americans had filed for unemployment [Tappe, 2020] and several major companies have announced that their employees may continue working from home indefinitely [Kelly, 2020], eliminating the need for millions of commuting trips that would have occurred on public transit. Additionally, as cities begin to slowly reopen, many urban dwellers are likely to opt for modes of travel that do not require being in close proximity to strangers, such as personal vehicles and biking.

The challenges to recuperating public transit ridership losses are immense. Having a deep understanding of who public transit riders were and how they responded to the COVID-19 crisis is critical as agencies attempt to chart a path forward. With such steep obstacles to recovering ridership, agencies will be well-served to learn what they can about their riders and craft policies with their needs in mind.

This chapter uses the customer segmentation methodology presented in the previous chapter and the city of Chicago as a case study for how a transit agency might analyze the impacts of COVID-19 on ridership and use this to inform policies geared at recovering lost riders and trips. First, we present context on the impact of COVID-19 on the CTA system as a whole. Then we establish the baseline behavior of CTA riders and examine the ridership responses to COVID-19 of each of the behavioral groups, highlighting findings related to ridership characteristics that are particularly predictive of COVID-19 ridership and what this may mean for policy going forward. Lastly, we offer policy recommendations for the revival of transit usage in Chicago, considering several behavioral groups in turn and developing policy suggestions that pay specific attention to the circumstances and needs of each group.

4.1 Structure of the Analysis

To study the impacts of COVID-19 on CTA ridership, we use Ventra fare card tap-in data. We establish the pre-COVID baseline behavior of riders on the system based on the eight complete weeks between Monday, January 13 and Sunday, March 8, 2020. To differentiate baseline behavior across different types of riders, we assign all Ventra

cards that were used at least once during the baseline period (about 1.3 million cards) to one of fourteen behavioral clusters.

Two of these clusters are define heuristically instead of algorithmically. The first such cluster includes all cards that were used only for a single day in the baseline period. Because these riders have often taken only a single trip, their extremely brief presence on the system is their behavioral attribute of the most interest and the other attributes of interest, which are largely calculated as the percent of trips taken that meet some criteria, are forced to the extremes which could lead to outcomes from the clustering algorithm that are less robust.

The second group consists of cards with a type of pass that allows the user to ride free. Individuals holding these passes are significantly more likely to share their card with others, making it harder to stand by the assumption that one card equals one person, which underlies this analysis. Because many of these riders are lower income, however, we did not want to exclude them from consideration altogether, so they are assigned their own cluster heuristically, like the one-day riders. The remaining cards (about 900,000) are then clustered using the k-means algorithm on the scaled values of the input features seen in Table 4.1. The elbow method was used to settle on 12 clusters based on these input features.

Having established a pre-COVID baseline categorization of CTA riders and their travel, we then track how their travel patterns change through the COVID-19 pandemic period. We define an early stage COVID period using the two complete weeks from Monday, March 23 until Sunday, April 5 and a late stage COVID period using the four complete weeks from Monday, June 22 until Sunday, July 19. The early stage COVID period spans between the implementation of Chicago's stay-at-home order on Saturday, March 21, and the implementation of the CTA's rear-door boarding policy on all buses on April 9. The late stage COVID period comes after the lifting of Illinois's stay-at-home order at the end of May and after the CTA resumed front-door boarding on buses on Sunday, June 21.

To characterize the ridership response to COVID-19 from each group, we investigate the percent of riders that used the system even once during each of the during-

Feature	Description
Weeks Rode	Number of weeks in which the rider used the system at least once
Percent Peak	Percent of all rides taken between 6AM and 10AM or between 3PM and 7PM on weekdays
Percent Weekend	Percent of all trips taken on a weekend
Range	Number of days between the riders' first and last trip during the study period
Average Weekly Rides	The average number of trips taken in weeks where at least one trip was taken
Percent Bus	Percent of all trips taken on bus
Percent Transfer	Percent of all trips involving a transfer (rail to rail transfers not captured)

Note: Journeys involving a transfer are counted as one trip

Table 4.1: Description of Input Features for COVID Cluster Analysis

COVID-19 analysis periods. Because we are interested in individual-level behavior, it is possible we are missing people who rode during the rear-door boarding period but not either of our analysis periods. Additionally, in this analysis we do not deal with the cards that did not appear in our baseline period but did appear during one or both of the COVID analysis periods. This group is not insignificant – it is about 25% of the riders in the late stage period, although many could simply be previous riders who have started using a new Ventra card – but because we were not able to establish baseline behavior for them, we set them aside for this analysis.

4.2 Context: COVID-19 and Public Transit Ridership in Chicago

Before discussing individual behavior changes due to COVID-19, it is useful to understand at an aggregate level the decrease in trip volume observed during the COVID-19 pandemic. Figure 4-1 shows the daily count of Ventra card taps by mode from early January 2020 until mid July. We note that trip volume appears largely consistent starting in the second week of January until the week of March 9, in which we note slightly lower than normal trip volume on rail, especially in the early part of the week,

and a steep drop off for both modes on Thursday and Friday and into the weekend. The week starting on Monday, March 16 appears to be a transition week, in which transit trips continued to drop. The Saturday of that week, March 21, marks the start of Chicago's stay-at-home order. The following two weeks show consistently low trip volumes, appearing to plateau at a new normal. On April 9, the CTA implemented a rear door boarding policy, meaning that all riders were required to board the back door to provide some protection for their operators. Because the vast majority of CTA buses did not have Ventra card readers installed at the rear door at the time, this policy essentially equated to free bus rides. As a result, as can be seen in the figure, during this time there is virtually no smart card data from bus trips. Front door boarding was re-instated at the end of June.

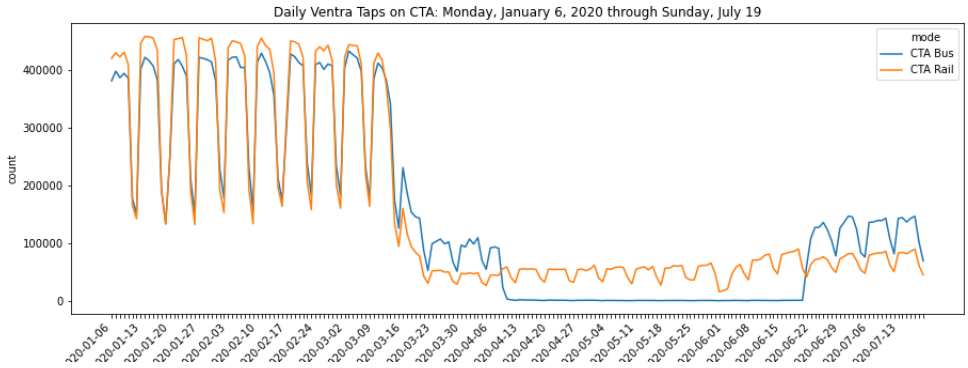


Figure 4-1: Daily Ventra Taps by Mode Since First Monday of 2020

The CTA saw a massive drop in trips across the board, down from almost 5 million average weekly trips to about 940,000 in the early stage and 1.3 million in the late stage. The drop was more pronounced on rail, which dropped from 2.5 million average weekly trips to 310,000 in the early stage, a drop of 88%. The count of rail trips has since risen to 490,000 in the late stage, or 20% of baseline volume. Bus, on the other hand, had baseline volumes just below those of rail (2.4 million average weekly trips), but early stage trip volumes more than double that of rail, at 630,000. Late stage bus ridership is at 840,000 average weekly rides, or 35% of baseline levels. Rail has seen a greater percentage increase in trips between the early and late stages of the pandemic compared with bus (+59% compared with +34%), but is still drawing trips

in numbers below even early stage bus trip counts.

4.2.1 Temporal Patterns

We also investigate the loss of trips along temporal and spatial dimensions. With regard to the temporal dimension, Figure 4-2 shows the hourly distribution of trips by mode for a typical weekday and weekend in each time frame. We note that trip volumes have decreased for every hour on both weekdays and weekends, but most dramatically during the weekday peak hours. COVID-19 has largely eliminated the strong peak pattern of weekday travel, with fully 50% of the initial lost trips in an average week coming from the hours of 7-10AM and 4-7PM on weekdays. This is likely due to a combination of these trips no longer being taken at all due to office jobs moving to remote work, as well as people shifting travel to other times for fear of crowded conditions on transit during these hours.

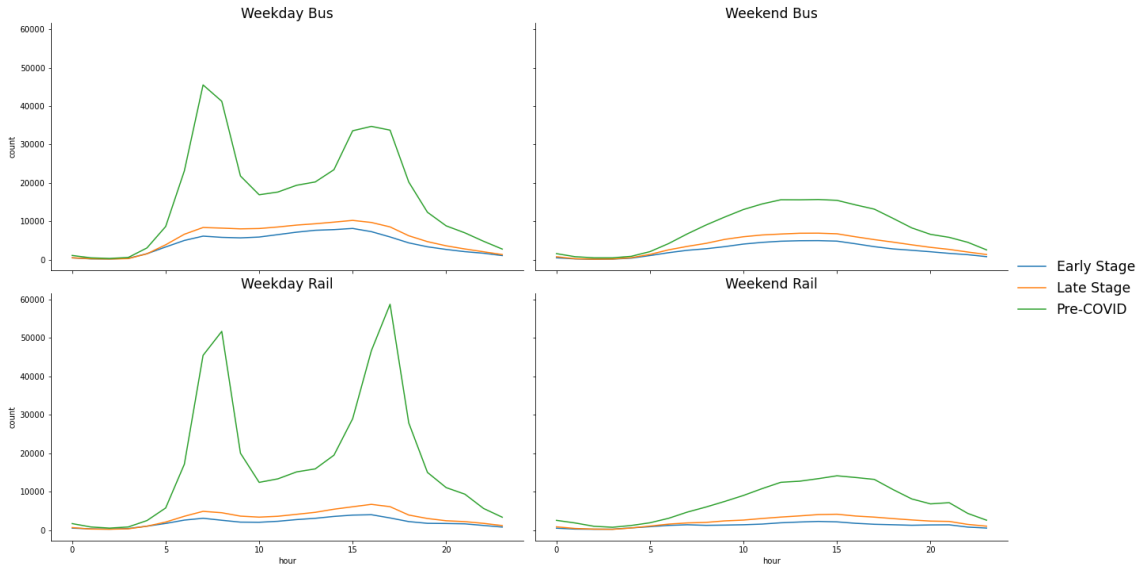


Figure 4-2: Temporal Distribution of Daily Trips by Mode, Weekend/Weekday, and Time Period

4.2.2 Geographical Patterns

When looking at the spatial distribution of trip loss, we see clear geographical patterns. Figure 3 shows the percentage decrease in average weekly trips by commu-

nity area in Chicago. The steepest declines are in the areas north of downtown, close to the coast of Lake Michigan. These neighborhoods have a greater percentage of white/Caucasian residents and are more affluent than the neighborhoods in the south and west of the city, which are majority minority and much lower income. The pattern of public transit usage dropping more in wealthier neighborhoods has been observed in other cities, and seems to be strongly related to the fact that "essential workers" are more likely to be lower income and people of color than the population as a whole [Valentino-DeVries et al., 2020, Goldbaum and Cook, 2020, Rho et al., 2020]. As Chicago has opened up, the percentage increase in rides from the early stage to the late stage has been greater on the north side, though the overall decline from the baseline remains much higher in the north (Figure 4-3).

This initial analysis allows us to see that the drop in trips due to COVID-19, while unprecedentedly large across all modes, time periods, and neighborhoods, was most pronounced on rail, during peak hours, and in wealthier, majority white communities. In later sections, we will show that this is a product of the distinct behavioral responses from different groups of riders.

4.3 Behavioral Baseline

We begin by establishing a behavioral baseline using data from the eight weeks leading up to the escalation of the pandemic and the response to it. Using the methodology described in detail in the previous chapter, we can describe the status quo of ridership behavior using 14 clusters, including one-day riders and free riders.

Table 4.2 gives the average value of each of the key features for the 14 clusters, including one-day riders and free riders, which can be interpreted as the re-scaled center of each cluster. We first note the clear delineation between riders who were active for only a small part of the baseline period (clusters 0, 1, 2, 3, 5, 12) and those that were active for the entirety (clusters 6, 7, 8, 9, 10, 11). We can deduce that the six clusters with mean ranges around 7 weeks consist of riders who live in Chicago. It is harder to draw definitive conclusions about the riders in one of the five clusters with a

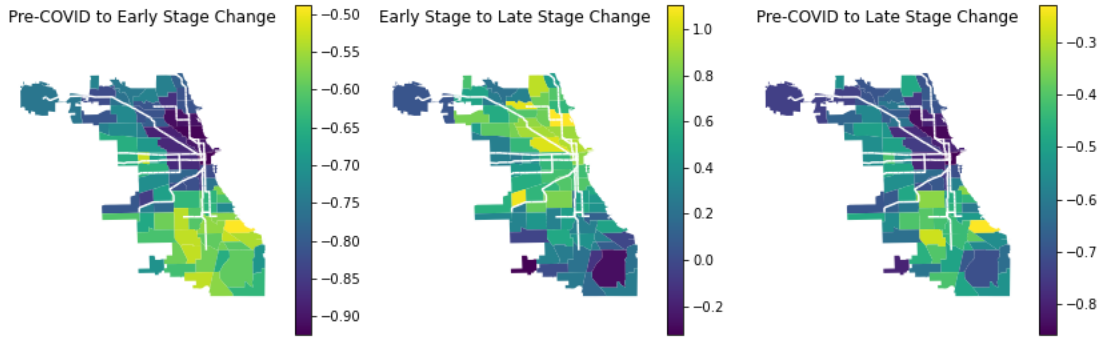


Figure 4-3: Percent Change in Average Weekly Trips Between Pre-COVID and Early Stage (Left), Between Early Stage and Late Stage (Middle), and Between Pre-COVID and Late Stage (Right) by Community Area

shorter range. It is possible that they were visiting the city, or perhaps made a lifestyle change toward the beginning or end of the study period that led to them appearing in or disappearing from the CTA system. This could also capture riders who had to replace an unregistered Ventra card, or riders who ride infrequently enough that they would not appear in the system for several consecutive weeks. These riders have on average fewer weekly rides than those who appear in the system for the duration of the study period, suggesting a combination of visitors and very low frequency riders.

The details of when they ride can provide some clues as to which clusters are which, with the first two clusters, which have a high percentage of weekend rides, being more likely to correspond to visitors, while those with a high percentage of trips taken at peak perhaps corresponding to very infrequent commuters.

Two slight exceptions to the general correlation between range and average weekly rides are clusters 4 and 5. The first of these, Medium Range Infrequent Semi-Peak Rail, consists of riders with a relatively long range on average, of over 5 weeks, but only about 2.5 rides per week. Additionally, despite having a range of over 5 weeks, typically ride during only 3 or 4 of those weeks. This suggests riders who live in Chicago but only use transit every once in awhile. These riders overwhelmingly opt for rail over bus and have a higher percentage of weekend trips and lower percentage of peak trips than the average rider. This group likely uses the CTA occasionally for leisure trips, for special purpose trips like getting to or from the airport, or when their primary mode is unavailable. When they do use the CTA, they avoid the bus and trips that require transfers.

The other exception is the Low Range Occasional Weekday Mixed Modes cluster, whose range is only a little more than two weeks on average, but typically takes around 4.5 rides per week. This group uses bus more than rail and has a high rate of transfers. This group is likely to be capturing some of those riders who are replacing unregistered Ventra cards, maybe because they purchase 7-Day Passes with cash.

The last six clusters are all high range clusters, with riders present in the system throughout the entire eight-week baseline period. The first of these, High Range Occasional Semi-Peak Bus, is classified as occasional because of the relatively low number of average weekly rides when compared with the remaining five clusters. All of the last five have similar values for average range, average weekly rides, and weeks rode. They differ by primarily by mode and the percentage of rides taken at peak, as well as their transfer rate.

The first of these five is characterized by the high percentage of trips taken on bus and during the peak. These riders take on average just over 6 trips per week, which are focused during the peak hours, likely for commuting, and do not involve transfers.

They almost never ride on the weekends. The next group is again characterized by a concentration of trips taken during peak hours, but these riders use both modes and transfer often. Like the other groups characterized by ridership at the peak, they take very few trips on the weekend. The third group rides primarily on rail and mostly in the off-peak. They take about a quarter of their trips on the weekend, which is more than the average rider. This behavior could capture rail commuters who work jobs that do not operate on a 9-5 schedule, students, or individuals who work from home but use rail for errands and other recreational needs, among other groups.

For our analysis of COVID-19 impacts by rider type, we will focus on the final two of these clusters. This is because, aside from one-day riders and free riders, these two groups represent the largest percentages of pre-COVID trips and riders on the entire CTA system. Further, apart from the frequency with which the riders use the system, they are different in every way and thus, as we will see, have very different responses to COVID-19. The High Range Frequent Off-Peak Bus Transfer cluster represents 7% of riders and 15% of trips. Only about one-third of these riders' trips are taken during peak hours, while nearly three-quarters are on bus and most involve a transfer. This group has the highest mean value among all the clusters for average weekly rides, at nearly eight. The High Range Frequent Peak Rail cluster, on the other hand, travels nearly exclusively via rail on weekdays during peak hours and almost never transfers. Of the twelve algorithmically defined clusters, this last group represents the largest share of riders (10%) and trips (20%) in the baseline period by a significant margin.

We can also examine the spatial distribution of the inferred home locations of riders (Figure 4-4, left panel). We note that the system's riders in general are concentrated largely along the north coast. Mapping the spatial distribution of the High Range, Frequent clusters separately mostly mirrors this trend—with the notable exception of the Off-Peak Bus Transfer group, which is spread more evenly among the community areas, with more riders living in the south and west parts of the city than we see in the other clusters (Figure 4-4, right panel). These areas of Chicago are disproportionately low income and overwhelmingly black. This suggests that riders from

No.	Name	Range	Avg Weekly Rides	Weeks Rode	%Peak	%Weekend	%Bus	%Transfer
0	LR Inf Weekend Bus	17.44	2.88	2.38	0.13	0.62	0.76	0.4
1	LR Inf Weekend Rail	10.36	3.16	1.87	0.16	0.51	0.10	0.1
2	LR Inf Peak Bus (No Transfer)	16.39	2.96	2.36	0.63	0.06	0.90	0.1
3	LR Inf Peak Rail	14.54	3.13	2.22	0.76	0.03	0.07	0.1
4	MR Inf Off-Peak Rail	37.95	2.67	3.86	0.36	0.23	0.15	0.2
5	LR Occ. Weekday Mixed Modes	17.23	4.42	2.56	0.46	0.11	0.69	0.7
6	HR Occ Off-Peak Bus	46.73	4.29	6.13	0.32	0.23	0.86	0.3
7	HR Freq Peak Bus (No Transfer)	50.47	6.10	7.34	0.83	0.06	0.90	0.1
8	HR Freq Peak Mixed Modes	50.54	7.10	7.37	0.81	0.06	0.61	0.7
9	HR Freq Off-Peak Rail	51.26	6.73	7.50	0.34	0.24	0.17	0.2
10	HR Freq Off-Peak Bus Transfer	50.43	7.71	7.24	0.32	0.22	0.73	0.7
11	HR Freq Peak Rail	50.73	6.86	7.42	0.86	0.04	0.07	0.1
12	One Day Riders	0.00	1.44	1.00	0.41	0.24	0.40	0.2
13	Free Riders	43.93	6.08	6.27	0.41	0.18	0.61	0.5

LR, MR, and HR refer to "Low Range", "Medium Range", and "High Range", respectively, referring to the average value for the Range feature and indicating the amount of the 8 week study period the riders were present on the system for. Inf, Occ, and Freq refer to "Infrequent", "Occasional," and "Frequent," respectively, and refer to the mean value for Average Weekly Rides, indicating the frequency with which the riders rode during the weeks in which they were active.

Table 4.2: Pre-COVID Baseline Behavior Cluster Centers

these neighborhoods, who are more likely to be lower-income and black, have travel behavior characteristics — off-peak rather than peak, bus rather than rail, frequent transfers—which are typically associated with lower levels of service. Even apart from COVID-19 responses, this suggests that a system that allocates resources according to where and when the majority of trips occur could overlook one of the largest blocs of riders responsible for 15% of all trips in pre-COVID times. This speaks to the importance of analysis that keeps the rider rather than the trip at the center, as it allows us to recognize a group of riders that is crucial to overall ridership numbers but typically uses the system at times when overall trip volume is relatively low.

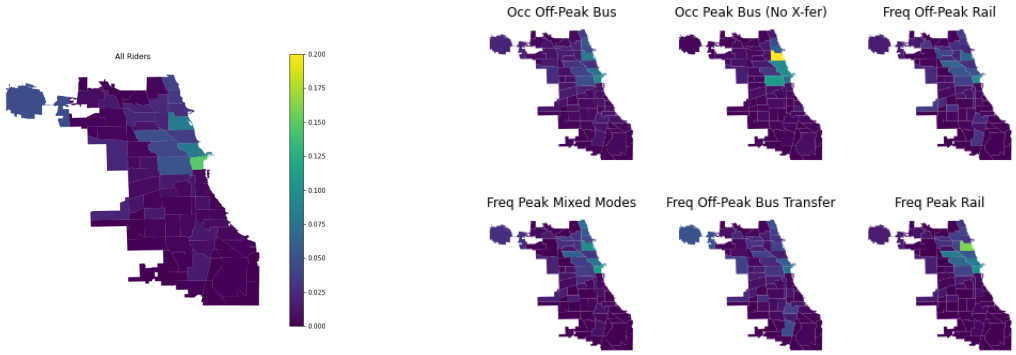


Figure 4-4: Inferred Home Locations for All Riders (Left) and by Cluster for Most Frequent Clusters (Right)

The Peak Rail group offers the other side to this story. These riders are heavily concentrated in the wealthiest and majority white areas of the city, and they exclusively use the system where and when its service levels are highest — on rail and in the peak hours. These travel patterns, coupled with demographics based on their inferred home locations, suggest that these riders opt for other modes when it comes to their non-commuting trips.

4.4 COVID-19's Impact on Ridership Behavior

The acceleration of the COVID-19 pandemic in America and the subsequent public health measures including the enactment of the stay-at-home order in Chicago led to an over 80% decrease in the number of CTA trips occurring during a typical week. This was closer to 90% for rail, and around 75% for bus. The weekday peak hours alone were responsible for about half of the lost trips. These statistics tell us a good deal about what types of trips were no longer considered essential, but they are not the complete story. Using the lens of the behavioral clusters, we can understand more about who in the city was making these essential trips, who was able to abandon public transit altogether, and what that means for the road to recovery. Most importantly, this knowledge can inform policies that will not only bring riders back onto the system but also make the system better than ever for the people who need it the most.

4.4.1 Ridership Churn

In this section, we aim to answer the question of who ceased riding public transit altogether during the pandemic (i.e., "churned").

Figure 4-5 gives a bar chart of the count of riders in each cluster pre-COVID, colored by whether they rode only during the early stage COVID period (Eventually Churned), only the late stage COVID period (Returned - July), both (Continual Rider), or neither (Completely Churned). Furthermore, Table 4 gives the percent of riders from each group riding during each of the COVID analysis phases. We note right away large variation in the percent of riders who completely churned from each group. Churn occurred in higher rates in clusters characterized by more infrequent or shorter term ridership. A glaring exception to this, however, is the Frequent Peak Rail cluster, whose riders completely churned at a rate of about 80%. The lowest complete churn rate belonged to the Frequent Off-Peak Bus Transfer group, which had only a third of its riders abandon the system altogether.

When looking at the system as a whole, we see a complete churn rate of 73%, with another 13% of riders not appearing in the early stage but riding in the late

stage, 10% riding in both, and the remaining 4% riding in the early stage but not the late stage. The clusters with churn rates significantly lower than the average are High Range Occasional Off-Peak Bus, High Range Frequent Off-Peak Rail, High Range Frequent Off-Peak Bus Transfer, and Free riders. It is notable that all four of these are characterized by High Range, Off-Peak travel. This further corroborates our finding that while off-peak hours see significantly fewer trips overall compared with peak hours, off-peak trips are taken by individuals who rely on the CTA for much of their travel and likely do not have other options for getting around. This is evident from their continued use of the service even during a global pandemic when use of public transit systems was discouraged.

The clusters with churn rates significantly lower than the average are Low Range Infrequent Weekend Rail, Low Range Infrequent Peak Bus (No Transfer), Low Range Infrequent Peak Rail, High Range Frequent Peak Rail, and One-Day riders. Again, we note that the unifying characteristic of these clusters, except for the High Range Frequent Rail cluster, is their infrequent usage of the system. This is unsurprising, as we expect visitors to the city to be captured within these groups, as well as people who use CTA one in awhile to supplement their primary travel modes, or for specific purposes. The fact that the High Frequency Peak Rail group churned at rates on par with the infrequent groups, and higher than some, suggests a fundamental difference between this group and the other high range or frequent groups of riders that goes beyond simply differing typical travel patterns. These individuals were almost entirely able to stop taking trips altogether or replace all transit trips with another mode.

4.4.2 Initial Ridership Recovery

All clusters saw an increase in the percent of their riders using the system between the early and late COVID analysis periods, as we would expect given that the city was under a stay-at-home order during the early phase but had begun phased re-opening of economic activities by June and July (Table 4.3). Among the Frequent rider clusters (clusters 7-11), all had returned at least a quarter of their riders to the system, except for the Peak Rail group, which remained at 17%, a rate more in line

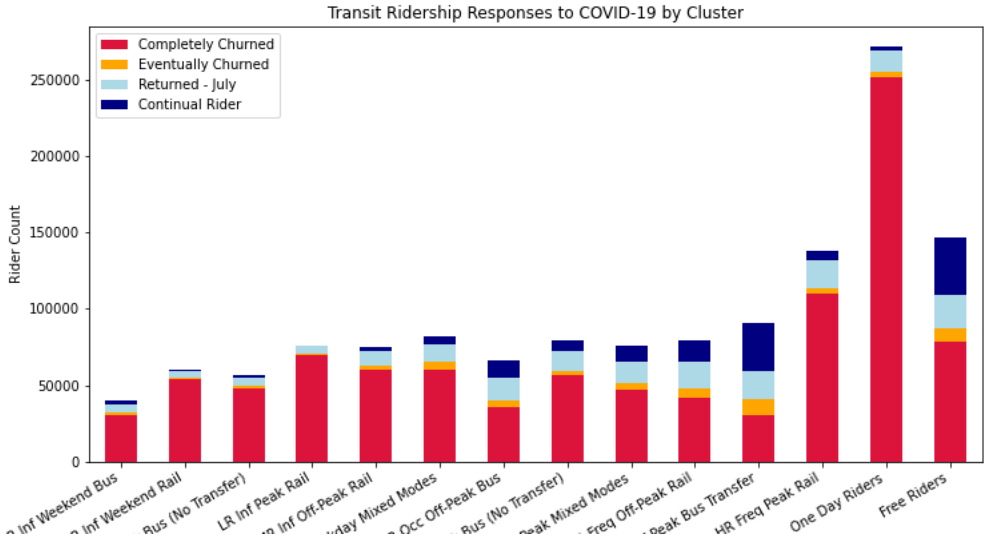


Figure 4-5: Number of Riders in Each Cluster Group by 2018 to 2017 Behavioral Shift

with some of the lowest frequency groups. These numbers help identify which groups will be most challenging to get back on transit. They also hint at which groups are responsible for the trips currently being taken on the CTA's system. In fact, during the early stage of the pandemic, just clusters 10 (High Range Frequent Off-Peak Bus Transfer) and 13 (Free Riders) accounted for over half of all trips on the system, with each accounting for about an equal proportion. In the late stage, their share has lessened somewhat as more other riders have returned. During this period, cluster 10 accounted for 19% of all trips and cluster 13 for 22%. Cluster 11, meanwhile, accounted for 4% of trips in the early stage and 5% in the late stage.

