2017

TRANSPORTATION NETWORK COMPANIES: INFLUENCERS OF TRANSIT RIDERSHIP TRENDS

Richard A. Mucci
University of Kentucky, muchi25@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-7217-6836
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.475

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Richard A. Mucci, Student

Dr. Gregory Erhardt, Major Professor

Dr. Yi-Tin Wang, Director of Graduate Studies

TRANSPORTATION NETWORK COMPANIES: INFLUENCERS OF TRANSIT RIDERSHIP TRENDS

THESIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Civil Engineering in the College of Engineering at the University of Kentucky.

By

Richard Alexander Mucci

Lexington, Kentucky

Director: Dr. Gregory Erhardt, Professor of Civil Engineering

Lexington, Kentucky

November 2017

ABSTRACT OF THESIS

TRANSPORTATION NETWORK COMPANIES: INFLUENCERS OF TRANSIT RIDERSHIP TRENDS

The major transit systems operating in San Francisco are San Francisco Municipal (MUNI), Bay Area Rapid Transit (BART), and Caltrain. The system of interest for this paper is MUNI, in particular the bus and light rail systems. During the past decade transit ridership in the area has experienced diverging growth, with bus ridership declining while rail ridership is growing significantly (Erhardt et al. 2017). Our data show that between 2009 and 2016, MUNI rail ridership increases from 146,000 to 171,400, while MUNI bus ridership decreases from 520,000 to 450,000. Direct ridership models (DRMs) are used to determine what factors are influencing MUNI light rail and bus ridership. The DRMs predict ridership fairly well, within 10% of the observed change. However, the assumption of no multi-collinearity is voided. Variables, such as employment and housing density, are found to be collinear. Fixed-effects panel models are used to combat the multi-collinearity issue. Fixed-effects panel models assign an intercept to every stop, so that any spatial correlation is removed. A transportation network company variable is introduced (TNC) to the panel models, to quantify the effect they have on MUNI bus and light rail ridership. The addition of a TNC variable and elimination of multi-collinearity helps the panel models predict ridership better than the daily and time-of-day DRMs, both within 5% of the observed change. TNCs are found to complement MUNI light rail and compete with MUNI buses. TNCs contributed to a 7% growth in light rail ridership and a 10% decline in bus ridership. These findings suggest that the relationship TNCs have with transit is complex and that the modes cannot be lumped together.

KEYWORDS: Transportation Network Company, Direct Demand Model, Transit Ridership, San Francisco, Bus, Rail

Alex Mucci

12/1/2017

Transportation Network Companies: Influencers of Transit Ridership Trends

By:

Richard Alexander Mucci

Dr. Greg Erhardt

_____

Director of Thesis

Dr. Yi-Tin Wang

_____

Director of Graduate Studies

November 17, 2017

_____

Date

Acknowledgements

I would like to thank my advisor, Dr. Greg Erhardt, for his guidance, wisdom, and never ending support throughout my pursuit of a M.S. Whenever I had a question he always made time to answer them properly. The opportunity he provided to work on such exciting research is priceless. The timing is impeccable and fate was on our side.

Next, I want to thank my entire Thesis Committee including Dr. Greg Erhardt, Dr. Mei Chen, and Dr. Reginald Souleyrette. Thank you for reviewing my thesis, and providing valuable insights about the research. All of you have inspired me, in one way or another, to love transportation and enjoy what research can bring.

I would also thank my co-worker Jacob Brashear, for being there to discuss my many confusions. Staggering through an all-nighter is not nearly as bad when you have good company. I could not have asked for a better person to bounce ideas off of, and occasionally play Halo with. I would like to thank Conner Meyer and Ryan Sturm for helping to motivate me. The gym can be a valuable tool during stressful times.

Last, but most definitely not least, I would like to thank my family for the never-ending support. I have always looked to my dad for proper guidance, and he has been called upon many times throughout my pursuit for a M.S. My mom never fails to show support, the many Facebook posts can boost my morale at any

given time. I would like to thank my sisters for their support as well. Having you all look up to me, helps me to hold myself to high standards, as a role model for you all. My brother Will is a bit too young to express himself, but I know he undoubtedly enjoys me babbling on about my research. I would like to thank my grandparents and the many other family members who have supported me along my journey.