4.4.3 Bringing in Geographic, Pass, and Payment Information

Investigating the Free Rider group poses an opportunity to learn a little more about who has continued riding during the pandemic, as we can examine the pass makeup of this group before and after mid-March. We see that whereas the pre-COVID group of Free Riders was comprised of 24% Disabled Ride Free passes, 25% Senior Ride Free passes, and 48% University student passes, in the COVID period, this group was 49% Disabled Ride Free, 35% Senior Ride Free, and only 12% U-Pass holders. Again we

Cluster No.	Cluster Name	% Riding in COVID Early Stage	% Riding in COVID Late Stage	% Completely Churned
0	**LR Inf Weekend Bus**	10%	19%	76%
1	**LR Inf Weekend Rail**	4%	9%	89%
2	**LR Inf Peak Bus (No Transfer)**	6%	13%	84%
3	**LR Inf Peak Rail**	2%	7%	91%
4	**MR Inf Off-Peak Rail**	6%	17%	80%
5	**LR Occ Weekday Mixed Modes**	13%	20%	74%
6	**HR Occ Off-Peak Bus**	24%	39%	54%
7	**HR Freq Peak Bus (No Transfer)**	13%	25%	72%
8	**HR Freq Peak Mixed Modes**	20%	32%	62%
9	**HR Freq Off-Peak Rail**	25%	39%	53%
10	**HR Freq Off-Peak Bus Transfer**	47%	54%	33%
11	**HR Freq Peak Rail**	7%	17%	80%
12	**One-Day Riders**	2%	6%	93%
13	**Free Riders**	31%	41%	54%

Table 4.3: Percent of Riders from Each Cluster Active by COVID Analysis Period

see a clear trend of more disadvantaged riders being the ones who need to continue using the system during the pandemic.

This is borne out when examining the churn rate by community area of inferred home location. Riders living in the south and west parts of the city continued riding at rates of up to 40% by community area, while only about 10% of riders living along the north coast continued to ride. These basic geographic patterns hold regardless of the cluster, showing that not only do clusters that have more disadvantaged riders exhibit lower pandemic-related churn rates, within these clusters, the more disadvantaged riders are the ones more likely to continue riding.

Lastly, we also examine the change in the makeup of riders before and during COVID along some dimensions not included in the clusters, namely pass usage and history of paying with cash. We see that COVID riders are more likely to use a pass and to have paid with cash than pre-COVID riders.

4.5 Policy Implications

4.5.1 Universal Measures

First, we acknowledge that there are some policies which, during a pandemic, benefit riders and the system universally. These include the continuation of public health guidelines already in place, namely the requirement that all operators and riders wear masks, the regular cleaning of vehicles and enforcement of vehicle capacity caps, as well as the effective communication of these policies to all agency staff and riders. Additionally, as public health officials continue to advise maintaining several feet of distance between people, adding capacity where possible is an important goal to have, regardless of the population in mind. For rail, this means the addition of cars to trains that typically run with fewer than the maximum number of cars and the addition of trains where there is the signal and track capacity for added service. For bus, this means running more vehicles and improving the efficiency of service via dedicated bus lanes, queue jumps, and traffic signal priority. While these public health measures

and capacity increases are important for everyone, they should be seen as baseline needs rather than panacea policies. Absent other interventions crafted with specific populations in mind, they will likely not be enough to bring all the necessary riders and trips back to the system. To accomplish this, the path forward must consider policies tailored towards key groups of riders.

4.5.2 Targeted Measures

High Range, Frequent Peak Rail Riders

The first group of riders that we consider is the High Range, Frequent Peak Rail Riders. We have seen that these riders are concentrated along the north coast of the city in neighborhoods that are largely higher incomes and majority white. They nearly entirely abandoned the system once the pandemic hit and have yet to recover significant ridership during the initial reopening of the city. These facts suggest that these riders have the means to opt for non-transit modes and the flexibility to work from home. As the economy re-opens, this group will be difficult to get back on transit, as they typically only used the system at times and places where it was particularly crowded and impossible to maintain distance from fellow riders. These riders will likely be aware of the health risks associated with riding transit, and be tempted to choose another mode or continue working remotely if their employer allows it, as early evidence suggests many will continue to do [Akala, 2020].To get these riders back on transit will require an acknowledgement of their situation and creative thinking. This group uses the mobile Ventra app at particularly high rates, meaning that tech-based interventions may be particularly effective in reaching them. This fact can be leveraged; smart design and use of mobile notifications letting riders know what to expect and how to prepare for transit trips, may be able to make these riders feel comfortable returning to the system. Additionally, accurate information about the crowding level of trains, or even specific rail cars, communicated via the Ventra app would likely help bring back riders in this group. Particularly effective would be prediction of crowding levels at their station of origin sufficiently far enough

in advance, so that they could plan a trip before leaving their home. If they feel that they have the proper information to make smart choices about how and when to use public transit in a way that makes them feel safe, they are more likely to do so than if they feel they are taking a big risk each time.

We also know that these High Range, Frequent Peak Rail Riders very rarely ride the bus, despite the fact that many live in areas that are well-served by bus. It is quite possible that these riders are unaware of bus routes that would serve them just as well as rail and would be open to using them if they felt it was a safer (less crowded) option during the pandemic. The CTA could inform residents who typically use rail of these alternate routes, using posters or announcements at rail stations or via targeted app notifications based on riders' travel history and inferred origins and destinations of their historical trips. These interventions would be most effective if combined with some of the other interventions already mentioned, such as dedicated bus lanes and more buses to increase capacity on those routes, as well as accurate information on the crowding level of buses and trains.

High Range, Frequent, Off-Peak Bus Transfer

When considering the High Range, Frequent Off-Peak Bus Transfer group, however, the objectives, challenges, and opportunities are somewhat different. This group did not abandon transit like the peak rail riders, suggesting a deeper reliance on the system for their travel needs. This is likely a group that largely overlaps with what has often been referred to in the literature as "captive riders." When defining policies aimed at ridership recovery, one might be tempted to ignore this group, as they will have few other options for how to make trips, and are likely to return without much enticement. But this ignores the fact that, as this analysis has revealed, this group — which was responsible for 15% of pre-COVID trips, a proportion smaller than only the Frequent Peak Rail riders and the Free riders—typically uses the aspects of the system associated with lower levels of service. They are reliant on buses that run less frequently at the times when they need them and often at low speeds [Wisniewski, a]. Furthermore, they regularly need to transfer between two such buses.

The CTA and the city of Chicago had already begun significant work to speed up their buses [Wisniewski, b] but the CTA's ability to increase service frequency is limited by laws such as an antiquated farebox recovery mandate, currently being fought by activists [Whitehead, 2020]. Despite obstacles, the CTA should make sure to specifically consider this group of riders when prioritizing investment to the system, doing what they can to more fully orient the system around the mobility needs of its riders. Furthermore, research has shown that joblessness resulting from the pandemic has hit lower income, minority communities the hardest [Mohammadian et al., 2020] and this analysis has shown that these riders are disproportionately located in such communities. More so than in the peak rail group, the loss of riders within this group may be attributed to a loss of jobs and therefore trip purposes. The re-employment of these groups is key to an equitable economic rebound for the city, and thus, better connecting these riders to jobs is a practical aim for the city of Chicago and the CTA.

In the short term, making sure these riders have access to Ventra cards is an important step. This group purchases and refills Ventra cards using cash and from vendors at higher rates that the average rider, meaning that during a global pandemic when many vendor shops are closed, their access to Ventra tickets may be cut off. Working with local businesses to distribute Ventra cards, or stocking them on buses, could help get them to the riders who need them. Furthermore, the CTA could make it cheaper for these riders to use the system, at least for the time being. Many of those still riding regularly are essential workers. Eliminating the transfer fee and discounting 7-Day passes, which are used by this group at higher than average rates, would help ease the financial burden on disadvantaged riders already hit hardest by the pandemic.

Additionally, better bus service is particularly important for this group. Many of these riders live in areas of the city not served by rail, leaving bus as the only option. Leveraging dedicated bus lanes, traffic signal priority, and additional vehicles to increase the efficiency of these routes would have a compounding effect of improving service for this group, as it would improve not only each leg of their bus travel, but decrease transfer times as well.

Lastly, the travel patterns of this group should be explicitly considered when determining where to prioritize bus infrastructure improvements. We have seen that these riders don't necessarily travel when overall volume is highest, and thus are at risk of being overlooked when ridership analysis is done at the trip level only and service improvements prioritized accordingly. This pandemic, and what it has revealed about who truly keeps Chicago running, should lead to more explicit consideration of the needs of these riders when designing policies and system investments.

4.6 Conclusion

Analyzing the differential impacts on transit ridership by key rider groups, as defined by pre-pandemic behavior, reveals significant heterogeneity in how Chicago transit riders changed their use of the system in response to COVID-19. Notably, frequent peak rail riders stopped riding the CTA altogether at rates on par with some of the lowest frequency pre-pandemic riders, while nearly half of regular off-peak bus riders with frequent transfers continued to ride the system, accounting for 20% of trips in July. While individuals' travel needs are likely to change in dramatic ways going forward, knowledge of how riders previously used the system can provide valuable insight into the distinct challenges facing different groups, and this should inform policies aimed at helping transit agencies recover ridership. In the case of the CTA, targeted policies at the two groups mentioned above will be more effective than only pursuing broad tactics for welcoming riders back to the system.

COVID-19 has affected transit agencies in unprecedented ways, and as such, there is no clear roadmap for recovery. As transit agencies develop and implement strategies, rather than taking for granted the riders that have continued to use the system even throughout the pandemic, they must ask what this says about their system and who it prioritizes. Those who continue using the CTA's system at the highest rates during the pandemic were much more likely to be disadvantaged riders. Any path forward must use this knowledge to aid these riders in improving the level of service they receive. This should be a focus not only of pandemic-time policies, rec-

ognizing that these are the essential workers helping to keep cities running, but also of continuing policies that focus on recovery and beyond. This will require support from lawmakers, who must recognize the limitations of current public transportation funding mechanisms and revise them in ways that acknowledge the crucial role public transit has to play in our societies.

Across the country, the COVID crisis has laid bare the fact that those most reliant on public transit are too often those who are not always provided the highest levels of service. A failure to consider the people making transit trips during such a critical time along with their distinct challenges and situations, will lead to a recovery plans that are short-sighted at best and harmful at worst. The CTA benefits from having a fare card system such as Ventra, which allows the individual pass holders to be used as a fundamental unit of analysis. The approach used in this work should be adopted by agencies and cities who have the requisite data to guide policy formation during this critical time.

Chapter 5

Determining Factors Related to COVID-19 Transit Ridership: A Linear and Spatial Regression Approach

The previous chapter provided an in-depth example for how to apply a clustering framework, rooted in the desire to keep riders as the focal point of analysis, to understand a major shock to the public transit system and guide policy evaluation. The findings indicated that individuals whose typical ridership behavior consists of predominantly bus trips, with transfers, taken during off-peak hours were significantly more likely to continue using the CTA during the stay-at-home order than individuals whose ridership is largely limited to rail trips at peak hours. A geographical analysis also suggested that those in the former group had inferred home locations in the South and West parts of the cities at far higher rates than the latter group, indicating that these riders are more likely to be Black and Hispanic, as well as lower income. These findings were consistent with what reports on transit ridership during this period from across America have shown— that much of the remaining transit ridership is from people serving as essential workers, who tend to be lower income

and non-white.

Having uncovered a clear relationship between pre-COVID ridership behavior, sociodemographics, and ridership behavior during the pandemic, in this chapter, we aim to tease out the relative importance of individual variables in predicting COVID-related ridership loss. Specifically, we employ linear regression techniques, using the percent change in average weekly trips at the census tract level as the dependent variable. Our main explanatory variables of interest here are demographics and pre-COVID ridership behavior of residents, which can be aggregated to the census tract and transit stop respectively. Because the latter nests within the former, we simply choose census tracts as the unit of analysis.

The goal of this analysis is to demonstrate first the benefit of including typical ridership characteristics of an area along with sociodemographics in explaining the change in trip counts observed after the stay-at-home order was issued, using the baseline and early COVID stage from the previous chapter. The second goal is to illustrate how a spatial regression approach can be used in this analysis to take into account the spatial autocorrelation present in the data.

5.1 Background

Estimation of transportation demand at the level of a spatial unit, for example a city, neighborhood, or station, is one objective of a large family of models often called "direct demand models," which typically rely on linear regression techniques. They gained prominence in the public transportation realm in part as a response to the industry standard four-step model's failure to capture or consider neighborhood-level characteristics, such as walkability and density, and their impacts on transit ridership [Cervero, 2007]. They grew in popularity due to their relative simplicity, compared with the four-step model or discrete choice models, in terms of implementation and interpretation, and have been used numerous times since in contexts such as determining drivers of BRT ridership in Los Angeles [Cervero et al., 2010] and estimating the role of TOD on rail ridership in Taipei [Lin and Shin, 2008]. The work in this

chapter inherits from this body of work, as it investigates ridership at a spatial unit and uses attributes of that space as the dependent variables. Unlike direct demand models, however, this work is interested in the percent change of ridership due to a particular event — the implementation of the stay-at-home order in Chicago — instead of the absolute volume of transit ridership. As such, variables typically included in direct demand models of the type discussed above, such as physical qualities of the neighborhoods, are excluded from this analysis due to the fact that, while they have been shown to impact transit ridership in general, they are unlikely to impact the magnitude of the transit ridership response to a global pandemic.

In addition to land use characteristics, sociodemographics have proven to be predictive of transit ridership in a wide variety of studies, and, as suggested by the previous chapter, are likely to play a role in explaining which groups were more likely to continue using public transit in Chicago during the pandemic. While the demographic traits in and of themselves, such as primary language, for example, do not impact travel behavior, these aspects of identity are closely associated with variables that are harder to capture, such as type or location of job, working hours, and parental obligation [Lu and Pas, 1999]. Therefore, the sociodemographics of an area have long been considered an important component of understanding travel demand, especially due to the wide availability of such data.

Demographic data has proven to be predictive in both cross-sectional studies of transit ridership and in studies that have modeled changes in transit ridership over time. Dill et al investigated stop-level bus and rail ridership in three Oregon cities and conclude that being white and college educated corresponds with less transit ridership [Dill, 2013]. Mucci and Erhardt find that high incomes are associated with lower ridership in San Francisco [Mucci and Erhardt, 2018], and Pasha et al. have a similar finding in Calgary [Pasha et al., 2016]. Giuliano finds that African-Americans use transit at higher rates, though mainly via their lower levels of access to vehicles [Giuliano, 2005]. Studies have also explored the role of demographics in the ridership decreases that occurred in the second part of the 2010s. Manville et al. found that increased vehicle ownership was the primary determinant of declining transit ridership

in Southern California [Michael Manville et al., 2018], while Berrebi et al. found an increased percentage of white residents in the vicinity of a bus stop corresponded with a reduction in ridership at that stop [Berrebi and Watkins, 2020]. When investigating the correlations between demographics and ridership at a fixed point in time, Berrebi et al. found that high proportions of non-white, carless, and high school educated riders were associated with higher levels of transit ridership. Looking at ridership changes across 14 years in 25 cities, Boisjoly et al. also find car ownership to be a primary driver of transit ridership loss [Boisjoly et al., 2018].

While there is significant precedent for using demographics to explain transit ridership by spatial unit— at one point in time as well as changes to ridership— I could find no examples of the ridership behavior typical of residents of an area being used as explanatory variables for changes in ridership. There are examples of habitual ridership behavior being included in mode choice models for emerging modes using stated preference data [Asgari and Jin, 2020], but in terms of modeling ridership changes or future demand at a set of locations, details on existing ridership behavior are absent.

It would of course be unnecessary to use transit ridership behavior as explanatory variables in a model of current trip volumes, but the question of whether the ridership behaviors typical of an area can tell us something about the trajectory of ridership trends seems to be a question that would be of interest to transit agencies. It is possible that this has not yet been studied because, when considering ridership changes over a longer period of time, one must consider the migration of urban residents, and thus the establishment of a baseline behavioral profile for each area may be less meaningful for changes studied over years, by which time the baseline population may have changed substantially. In our case, however, we are concerned with the behavioral response to an event that was extreme and abrupt. We can safely assume that people did not move to another location in the city within the one-week transition period separating our baseline period and early COVID stage as outlined in CHapter 4. Furthermore, in the previous chapter we demonstrated the strong relationship between behavior and ridership response to the COVID pandemic. By

including baseline ridership behavior alongside demographic variables, we can deepen our understanding of the dynamics at play by teasing out which individual variables prove to be most predictive of ridership response while controlling for all other variables. In doing so, we can understand if the patterns we saw in Chapter 4 were just an artifact of the correlation between certain behaviors and a set of demographic traits, or if each independently has predictive power in the question of who continued to use transit during the pandemic.

Models predicting ridership or ridership changes at the level of a spatial unit often employ linear regression techniques, as mentioned before. One assumption of linear regression models is the independence of error terms. Spatial correlation in OLS residuals can be indicative of an omitted spatially lagged explanatory variable, in which case the OLS estimates will be biased, failure to account for spatial correlation in the error structure, in which case the OLS standard errors will be wrong, or both, in which case the model will have both issues [Anselin, 1988b]. This concern is often not addressed in the transportation literature, perhaps in the hopes that the demographic data or other data associated with each location will account for all spatial variation in the data and result in residuals that are, in fact, random in space. Despite this being easy to check by investigating the spatial distribution of the OLS model residuals, this part is often skipped in transportation demand literature that employs linear regression. The second part of this analysis concerns the exploration of spatial dependencies in our model and the appropriateness of the spatial lag model in particular.

In recent years, some studies have begun to explore the role of space more explicitly when modeling transit demand. Gan et al. compared an OLS model estimating rapid transit ridership in Nanjing to a spatial lag model, a spatial error model, and a geographic weighted regression model using the same data, and found that all the spatial models fit the data better than the OLS model [Gan et al., 2019]. In addition, Chow et al., Cardozo et al., Zhao et al., and Ma et al. have all used geographically weighted regression or, in the latter case, geographic and temporally weighted regression, to explore the ways in which coefficients on explanatory variables

may vary as a function of space [Chow et al., 2006, Cardozo et al., 2012, Zhao et al., 2013, Ma et al., 2018]. As geographically weighted regression is more controversial than spatial lag or spatial error models in the research community, we limit our consideration in this chapter to the latter two [Chi and Zhu, 2020].

5.2 Data

The data used in this section comes from two sources: The CTA's Ventra database and the 2013-2018 American Community Survey [United States Census Bureau, 2020]. The former was used to calculate the average weekly public transit trips before and after the stay-at-home order, using the same time frames for baseline and COVID analysis as seen in Chapter 4. The Ventra database was also used to calculate typical ridership features for each Ventra card in the baseline period, and match each Ventra card to an inferred home location, defined, as in Chapters 3 and 4, as the stop most often used for the first trip of the day. Based on this stop assignment, each card in the baseline period is matched to a "home" census tract, and the ridership features for these cards are averaged to summarize the typical ridership behavior of riders associated with each census tract. While the use of the data is different in this chapter compared with previous chapters — the cards are not clustered at all in this analysis — the data used for this portion of the analysis is the same as the data used in Chapter 4.

We also use information on the location of rail stations to indicate if the census tract contains a stop on a rail line. The locations of rail stops were also obtained from the Ventra database, which holds information on the location of all stops. Lastly, we employ dummy variables indicating the membership of each census tract to one of nine regions in the city of Chicago. We drew the definitions of the regions from delineations used sometimes in the real estate market [The Chicago 77, 2008]. These region definitions are useful because each of the 77 community areas belong to exactly one region, and each census tract belongs to exactly one community area. The regions are shows in Figure 5-1.

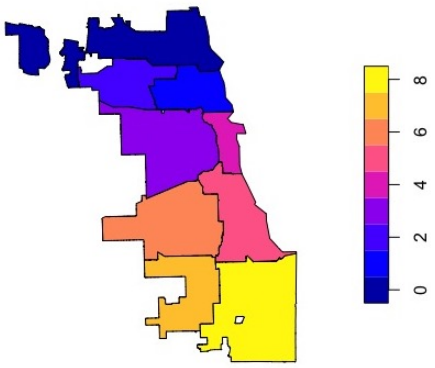
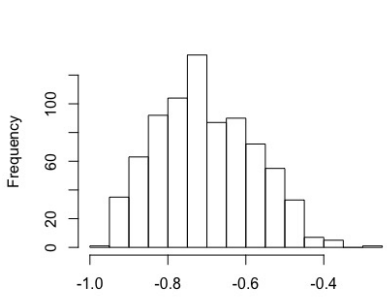


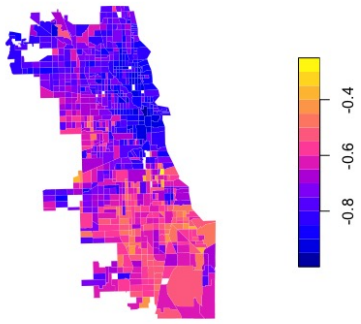
Figure 5-1: Chicago Regions

5.3 Descriptive Statistics

The unit of analysis for this section is the census tract. There are 801 census tracts in the city of Chicago. After removing tracts that have no public transit stops or are missing data for any of our attributes, we were left with 779 tracts on which to perform are analysis. Figure 5-2 shows a histogram of the percent changes by tract, as well as the map of values for all tracts. We see that the distribution of percent changes is roughly normally distributed, justifying our use of linear regression to model this as the dependent variable. When looking at the map, however, we note clear spatial patterns in the value of the dependent variable, which motivates our exploration of spatial models for this problem.



(a) Histogram of Tract-level Percent Changes



(b) Percent Change by Tract

Figure 5-2: Percent Change in Average Weekly Trip Volume by Tract after Stay-at-Home Order

Regarding explanatory variables, as mentioned above, these analyses use two main

categories of explanatory variables: sociodemographics and baseline ridership behaviors, along with dummies indicating the presence of a rail station and the region of the city. We saw in Chapter 4 indications that ridership behavior was not independent of demographics. Notably, one of the behavior clusters that we focused our analysis around — frequent off-peak bus riders — seemed to contain a disproportionate number of riders from areas of the city with lower incomes and higher minority populations. Entering our ridership attributes alongside demographic information in a linear regression model will enable us to determine the extent to which each of the factors is significant while controlling for the others. In other words, while the clustering approach in Chapter 4 enabled us to tell a story about the different groups that constitute CTA's ridership and their distinct needs and challenges during and as ridership recovers, this approach will enable us to quantify the relevance of the various sociodemographic and ridership attributes in explaining the ridership dropoff after the stay-at-home order.

We first explore the correlation levels among our potential variables of interest. While it is accepted practice and indeed, even the goal, to include variables which are correlated with one another in multiple linear regression models so as to determine the unique explanatory contributions of each and avoid omitted variable bias, it is useful to explore extreme correlations to determine pairs of variables that may cause multicolinearity issues. If we decide to include variables that are highly correlated, it is important that we feel they are measuring different things. The correlation heatmap is shown in Figure 5-3

First, we note that correlations are stronger between demographic variables and between behavioral variables than across these two groups in general. A few values stick out as being particularly high in magnitude: the correlation between the percent of black residents and the percent of white residents is -0.92, the correlation between average weekly rides and range is 0.87, and the correlation between the presence of a rail station and the percent of rides typically taken on bus is -0.97.

Regarding the first, because the correlation is so strong and they are both capturing the racial makeup of the tract, we opt to include only the percent of black

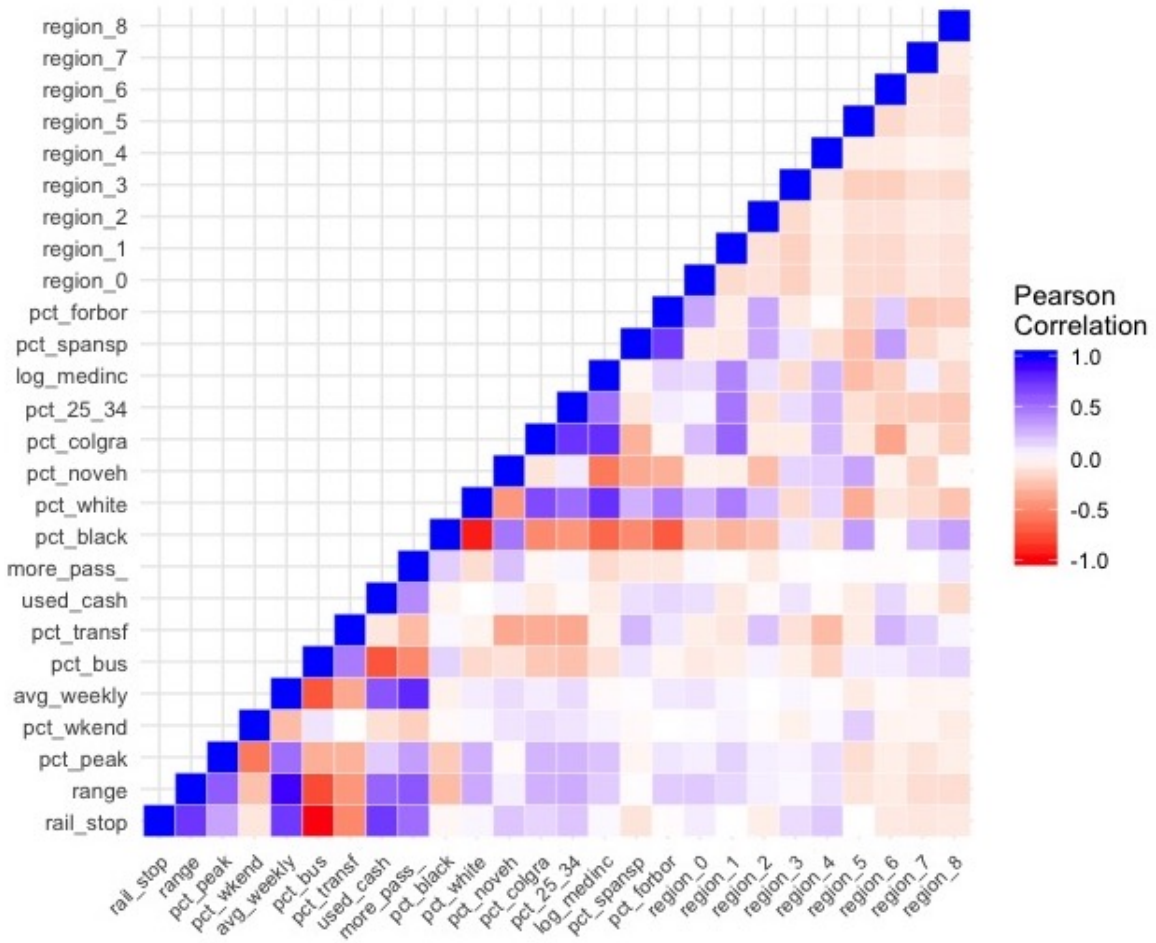


Figure 5-3: Pearson Correlations Among Explanatory Variables

residents as an explanatory variable. For the second, we opt to include only average weekly rides, as regularity of use is the behavioral attribute of more interest and less subject to arbitrary values based on the definition of the study period. The final high correlation value poses a particular problem, in that these values should, in theory, measure very different things, and we would like to be able to control for the presence of a rail station when evaluating the importance of the typical percent of rides taken on bus among residents. The fact that the negative correlation is so strong is informative in itself, however, suggesting that people whose typical first ride of the day occurs on a rail station almost exclusively ride rail rather than bus. We opt to include only the pct_bus explanatory variable and, when interpreting our results, keep in mind that high values of this variable are strongly associated with the lack of

a rail station in that census tract.

Table 5.1 gives the final set of independent variables included in the regressions along with their descriptions.

Category	**Variable Name**	**Description**
Behavior	pct_peak	Average share among riders of rides taken during peak hours
	pct_wkend	Average share among riders of rides taken on the weekend
	avg_wkly_rides	Mean value among riders of average weekly rides
	pct_bus	Average share among riders of rides taken on bus
	pct_transfer	Average share among riders of rides involving a transfer
	used_cash	Percent of riders who used cash for a ticket or pass transaction during the baseline period
	pct_pass	Percent of riders who spent more money on pass products than pay per use rides during the baseline period
Demographic	pct_black	Percent of residents who are black only
	pct_colgrad	Percent of residents with a college degree
	pct_25_34	Percent of residents between the ages of 25 and 34
	log_medinc	Logged median household income
	pct_speakspan	Percent of residents speaking Spanish at home
	pct_forborn	Percent of residents that are foreign born
	pct_noveh	Percent of households without a vehicle
Other	region_X	Boolean equal to 1 if the tract is in Region 1 through 8, leaving 0 as the base

*Behavior variables take the average value of the variable among all riders with inferred home locations in the given tract.

Table 5.1: Independent Variable Descriptions

5.4 OLS Regressions

5.4.1 Model Formulation

In the first part of this analysis, we ignore any issues that may arise from spatial autocorrelation in our data, and instead run traditional OLS regression models. Here, we are assuming that the census tracts represent independent observations, where the values of our variables in one census tract exert no influence on the dependent variable in a nearby tract, and there is no correlation in the error terms across tracts.

The models can be formulated generally as follows:

$$PctChange_j = \alpha R_j + \beta X_j + \gamma Z_j + \epsilon_j$$

where $PctChange_j$ is the percent change in average weekly trips observed between the baseline period and the early analysis period in census tract j, R_j is the set of region dummies, X_j is a vector of sociodemographics associated with census tract j, Z_j is a vector of average ridership behavior characteristics among CTA riders in census tract j, and ϵ_j is a normally distributed error term. In our first model, we include only the region dummies, restricting $\beta = \gamma = 0$. In the second model we keep the regional dummies and investigate the impact of sociodemographics only on ridership change, restricting $\gamma = 0$. In our third model, we investigate the impact of typical ridership behavior only, including the region dummies and restricting $\beta = 0$. Our final model allows α, γ, and β to be nonzero.

Aside from the assumption that our error terms are independent and identically distributed, which we will address later, we are also ignoring the fact that that our dependent variable is limited. Because we are modeling the percent change, the dependent variable cannot, in reality, assume a value below -1. While this could lead to some predicted values that are infeasible, since we are not concerned with prediction accuracy but rather capturing the relationship among variables, we set this issue aside. Furthermore, despite being limited, the distribution of the dependent variable does appear to be approximately normal as shown in Figure 5-2a, rather than

having many variables clustered around -1, which reassures us that a true relationship will be captured by the OLS model.

5.4.2 Results

Table 5.1 gives the results from the four regressions described above. We note that even the regression containing only the regional dummies explains about half of the variation in the dependent variable, confirming the strength of the geographical patterns observed in the reaction to COVID.

In the second regression, which controls for region and examines sociodemographic characteristics of census tracts as explanatory variables for ridership loss due to COVID, we see that the percent of black residents and percent of residents who speak Spanish at home both have a positive impact on ridership change, meaning that higher values of those variables are associated with smaller (less negative) declines in trip numbers. The percent of residents between the ages of 25 and 34 and the percent of foreign born residents both negatively impact the change in trip volume, with younger tracts and tracts with larger immigrant populations seeing a steeper decline in ridership. This may suggest that, when controlling for Spanish speakers, more foreign residents were more able to stop traveling or use other modes after the stay-at-home order. Lastly, the percent of college educated residents, the logged median income of residents, and the percent of households without access to a vehicle all have impacts on ridership changes that are indistinguishable from zero. This is surprising, as we would have expected these variables to explain one's ability to work from home or use other modes during the pandemic.

Regression 3 also maintains the region dummies but considers only average ridership characteristics as explanatory variables. As described above, the values for the variables associated with each census tract come from the average among all riders with that tract as an inferred home census tract. Our method for inference leaves room for some error, especially in the case of infrequent riders. As a result, it is likely that infrequent riders, including tourists, are largely assigned to tracts that see a lot of transit volume typically, such as tracts in and around the Loop.

We see that riders' share of trips taken on bus and share of trips that involve a transfer, as well as the percent of riders that use cash and the percent that spent more money on a pass than on pay per use rides are all associated with smaller ridership declines. This largely fits with what we saw in Chapter 4: the group of riders most likely to continue riding in COVID were those whose ridership patterns were typified by frequent bus trips with high transfer rates. We note that the percent of rides taken at both peak times and weekends by riders are associated with larger drops in the number of trips for a census tract. The former is also likely explained in part by the near-complete abandonment of the system by peak rail riders, who also tend to be geographically concentrated. The latter is likely due to the fact that a high percentage of trips of the weekend corresponds to tourists or other leisure riders who are likely to drop off after a stay-at-home order. Lastly, and most surprisingly, the mean value of riders' average weekly trips by census tract is not significant in the model, suggesting that how frequently riders in an area typically used transit was not predictive of how much ridership dropped during the early COVID stage. This is at odds with our finding in the previous Chapter that in general, clusters characterized by lower frequencies saw more churn than those characterized by high usage, with the exception of the Peak Rail group. This may be due to the fact that the distribution of values for typical average weekly rides by census tract skews toward the low end with census tracts with larger values tending to be located near rail lines. Perhaps, after controlling for modal split, this variable, at least as aggregated here to the census tract, was not predictive of ridership changes.

Finally, we examine the results when both sets of explanatory variables are included together. We note that the percent of rides typically taken on a weekend becomes insignificant and the percent of riders using cash becomes only marginally significant. On the other hand, the percent of households without a vehicle becomes significantly predictive of a smaller drop in ridership, as we would expect.

The significant demographic attributes include vehicle ownership, the percent of black residents and the percent of Spanish speakers, all associated with lower ridership drops, and the percent of residents between the ages of 25 and 34 and the share of

foreign born residents, which both predict larger drops. In terms of ridership behavior attributes, percent of trips taken at peak remains the only significant predictor of larger drops in ridership. The percent of trips taken on bus, percent of trips involving a transfer, and percent of riders spending more money on passes than pay per use are predictive of smaller drops in ridership.

We can also examine the change in the coefficient estimates associated with the Region dummies as each set of explanatory variables was added. These are given in Table 5.3. We see that regions one through 4 are associated with steeper drops in ridership, while regions 5-8, which are located in the Southern half of the city, are associated with smaller drops in ridership. This is consistent with Figure 5-2. The fact that many of the regional dummies remained significant motivates further explorations of spatial dependency in the data.

5.4.3 Conclusion

The sustained significance of most sociodemographic and ridership behavior attributes when combined into a single model, along with the increase in the adjusted R^2 value suggests that it is worth exploring including both groups of explanatory variables together when seeking to understand the impact of COVID on transit ridership. It also implies that the type of transit ridership that is typical in an area is worthwhile to include in analyses seeking to understand ridership changes, though issues with multicolinearity must be considered. It is possible that some of our non-intuitive results, such as lack of significance on the part of income and typical ride frequency, may be explained by their relationships to other variables in the model. Further work on this front should explore other ways of assigning behavioral attributes to census tracts, for example, as the one employed here is simplistic and may over-assign infrequent riders to rail stations, for example.

Regardless, some variables stand out as being clearly predictive of transit ridership after the stay-at-home order. If we view continued riding during COVID as a rough proxy for transit reliance, this study is illuminating because it reveals that, even when controlling for other factors, a high percentage of peak travel is indicative of

	(1)	(2)	(3)	(4)
Region Dummies	Yes	Yes	Yes	Yes
pct_black		0.144***		0.125***
		(0.020)		(0.018)
pct_colgra		−0.081		0.002
		(0.054)		(0.049)
pct_25_34		−0.149***		−0.107**
		(0.050)		(0.044)
log_medinc		−0.0004		0.001
		(0.014)		(0.012)
pct_spansp		0.122***		0.102***
		(0.023)		(0.021)
pct_forbor		−0.109***		−0.081**
		(0.037)		(0.033)
pct_noveh		−0.002		0.066**
		(0.034)		(0.032)
pct_peak			−0.382***	−0.227***
			(0.053)	(0.047)
pct_wkend			−0.202***	−0.094
			(0.065)	(0.057)
avg_weekly			−0.009	−0.003
			(0.007)	(0.006)
pct_bus			0.167***	0.104***
			(0.026)	(0.024)
pct_transf			0.212***	0.255***
			(0.037)	(0.034)
used_cash			0.148***	0.067*
			(0.045)	(0.039)
more_pass_			0.720***	0.393***
			(0.098)	(0.088)
Constant	−0.767***	−0.709***	−0.835***	−0.865***
	(0.009)	(0.155)	(0.050)	(0.143)
Adjusted R^2	0.459	0.630	0.608	0.709
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5.2: OLS Regression Results on Percent Change in Average Weekly Trips

	(1)	(2)	(3)	(4)
region_1	−0.067***	−0.047***	−0.048***	−0.048***
	(0.013)	(0.013)	(0.011)	(0.012)
region_2	0.014	−0.025*	−0.003	−0.023*
	(0.014)	(0.013)	(0.013)	(0.012)
region_3	0.051***	−0.032***	0.049***	−0.015
	(0.012)	(0.012)	(0.010)	(0.011)
region_4	−0.088***	−0.065***	−0.015	−0.026
	(0.019)	(0.019)	(0.018)	(0.017)
region_5	0.138***	0.029**	0.120***	0.034***
	(0.013)	(0.012)	(0.011)	(0.011)
region_6	0.128***	0.026*	0.086***	0.024**
	(0.012)	(0.013)	(0.011)	(0.012)
region_7	0.172***	0.050***	0.117***	0.043***
	(0.016)	(0.016)	(0.014)	(0.014)
region_8	0.183***	0.035**	0.139***	0.042***
	(0.014)	(0.015)	(0.013)	(0.014)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5.3: Region Dummies for OLS Regression

less reliance on this system. This underscores our conclusions from the Chapter 4, where we determined that systems for allocating transit funds that focus only on the highest absolute volume of ridership will systematically miss the riders that need the system the most.

5.5 Spatial Regression

Based on Figure 5-2, we can see clear spatial patterns to the data, meaning that it is likely that classic OLS assumptions are violated, specifically the assumption that error terms are independent of one another. We can check this directly by mapping the residuals of the OLS regression and using Moran's I statistic to test for spatial autocorrelation.

For the remaining portion of this analysis, we will be comparing the OLS model to models that take into account spatial dependence. The spatial models require

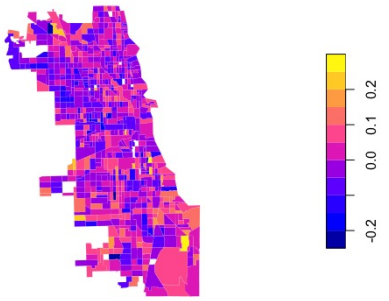
the specification of spatial weights matrix, which specifies how the "neighborhood" of a census tract might be defined. In other words, it dictates what relationship a census tract has to have to another tract to exert an influence on it. We will use a Queen weighting mechanism, which says that any tract sharing an edge or vertex with another tract is considered a neighbor, and each neighbor for a given census tract is weighted equally. This form of weighting matrix is very common, though future work may explore other varieties, such as one that counts tracts as neighbors if they lie on the same transit line.

5.5.1 OLS Residual Analysis

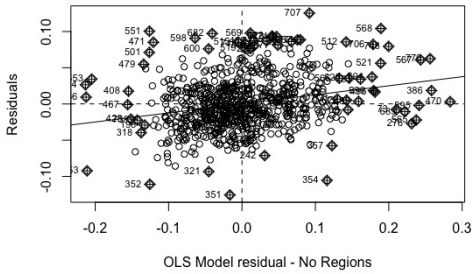
As mentioned before, the maintained significance of the regional dummies in the OLS regression suggests spatial dependence in the model not explained away by the demographic variables alone. We take this as one motivation for exploring spatial regression models. Additionally, we explore the level of spatial autocorrelation in the residuals from two OLS models: The one presented above in column 4 of Table 4.2 and the same regression but without the regional dummies. Because we would not include regional dummies in our spatial models, we want to have a baseline comparison for spatial autocorrelation of residuals when spatial regression is employed. This will allow us to gauge how much better our spatial regression model is at eliminating spatial autocorrelation in the residuals than simply including regional dummies in the OLS model.

Figure 5-4 shows the map of residuals from the OLS model without regional dummies, along with a graph which plots each residual by the weighted combination of the neighboring residuals. While spatial clustering of residuals is not very evident from the map, the plot shows the small but positive correlation between a tract's residual and that of its neighbors. Finally, we also calculate Moran's I statistic for the map, which is 0.13, and test it against the null hypothesis that the data is randomly distributed across space. We find our test to soundly reject the null with a p-value of 2.76×10^{-10}.

Turning to the model with regional dummies, we see that adding the dummies



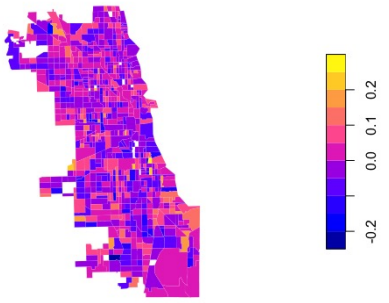
(a) OLS Residual by Census Tract



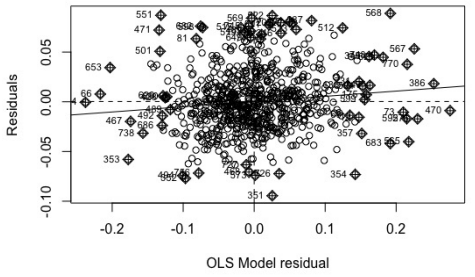
(b) OLS Residual vs. Spatially Lagged OLS Residuals

Figure 5-4: Residual Analysis for OLS Model with No Region Dummies

has significantly reduced but not eliminated the spatial autocorrelation of the residuals. Figure 5-5 shows the residual map and Moran plot for the model with regional dummies. Calculating the Moran's I statistic, we find that it is 0.05 but that a statistical test still rejects the null hypothesis of no autocorrelation in the residuals with a p-value of 0.004. Thus, we accept the OLS model with the regional dummies as a good approximation of a model with independent error terms, but we turn to spatial regression models to see if we can improve on this.



(a) OLS Residual by Census Tract



(b) OLS residual vs. Spatially Lagged OLS Residuals

Figure 5-5: Residual Analysis for OLS Model with Region Dummies

5.5.2 Spatial Lag vs. Spatial Error Model

Spatial Lag Model

The two most popular spatial regression models are the spatial lag model and the spatial error model. The spatial lag model in our case would take the following form:

$$y = \rho W y + X\beta + Z\gamma + \epsilon$$

where y is the dependent variable of interest, in this case percent change of average weekly trips, W is the spatial weights matrix, ρ is the spatial autoregressive parameter, to be estiamted, X is an $n \times k$ matrix of k behavioral explanatory variables for n observations, β is a $k \times 1$ vector of the parameters to be estimated for each behavioral explanatory variable, Z is an $n \times m$ matrix of m demographic explanatory variables for n observations, and γ is a $m \times 1$ vector of the parameters to be estimated for each demographic explanatory variable. Lastly, ϵ is a vector of random error terms.

The inclusion of the spatially lagged dependent variable in the model implies a relationship between the explanatory variables of neighboring tracts and the dependent variable of the tract of interest, because the lagged dependent variables can instead be written as the linear combination of their explanatory variables and their lagged dependent variables. This makes particular intuitive sense in our case, specifically for demographic explanatory variables, since it is quite possible that someone we have assigned to a given census tract due to their use of a transit station in that tract may very well actually reside in a neighboring tract. Thus, relating the demographics of neighboring tracts to ridership in a given tract makes sense for our purposes.

This also makes sense for the behavioral variables, as transit riders may occasionally use other transit stops around their home other than their most frequent first origin stop, thus impacting the change in ridership in neighboring tracts.

Spatial Error Model

Spatial error models, on the other hand, posit that spatial correlation in the OLS residuals is not due to the omission of spatially lagged dependent variables, but rather due to some unobserved factor leading to residuals that demonstrate spatial patterns. The model can be written as

$$Y = X\beta + Z\gamma + u$$

$$u = \lambda W u + \epsilon$$

where Y, X, W, Z, γ, and β are as before, and the error term is decomposed into a random component, ϵ, and a weighted combination of neighboring error terms multiplied by some error autocorrelation parameter λ, which is estimated by the model.

For our purposes, it is logical to test a spatial error model as well as a spatial lag model. In spatial econometrics, in the presence of clear spatial patterns to the data, it is common to test both models to ascertain whether the patters can be explained by spatial relationships among variables for which we have data (spatial lag model) or variables for which we do not have data (spatial error model).

Comparison

There are a couple ways to compare which spatial regression structure is a better fit for our data. The first is to use the log-likelihood or the Akaike Information Criteria (AIC). In our case, to do both would be redundant, as the AIC is a measurement of the log-likelihood that adjusts for the number of parameters. Specifically, the formula for AIC is given as follows:

$$AIC = 2k - 2\ln(\hat{L})$$

where k is the number of parameters estimated in the model and $\ln(\hat{L})$ is the log-likelihood. The lower (more negative) the AIC, the better the model fits the data.

Calculating the AIC for each, we get a value of -1978.718 for the Spatial Lag

model and a value of -1953.581 for the Spatial Error model. By this criteria, the Spatial Lag model is a better fit.

We can also perform Lagrange Multiplier tests to compare the two models. The Lagrange Multiplier Test tests different spatial regression formulations against the traditional OLS formulation. In our case, this is our OLS model without regional dummies. Remember that we can write the formula for a regression model accounting for spatial dependence as follows:

$$Y = \rho W y + X\beta + u$$

$$u = \lambda W u + \epsilon$$

where ϵ is a serially uncorrelated error term. The spatial lag model is obtained by setting $\lambda = 0$ and the spatial error model is obtained by setting $\rho = 0$. We can perform 4 different Lagrange Multiplier Tests to determine the best model to capture the spatial dependence in our data:

- A simple LM test for error dependence, which restricts $\rho = 0$ and then tests the alternative hypothesis $\lambda \neq 0$ against the null hypothesis $\lambda = 0$ (LMerr)

- A simple LM test for a missing spatially lagged dependent variable, which restricts $\lambda = 0$ and then tests the alternative hypothesis $\rho \neq 0$ against the null hypothesis $\rho = 0$ (LMlag)

- A robust LM test for error dependence robust to the presence of a missing spatially lagged dependent variable (RLMerr)

- A robust LM test for a missing spatially lagged dependent variable robust to the presence of error dependence (RLMlag)

The test statistic is given by $d'(\theta)I'(\theta)d(\theta)$ where $d(\theta) = \frac{\partial L}{\partial \theta}$ evaluated at the null with L the log-likelihood for the spatial model and θ the parameter of interest, and $I(\theta)$ is the information matrix for the spatial model evaluated at the null [Anselin,

1988a]. We perform these tests using R's lm.LMtests command from the "spdep" package. The results are in Table 5.4.

Test	*Test Statistic*	*p-value*
LMerr	37.336	9.945×10^{-10}
LMlag	67.975	2.2×10^{-16}
RLMerr	0.47	0.4909
RLMlag	31.114	2.433×10^{-8}

Table 5.4: Lagrange Multiplier Test Results

We see that the null is soundly rejected in both the simple tests, so we look to the robust tests. The Robust LM error test fails to reject the null, suggesting that, allowing for the possibility of a missing spatially lagged dependent variable, we cannot reject the hypothesis that the error terms are not spatially correlated and indeed random in space. However, even allowing for potential spatial correlation of the error terms, we cannot reject the hypothesis that the spatially lagged dependent variable should not be included in the model. Thus, in accordance with the AIC, we determine that the spatial lag model is the best fit for this data.

5.5.3 Spatial Lag vs. OLS with Regional Dummies

Spatial Lag Model Results

Table 5.5 gives the results of the Spatial Lag model next to the results from our OLS model. We see that ρ is estimated at 0.29, indicating positive correlation between a tract's ridership loss and the ridership loss of surrounding tracts. This value is also used to calculate the direct and indirect effects of each of the explanatory variables in the spatial lag model. Because of the lagged dependent variable within the spatial lag model, the interpretation of the reported coefficients is not as straightforward as the interpretation of OLS coefficients. The direct, indirect, and total impacts of each explanatory variable is shown in Table 5.6. Direct effects should be very close to those reported in Table 5.5.

	OLS	*Spatial Lag*
Region dummies	Yes	No
ρ		0.29^{***}
pct_black	0.125^{***}	0.100^{***}
	(0.018)	(0.018)
pct_colgra	0.002	0.003
	(0.049)	(0.045)
pct_25_34	-0.107^{**}	-0.122^{***}
	(0.044)	(0.042)
log_medinc	0.001	0.0001
	(0.012)	(0.011)
pct_spansp	0.102^{***}	0.053^{***}
	(0.021)	(0.019)
pct_forbor	-0.081^{**}	-0.024
	(0.033)	(0.029)
pct_noveh	0.066^{**}	0.062^{**}
	(0.032)	(0.029)
pct_peak	-0.227^{***}	-0.231^{***}
	(0.047)	(0.046)
pct_wkend	-0.094	-0.097^{*}
	(0.057)	(0.056)
avg_weekly	-0.003	-0.005
	(0.006)	(0.006)
pct_bus	0.104^{***}	0.101^{***}
	(0.024)	(0.023)
pct_transf	0.255^{***}	0.232^{***}
	(0.034)	(0.033)
used_cash	0.067^{*}	0.069^{*}
	(0.039)	(0.038)
more_pass_	0.393^{***}	0.380^{***}
	(0.088)	(0.086)
Constant	-0.865^{***}	-0.592^{***}
	(0.143)	(0.136)
Akaike Inf. Crit.	-1,988.043	-1,978.718
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 5.5: Spatial Lag and OLS Model Results

	Direct	*Indirect*	*Total*
pct_black	0.102	0.048	0.151
pct_colgrad	0.003	0.002	0.005
pct_25_34	−0.124	−0.058	−0.183
log_medinc	0.000	0.000	0.000
pct_spansp	0.055	0.026	0.080
pct_forbor	−0.024	−0.011	−0.036
pct_noveh	0.063	0.030	0.093
pct_peak	−0.236	−0.111	−0.347
pct_wkend	−0.099	−0.047	−0.145
avg_weekly	−0.005	−0.002	−0.007
pct_bus	0.103	0.048	0.152
pct_transf	0.236	0.111	0.358
used_cash	0.070	0.033	0.103
more_pass_	0.388	0.182	0.571

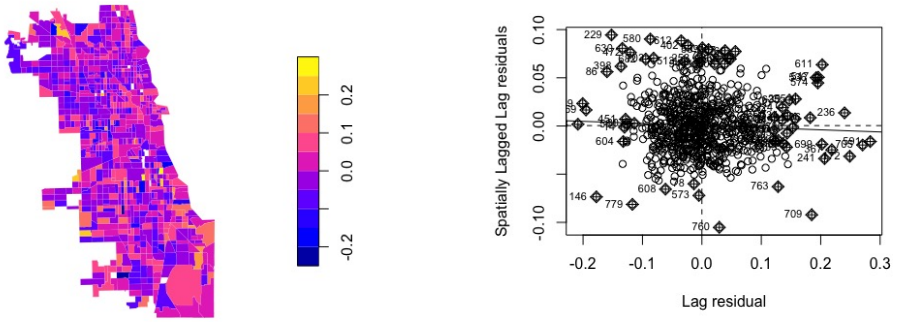
Table 5.6: Spatial Lag Variable Impacts

All in all, the results of the spatial lag model are very similar to that of the OLS model in terms of significance and magnitude of the direct impact of the explanatory variables. The percent of foreign born residents ceases to be important in the spatial lag model, but other than this and an intensification of the effect of the percent of young people in a tract, the spatial lag presents direct effects that are slightly lesser in magnitude than the OLS variables, with some of the impact of each variable being handed off to neighboring tracts.

Model Comparison

Figure 5-6 gives the map of residuals from the Spatial Lag Model and the Moran plot. The Moran's I statistic for the spatial autocorrelation of the residuals is −0.03 with a p-value of 0.925, showing that this model does succeed in eliminating spatial autocorrelation from the model residuals. However, looking at Table 5.6, we note that the AIC for the OLS model is actually better than for the spatial lag model, indicating that the OLS model with regional dummies fits the data slightly better than a spatial lag model without regional dummies.

It is unclear which of these models is objectively better than the other. While



(a) Spatial Lag Residual by Census Tract

(b) Model Residual vs. Spatially Lagged Residuals

Figure 5-6: Residual Analysis for Spatial Lag Model

the Spatial Lag model successfully eliminated concerns about autocorrelation of the residuals, the OLS model with regional dummies exhibited residual spatial autocorrelation that was very small, albeit significantly non-zero. Furthermore, the OLS model demonstrated a slightly better fit for the data, and has an interpretation that is more straightforward.