This work was funded, in part, by the San Francisco County Transportation Authority (SFCTA). Thanks to the staff at SFCTA and the San Francisco Municipal Transportation Agency, who supplied data and provided valuable input and advice.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1 Introduction

The purpose of this chapter is to outline this thesis. First, background is provided to introduce the research topic. Then the structure and research objectives of this thesis are presented.

1.1 Background and Thesis Structure

The 9-county San Francisco Bay area covers approximately 7,000 square miles with a population of 7.5 million residents, and covers 50 square miles. San Francisco is one of the core counties and has a population of 850,000 residents. The major transit systems operating in San Francisco are San Francisco Municipal (MUNI), Bay Area Rapid Transit (BART), and Caltrain. The system of interest for this paper is MUNI, in particular the bus and light rail systems.

The MUNI system experienced significant service cuts in 2010 due to budget constraints. Since then, there have been incremental service improvements, which have accelerated from about 2014 as the agency implements its MUNI Forward program (San Francisco Municipal Transportation Agency 2014), focused on deploying Rapid (skip-stop service with associated operational improvements) bus routes in high-use corridors.

During the past decade transit ridership in the area has experienced diverging growth, with bus ridership declining while rail ridership is growing significantly (Erhardt et al. 2017). Our data show that between 2009 and 2016, Muni rail ridership increases from 147,500 to 161,400, while MUNI bus ridership decreases from 515,000 to 450,000. The employment and population in San Francisco has increased significantly, so it would be natural to expect a corresponding growth in transit ridership. San Francisco is not the only city to experience a diverging transit ridership trend, chapter 2 discusses the national and local trends and previous research on the topic in more detail.

Direct ridership models (DRMs) are used in chapter 3 to determine what factors are influencing MUNI light rail and bus ridership. The model is applied and sensitivity tests are used to understand how much each variable is contributing to the change in bus ridership.

Fixed-effects panel models are used in chapter 4 to combat the multi-collinearity issue. Fixed-effects panel models assign an intercept to every stop, so that any spatial correlation is removed. The spatial correlation is what was causing the multi-collinearity in the DRMs. A transportation network company variable is introduced (TNC) to the panel models to quantify the effect they have on MUNI bus and light rail ridership.

1.2 Motivation and Research Objectives

Previous research on the topic of transit ridership trends is typically more focused on what factors are influencing transit ridership in specific regions, outside of San Francisco. Erhardt does analyze what factors are influencing bus ridership in San Francisco, but has a negative 11% trend left unexplained (Erhardt et al. 2017). Clewlow

and Mishra (2017) use survey data on TNC users to determine that TNCs have a negative effect on bus ridership. Explaining the unknown trend with TNC trip data, rather than survey data, is the motivation behind this research. This thesis aims to understand what is causing the unexplained trend by accomplishing the following research objectives:

1. Determine what factors influence MUNI bus and light rail ridership in San Francisco
2. Determine what factors are contributing to changes in MUNI bus and light rail ridership.
3. Determine the effect TNC growth has on MUNI bus and rail ridership in San Francisco.

The first question is focused on understanding what variables should be included in the final model. The output is a list of variables that are significantly correlated with transit ridership. This is completed by estimating cross-sectional DRMs for MUNI light rail and bus ridership.

The second and third questions are addressed by estimating fixed-effects panel data models with a TNC variable. Sensitivity tests are used to quantify the effects that each variable has on MUNI light rail and bus ridership.

Chapter 2 Understanding Transit Ridership Trends

Transit agencies have operated for over a century. The first forms of bus public transportation, the horse drawn omnibus, began around the 1820s. They were followed by streetcars towards the end of the 19th century. Rail transportation as we know it today began around the same time. The Chicago L train launched in 1895. The New York subway was the first underground railway in 1904. Transit ridership has ebbed and flowed throughout the years, with sparse analysis on what is causing changes. Recent automated passenger count (APC) data has helped, in the past few decades, strengthen our understanding of what influences transit ridership.

The purpose of this chapter is to provide a background on recent transit ridership trends and to understand what is influencing them. Section 2.1 discusses national transit ridership trends and how San Francisco compares to other cities. Section 2.2 zooms in, describing San Francisco and the local transit ridership trends. Then the local trends in San Francisco are compared to the national trends. Lastly, section 2.3 discusses previous research on what factors influence transit ridership.