This exercise demonstrates that, if spatial correlation of model residuals is of significant concern, spatial regression models offer a way of accounting for this. Specifically in the case of spatial lag models, they offer a possible way to eliminate some bias from coefficient estimates by accounting for variables that would be omitted in a straightforward OLS application. The usefulness and interpretability of the model depends on the types of variables used and the structure of the spatial weights matrix. Here, we have a model that indicates that neighboring demographics and ridership behavior attributes explain ridership decline due to COVID. This does make intuitive sense. Transit riders do not ride only from one single transit station, and we would expect the attributes associated with a census tract to spillover and impact the ridership numbers of a neighboring tract.

Because the estimates of the impacts of each variable did not change much between the spatial lag model and the OLS model, however, we can draw many of the same high-level conclusions from either of them. This experiment was intended to tease out which demographic and behavioral variables were most predictive of ridership

changes due to COVID in light of our findings from the previous chapter, and the conclusions are much the same regardless of the model one chooses to examine.

5.5.4 Discussion of Findings

While some of the variables we expected to be highly predictive of ridership decline due to the COVID stay-at-home order proved insignificant in our models, there are still important takeaways we can draw. First, we have identified several strong predictors of smaller levels of ridership drop. The largest in terms of magnitude is the percentage of riders living in that tract who spend more on passes than on pay-per-use rides. This is likely explained in a couple different ways. First, people who purchase passes may do so because of their lack of other options for getting around and subsequent certainty that they will ride enough to warrant a pass. If a pass is an indication of transit reliance, it makes sense that it would be correlated with higher ridership during COVID. Secondly, for people who needed to make a trip during the early COVID analysis phase, having a pass may have been an incentive to use transit for that trip since they had already paid for it. Regardless of the reason, a history of buying transit passes being predictive of ridership after shocks to the system may be useful for transit agencies hoping to understand ridership repercussions of serious delays or other incidents on transit networks.

Rate of transfers and percent of trips taken on bus are also strongly predictive of higher levels of COVID ridership. Because these remained predictive despite controlling for several demographic variables, it suggests that the behaviors themselves tell us something about the riders not captured by their home tract's racial and socioeconomic makeup. While it is harder to connect the dots and explain the particular thread between these behavioral explanatory variables and the ultimate outcome of continued ridership during COVID, it is nonetheless useful to recognize that riders on the system who use the bus and take routes requiring transfers have needed to continue using the system during the pandemic, and underscores the necessity to invest in these parts of the system, as was stressed in the previous chapter.

These variables in particular may also tell us, as we discussed earlier when ex-

ploring the correlation between percent of trips taken on the bus and presence of a rail station, something about the geography of where these tracts are. The fact that high levels of bus use indicate distance from rail stops may also indicate distance from other amenities, such as grocery stores. These people may have continued to use transit during the pandemic both because they have no other option for how to get around and because they are not within walking distance of the essential activities they need to undertake. Further work should seek to include information on the density of errand destinations.

On the other side of the coin, peak ridership was strongly predictive of steeper drops in ridership, confirming what we saw in the previous chapter. This may be because high levels of peak usage communicate both 9-5 jobs that may have easily switched online but also historic choices to use other modes at off-peak times, indicating more choice in how these individuals get around. This again should speak to the importance of investing in transit systems at times other than the peak if the goal is to reach those most in need of the system.

Finally, we see that high proportions of black residents, Spanish speakers, and households without cars are predictive of smaller drops in ridership. While it is surprising that income was insignificant when including these controls, it is no question that income is correlated with each of these. Leaving these variables out of the model leads to a significant negative coefficient on income, as we would expect, so it is possible that the effect of income is obscured by that of the percent of black residents, for example, which tends toward the extremes on the unit scale. This finding suggests that special attention should be paid to these communities to ensure that they have the ability to get around the city during this time. As was suggested in the previous chapter, dissemination of Ventra cards is an important step, as is ensuring that important information about service updates and public health on the system is readily available in Spanish.

5.6 Conclusion

This chapter expanded upon the previous and attempted to parse out the important variables in explaining COVID ridership decline. We demonstrated that combining demographic data and aggregate behavioral traits led to a more powerful model and suggested that behavior was an important predictor of ridership changes even when controlling for sociodemographics.

The findings reinforced many of those from the previous chapter and could serve as added justification for any of the policy recommendations outlined in depth at the end of Chapter 4. They are also consistent with patterns that have been observed in other cities.

This section shows that including baseline behavior can be helpful in understanding changes in ridership. In this case, the changes were the result of a sudden and extreme event, but similar data could be used to understand more gradual changes, or changes that are smaller in magnitude. Such an approach could help transit agencies target policies or interventions more appropriately by relying on some key aspects of past behavior.

Further research should refine models such as these in a few ways. First, deeper thought to how behaviors are assigned to an area is warranted. The approach taken here is simplistic and only truly appropriate for regular riders, though it was applied to all riders. Assigning each stop a behavioral profile based on the type of riders that typically use it, perhaps even broken down by time of day, might be one way to approach the problem. Second, depending on the question, including more land use attributes could be worthwhile. This particular model would have benefited from information on the quantity and quality of grocery stores in each area, for example.

Next, other spatial weight matrix constructions should be explored to reflect the structure of the transit network. There are likely insights to be had about how changing ridership levels are linked across the system in a way that cannot be captured with weights matrices that consider only local influence. Lastly, different spatial weight matrices for different types of variables may be an avenue worth pursuing, as

we may have enough knowledge about how different variables interact to impose more complex structures that reflect this.

Chapter 6

Exploration of Application of Spatial Regression Frameworks to High Dimensional Data

Urban areas are complicated places upon which countless attributes can be measured and mapped and the relationships among them modeled. The question of the impact of the COVID-19 pandemic on public transit trips lent itself to a model that could use cross-sectional data to explain the magnitude of the change in trip volumes due to a particular event in different areas. In many urban mobility questions, however, the temporal dimension is just as if not more important that the spatial dimension, as travel needs differ significantly throughout the course of the day. Significant aggregation in the temporal dimension, while often necessary due to data or computational limitations, loses information about the ebb and flow of travel demand over the course of a day, which may be important for the design of policy, infrastructure, or technology, depending on the circumstance. In this chapter, we offer preliminary thoughts on and exploration of how rich spatial and temporal urban mobility data can be mined for insights within a flexible framework that allows for any spatial, temporal, or spatio-temporal data the modeler may have access to. We leverage some of the tools and lessons from spatial regression to capture some parameters that describe urban mobility dynamics in Chicago. The applications in this chapter focus

on modeling TNC use, but could be adapted for any mode.

One motivation for this chapter comes from the desire to leverage high-dimensional data to understand the interplay between usage patterns of different travel modes. The analysis in Chapter 3 revealed that year over year losses in trip volumes on the CTA were due more to people using the system less frequently than to a decrease in the number of riders. A natural follow-up would be to ask why this was occurring—was it that people were taking fewer trips? Were they driving more, or walking more? A popular hypothesis for the decline in public transit trips across US cities is that people are opting for TNCs in place of public transit, specifically in the evenings and on weekends. Many studies have examined this, either using surveys or by examining the change in the number of transit rides after the entrance of TNCs into a city [Feigon et al., 2018, Murphy et al., 2016, Michael Graehler et al., 2019]. The exploration in this chapter is motivated in part by a desire to understand the extent to which TNCs and public transit exhibit similar spatio-temporal patterns in a city and identify the times and places when usage of each are most in line or most divergent. While this would not allow us to know which Ventra cards were replacing public transit rides with TNC rides, it could shed light on times and places public transit could be improved to be more competitive with TNCs, or what areas and time frames should be targeted in the formulation of a TNC fee, for example. This work does not answer these questions specifically, but potential extensions of the initial work presented here that may achieve such aims are discussed at the end.

The exploration in this chapter does not explicitly build off the rider-centric analysis in the previous chapters, as the public transit data used here is trip counts disaggregated across space and time. It can, however, aid in the understanding of public transit rider behaviors by filling in details on the context in which those behaviors are occurring. Customer segmentation using smartcard data is by definition limited to capturing behaviors only on public transit, but many urban-dwellers live multi-modal lives, and many public transit users also use TNCs [Gehrke et al., 2018]. TNCs are unlikely to release data to cities in which multiple trips are tied to an individual, so similar customer segmentation for TNC riders is infeasible for an outside

party. Using trip counts for both modes, however, to obtain a rich picture of how the usage patterns of these modes relate to one another and identify axes of high spatio-temporal correlation could point to public transit behaviors that are regularly supplemented by TNC use.

6.1 Context

This chapter can be framed as the beginning of an exploration into bringing together the realms of spatial regression and machine learning in order to capture complicated dynamics occurring in the world of urban mobility. Linear regression in general, as well as the expansion of linear regression methods to deal with spatial dependency, have the distinct benefit of offering parameter estimates that purport to describe something about how the world operates. For example, in the previous chapter, our parameter estimates communicated the relative importance of each of the independent variables in explaining the level of decline in trips seen across the city of Chicago in the early stages of the COVID-19 pandemic. The ability to glean such insight is the ultimate objective of models from the perspective of the policy-maker.

On the other hand, the volume of data now available from smartcard systems such as Ventra has made urban mobility into a playground for some researchers focused on the other possible objective of models— prediction. Prediction of public transit usage has become a common task in the development of new architectures for machine learning algorithms and deep neural networks [Cheng et al., 2019, Ma et al., 2017, Ma et al., 2019]. These models take full advantage of the rich data available, and often have incredible prediction power, but are of little use to people hoping to uncover insight into urban mobility patterns.

This chapter offers preliminary thoughts and examples for how a city or transit agency with access to rich, high-dimensional data could leverage it using models that lend themselves to interpretation but taking advantage of flexible, machine-learning based estimation frameworks.

6.2 Spatio-Temporal Regressions, Data, and Experiment Setup

In this section, I outline the types of data I compiled for these experiments as well as the framework for building the models. As mentioned earlier, this is preliminary work, and I do not have robust results for all of the model configurations outlined here. This section is meant as a methodological explanation that can serve as a guide for furture work.

6.2.1 Data

The data for this study can be categorized into 3 main groups: temporal, spatial, and spatio-temporal. In order to perform the regression, each variable was transformed into a $T \times I \times J$ array, representing a grid of the city of shape $I \times J$ at each of the T total timesteps. We will now explain each group of data in further detail.

Temporal data

The first group of data is temporal data— data that varies only along the temporal dimension. Variables of this type take the form of an array in which each $I \times J$ slice of the $T \times I \times J$ array contains the same value in each cell. In other words, if $c^{z_{d_z}}_{t',i',j'}$ is the value of $\mathbf{z}_{d_z,t'}$ in grid cell i', j', then $c^{z_{d_z}}_{t',i',j'} = c^{z_{d_z}}_{t',i,j} \forall i \in \{1, 2, ..., I\}$, $j \in \{1, 2, ..., J\}$, and $t' \in \{1, 2, ..., T\}$. The spatial data in this study include hour of day and day of the week dummies, as well as dummies indicating if it snowed on that day or if it rained on that day. The weather data was provided by the CTA for each day of the study, so no finer temporal granularity could be achieved.

Spatial data

The spatial data group contains variables that vary over space but not time, such as land use and demographic data that do not change over the course of our one month study period. Each of these variables takes the form of a $T \times I \times J$ array in

which the same $I \times J$ grid is repeated T times. Mathematically, $c^{x_{d_x}}_{t',i',j'} = c^{x_{d_x}}_{t,i',j'} \forall t \in \{1, 2, ..., T\},\ i' \in \{1, 2, ..., I\},\ j' \in \{1, 2, ..., J\}$. Data of this type comes from a variety of sources. Demographic data other than job counts comes from the 2014-2018 American Community Survey (ACS), which provides sociodemographic information at the block group level. These estimates are then mapped onto the grid by allocating counts of people proportionally by area. Slightly more consideration had to be given to non-count data, such as median income per capita and travel time to work. In the former case, the value for a cell is achieved by taking the population-weighted average of the median incomes for the overlapping block groups, where the population used for the weights is that of the overlapping area, determined by the count allocation method described above. In the latter case, data is provided in the ACS in the form of counts of people whose commute time falls within each of several buckets of travel time ranges. Here we took a population-weighted average of the median value of each bucket, with the population weights determined once again by the count allocation.

We took information on the number of jobs from the Longitudinal Employer-Household Dynamics (LEHD) data, which is compiled yearly by states from unemployment insurance earnings and available via the US Census. The job counts are at the block level, and we mapped them to the grid proportionally by area, in the same way that population counts from the ACS were mapped to the grid. We used the most recent jobs LEHD data available, which was from 2017.

Land use information was obtained in the form of points of interest data from Open Street Maps. Data on restaurants, bars, shops, public attractions, hotels, schools, and services were retained and their point locations mapped to the corresponding grid cell. Additionally, information on the location of bus and rail stations came from the city of Chicago's Open Data portal.

Spatio-temporal data

The last group of data is that which varies along all dimensions. In this category are public transit and TNC usage counts, as well as measures of the frequency of public transit, which is used as a level of service control.

The public transit data comes from the Ventra database. The CTA rail and bus networks are tap-on only, so the activity captured in this data are only boardings. For each tap, we know the exact time stamp and boarding station, as well as the mode, so public transit usage can be mapped precisely to the $T \times I \times J$ array, though observations are constrained to those grids cells containing a public transit stop.

The TNC data comes from the city of Chicago, which requires quarterly reports from all TNCs operating in the city. The city aggregates the trip information to 15-minute intervals and census-based origin and destination geographic areas and make this data publicly available. Trip counts were allocated to the grid in the same way as the other census block and block group-based count data in this study, and 2 separate $T \times I \times J$ arrays were created— one for origins and one for destinations.

The final spatio-temporal piece of data is that which captures the frequency of transit service, measured as the length of the headways between transit trips. This was calculated using public General Transit Feed Specification (GTFS) data which provides information on scheduled trips. To deal with the fact that headways in many parts of the city are of similar length or longer than the smallest time interval used for our analysis (15 minutes), we calculated average headways separately for weekdays, Saturday, and Sunday, and within these days for 4 time frames whose headways differ from one another but are largely consistent within groups: Peak service (6AM - 10AM, 4PM - 8PM), Midday (10AM - 4PM), Evening (8PM - 11PM), and Night (11PM - 6AM). Each $t \in \{1, 2, ..., T\}$ is mapped to the corresponding time of day and day of week and given that average headway. Average headways within grid cells were weighted by the number of trips occurring at that headway, which we use as a proxy for demand.

6.2.2 Spatio-Temporal Regressions

This study uses three types of variables: spatial, temporal, and spatio-temporal variables. Formally, let i, j indicate the geographical index $i \in \{1, 2, ..., I\}$ and $j \in \{1, 2, ..., J\}$. Hence IJ is the total number of urban grids. Let $\mathbf{X} \in \mathrm{R}^{IJ \times D_x}$ denote the spatial variables, in which D_x represents the number of spatial variables.

$\mathbf{Y}^{(o)}, \mathbf{Y}^{(d)} \in Z^{T\times IJ}$ and $\mathbf{y}_t^{(o)}, \mathbf{y}_t^{(d)} \in Z^{IJ\times 1}$ denote the spatio-temporal variables. In this study, the OD counts and the public transit level of service metrics are the only inherently spatio-temporal variables, but we can create others by interacting a purely temporal variable, such as hour of day, with a purely spatial variable, such as presence of a rail station. $\mathbf{Z}_t \in \mathrm{R}^{IJ\times D_z}$ denote the temporal variables, in which D_z represents the number of temporal variables. By using $vec()$ as an operator that vectorizes an urban grid from $\mathrm{R}^{I\times J}$ to $\mathrm{R}^{IJ\times 1}$, then

- We use $Z_t = [vec(\mathbf{z}_{1,t}), ..., vec(\mathbf{z}_{d_z,t}), ..., vec(\mathbf{z}_{D_z,t})]$ to represent the temporal variables, where each $\mathbf{z}_{d_z,t} \in \mathrm{R}^{I\times J}$ represents a matrix of a temporal variable. Note that the temporal variables are often dummy variables and don't have spatial variations, so the $vec(\mathbf{z}_{d_z,t})$ is typically full zero or one vectors: $[0, 0, ..., 0]^T$ or $[1, 1, ..., 1]^T$. The temporal variables include the weekday vs. weekend dummies, time of day dummies, and weather dummies.

- Spatial variables can be represented as $\mathbf{X} = [vec(\mathbf{x}_1), ..., vec(\mathbf{x}_{d_x}), ..., vec(\mathbf{x}_{D_x})]$, in which each $\mathbf{x}_{d_x} \in \mathrm{R}^{I\times J}$ represents a matrix of a spatial variable, such as the average incomes of the urban grids. Note that each $\mathbf{x}_{d_x}$ has spatial variation since $\mathbf{x}_{d_x,ij}$ varies with the spatial indicators i, j. The spatial variables include the socio-economic, land use, built environment, and locations of subway and bus stations.

- The spatio-temporal OD counts $\mathbf{Y}^{(o)}, \mathbf{Y}^{(d)}, \mathbf{y}_t^{(o)}, \mathbf{y}_t^{(d)}$ vary across time and space. Note both $\mathbf{y}_t^{(o)}$ and $\mathbf{y}_t^{(d)}$ have been vectorized from the urban grids $\mathrm{R}^{I\times J}$ to a vector $\mathrm{R}^{IJ\times 1}$. The target variables to explain are the OD counts $\mathbf{y}_t^{(o)}$ and $\mathbf{y}_t^{(d)}$.

Using the origin trip counts for an arbitrary mode as an example dependent variable, a general regression form can be established for investigating the spatiotemporal patterns of usage:

$$\mathbf{y}_t^{(o)} = \beta_0 1_{IJ} + \underbrace{\mathbf{Z}_t\boldsymbol{\beta}_z}_{\text{temporal}} + \underbrace{\mathbf{X}\boldsymbol{\beta}_x + \mathbf{W}\mathbf{X}\boldsymbol{\beta}_{xw}}_{\text{spatial}} + \underbrace{\mathbf{Y}_{t-1}\boldsymbol{\beta}_y + \mathbf{W}\mathbf{Y}_{t-1}\boldsymbol{\beta}_{yw}}_{\text{spatio-temporal}} + \epsilon_t \tag{6.1}$$

In equation 6.1, 1_{IJ} represents an one vector with IJ length, $\mathbf{Y}_{t-1} = [\mathbf{y}_{t-1}^{(o)}, \mathbf{y}_{t-1}^{(d)}]$ is a $IJ \times 2$ matrix that combines the OD counts at time $t-1$. The coefficients to be estimated include a constant β_0, the coefficients for temporal variables $\boldsymbol{\beta}_z \in \mathrm{R}^{D_z}$, the coefficients for spatial variables $\boldsymbol{\beta}_x, \boldsymbol{\beta}_{xw} \in R^{D_z}$, and the coefficients for spatio-temporal variables $\boldsymbol{\beta}_y, \boldsymbol{\beta}_{yw} \in R^2$. The $\mathbf{W} \in \mathrm{R}^{IJ \times IJ}$ is defined as the spatial weighting matrix that captures the spatial correlation. There are many ways to define the spatial weighting matrix, and we only use the simplest version that uses ones to denote the connectivity of neighbouring four urban cells. To predict $\mathbf{y}_t$, the lags of the spatio-temporal variables can be more than degree 1, but here we write the general regression formula with only one lag for simplicity. The last term ϵ_t represents a random Gaussian vector with $IJ \times 1$ dimension.

To train the spatio-temporal regression model, we use the least square estimation. Particularly, the empirical risk function is represented as:

$$\mathrm{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \frac{1}{I \times J \times T} \sum_{t=1,\ldots,T} (\mathbf{y}_t^{(o)} - \hat{\mathbf{y}}_t^{(o)})'(\mathbf{y}_t^{(o)} - \hat{\mathbf{y}}_t^{(o)}) \tag{6.2}$$

in which $\hat{\mathbf{y}}_t^{(o)} := \mathbf{y}_t^{(o)} - \epsilon_t$ is the predicted vector of origin counts. The trained coefficients are defined as the minimum to the empirical risk:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \mathrm{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}) \tag{6.3}$$

Computationally, we use the stochastic gradient descents to train the model. The variance $Var(\hat{\boldsymbol{\beta}})$ can be estimated by using analytical methods or bootstrap.

6.2.3 Experiment Design

For computational expediency, the dimensions of the urban grid used here are 86×76 with each cell's size roughly 500m in length. When running models employing spatial weight matrices, we focus on an inset of the grid that is 30×25 and contains the downtown area. We use data from the month of October 2019 to model the relationship between various temporal and spatial factors and the volume of TNC

trip origins across the city.

As mentioned previously, this chapter offers a framework for analysis and preliminary results. Though not completed here, one could use the general regression formula seen in equation 6.1 to successively explore the temporal, spatial, and spatio-temporal patterns of TNC usage, answering such questions as:

- **Temporal.** What are the temporal patterns of the relative prevalence of TNCs? When during the week are they most favored to public transit? How is this affected by the weather or special events?

- **Spatial.** What are the spatial patterns of the relative prevalence of TNCs? Specifically, how does their prevalence relate to aspects of the built environment, sociodemographics, and the location of the transit network?

- **SpatioTemporal.** How do the answers of each of the previous questions vary with the other? Are there areas that exhibit different temporal patterns than others? How strongly do the spatiotemporal patterns of one mode relate to another?

Six models, each suited for a different question, can be derived from the general regression formula and are outlines in Table 6.1 .

Models	**Restrictions**	**Goals**
Model 1	$\boldsymbol{\beta}_z \neq 0$; $\boldsymbol{\beta}_x = \boldsymbol{\beta}_{xw} = \boldsymbol{\beta}_y = \boldsymbol{\beta}_{yw} = 0$	temporal
Model 2	$\boldsymbol{\beta}_x \neq 0$; $\boldsymbol{\beta}_z = \boldsymbol{\beta}_{xw} = \boldsymbol{\beta}_y = \boldsymbol{\beta}_{yw} = 0$	spatial
Model 3	$\boldsymbol{\beta}_x, \boldsymbol{\beta}_{xw} \neq 0$; $\boldsymbol{\beta}_z = \boldsymbol{\beta}_y = \boldsymbol{\beta}_{yw} = 0$	spatial
Model 4	$\boldsymbol{\beta}_y \neq 0$; $\boldsymbol{\beta}_z = \boldsymbol{\beta}_x = \boldsymbol{\beta}_{xw} = \boldsymbol{\beta}_{yw} = 0$	spatio-temporal
Model 5	$\boldsymbol{\beta}_y, \boldsymbol{\beta}_{yw} \neq 0$; $\boldsymbol{\beta}_z = \boldsymbol{\beta}_x, \boldsymbol{\beta}_{xw} = 0$	spatio-temporal
Model 6	$\boldsymbol{\beta}_z, \boldsymbol{\beta}_x, \boldsymbol{\beta}_{xw}, \boldsymbol{\beta}_y, \boldsymbol{\beta}_{yw} \neq 0$	a joint model

Table 6.1: High Dimensional Spatio-Temporal Regression Experiment Design

6.3 Preliminary Data Analysis

We begin by exploring some of the spatial and temporal patterns evident in the data.

6.3.1 Public Transit and TNC Usage

Temporal

First, it is helpful to get a sense of the scale of usage for the modes we are interested in. Figure 6-1 shows the hourly count of trips broken down by mode (rail, bus, or TNC) for a typical week in October. We note that on weekends there are similar numbers of trips on each of these modes. During the week, except in the very early and late parts of the day, both public transit modes are used in much greater numbers than TNCs. Rail in particular, during the peak periods, sees almost twice the trip count as bus and nearly five times the trip count of TNCs. TNCs, on the other hand, are used in much greater volume on Friday and Saturday evenings. This confirms the importance of the temporal dimension when exploring the demand of an individual mode or the dynamics of demand of among multiple modes.

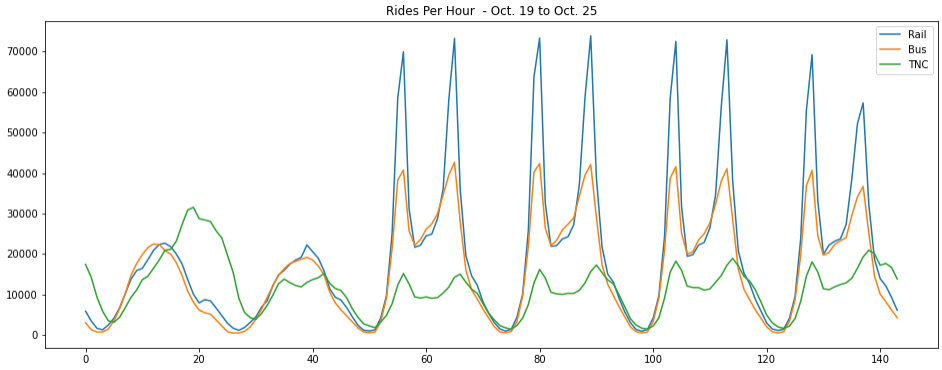


Figure 6-1: Hourly Trips by Mode: Oct. 19 - Oct. 25

Spatial

Second, we can investigate the spatial distribution of demand for each mode. Figure 6-2 shows the maximum hourly value for each grid cell, separated by mode and, in the case of TNCs, whether origins or destinations are being counted. This figure's scale is capped at 2000 trips so that the distribution is visible on each image. Rail has values that are much higher than this, due to the fact that it provides more trips

than the other modes, it is far more concentrated spatially due to the limited number of rail stations, and, as we saw above, it is more concentrated temporally as well. The grid cells in the loop all have max usage values higher than the threshold, as do a few in the north that hold stations for the Brown, Purple, and Red lines. Figure 6-3 shows the average value for each grid cell, separate by mode. This image's scale is capped much lower, at 200 trips. The story is much the same for rail in this figure, with high values in the loop, to the north and northwest, and at the ends of lines.