2.1 National Transit Ridership Trends

The purpose of this section is to discuss trends in transit ridership at a national level. This provides a macro-level understanding of ridership trends.

Transit ridership, across all modes, has decreased slightly in recent years, but is trending north if you include the past couple of decades (Federal Transit Administration 2016). Figure 1 shows the annual boardings for the 2000s, across all modes of transit. The boardings are found using the unlinked passenger trip, refers to data from the National Transit Database (Federal Transit Administration 2016). Unlinked passenger trips are the number of times a passenger boarding a vehicle. Meaning a trip with a transfer would count as two unlinked passenger trips. All graphs in this section are referring to annual boardings, and are obtained the same way. The effects of the recession in 2008 can be observed by the drop in boardings right after 2008. There was a slight increase following the recession, and then ridership tapered off again.

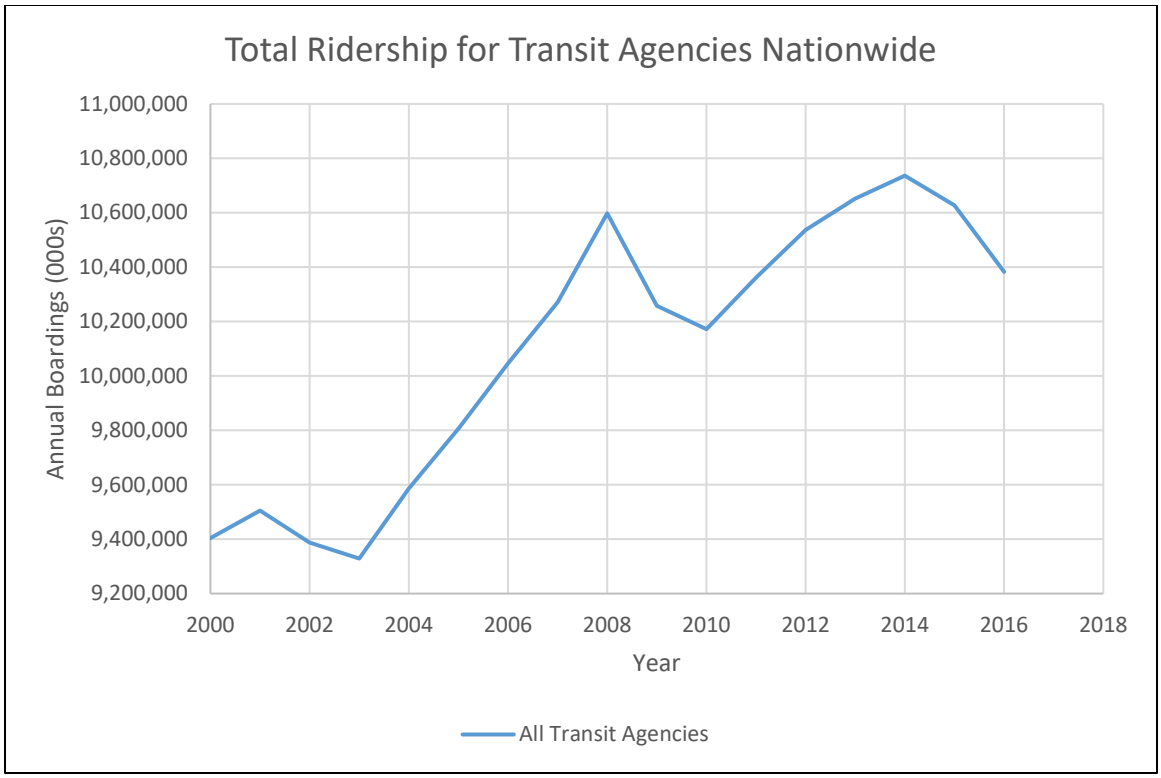

Figure 1: National Transit Ridership (All Modes)

Figure 2 shows a comparison of rail and bus ridership for the 2000s as well. Instead of total ridership, the annual boardings percent change, from 2000, is compared. Bus refers to rubber-tired vehicles operating on fixed routes and schedules over roadways, which includes diesel and electric powered buses. The rail modes compared in Figure 3 are light rail, heavy rail, and commuter rail. Light rail cars are powered by overhead wires and have "lighter" capacities than heavy rail cars. Heavy rail cars are powered by an electrified "third rail" and operate on elevated platforms or in tunnels to keep pedestrians safe. Commuter rail is used for long distance travel, typically inter-city trips. Light rail and bus ridership are the two modes that this thesis analyzes. All rail ridership has grown notably, while bus ridership has declined. This trend has occurred in many cities including San Francisco, Boston, Houston, and New York.

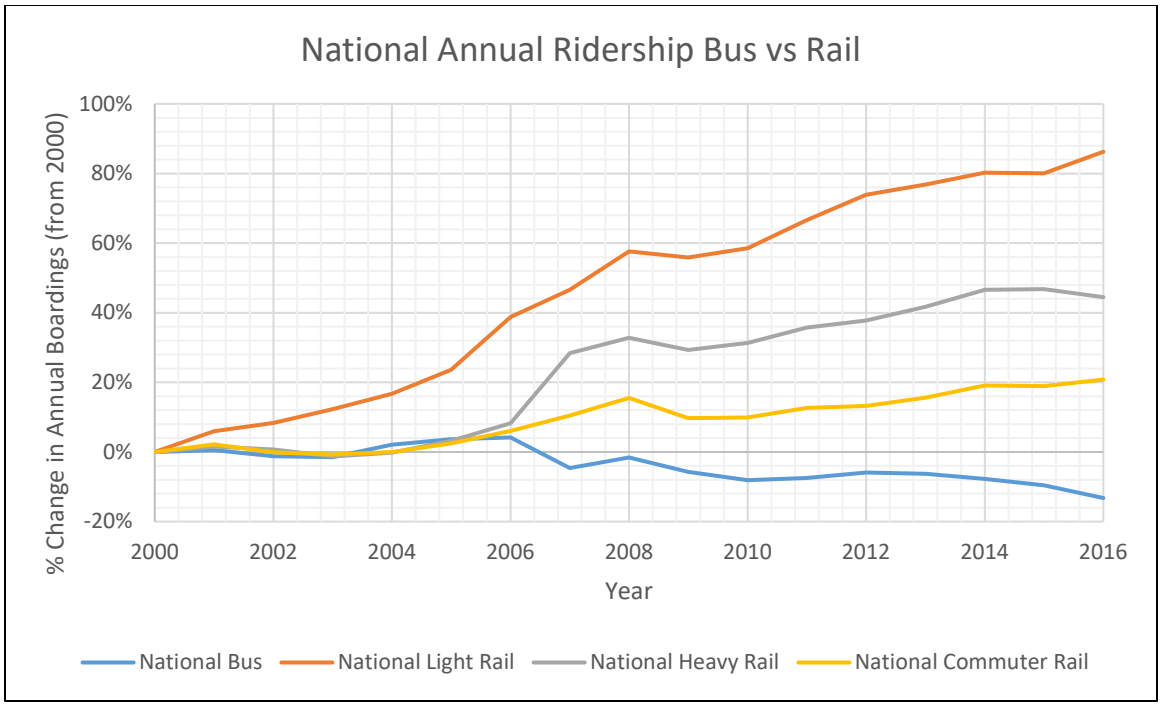

Figure 2: Percent Change in National Annual Ridership (Bus vs. Rail)

2.2 Local Trends

The purpose of this section is to zoom in from a macroscopic view to a microscopic "local" view. San Francisco is discussed in detail, to understand the "local" transit ridership trends. Then the national and local trends are compared. This helps to determine how San Francisco fits in with cities nationwide.

Figure 3 shows the ridership changes for the bus and rail systems in San Francisco. Bay Area Rapid Transit (BART) is the heavy rail agency and San Francisco Municipal Railway (MUNI) is the bus and light rail agency, with MUNI Bus being the bus system and MUNI Metro being the light rail system. The rail ridership shown in Figure 3 consists of MUNI Metro and BART's ridership. The data provided by NTD for Caltrain ridership did not match up with reports published by Caltrain.(Caltrain 2017) Average weekday ridership, found in the Caltrain 2017 report, is used instead of the NTD data. The bus ridership shown in Figure 3 consists of the MUNI Bus fleet. NTD historical ridership data for specific transit agencies, rather than national totals, is only available through 2002. This is why 2002 was chosen as the base year, rather than 2000.