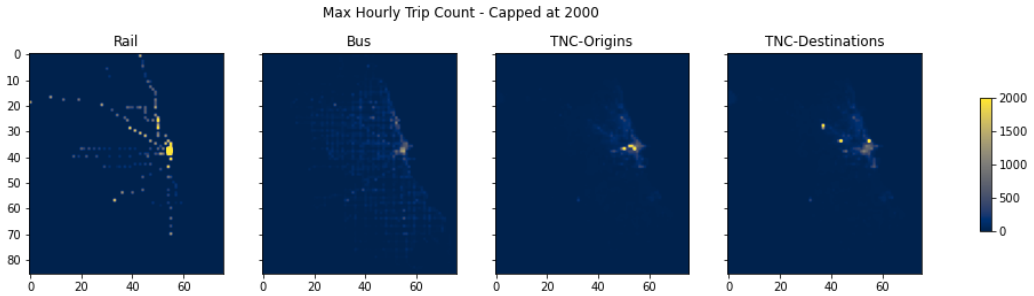


Figure 6-2: Maximum Hourly Usage by Grid Cell for Each Mode

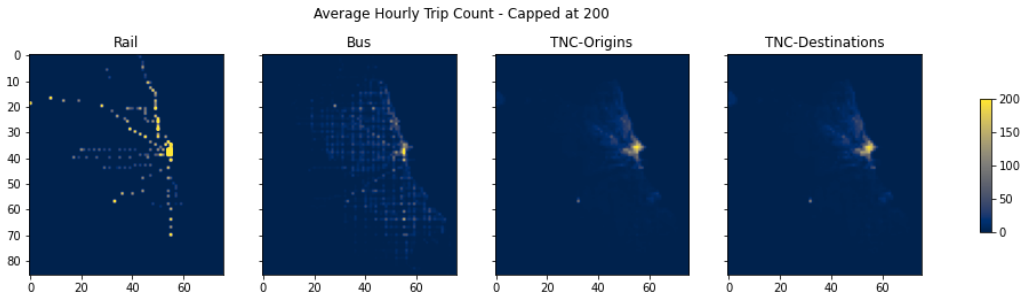


Figure 6-3: Average Hourly Usage by Grid Cell for Each Mode

Looking at the pair of images for the bus, as with rail, we see the shape of the network, though the stations cover much more of the city than the rail network. They also see much lower trip volumes, with the highest values for both maximum and average trips in the loop, along the north coast, and in grid cells that also contain rail stations.

The images depicting TNC usage show, as we would expect, a general region with significant activity as opposed to a network. The heart of the activity is the loop,

and it radiates outward, primarily in two directions: to the north along the coast, and to the northwest along Milwaukee Avenue and the Blue Line. There is much less activity in the south. The image showing the max values for origins and destinations each show a small number of cells that meet the maximum threshold. Interestingly, these are different for origins and destinations. This may suggest that large spikes in demand for TNCs are more associated with events than for rail or bus, at least for destinations. The places with very high values for maximum usage in terms of origins largely seem to be in the West Loop. This may correspond to events, or to the fact that this is where many of the restaurants and bars are located in the city, and we know from the temporal analysis that TNCs are particularly popular on Friday and Saturday evenings.

Spatio-Temporal

Next, we can investigate the typical spatial distribution of trips on each mode throughout the course of a weekday or a weekend. Figures 6-4 - 6-7 show the average count of trip origins per grid cell by hour of the day for TNCs on a weekday, TNCs on a Saturday, public transit on a weekday, and public transit on a Saturday. First, we note that the concentration of public transit origins on weekdays in the downtown area during the PM peak drowns out all other demand, rendering it invisible. On the graph of public transit origins on Saturdays, we can see trips beginning along the rail lines especially in the north in the afternoon and evening, but the few grid cells in the heart of the loop still dominate the visualization. In the TNC visualizations, the maximum value is not quite so high to utterly drown out all other areas of activity, and we see that the area radiating out from downtown in the northwestern direction remains consistently active, especially in the very early morning (12AM - 3AM). This mode too is dominated by activity in the downtown core, however.

To get a better sense of how these mode are operating simultaneously to one another, we can map the distribution of trip origins for each mode at any given 15 minute increment in the month of October 2019. A few select 15 minute intervals are shows in Figures 6-8 - 6-12 First, we note that in general, the volume of trips

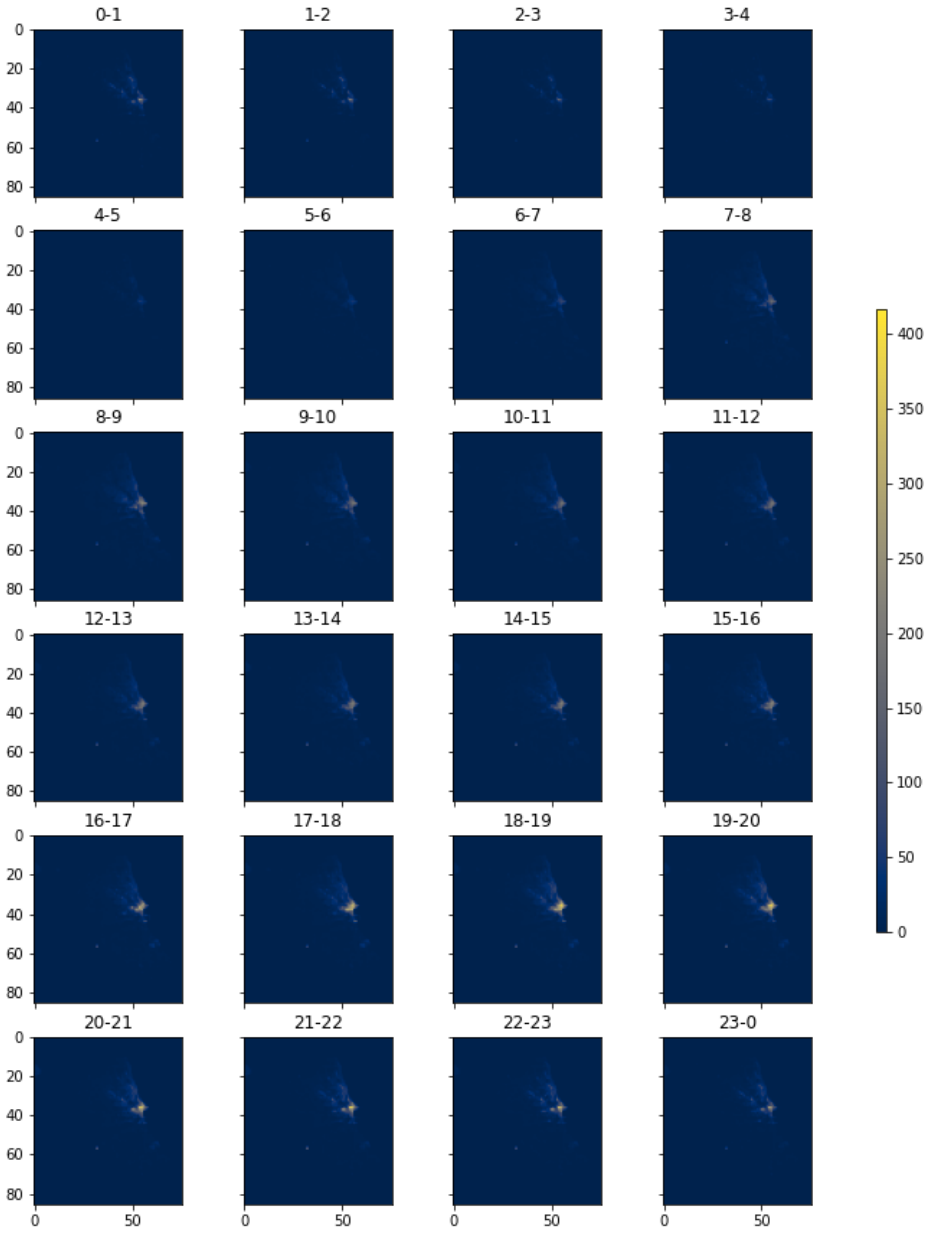


Figure 6-4: Average TNC Trip Origins by Hour for Weekdays in October 2019

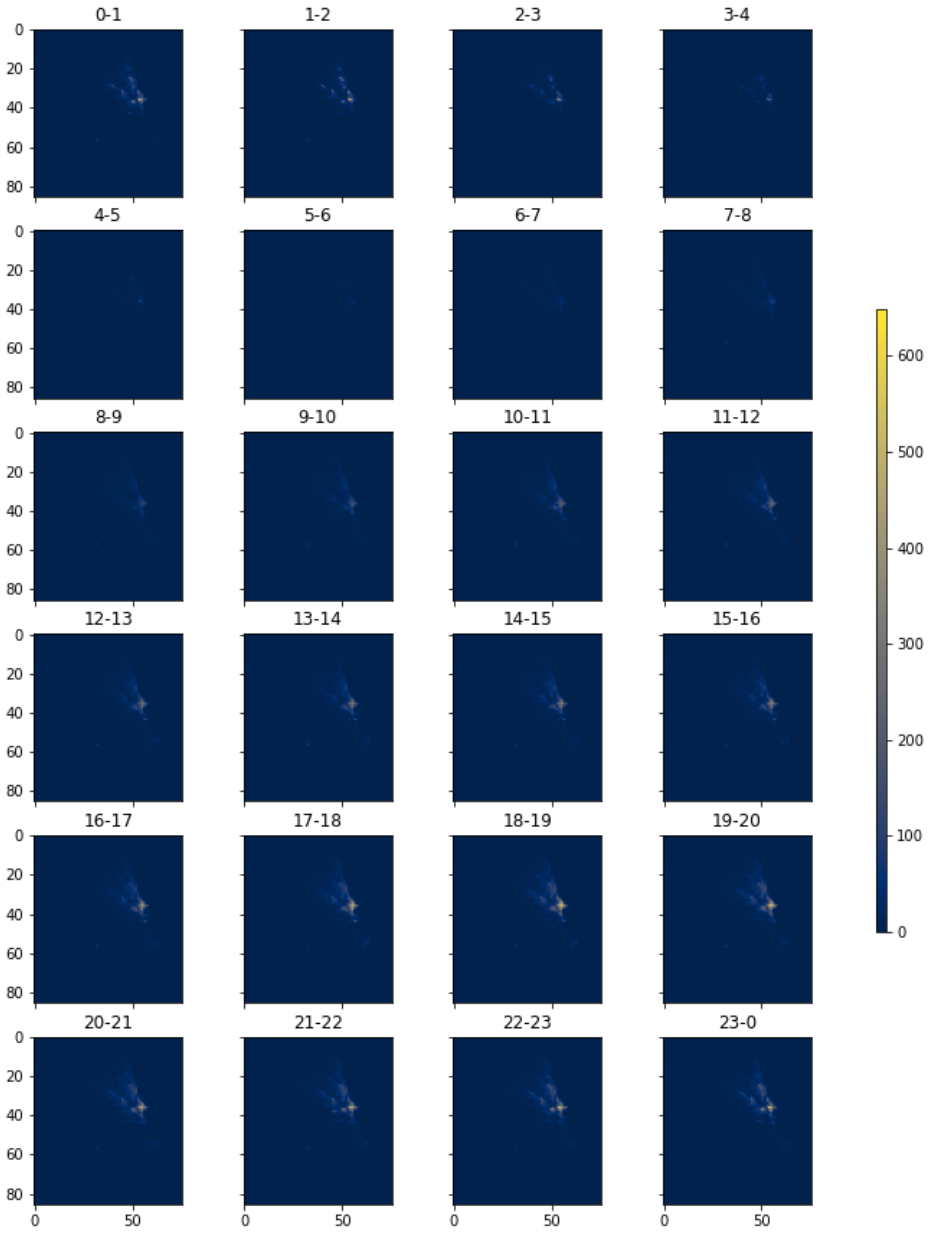


Figure 6-5: Average TNC Trip Origins by Hour for Saturdays in October 2019

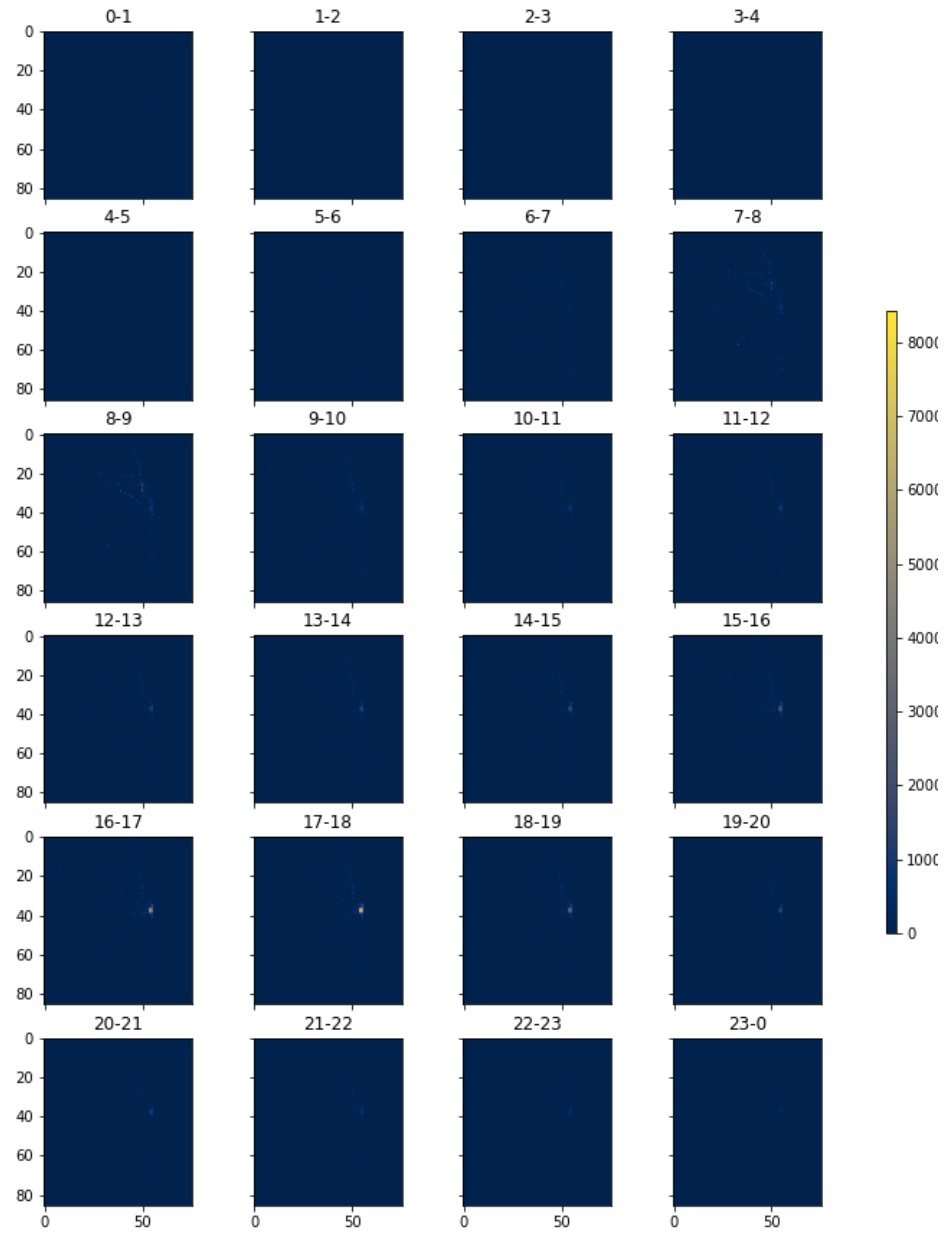


Figure 6-6: Average Public Transit Trip Origins by Hour for Weekdays in October 2019

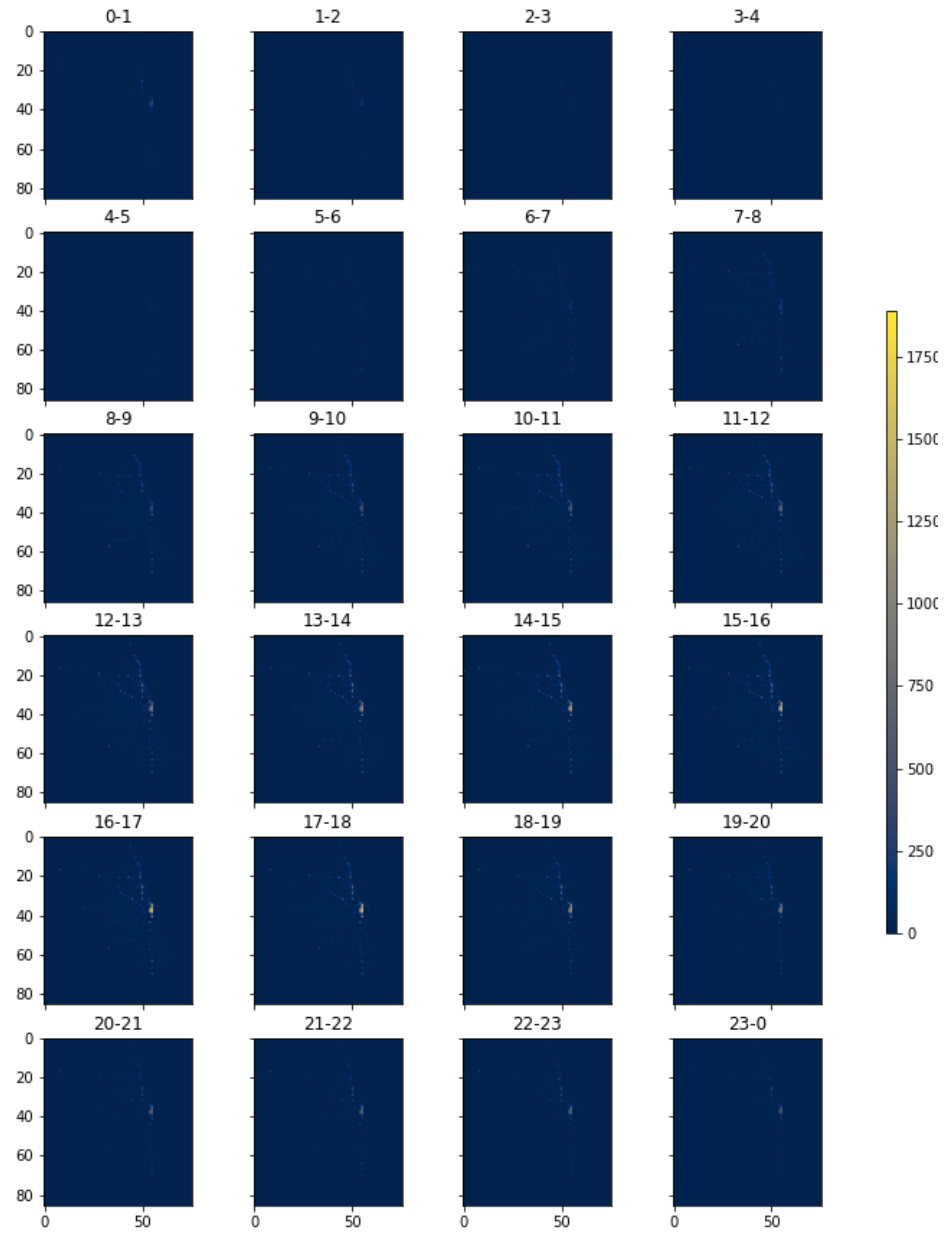


Figure 6-7: Average Public Transit Trip Origins by Hour for Saturdays in October 2019

occurring on public transit is significantly higher than the volume of trips occurring on TNCs. This is particularly noticeable during peak periods on weekdays, such as Wednesday, October 2 from 8:00-8:15AM and the same day from 6:00-6:15PM. During each of these, there is only a small pocket of noticeable TNC trips clustered around the downtown area. On public transit, from 8:00-8:15AM we see significant demand further out on the rail lines, especially in the north, while from 6:00-6:15PM we note the heavy concentration in the loop as people head home from work. During these times, it seems that TNC trips represent a small fraction of the trips occurring in Chicago, though the ones that are occurring are mainly happening where public transit trips are also happening. The story is a little different if we look much later in the evening, specifically between 2:00-2:15AM on Thursday, October 3. Here we see TNC activity around downtown and to the north and west of the loop. We see public transit usage there as well, though the highest public transit trip volume is occurring in the south, where we see no TNC trips.

Continuing through the week, we look at the TNC and public transit usage on Friday at 6PM, and notice markedly fewer public transit trips occurring in the loop than at the same time on Wednesday. Furthermore, the number of TNC trips is much more comparable to public transit volume than on Wednesday, suggesting a different dynamic between the two modes depending on the day of the week. Later that evening, at 11PM, we see comparable volumes of public transit and TNC trips, occurring in largely the same locations in the city, except for the usage along the red line in the south that has no TNC counterpart. Finally, investigating demand on Saturday morning at 2AM, we see much more TNC demand as people likely leave bars downtown and in neighborhoods in the north, while public transit usage is very low on the lines that are still running. We also note that public transit usage at this time seems to be pretty similar to usage at the same time on Thursday morning, but TNC usage is much higher here. This suggests that public transit is not being used for the same purposes as TNCs at this time.

Lastly, we can explore the share of public transit and TNC rides taken on public transit rides each hour of each day. We note that many grid cells do not contain any

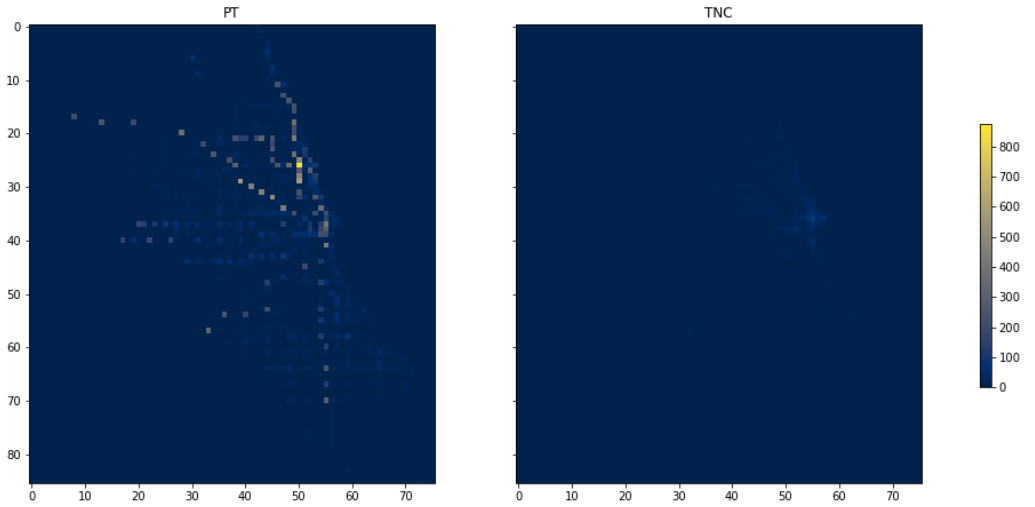


Figure 6-8: Public Transit and TNC Trip Origin Volumes on Wednesday, October 2, from 8:00-8:15AM

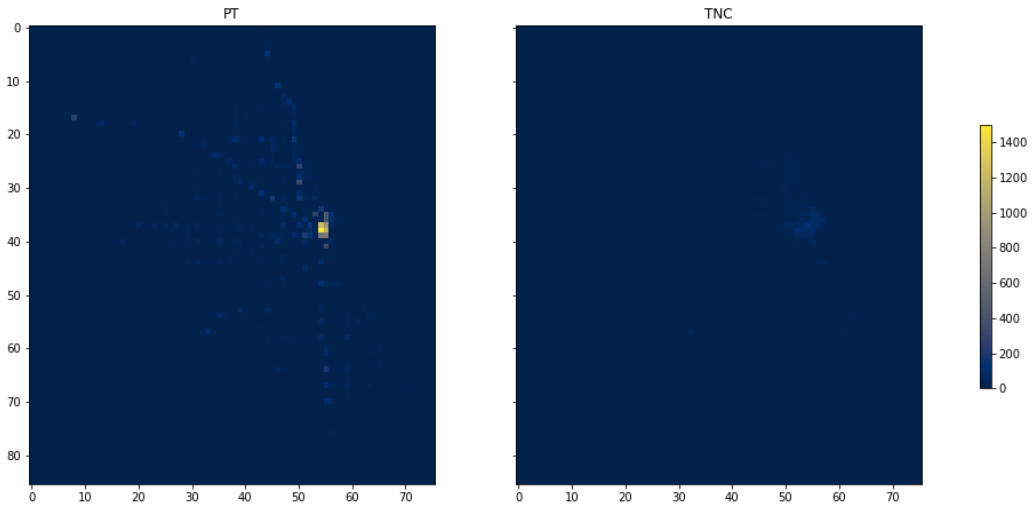


Figure 6-9: Public Transit and TNC Trip Origin Volumes on Wednesday, October 2, from 6:00-6:15PM

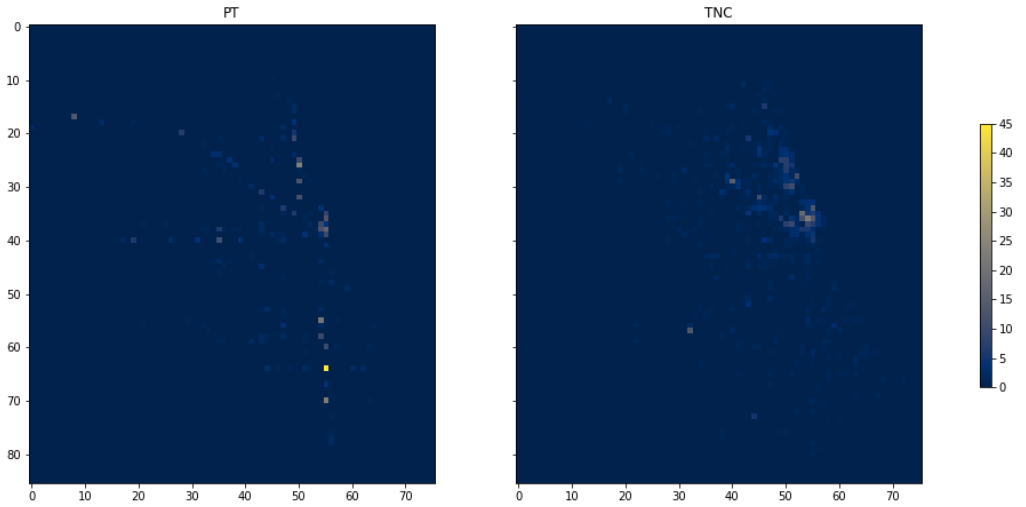


Figure 6-10: Public Transit and TNC Trip Origin Volumes on Thursday, October 3, from 2:00-2:15AM

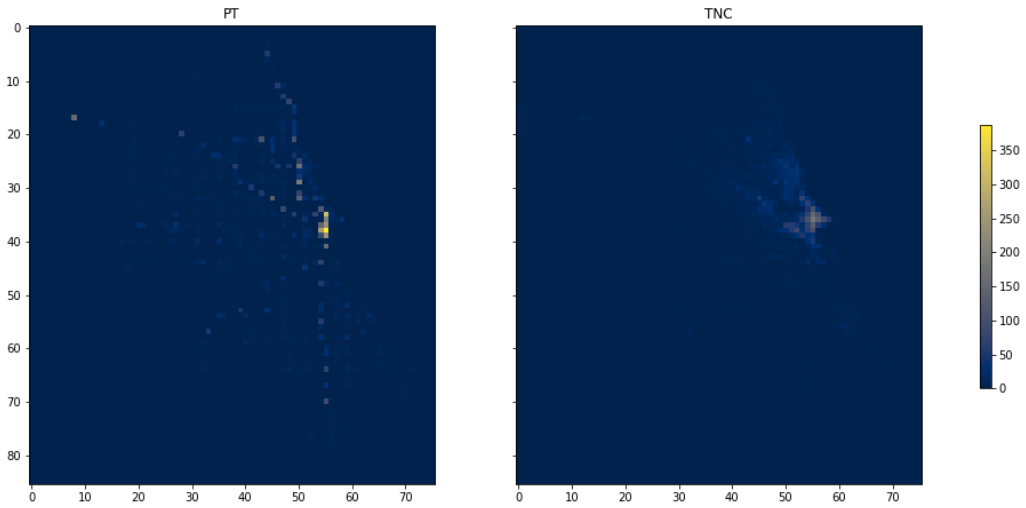


Figure 6-11: Public Transit and TNC Trip Origin Volumes on Friday, October 4, from 6:00-6:15PM

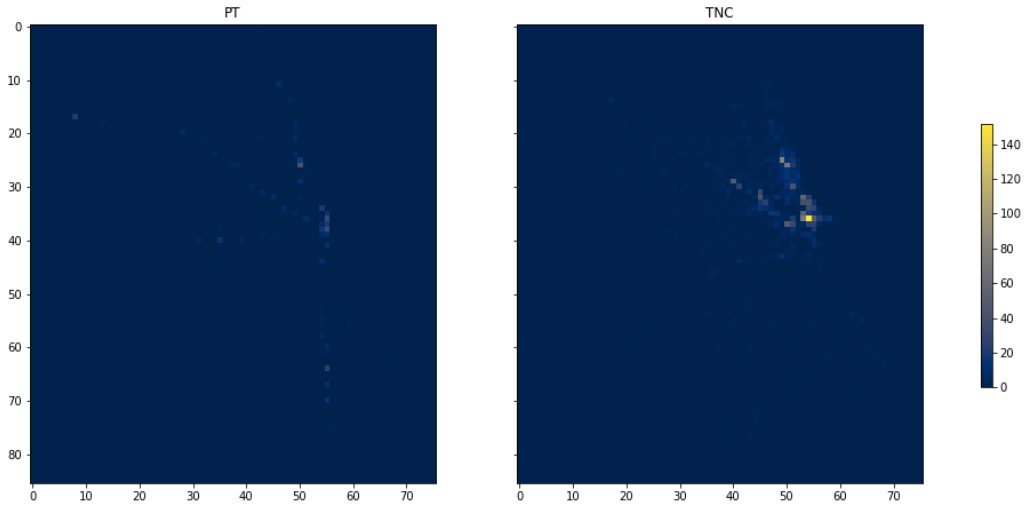


Figure 6-12: Public Transit and TNC Trip Origin Volumes on Saturday, October 5, from 2:00-2:15AM

public transit stops at all, so the share of public transit trips originating from these cells will always be zero. If there are no TNC trips either, the cell will not be graphed in the figure. Furthermore, because bus stops provide good coverage of the city, there is a dense enough grid of transit stops to generally communicate via the visualization whether public transit trip volumes are dominating TNC trips in an area, or if TNCs are providing more service.