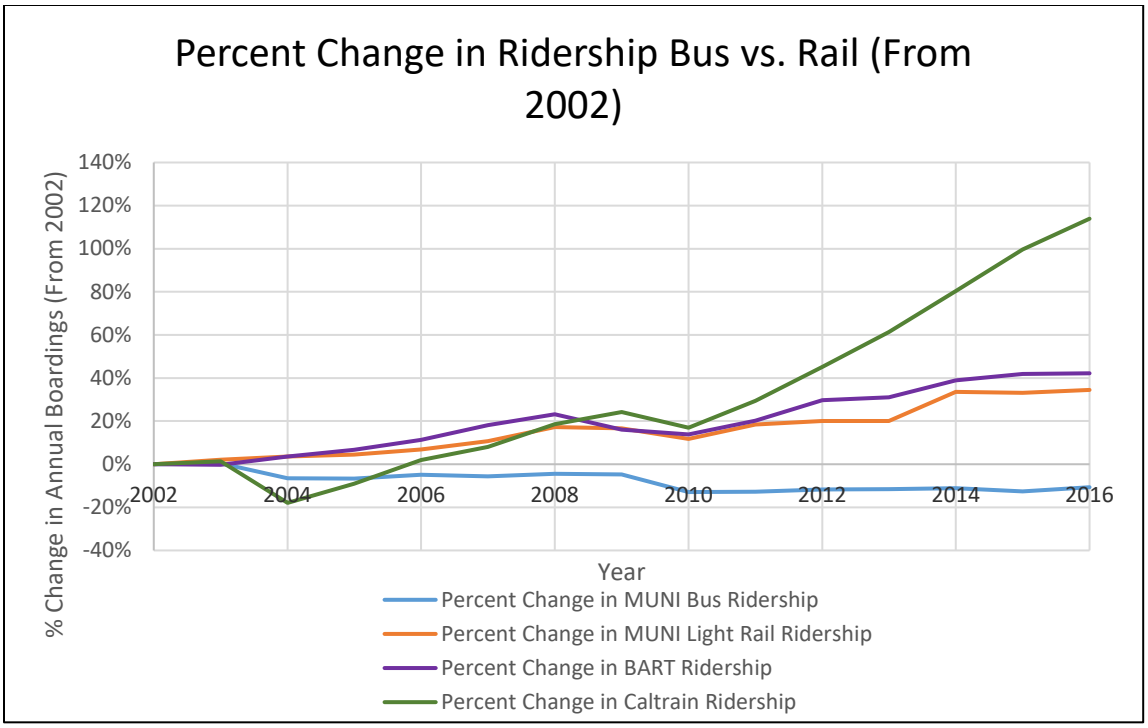

Figure 3: San Francisco Bus and Rail Ridership Trends

Similar to the national ridership, there is a dip after the 2008 recession for all modes. Rail ridership has grown while bus ridership has stagnated following the recovery after the 2008 recession, which also follows the national trend.

The San Francisco—Oakland, CA urbanized area is the thirteenth most-populated, housing roughly 3,281,212 people. (Bureau 2016) Table 1 shows the ten urbanized areas in 2016 that have the highest number of motorized bus boardings. (Federal Transit Administration 2016) UZA refers to the name of an urbanized areas as defined by the U.S. Census.

Table 1: Bus Ridership Ranking (Top 10 Cities Nationwide)

UZA Name	Bus Boardings	Rank
New York-Newark, NY-NJ-CT	1,103,280,638	1
Los Angeles-Long Beach-Anaheim, CA	441,834,026	2
Chicago, IL-IN	288,495,190	3
San Francisco-Oakland, CA	175,419,546	4
Washington, DC-VA-MD	173,806,234	5
Philadelphia, PA-NJ-DE-MD	172,699,663	6
Seattle, WA	118,899,852	7
Boston, MA-NH-RI	114,144,422	8
Miami, FL	105,042,540	9
Baltimore, MD	73,319,247	10

The San Francisco – Oakland urbanized area is ranked fourth, and raises the question of whether or not the regionally-specific trends are representative of trends in only "big" cities. Figure 4 compares the annual percent change in boardings for cities of various sizes. Percent change is used instead of total boardings to allow cities of various sizes to be compared. However, cities such as Lexington that start with low ridership totals in 2002 can have a large percent increase, but a small absolute increase in ridership. San Francisco is the bold black line, and should be the focal point. San Franicsco trends below the pack slightly, but is not a noticeable outlier. This provides hope that the findings in this paper would be applicable in other cities, but ultimately future research is needed on other cities to validate this claim. Light rail ridership is not shown due to some of the cities not having light rail ridership data. This is mostly due to the lack of a light rail transit agency being present in the city.