We turn first to Thursday, October 3, and look at the share of trips on public transit across the city each hour, as shown in Figure 6-13. We see that until 4am, TNCs are providing more trips in the areas where any trips are occurring via either mode. After that, public transit starts to take over in the outskirts of the city, and then moves inward until public transit is the dominant mode during the AM peak. The one area that is an exception is the area just surrounding and to the north of the loop, where TNCs have a strong presence. In the middle of the day, this area of TNC presence gets slightly larger, while public transit remains dominant in the south and west. In the PM peak, we see TNCs continue to have a presence surrounding downtown, which grows stronger in terms of share of rides and larger in terms of

geographic area as the night progresses. This hour by hour picture of the share of public transit rides illustrates the spatial and temporal dimensions to the dynamics between the modes.

Looking at Saturday, October 5, as seen in Figure 6-14, we see a slightly different picture. As with Thursday, the early hours show TNC being the dominant mode until the public transit system starts to run fully in the morning. Unlike Thursday, however, public transit never completely overtakes TNC trips, especially on the northside. Public transit is the primary mode in the south and west, but there is more of an overall balance on Satruday than on Thursday. Starting around 5PM, TNCs become more dominant downtown and in the north, and the area of their prevalence widens as the hours go on. In general, public transit never comes to dominate the TNC-PT demand on this Saturday as it did two days previously.

6.3.2 Spatial Covariates

In our models, we will seek to capture relationships between spatial attributes of the city of Chicago and some measures of these modal usages. Aside from information about the location of rail and bus stations in the CTA system, our spatial variables of interest fall into two main categories: demographic and land use. In this section, we will explore the spatial distribution of each of our variables as well as their correlations with one another and with our dependent variables of interest to determine which to prioritize as we build our models.

Figures 6-15 and 6-16 shows the spatial distribution of the 12 demographic variables and ten land use variables, respectively. Within the demographic variables, many seem potentially useful, particularly those relating to education, race, and income. The percent of residents between the age of 35 and 50 and the percent of residents over 65 seem potentially too uniform across the city to be useful, as does the average travel time to work.

Fewer of the land use variables collected seem promising, and future work should look into other sources of this data. Furthermore, many seem to be highly correlated with one another. We can examine this more deeply by calculating the pairwise

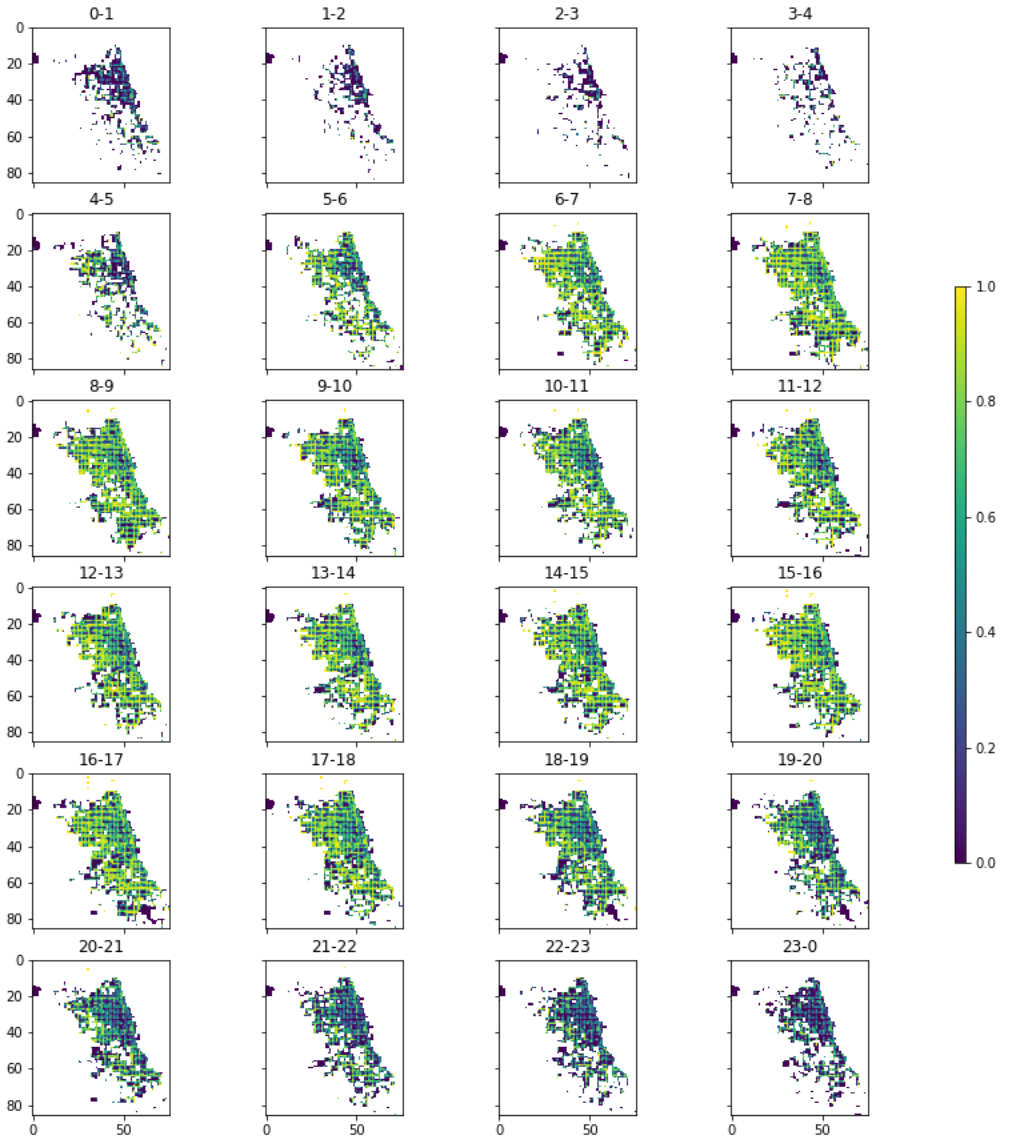


Figure 6-13: Public Transit Share of Public Transit and TNC Trip Origins on Thursday, October 3

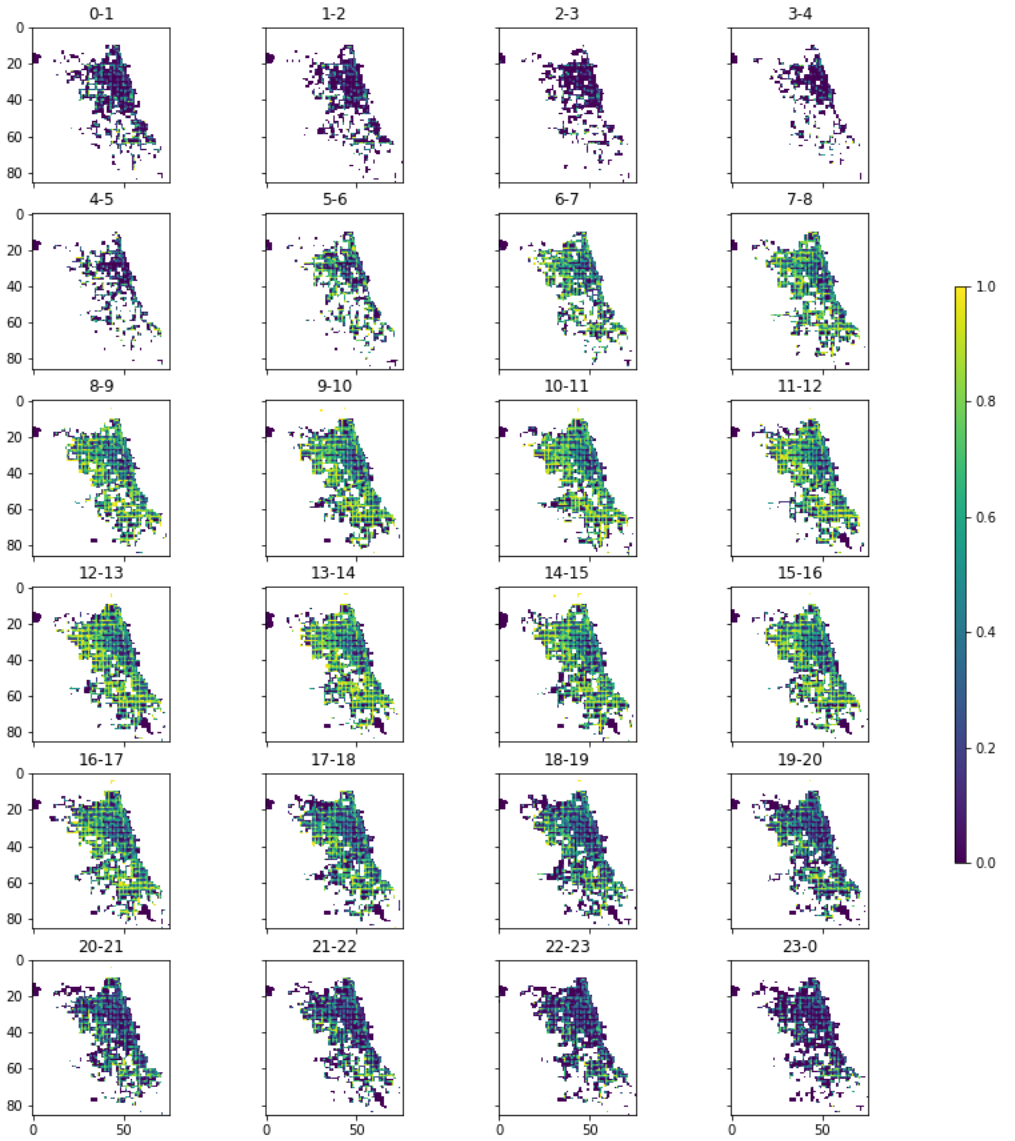


Figure 6-14: Public Transit Share of Public Transit and TNC Trip Origins on Saturday, October 5

Pearson's correlation coefficient for all demographic and land use variables. The heatmap of these correlation coefficients is shown in Figure 6-17.

Indeed, within the land use category, we note that the count of bars is highly correlated with the count of restaurants and the count of all points of interest. In our models, we will use only the count of all points of interest as a control variable to isolate other relationships as best we can.

We note several correlations among the demographic variables of which we should be wary. Specifically, the correlation of the percent of white alone residents and the percent of black alone residents is nearly -1, so we will use only one of these. Income per capita and percent of residents with college graduates are also highly correlated.

6.3.3 Relationships between Spatial Covariates and Trip Volumes

We also explore the fundamental relationship between these spatial covariates and trip volumes by calculating the correlation coeffiecients between each covariate and, in turn, average weekday public transit trips per grid cell, average Saturday public transit trips, average weekday TNC origins, average Saturday TNC origins, average weekday TNC destinations, and average Saturday TNC destinations. The correlations for each are shown in Figure 6-18.

We note the highest correlations between bars, restaurants, and all points of interest and TNC trip counts (origins and destinations). Interestingly, the only correlations that approach being negative are that of the percent of black residents and all the trip volume measures, specifically TNC counts. Among the demographic variables, income, percent of college educated residents, percent of peopled between 25 and 34, and total population are most highly correlated with TNC trip volume counts. For each spatial attribute, correlation values tend to be similar for all mode trip count variables, though there is often separation between the TNC measures and the public transit measures.

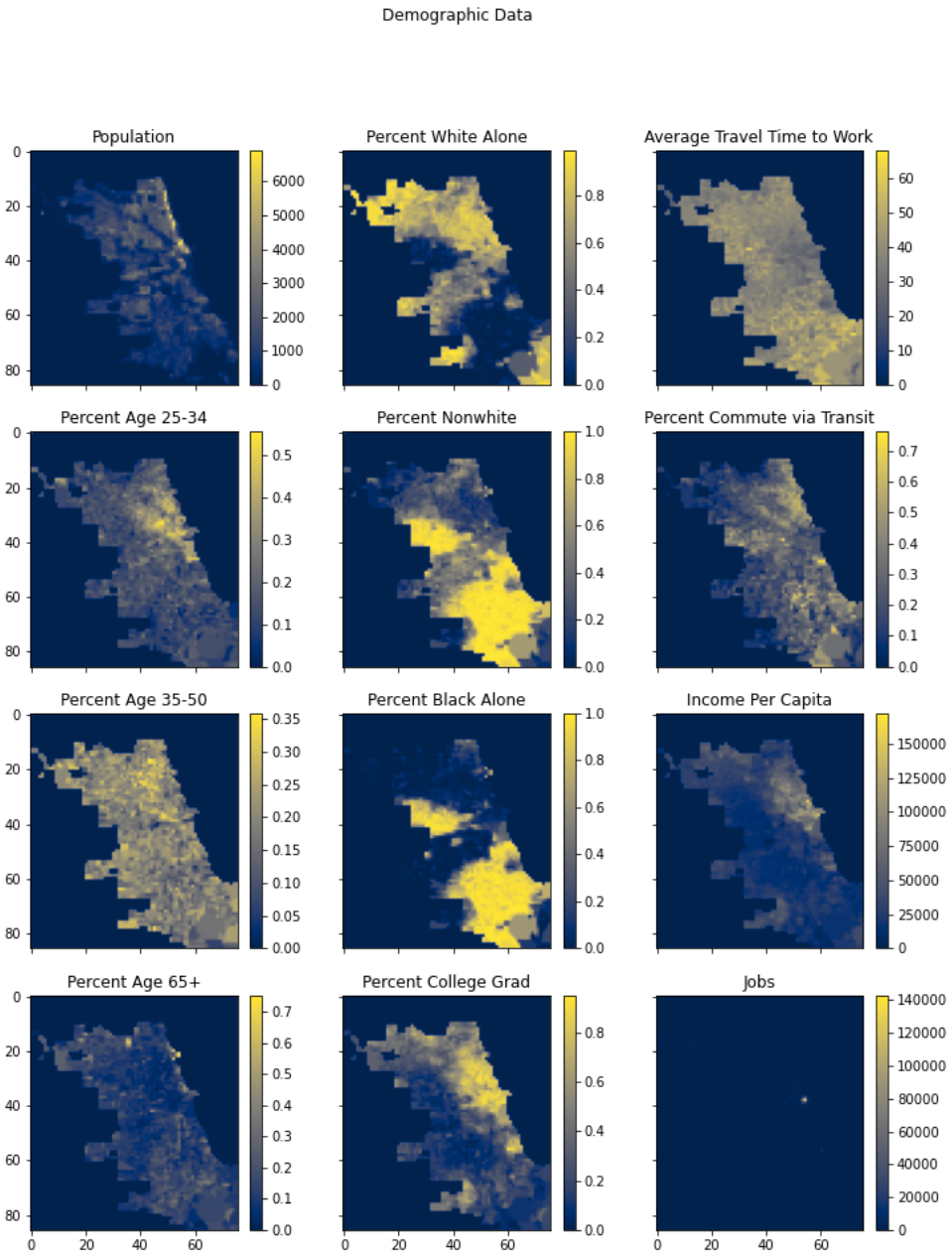


Figure 6-15: Spatial Distribution of Demographic Variables

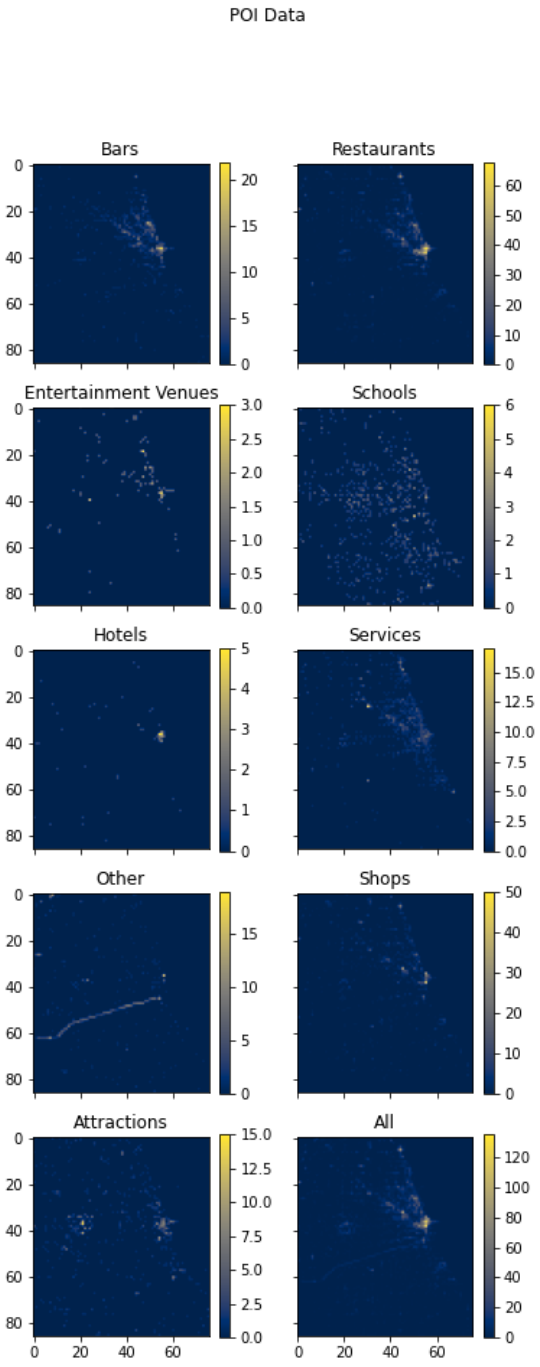


Figure 6-16: Spatial Distribution of Land Use Variables

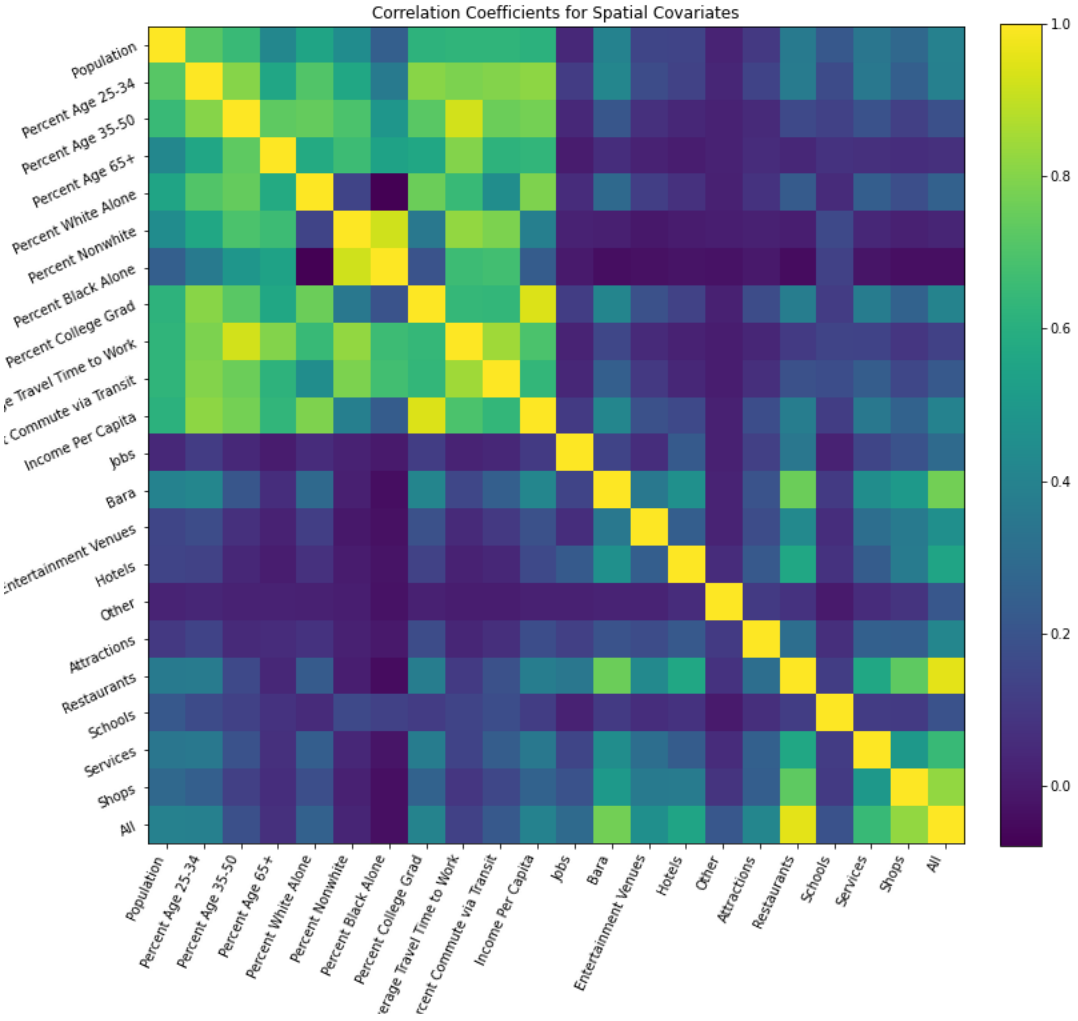


Figure 6-17: Correlation Heatmap for Demographic and Land Use Variables

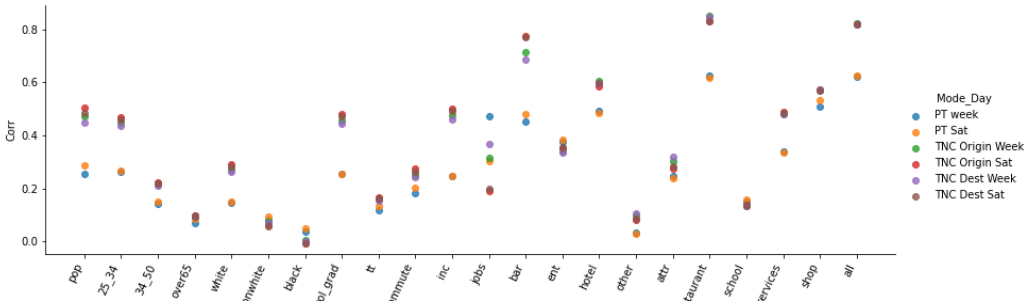


Figure 6-18: Correlations Between Spatial Covariates and Trip Count Volumes by Mode and Time of Week

6.4 Initial Model Results

In this section, we will present results for a set of initial models run with TNC origin counts as the dependent variable. The aim of this section is to demonstrate the flexibility of this framework and present some preliminary findings about the spatio-temporal patterns of the origins of TNC trips.

6.4.1 Temporal Model

In this first model, we regress TNC origin counts at each hour and location on only temporal data. Specifically, we include dummies for each hour of the day and dummies for each hour of the day multiplied by a dummy indicating if it is a weekday or not. The model is as follows:

$$TNC_{i,j,t} = \beta_0 + \sum_{h=1}^{23} w_h HOD_t + \sum_{h=1}^{23} w_{wh}(HOD_t \times WK_t) \tag{6.4}$$

In equation 6.4 TNC_t is the number of TNC origins at hour t. Because there are no spatial inputs to this model, the output will be the same value for every grid cell. HOD_h is a dummy variable indicating if it is the h-th hour of the day, leaving the 0th hour (12AM-1AM) as the base, WK_t is a dummy variable indicating if the t-th hour occurs on a weekday, and w_h and w_{wh} are the parameter estimates. Specifically, w_h is a 23×1 vector giving the relative change in mean trip totals by each hour of the day, and w_{wh} is a 23×1 vector showing the impact that it being a weekday has on the corresponding hour parameter for w_h.

The model results are given in Table 6.2. The model output for each time point is the average count of TNC origins across all cells in the grid. This explains the small magnitudes, as most grid cells have a count of 0 TNC trip origins at any given time, as well as the very low R-squared, which is extremely close to 0 because the spatial dimension of the data ($86 \times 76 = 6,536$) is so much higher than the temporal dimension (31 days times 24 hours = 744). We have confidence that the temporal patterns detected by the model represent truth, however, because of how closely the

predicted trip volume summed over space matches with true trip volume summed over space (Figure 6-19) We can still learn from this model, even though the prediction performance would be very poor. We can interpret the β_0 coefficient as the average number of trip origins across all cells between 12AM and 1AM, the w_h vector as the change in mean relative to β_0 by hour for weekends, and the w_{wh} vector as the change to w_h when it is a weekday.

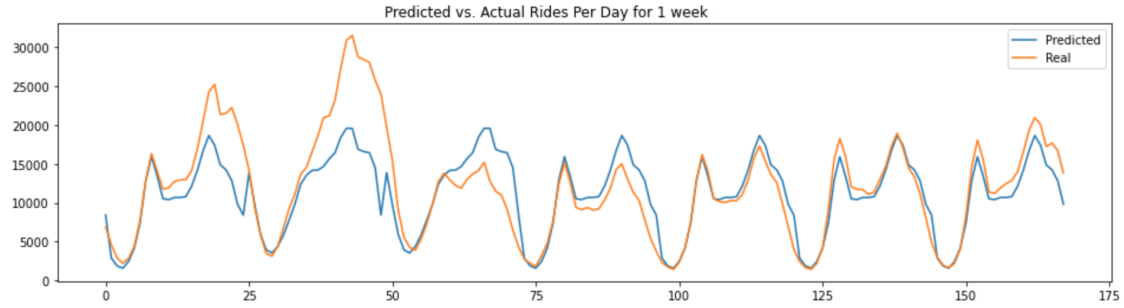


Figure 6-19: Predicted vs. Real Total TNC Trip Counts for 1 Week - HOD dummies for Week and Weekend

We see that, on weekends, volume is lowest between 4AM and 7AM and highest from the afternoon into the evening and night, with the peak between 7PM and 8PM. On weekdays, volume is lower at most hours, except between 5AM and 10AM and between 6PM and 7PM. This biggest difference is between 12AM and 2AM.

6.4.2 Spatial Model

Next, we run a model where the only inputs are spatial data. The result is that the outputs vary only along the spatial dimension, predicting the same trip count at every time point. More specifically, the outputs are the average TNC trip counts for each grid cell. The model is given in equation 6.5.

$$TNC_{i,j,t} = \beta_0 + D_{i,j}w_d + POI_{i,j}w_{POI} + PTDENS_{i,j}w_{PTdens} \tag{6.5}$$

In Model 6.5, $\beta_0 1_{IJ}$ is the same as before, $D_{i,j}$ in this case is a 6536×3 matrix representing the vectorized values of three demographic variables— population,

$\beta_0 = 0.82$		
Hour	w_h	w_{wh}
1	1.30	−1.68
2	0.70	−1.23
3	0.11	−0.69
4	−0.19	−0.26
5	−0.23	0.04
6	−0.12	0.43
7	0.10	1.04
8	0.39	1.23
9	0.76	0.47
10	1.14	−0.35
11	1.29	−0.51
12	1.39	−0.58
13	1.37	−0.55
14	1.47	−0.63
15	1.63	−0.60
16	1.66	−0.29
17	2.09	−0.35
18	1.92	0.18
19	2.21	−0.37
20	1.80	−0.34
21	1.76	−0.40
22	1.73	−0.59
23	1.44	−0.76
R^2	0.004	

Table 6.2: Temporal Model Parameter Estimates

percent of black residents, and income— in each grid cell, and w_d is a 3×1 vector of parameter values for each of those variables. $POI_{i,j}$ is a 6536×1 vector of the normalized count of all types of points of interest variables by grid cell, with w_{POI} the scalar parameter estimate for this. Lastly, $PTDENS_{i,j}$ is a 6536×2 matrix giving the density of rail and bus stops respectively in each grid cell, and w_{PTdens} is the 2×1 parameter vector.

$\beta_0 = -1.11$	
Variable	Estimate
Population	7.57
Percent Black	−2.29
Median Income	10.49
POI count	111.30
Rail stop density	30.61
Bus stop density	32.24
R^2	0.425

Table 6.3: Spatial Model Parameter Estimates

The results are shows in Table 6.3. We note a negative value of β_0, while hints at one of the issues of these models. Every grid cell is included in the model, even though we only have data for a fraction of them, with several being located outside the city bounds or even in Lake Michigan. Because most of them have a value of zero for all variables, we are still able to capture relationships between our independent and dependent variables, but it obscures the meaning of some of the spatial variables. For example, the coefficient on percent black here is negative, as we would expect, but very small, which is likely because cells with a 0 value for this variable are often places with no trips, because there is no data at all, but are also frequently places where all or nearly all residents are not black, and these may be places with high volumes of trips. This strongly biases the parameter estimate toward zero. The same will happen for other demographic variables where 0 values may occur within the city bounds in a way that is meaningfully associated with trip volumes. Future work should make sure to correct this, perhaps by only including cells within city bounds.

This model does reveal the importance of the concentration of points of interest, however. In future models we will remember to control for this value. Furthermore, both rail and bus stop density are associated with significantly more TNC trips. The demographic variables are less important.

6.4.3 Spatio-Temporal Model

Next, we can combine spatial and temporal variables to capture some of the more complicated dynamics in the mobility data. In this example, we control for the number of points of interest and whether or not in rained on the day in question and investigate how TNC trips vary by hour of day and how this changes based on whether a rail or bus station is nearby. The model is given in equation 6.6.

$$\begin{aligned} TNC_{i,j,t} =& \beta_0 + POI_{i,j} w_{POI} + RAIN_t w_{rain} \\ &+ \sum_{h=1}^{23} HOD_t w_h + \sum_{h=1}^{23} HOD_t FRI_t w\_fri_h \\ &+ \sum_{h=1}^{23} HOD_t WKEND_t w\_wkend_h \\ &+ \sum_{h=1}^{23} HOD_t RAIL\_BOOL_{i,j} w\_r_h \\ &+ \sum_{h=1}^{23} HOD_t RAIL\_BOOL_{i,j} FRI_t w\_r\_fri_h \\ &+ \sum_{h=1}^{23} HOD_t RAIL\_BOOL_{i,j} WKEND_t w\_r\_wkend_h \\ &+ \sum_{h=1}^{23} HOD_t BUS\_BOOL_{i,j} w\_b_h \\ &+ \sum_{h=1}^{23} HOD_t BUS\_BOOL_{i,j} FRI_t w\_b\_fri_h \\ &+ \sum_{h=1}^{23} HOD_t BUS\_BOOL_{i,j} WKEND_t w\_b\_wkend_h \end{aligned} \tag{6.6}$$

Once again, HOD_t is a dummy variable that is 1 if the time period t corresponds to the hour of day h from the summation. FRI_t is a dummy variable that is 1 if the time period t falls on a Friday, and $WKEND_t$ is a dummy variable that is 1 if the time period t falls on a weekend. $RAIL\_BOOL_{i,j}$ is a dummy variable that is 1 if grid cell (i, j) contains a rail station, and $BUS\_BOOL_{i,j}$ is a dummy variable that is 1 if grid cell (i, j) contains a bus station. $POI_{i,j}$ is as before, and $RAIN_t$ is a dummy variable equal to 1 if t fell on a day where rain was recorded.

This model allows us to estimate, while controlling for points of interest and rain, the impact that each hour of the day has on the number of TNC trip origins observed in the following cases:

- On a Monday-Thursday when there are no public transit stations in the same grid cell (w_h)
- On a Friday when there are no public transit stations in the same grid cell (w_fri_h)
- On a Saturday or Sunday when there are no public transit stations in the same grid cell (w_wkend_h)
- On a Monday-Thursday when there is a rail station in the same grid cell (w_r_h)
- On a Friday when there is a rail station in the same grid cell ($w\_r\_fri_h$)
- On a Saturday or Sunday when there is a rail station in the same grid cell ($w\_r\_wkend_h$)
- On a Monday-Thursday when there is a bus station in the same grid cell (w_b_h)
- On a Friday when there is a bus station in the same grid cell ($w\_b\_fri_h$)
- On a Saturday or Sunday when there is a bus station in the same grid cell ($w\_r\_wkend_h$)

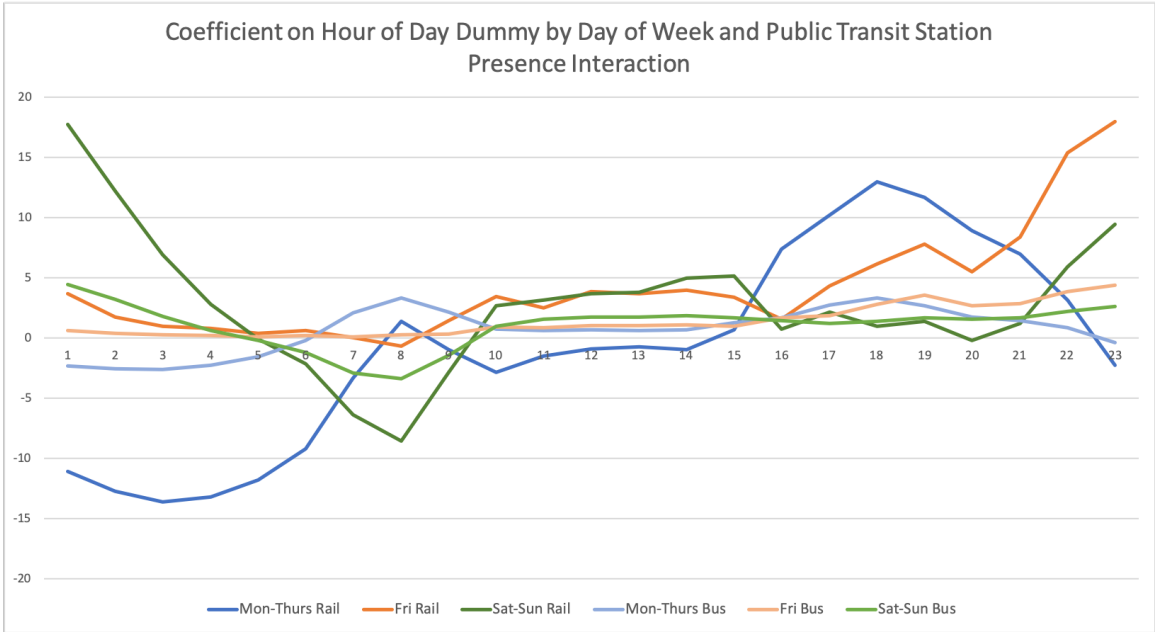


Figure 6-20: Hour of Day Coefficients for Each Day of Week and Station Presence Interaction

The resulting model has an R-squared of 0.43 and estimates β_0 at -0.67, w_{POI} at 170.7, and w_{rain} at 0.17. The estimates for w_h, w_fri_h, and w_wkend_h are all under 0.25. The remaining coefficients are plotted by the value of h in Figure 6-20

We note the biggest change throughout the course of the day for grid cells with rail stations on Monday-Thursday. In the morning on these days, grid cells with rail stations see fewer TNC trips. As the morning peak approaches, the effect becomes neutral and remains slightly negative through the midday until the beginning of the afternoon peak, and which point grid cells with rail stations see significantly more TNC trips, suggesting that people may be opting for TNC trips over public transit at the end of the workday.

On Friday, grid cells with rail stations see a different pattern, not seeing significantly more TNC trips until around 5PM, at which point the number of trips grows for a few hours, drops around 8PM, and then climbs sharply through midnight, revealing that throughout Friday nights, TNC trips become more and more concentrated around rail stations.

The early morning Saturday and Sunday coefficients for grids with rail stations

picks up very close to where the Friday evening one drops off, then drops steadily until becoming associated with fewer TNC trips around 8AM, and then remaining mostly neutral for the remainder of the day until the late evening.

The coefficients for grid cells with bus stations are smaller, but we do observe more subdued versions of the same patterns as rail for each set of days. All in all, this model shows that, even when controlling for points of interest, TNC trips tend to congregate around rail stations in the early evenings during the week and the late evenings on weekends. While we cannot conclude from this model alone that these trips could be taken on rail, their origins being so close to a rail station makes it likely. This model provides some evidence that particularly in the evenings, TNCs are being used to replace rail trips. On the other hand, there is also evidence that people opt for TNCs over public transit particularly when frequency is likely to be low, or perhaps service has stopped completely. Further investigation controlling for frequency may explain this further.

6.4.4 Spatio-Temporal Models with a Spatial Lag

We can also investigate the extent to which TNC demand is related to TNC demand in neighboring areas or public transit demand in neighboring areas. For these models, we use only a 30×25 inset of our original grid containing the downtown area for computational efficiency.

The first model uses only the spatial lag of the dependent variable as the explanatory variable. Mathematically, the model is given in equation 6.7

$$TNC_t = \beta_0 + \rho W * TNC_t + \epsilon_t \tag{6.7}$$

where TNC_t is the $I \times J$ grid of TNC origin counts at time t, W is the spatial weighting matrix (in this case the normalized Rook's spatial weights matrix, where each grid cell has as its neighbors the at most four with which it shares an edge), and ρ is the spatial correlation coefficient estimated by the model.

This simple model performs rather well, yielding an R-squared of 0.74. The estimated value of β_0 is -0.50 and the estimated value of the correlation coefficient ρ is 1.04, suggesting that the number of TNC trips originating from a grid cell at any given hour is typically about the average of the trip origin counts in neighboring cells. The predicted and real values for total trip counts in this inset for a week are shown in Figure 6-21.

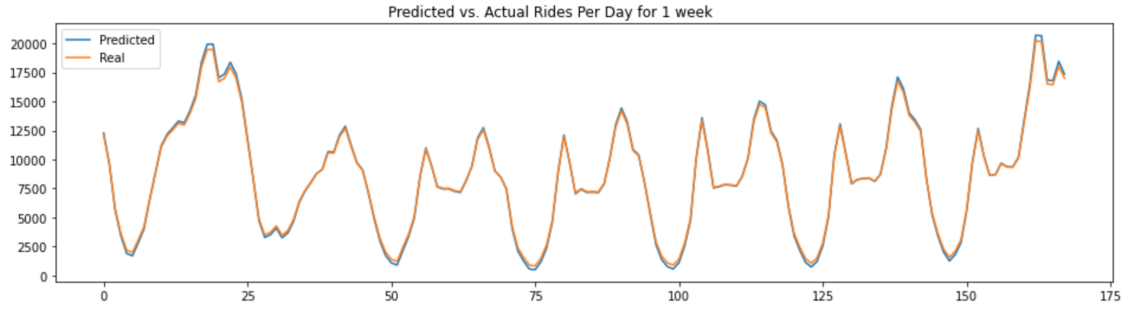


Figure 6-21: Predicted vs. Real Total TNC Trip Counts for 1 Week - Spatially Lagged Dependent Variable

We can furthermore explore the relationship between TNC origin trips counts in a cell and public transit trip counts in that cell and neighboring cells. The model is as follows:

$$TNC_t = \beta_0 + w_1 PT_t + \lambda W * PT_t + \epsilon_t \tag{6.8}$$

where TNC_t is the $I \times J$ grid of TNC origin counts at time t, PT_t is the $I \times J$ grid of public transit (rail + bus) origin counts at time t, W is the spatial weighting matrix, and λ is the spatial correlation coefficient estimated by the model.

This model performs markedly worse than the model with a lagged dependent variable, yielding an R-squared value of only 0.26. The estimated value for β_0 is 6.67, and the estimated values for w_1 and λ are 0.012 and 0.107 respectively. These parameter estimates indicate that TNC demand is positively correlated with the public transit demand occurring around it, but the relationship only explains about a quarter of the variation in TNC trip count usage, and TNCs see typically much smaller

volumes of trips that public transit does. Figure 6-22 illustrates the performance of this model on a spatially aggregated level for a week, revealing that this model performs much better on weekdays than on weekends. Further exploration of the prediction power of models interacting public transit demand and temporal variables would be interesting.

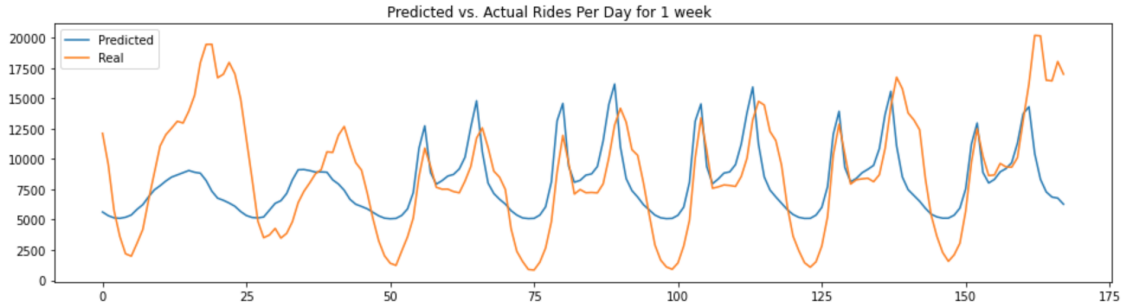


Figure 6-22: Predicted vs. Real Total TNC Trip Counts for 1 Week - Spatially Lagged Public Transit Usage

6.5 Thoughts on Future Directions

This chapter offers only a few preliminary models exploring the spatio-temporal patterns of TNC trips in Chicago and how those relate to public transit usage. There are many potential extensions to this work, including improvements to data quality and handling as well as extensions of the model formulations.

6.5.1 Data Improvements

There are several opportunities to improve the data used in these models, which we will list here.

- **Land Use Data** The information on land use obtained from Open Street Maps for these analyses is sparse. While the total counts of points of interest do seem to capture the activity centers of the city and provide substantial predictive power, other land use data regarding design elements such as the width of sidewalks, amount of green space, and number of intersections could be included

to understand how these variables impact TNC ridership or TNC ridership vis a vis public transit ridership.

- **Special Events** Information on major events in the area would be important to include in these models to capture modal responses to irregular occurrences. Included in this should be data on public transit service disruptions.

- **Public Transit Level of Service** Frequency is only one element of the level of service of the public transit system at any given point in space and time. Furthermore, scheduled frequency, as is used here, is less useful than actual frequency in summarizing how a rider likely perceives the level of service to be. Reliability and capacity would also be valuable dimensions to add to a level of service metric. Understanding the role of public transit service levels in the relative volume of usage on TNCs would be a valuable question to use this framework to answer.

- **Handling of Grid Cells Outside City** A way to leverage the matrix-based implementation of these models while ignoring the cells in the grid with missing data would greatly benefit the performance of the model by not falsely correlating zero TNC rides with zero values for other explanatory variables. Because the grid is vectorized at each time point for the model estimation anyway, this likely amounts to including in the vector only those cells within the city bounds.

6.5.2 Model Formulations

In addition to improving data quality and handling, certain adjustments to the model formulations may allow for deeper insights into how public transit and TNCs coexist in a given city. I outline a few ideas here.

Temporal and Spatial Lags

Spatial Lags

In the models presented here that use a spatial lag, only a very simple version of the spatial weights matrix is used: namely, one that considered the (typically) four grid cells with which any individual cell shares an edge to be a "neighbor" and thus capable of exerting direct influence on the dependent variable of the cell of interest. While this likely approximates local effects across variables, there is no reason to believe it is the best representation of the spatial dependencies among any set of variables.

A worthwhile expansion of these models would explore various formations of spatial weighting matrices, including perhaps different weights matrices for different independent variable. One example would be a weights matrix that captures the network structure of the transit system, considering cells to be "neighbors" if they are on the same transit line, for example, or weighting the correlation by the travel time between the two cells on public transit. Such a formulation would be particularly interesting for independent variables related to the transit network.

Temporal Lags

This chapter did not offer results for any models containing temporal lags, but they have a clear role to play within the goal of exploring the relationship between demands for different modes. Further work should explore various lengths of temporal lags on the dependent variable or the usage of other modes.

Additionally, exploration of the combination of spatial and temporal lags could protentially prove fruitful. A systematic exploration of various formulations in this fashion could lead to significant insight into the ways one or many modes' demands exhibit spatio-temporal correlation, and perhaps suggest ways in which people use TNCs to supplement public transit.

Longitudinal Analysis

The models presented here used data from a single month to describe dimensions of TNC usage patterns as they existed on average in October 2019. They could be extended to investigate longitudinal changes in parameters by incorporating data from multiple time periods and using dummy variable to indicate from which period

each data point comes. This could be used year over year to identify times and areas where public transit is losing out to TNCs, for example, or the time frames could be from before and during the COVID pandemic to capture how the dynamics of a single mode or the relationship between modes changed as a result of the virus.

Modeling Origin-Destination Flows

Instead of using just origin or destination counts as the dependent variable, the model could be structured so that the dependent variable was O-D flows. Depending on the question one hoped to answer, this would require reliable inferences of destinations to public transit origins, but such algorithms exist. Furthermore, it would greatly increase the size of the input data, and the modeler would need to have the requisite computing power. Nevertheless, it would allow for a much more complete picture of the interaction between modes, as it would highlight the corridors and times where usage patterns were most in line and where and when they were most divergent. Specifically, it could illuminate when and where and to what extent TNC trips occur on corridors served by transit, and what variables lead to more or fewer TNC trips along such routes. This knowledge could inform policy in a powerful way.

Machine Learning

Many of the examples mentioned above require exploration of dozens, if not hundreds of model formulations as one begins to consider the number of combinations of spatial lags and temporal lags and included variables. To systematically explore them all, one should employ machine learning methods to learn the most appropriate structure of the model, as measured by some criteria set by the modeler.

6.6 Conclusion

This chapter demonstrates the potential for insights on urban mobility patterns in an urban context given appropriate data. All spatial and temporal data in this chapter is available to the public, as is information on public transit scheduled frequencies. The

granularity of the public transit ridership data is easily accessible to analysts within a transit agency, though access to TNC data at this level is still relatively uncommon. It is likely that more cities will follow Chicago's lead and require data from TNCs in exchange for the opportunity to operate in those cities, which would open up this type of analysis to many more agencies. The modes in consideration need not be public transit and TNCs, however, and similar analysis could be conducted with scooter or bike share data.

The analysis conducted here regarding TNC origins in Chicago indicated that they are largely concentrated in the downtown, with significant numbers also originating along the coast in the north and along Milwaukee Avenue/the Blue Line. They typically occur in much lower numbers than public transit trips, except late on Fridays and in the afternoon and evening on Saturdays. TNC trip origins are largely co-located with points of interest, and are more likely to occur where there is high rail or bus station density. Compared with grid cells with no public transit stations, the presence of a rail station is associated with fewer TNC trip origins in the early morning Monday-Thursday and more TNC trips between 4PM and 9PM. The times and places in the week most associated with higher numbers of TNC trips are late Friday night into Saturday morning in grid cells containing rail stops, suggesting that TNCs may prove particularly popular when public transit frequency is low.

When exploring the relationship between TNC trip origin counts and the counts from neighboring grid cells, we find that TNC trip origins are very spatially correlated, and that taking the average value of trips from the neighboring cells alone provides very strong predictive power. TNC trips are less correlated with public transit trips from the same or neighboring grid cells, but the correlation is still positive. The relationship appears stronger during the week than on weekends.

This framework can be used, as it was here, to explore parameters describing the dynamics of TNC usage patterns in Chicago, or the formulation and variables can be selected to answer a specific question. Regardless, there is significant potential to apply this framework in new ways to deepen our collective understanding of the usage patterns of new mobility modes.

Chapter 7

Conclusion

The initial impetus for this work was to develop a framework that would allow transit agencies to leverage the rich data available to them in order to develop a deeper understanding of their riders and use this to inform policy. The underlying philosophy is that analysis based solely on counts of trips, be it by hour or day or mode or route, only tells part of the story. With this as the only analytical tool, a dip in trip counts is just that. There is no differentiation between a drop due to riders leaving the system entirely, or one due to riders remaining in the system by decreasing their use of it. Despite the result being the same, these types of ridership loss could lead to quite distinct policy interventions. In truth, the behavioral dynamics behind a drop in ridership on a major urban mass transit system are likely quite complex, but with the proper data and tools, the dominant forces can be teased out and policies tailored accordingly. Thus, this work makes the case that, in conjunction with an understanding of aggregate trip volumes must come a partner analysis that keeps the *rider* as the fundamental unit of analysis, as the rider will ultimately be the one observing and responding to any policy intervention.

After the COVID-19 virus began to spread quickly in America, cities and states imposed stay-at-home orders across the country and workplaces, schools, and non-essential businesses were ordered to close. The result was a sudden and sustained drop in public transit ridership across the country to a degree never before seen. As transit agencies continue to learn about the new ridership dynamics on their system

and craft creative policies to entice riders safely back onto trains and buses, the framework offered here, specifically in chapter 4, can be used as a model for how an agency can orient policy around the needs of its riders.

7.1 Summary of Findings

7.1.1 Year Over Year Ridership Behavior Changes

Chapter 3 of this work addresses the question of who is driving the top-level year to year trip loss observed on the CTA. We cluster data from four months of 2017 and 2018 each and observe changes in the number of cards exhibiting each of 10 key behaviors. We also investigate the number of cards exhibiting different kinds of behavior changes. The key findings are as follows:

- The decrease in the number of trips taken on the system between 2017 and 2018 is driven more by individuals taking fewer trips in 2018 compared with 2017 than it is due to riders leaving the system altogether.

- There is a trend of riders tending to ride less often over time, regardless of their initial behavior.

- Regular ridership during weekday peak hours has decreased to a lesser extent than during other periods of the day or week.

- The group of riders whose behavior is characterized by regular peak ridership was the only group of occasional or regular riders to grow between 2017 and 2018, and it was due to the number of riders new to the system with this behavior outnumbering the number of riders churning/leaving the system or changing their behavior.

- These regular peak riders were a key reason why the CTA outperformed ridership and revenue projections for the year despite implementing a fare increase in January of 2018.

7.1.2 Public Transit Ridership in Chicago During the COVID-19 Pandemic

After March of 2020 and the growth of the COVID-19 outbreak within US borders, the goal of explaining a 2% change in ridership between years began to seem quaint as transit agencies across the country saw ridership plummet by 70%-90%. Chapter 4 of this work applied the same customer segmentation framework established in Chapter 3 to analyze which groups of riders were more or less responsible for the precipitous drop in trips taken on the CTA by investigating rates of complete churn by behavioral group. Chapter 5 then built on this work by using linear and spatial regression to identify which behavioral and demographic variables were most predictive of transit ridership loss at the census tract level. The explanatory variables were the demographic traits of the census tracts and average behavioral attributes for riders with inferred home locations in that tract. The summarized behavioral attributes were largely the same ones used as inputs to the clustering algorithm in previous chapters. The key findings are as follows:

- Two key behavioral groups – frequent peak rail riders and frequent off-peak bus riders with high transfer rates – exhibited drastically different ridership responses to the pandemic and subsequent stay-at-home order. The former saw 93% of its riders stop riding initially, while only about half of the latter group stopped riding.

- Frequent peak rail riders were more likely to live in higher income, white neighborhoods, while frequent off-peak bus riders with high transfer rates were more likely to live in lower income, majority-minority neighborhoods.

- Including both summary behavioral and demographic attributes in a model predicting ridership drops due to COVID significantly improved the model fit compared with models using either but not both of these groups of explanatory variables.

- Even when controlling for demographics, the typical behavior of riders in a tract

was important in predicting the COVID ridership response.

- The decrease in ridership intensified as the share of rides typically taken at traditional peak times among riders grew. Conversely, the decrease in ridership lessened with higher usage of bus and higher rates of transfer, as well as a higher proportion of riders using a pass product as opposed to pay-per-use.

- Demographically, the percent of black residents, Spanish-speakers, and percent of households without a vehicle were associated with smaller drops in ridership, while the percent of young people and foreign-born residents was associated with larger drops in ridership.

7.1.3 Usage Patterns of TNCs in Chicago

The final analytical portion of this work offered initial thoughts on how to consider the potential influence of other modes on people's use of public transit and behavior changes over time when the available data are public transit usage and non-identified trips of an alternate mode, in this case TNCs. We offer a systematic way of capturing the spatial, temporal, and spatio-temporal attributes of usage patterns of a given mode and offer preliminary results describing TNC usage in Chicago in October of 2019:

- TNC trip origins in Chicago are heavily concentrated downtown, with significant numbers also originating along the north coast of Lake Michigan and Milwaukee Avenue/the Blue Line to the northwest.

- TNC trips only outnumber public transit trips late on Friday evenings and on Saturday in the afternoon and evening. During the weekdays there are far more public transit trips occurring at any given time between 6AM and 8PM than TNC trips.

- Even controlling for count of points of interest, which is consistently a very strong predictor of the number of TNC origins, density of rail and bus stops are associated with higher numbers of TNC trip origins.

- For Monday-Thursday, grid cells with rail stations see fewer TNC trips, but the presence of a rail station in the afternoon is associated with significantly higher numbers of TNC trips. Late night hours on Friday into Saturday mornings see many more TNC trips in areas where there is a rail station.

- TNC trips are highly spatially correlated and 74% of the variation is explained by averaging the counts of TNCs in neighboring cells. On the other hand, 26% of the variation is explained by the public transit usage in and around the cell.

The modeling framework presented in this chapter could be extended in numerous ways, such as modifying the spatial weights matrix and adding temporal lags to locate the strongest spatio-temporal correlations between TNC and public transit usage, or by adding data from a subsequent month or year to capture how the inter-modal dynamics have changed over time. Such information could inform simultaneous changes in public transit ridership behavior and suggest what types of changes might be related to competition with TNCs.

7.2 Recommendations

Out of this work emerges several recommendations for the CTA, who funded this research and offered its data for the myriad analyses presented here. The recommendations offered here, separated into analysis recommendations and policy recommendations, are focused on helping the CTA understand and recover ridership as Chicago and the world continue to deal with the COVID-19 virus, recognizing this to be the preeminent challenge facing transit agencies today.

7.2.1 Analysis Practices

The overarching suggestion for the CTA is to incorporate the customer segmentation framework into regular ridership analysis, when possible establishing stable behaviors on the system as was done in Chapter 3, or, as was done in Chapter 4, using baseline behavior groups to track ridership changes over time by segment. The latter option

is the most relevant now as transit agencies continue to face unprecedentedly low ridership, but can be supplemented by periodic clustering of riders using the most recent data to understand what pandemic-era ridership looks like as it evolves. This analysis will not only continue to offer valuable insights into how people are using the system and what behaviors are behind overall trip counts, but will also center people in such a way that facilitates the formation of policy geared at riders, as the connection between the results of the analysis and the person that is the target of a policy become stronger and more obvious, as was seen in Chapter 4.

Specific actions the CTA can take include the following:

- Using the baseline clustering results from Chapter 4 or a modified version of their choosing, assign all Ventra cards present during the baseline period a cluster label and store this information in a data table that can be linked to other tables on the account ID. This will facilitate continued monitoring of COVID ridership behavior rooted in knowledge of individuals' pre-pandemic behaviors.

- Using the same set of inputs as the baseline clusters, run the k-means clustering algorithm on all cards active during different phases of the pandemic. Because of the much smaller number of active cards, there may not need to be as many clusters. Use the resulting clusters to understand the new predominant behaviors on the system. Investigate the number of cards by pre-pandemic and pandemic cluster assignment to determine patterns in how people have altered their ridership behavior.

- Periodically re-cluster cards on the system based on data from more recent time frames. Investigate the extent to which the resulting clusters are similar to those from the previous time period. If they are, analysis like that in Chapter 3 can be done to gauge which behaviors are most prevalent among riders re-entering the system, and whether people who have been riding during the pandemic are exhibiting significant behavior changes.

- Overlay information on inferred home location and other data points of interest to identify geographic patterns to behaviors, as this will allow for more targeted policies.

Analysis along this vein will enable the CTA to understand how people first returning to the system are interacting with it and potentially glean information about their mobility needs. This could inform decisions on fare policy structures; for example, new passes could be designed that better reflect the behaviors of people using the system.

7.2.2 Policy Design

The analysis from Chapters 4 and 5 revealed two crucial but distinct ridership challenges facing the CTA going forward. The first is the need to bring people with other travel options back onto the system, such as the frequent peak rail group. This is important not only because of the size of this contingent but also because these individuals will likely be opting for less sustainable modes of transportation to replace public transit, especially as temperatures get colder and active modes of transportation become less appealing. Policy recommendations rooted in this analysis— specifically that this group is more likely to be younger, live on the northside, predominantly use rail, and make use of the Ventra app — include

- Outreach via smartphone notification or app-based information. Information on CTA's sanitation procedures, crowding level of trains, and the lack of evidence that riding transit puts one at significant risk of transmission may be particularly useful.

- Undertaking education campaigns about alternate routes available, specifically between the northside and downtown, which is well-served by bus as well as rail. These would be particularly effective if coupled with crowding information for these buses and trains.

- Exploring partnerships with local businesses and restaurants who may be eager to attract patrons and willing to offer discounts to people who ride the CTA.

The other major ridership challenge facing the CTA is to make the system feel safe and efficient for those who have needed it all along. A key finding of this work is that those most reliant on public transit—so much so that they continued to use it during a global pandemic when citizens were advised against taking mass transit – were riders who, despite riding often, did not use the system at its busiest times. A significant implication of this is that policies that direct resources to places and times when the system is busiest, or make it difficult to add service in the off-peak, systematically harm riders who are most reliant on the system. Thus, policies geared toward this group must not only seek to improve the system for them during the pandemic when they constitute the majority of riders, but going forward, as they no doubt continue to be reliable users of the system.

Specific policy actions the CTA can take include:

- Continuing to shift resources during the pandemic to provide as much capacity to routes that are seeing relatively high volumes.

- Work with the city of Chicago to capitalize on the low levels of car traffic during the pandemic to add more bus lanes in order to increase speed and reliability on bus routes. As many of the riders remaining on the system rely primarily on bus and frequently have to transfer among buses, improved service on bus routes will have a compounding effect for these riders.

- Ally with activists to lobby the state to revise outdated public transit funding mechanisms, particularly mandated recovery ratios that lead to significantly longer headways in the off-peak and on weekends. Use this work as evidence that the most frequent users of the system ride when few other people are on the network, so direction of resources away from these parts of the system hurts exactly those individuals who stand to benefit the most from increased investment.

If the CTA is able to improve bus speeds and reliability via bus lanes and offer more frequent service in the off-peak, the whole city stands to benefit tremendously, not only the riders who have historically used these aspects of the service. These are exactly the steps that need to happen for rail riders to be enticed to use the bus when it is available, or for occasional riders to increase the frequency with which they use the service. Closing the gap between the level of service on bus and rail will make transit more competitive with other travel modes and help ensure its continued place as an essential facet of urban life in America. While there are many aspects of the transit funding picture that are out of the CTA's control, opportunities for collective action with other stakeholders to demonstrate the necessity of these steps and lobby lawmakers should be sought after and capitalized upon.

7.3 Limitations and Future Work

7.3.1 Limitations

Despite the several strong findings highlighted above, there are several limitations to this study worth pointing out before offering thoughts on future work. First, all the customer segmentation in this work relies on the assumption that one Ventra card is equivalent to one person. We know that this is not universally true, and that there are likely patterns to where this assumption is more or less true. This study could be improved by a systematic plan for connecting multiple Ventra cards to the same person if possible or using all available knowledge to account for biases in levels of churn related to higher turnover of cards.

Furthermore, while the data from Ventra is generally very comprehensive and complete, the location of card taps on buses is occasionally undetermined, leading to a portion of primarily bus users having unidentified inferred home locations and being left out of analysis that required home locations of each rider. Because of this, bus riders would be undercounted in these analyses, or aggregations across riders would only include the bus riders with inferred home locations. If there was systematic bias

as to which buses logged locations and which did not, this could skew the results. Future work should, to the extent possible, use other trip information to infer a home location for each rider and determine if calculations need to correct for biases.

Lastly, in the COVID analysis, because of the rear-door boarding policy on buses, after two and a half weeks of the stay-at-home order, all bus Ventra tap data disappeared. As a result, our analysis of COVID ridership is limited to the two complete weeks immediately following the stay-at-home order and four additional weeks a few months later after front-door boarding was reinstated. Therefore, conclusions drawn about COVID-era ridership may be biased due to the limited time frame available for analysis.

7.3.2 Future Work

The Analysis Practices portion of the Recommendations section above outlines specific ways for the CTA to continue the work begun in this thesis. More broadly, the findings presented here suggest research questions that should be the focus of future work. These include:

- What are the driving forces behind the behavior changes observed due to COVID-19? How do different attitudes and changing life circumstances manifest in changed travel behavior, and do these vary by cluster? Surveys that can be linked back to cluster membership can address these research questions.

- What are the typical features of ridership behavior as one returns to the transit system after not riding for a significant duration of the pandemic? Does it happen gradually or all at once? What policies are successful in enticing people back to the system?

- What percent of behavior changes exhibited after the pandemic are/will be due to hesitancy to use public transit versus fundamental changes in one's mobility needs? Can past behaviors or other attributes of a rider predict which will be a more dominant factor in their changed behavior?

- How do land use characteristics relate to how much one reduced their travel on public transit during the pandemic, and in what ways?
- How did TNC use change due to the pandemic? In what ways was it similar or different to how public transit usage changed? Can this tell us anything about times and places absolute travel was down versus when people were more likely to be hesitant to use public transit and opt for a different mode?

The advent of the COVID-19 pandemic in the final quarter of the time frame for this work dramatically changed the public transportation landscape in America and shifted the goals of this analysis. What started as a framework for understanding the behavioral dynamics underlying a slow but steady dip in public transit usage each year became a way to capture the impact of the pandemic on public transit use in a major US city. At the time of this writing, America is still very much in the midst of grappling with the virus and its implications for the economy, schooling, transportation, and so many other things. A clear extension of this work should be the continued analysis of public transit ridership in a way that centers on the rider. Such analysis will not only help transit agencies craft policies aimed at helping their riders, but will also offer valuable information to society at large about the evolving mobility needs of different segments of the population and what this says about where urban life and mass transit ridership may be headed. Facing such uncertainty, there are a million things we can and should be doing to monitor the evolving situation and help bring into place versions of the future that are beneficial rather than harmful. In the case of public transportation, this framework offers a way to monitor which and how people are or are not re-entering the system, reach out to communities in need of extra resources, reassure riders wary of returning to mass transit, and inform policies that will promote a future where transportation is more sustainable and equitable than it was before.

Bibliography

[Agard et al., 2006] Agard, B., Morency, C., and Trépanier, M. (2006). Mining Public Transport User Behavior from Smart Card Data. *IFAC Proceedings Volumes*, 39(3):399–404.

[Akala, 2020] Akala, A. (2020). More big employers are talking about permanent work-from-home positions. *CNBC*. Library Catalog: www.cnbc.com Section: Workforce Wire.

[American Public Transportation Association, 2020] American Public Transportation Association (2020). *2020 Public Transportation Fact Book*. APTA Fact Book. 71 edition.

[Anselin, 1988a] Anselin, L. (1988a). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, 20(1):1–17. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1988.tb00159.x.

[Anselin, 1988b] Anselin, L. (1988b). *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Springer Netherlands.

[Asgari and Jin, 2020] Asgari, H. and Jin, X. (2020). Incorporating habitual behavior into Mode choice Modeling in light of emerging mobility services. *Sustainable Cities and Society*, 52:101735.

[Austrian Agency for Health and Food Safety, 2020] Austrian Agency for Health and Food Safety (2020). Epidemiologische AbklÃrung am Beispiel COVID-19.

[Basu, 2018] Basu, A. (2018). Data-Driven Customer Segmentation and Personalized Information Provision in Public Transit. Master's thesis, Massachusetts Institute of Technology.

[Berrebi and Watkins, 2020] Berrebi, S. J. and Watkins, K. E. (2020). Who's ditching the bus? *Transportation Research Part A: Policy and Practice*, 136:21–34.

[Berrod, 2020] Berrod, N. (2020). Coronavirus : pourquoi aucun cluster n'a été détecté dans les transports. *Le Parisien*. Library Catalog: www.leparisien.fr Section: /societe/.

[Bliss, 2020] Bliss, L. (2020). The New York Subway Got Caught in the Coronavirus Culture War. *Bloomberg.com*.

[Boisjoly et al., 2018] Boisjoly, G., Grisé, E., Maguire, M., Veillette, M.-P., Deboosere, R., Berrebi, E., and El-Geneidy, A. (2018). Invest in the ride: A 14year longitudinal analysis of the determinants of public transport ridership in 25 North American cities. *Transportation Research Part A: Policy and Practice*, 116:434–445.

[Briand et al., 2016] Briand, A.-S., Côme, E., El Mahrsi, M. K., and Oukhellou, L. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics*, 1(1):37–50.

[Briand et al., 2017] Briand, A.-S., Côme, E., Trépanier, M., and Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79:274–289.

[Cardozo et al., 2012] Cardozo, O. D., García-Palomares, J. C., and Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34:548–558.

[Center for Disease Control and Prevention, 2020a] Center for Disease Control and Prevention (2020a). Coronavirus Disease 2019 (COVID-19). Library Catalog: www.cdc.gov.

[Center for Disease Control and Prevention, 2020b] Center for Disease Control and Prevention (2020b). Coronavirus Disease 2019 (COVID-19) - Transmission. Library Catalog: www.cdc.gov.

[Cervero, 2007] Cervero, R. (2007). Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *Journal of the American Planning Association*, 72(3). Publisher: Taylor & Francis Group.

[Cervero et al., 2010] Cervero, R., Murakami, J., and Miller, M. (2010). Direct Ridership Model of Bus Rapid Transit in Los Angeles County, California:. *Transportation Research Record*. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[Cheng et al., 2019] Cheng, X., Zhang, R., Zhou, J., and Xu, W. (2019). DeepTransport: Learning Spatial-Temporal Dependency for Traffic Condition Forecasting. *arXiv:1709.09585 [cs]*. arXiv: 1709.09585.

[Chi and Zhu, 2020] Chi, G. and Zhu, J. (2020). Models Dealing with Spatial Heterogeneity. In *Spatial Regression Models for the Social Sciences*, number 14 in Advanced Quantitative Techniques in the Social Sciences. SAGE, Thousand Oaks. Library Catalog: us.sagepub.com.

[Chicago Transit Authority, 2020a] Chicago Transit Authority (2020a). Coronavirus Info (COVID-19) - CTA.

[Chicago Transit Authority, 2020b] Chicago Transit Authority (2020b). Cta - ridership - annual boarding totals. https://data.cityofchicago.org/Transportation/CTA-Ridership-Annual-Boarding-Totals/w8km-9pzd.

[Chow et al., 2006] Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., and Ubaka, I. (2006). Transit Ridership Model Based on Geographically Weighted Regression. *Transportation Research Record*, 1972(1):105–114. Publisher: SAGE Publications Inc.

[Clark, 2017] Clark, H. M. (2017). Who Rides Public Transportation. Technical report, APTA.

[Côme and Oukhellou, 2014] Côme, E. and Oukhellou, L. (2014). Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Vélib’ System of Paris. *ACM Transactions on Intelligent Systems and Technology*, 5(3):39:1–39:21.

[De la Garza, 2020] De la Garza, A. (2020). COVID-19 Has Been 'Apocalyptic' for Public Transit. Will Congress Offer More Help? *Time*.

[Dill, 2013] Dill, J. (2013). Predicting Transit Ridership at the Stop Level: The Role of Service and Urban Form. page 19, Washington, D.C.

[El Mahrsi et al., 2017] El Mahrsi, M. K., Côme, E., Oukhellou, L., and Verleysen, M. (2017). Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(3):712–728. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[Feigon et al., 2018] Feigon, S., Murphy, C., Transit Cooperative Research Program, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine (2018). *Broadening Understanding of the Interplay Between Public Transit, Shared Mobility, and Personal Automobiles*. Transportation Research Board, Washington, D.C. Pages: 24996.

[Fowler, 2020] Fowler, A. (2020). Starting March 30: New Muni Service Changes. Library Catalog: www.sfmta.com Publisher: San Francisco Municipal Transportation Agency.

[Gan et al., 2019] Gan, Z., Feng, T., Yang, M., Timmermans, H., and Luo, J. (2019). Analysis of Metro Station Ridership Considering Spatial Heterogeneity. *Chinese Geographical Science*, 29(6):1065–1077.

[Gehrke et al., 2018] Gehrke, S. R., Felix, A., and Reardon, T. (2018). Fare Choices Survey of Ride-Hailing Passengers in Metro Boston. Technical report, Metropolitan Area Planning Council. Library Catalog: www.mapc.org.

[Ghaemi et al., 2017] Ghaemi, M. S., Agard, B., Trépanier, M., and Nia, V. P. (2017). A visual segmentation method for temporal smart card data. *Transportmetrica A: Transport Science*, 13(5):381–404. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/23249935.2016.1273273.

[Giuliano, 2005] Giuliano, G. (2005). Low income, public transit, and mobility. *Transportation Research Record*, (1927):63–70.

[Goldbaum and Cook, 2020] Goldbaum, C. and Cook, L. R. (2020). They Can't Afford to Quarantine. So They Brave the Subway. *The New York Times*.

[Harris, 2020] Harris, J. E. (2020). The Subways Seeded the Massive Coronavirus Epidemic in New York City. page 22.

[He et al., 2020] He, L., Agard, B., and Trépanier, M. (2020). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. *Transportmetrica A: Transport Science*, 16(1):56–75. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/23249935.2018.1479722.

[Higashide, 2016] Higashide, S. (2016). Who's On Board 2016. Technical report, TransitCenter. Library Catalog: transitcenter.org Section: Reports.

[Higashide and Buchanan, 2019] Higashide, S. and Buchanan, M. (2019). Who's On Board 2019: How to Win Back America's Transit Riders. Technical report, TransitCenter, New York.

[Holshue et al., 2020] Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S. I., Kim, L., Tong, S., Lu, X., Lindstrom, S., Pallansch, M. A., Weldon, W. C., Biggs, H. M., Uyeki, T. M., and Pillai, S. K. (2020). First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine*, 382(10):929–936. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa2001191.

[Johns Hopkins University and Medicine, 2020] Johns Hopkins University and Medicine (2020). COVID-19 Map.

[Kelly, 2020] Kelly, J. (2020). Here Are The Companies Leading The Work-From-Home Revolution. *Forbes*.

[Kieu et al., 2013] Kieu, L. M., Bhaskar, A., and Chung, E. (2013). Mining temporal and spatial travel regularity for transit planning. In *Australasian Transport Research Forum*, Bisbane, Australia.

[Kieu et al., 2015] Kieu, L. M., Bhaskar, A., and Chung, E. (2015). Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1537–1548. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[Laura J. Nelson, 2019] Laura J. Nelson (2019). L.A. is hemorrhaging bus riders — worsening traffic and hurting climate goals. Library Catalog: www.latimes.com Section: California.

[Levy, 2020] Levy, A. (2020). The Subway is Probably not Why New York is a Disaster Zone. Library Catalog: pedestrianobservations.com.

[Lin and Shin, 2008] Lin, J.-J. and Shin, T.-Y. (2008). Does Transit-Oriented Development Affect Metro Ridership?: Evidence from Taipei, Taiwan. *Transportation Research Record*, 2063(1):149–158. Publisher: SAGE Publications Inc.

[Lloyd, 1957] Lloyd, S. (1957). Least Squares Quantization in PCM. *Bell Telephone Laboratories Paper*.

[Lu and Pas, 1999] Lu, X. and Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice*, 33(1):1–18.

[Ma et al., 2017] Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. (2017). Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors*, 17(4):818.

[Ma et al., 2013] Ma, X., Wu, Y.-J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12.

[Ma et al., 2018] Ma, X., Zhang, J., Ding, C., and Wang, Y. (2018). A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Computers, Environment and Urban Systems*, 70:113–124.

[Ma et al., 2019] Ma, X., Zhang, J., Du, B., Ding, C., and Sun, L. (2019). Parallel Architecture of Convolutional Bi-Directional LSTM Neural Networks for Network-Wide Metro Ridership Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2278–2288.

[Maciag, 2014] Maciag, M. (2014). Public Transportation's Demographic Divide. Technical report, Governing: The Future of States and Localities. Library Catalog: www.governing.com.

[Mahtani et al., 2020] Mahtani, S., Asia, c. M. c. S., Kim, H. K. J., South, c. J. K. i. S. c., and Rolfe, N. K. (2020). Subways, trains and buses are sitting empty around the world. It's not clear whether riders will return. *Washington Post*. Library Catalog: www.washingtonpost.com.

[Mallett, 2018] Mallett, W. J. (2018). Trends in Public Transportation Ridership: Implications for Federal Policy. Technical report, Congressional Research Service.

[Michael Graehler et al., 2019] Michael Graehler, Alex Mucci, and Gregory D. Erhardt (2019). Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes? In *ResearchGate*, Washington, D.C. Library Catalog: www.researchgate.net.

[Michael Manville et al., 2018] Michael Manville, Brian D. Taylor, and Evelyn Blumenberg (2018). Falling Transit Ridership: California and Southern California. Technical report, UCLA Institute of Transportation Studies.

[Mohammadian et al., 2020] Mohammadian, K., Shabanpour, R., Shamshiripour, A., and Rahimi, E. (2020). TRB Webinar: How much will COVID-19 Affect Travel Behavior?

[Morency et al., 2006] Morency, C., Trepanier, M., and Agard, B. (2006). Analysing the Variability of Transit Users Behaviour with Smart Card Data. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 44–49. ISSN: 2153-0017.

[Mucci and Erhardt, 2018] Mucci, R. A. and Erhardt, G. D. (2018). Evaluating the Ability of Transit Direct Ridership Models to Forecast Medium-Term Ridership Changes: Evidence from San Francisco. *Transportation Research Record*, 2672(46):21–30. Publisher: SAGE Publications Inc.

[Munks and Anderson, 2020] Munks, J. and Anderson, J. (2020). Illinoisâ stay-at-home order ends and restrictions lifted on churches as the state advances to next phase of reopening. *Chicago Tribune*. Section: Coronavirus, News, Breaking News.

[Murphy et al., 2016] Murphy, C., Feigon, S., and Firsbie, T. (2016). Shared Mobility and the Transformation of Public Transit. Technical report, Shared-Use Mobility Center, Washington, D.C. Pages: 23578.

[NBC Chicago, 2020a] NBC Chicago (2020a). Chicago Enters Phase 3 of Coronavirus Reopening Plan: Hereâs Whatâs Changing â NBC Chicago.

[NBC Chicago, 2020b] NBC Chicago (2020b). Illinois Enters Phase 4 of Reopening Plan: Here's What's Changing.

[Nigam et al., 2000] Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134.

[Pasha et al., 2016] Pasha, M., Rifaat, S. M., Tay, R., and De Barros, A. (2016). Effects of street pattern, traffic, road infrastructure, socioeconomic and demographic characteristics on public transit ridership. *KSCE Journal of Civil Engineering*, 20(3):1017–1022.

[Puentes, 2020] Puentes, R. (2020). COVID's Differing Impact on Transit Ridership. Technical report, Eno Center for Transportation. Library Catalog: www.enotrans.org.

[Rho et al., 2020] Rho, H. J., Brown, H., and Fremstad, S. (2020). A Basic Demographic Profile of Workers in Frontline Industries. Technical report, Center for Economic and Policy Research. Library Catalog: cepr.net.

[Sadik-Khan and Solomonow, 2020] Sadik-Khan, J. and Solomonow, S. (2020). Fear of Public Transit Got Ahead of the Evidence. *The Atlantic*. Library Catalog: www.theatlantic.com Section: Ideas.

[Siddiqui, 2018] Siddiqui, F. (2018). Falling transit ridership poses an 'emergency' for cities, experts fear. Library Catalog: www.washingtonpost.com.

[Tappe, 2020] Tappe, A. (2020). 30 million Americans have filed initial unemployment claims since mid-March - CNN. *CNN*.

[Templeton, 2020] Templeton, B. (2020). Will COVID-19 Sound The Permanent Death Knell For Public Transit? *Forbes*. Section: Business.

[The Chicago 77, 2008] The Chicago 77 (2008). Chicago Neighborhoods. Library Catalog: www.thechicago77.com.

[Transit, 2020] Transit (2020). How coronavirus is disrupting public transit.

[Transit, 2020] Transit (2020). Who's left riding public transit? Hint: it's not white people. Library Catalog: medium.com.

[Tribune staff, 2020] Tribune staff (2020). COVID-19 in Illinois, the U.S. and the world: Timeline of the outbreak. Section: Coronavirus, News, Breaking News.

[United States Census Bureau, 2020] United States Census Bureau (2020). 2014-2018 american community survey 5-year estimate. https://www.nhgis.org/.

[Vaishnav, 2019] Vaishnav, M. (2019). Ventra Card Use in Chicago.

[Valentino-DeVries et al., 2020] Valentino-DeVries, J., Lu, D., and Dance, G. J. X. (2020). Location Data Says It All: Staying at Home During Coronavirus Is a Luxury. *The New York Times*.

[Viallard et al., 2019] Viallard, A., Trépanier, M., and Morency, C. (2019). Assessing the Evolution of Transit User Behavior from Smart Card Data. *Transportation Research Record*, 2673(4):184–194. Publisher: SAGE Publications Inc.

[Washington Metropolitan Area Transit Authority, 2020] Washington Metropolitan Area Transit Authority (2020). Metro to reopen 15 stations, reallocate bus service to address crowding, starting Sunday | WMATA.

[Whitehead, 2020] Whitehead, K. (2020). Public transit is critical to Chicago's COVID-19 response. Library Catalog: activetrans.org Section: Blog.

[Wise, 2010] Wise, D. (2010). Public Transportation: Transit Agencies' Actions to Address Increased Ridership Demand and Options to Help Meet Future Demand. Technical report, United States Government Accountability Office.

[Wisniewski, a] Wisniewski, M. 'Report card' on CTA bus service gives poor grades to most wards, busy routes. *chicagotribune.com*. Library Catalog: www.chicagotribune.com Section: Business.

[Wisniewski, b] Wisniewski, M. Tired of being stuck on slow CTA buses? City awards $20 million to a program that aims to speed things up. *chicagotribune.com*. Library Catalog: www.chicagotribune.com Section: Transportation, Business, News.

[Zhao et al., 2013] Zhao, J., Deng, W., Song, Y., and Zhu, Y. (2013). What influences Metro station ridership in China? Insights from Nanjing. *Cities*, 35:114–124.