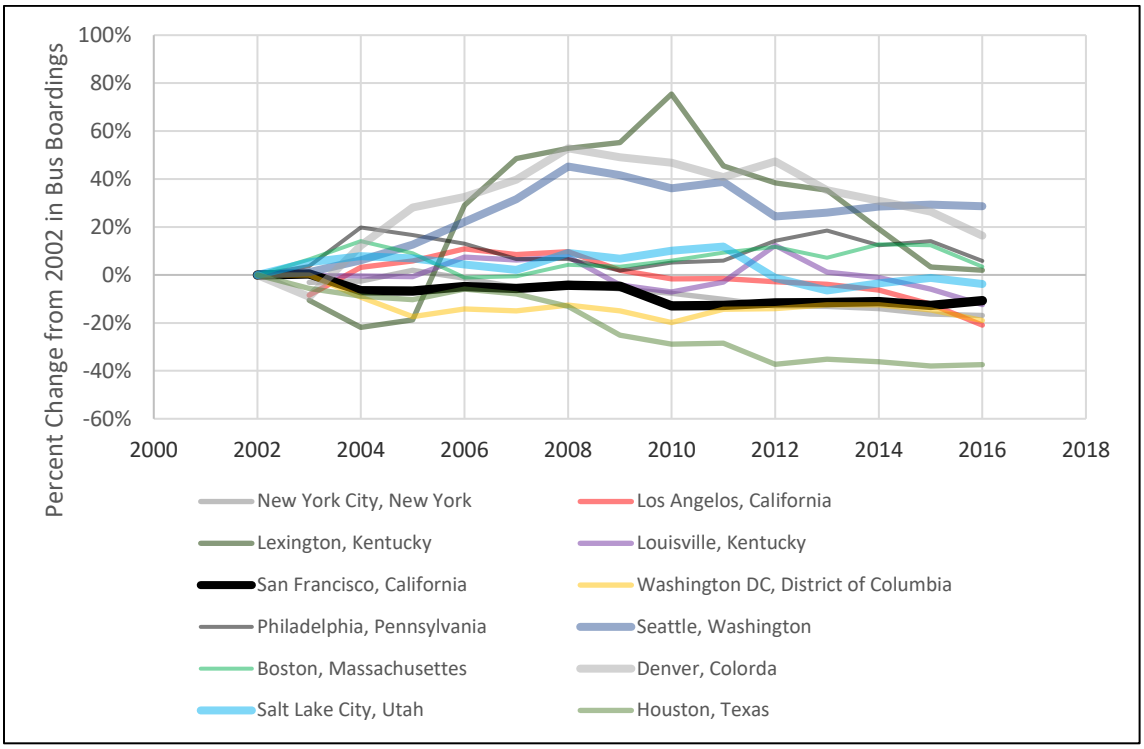

Figure 4: Yearly Bus Boardings Percent Change (From 2002)

2.3 Understanding the Factors Influencing Transit Ridership Trends

There is not an exhaustive amount of peer reviewed research on what factors are causing the trends at a national level. American Public Transportation Association (APTA) has released a report on the characteristics of people who use transit. (Clark et al. 2017) The highlights of the report are that 79% of the people who use transit are between the ages of 25 and 54, 78% are either employed or are students, 21% of transit users have an income of less than $15,000 and 13% of the U.S. population fall within that income range, people using transit for shopping grew from 9 to 21 percent between 2007 and

2017, and the most popular mode choice for accessing transit is walking. This helps paint the picture of who rides transit and their motivations behind using transit. These statistics are used as a guide when choosing which variables to include in the models. One example of research on the topic at a national scale is (Taylor et al. 2009). Taylor et al. (2009) lays out previous research on what factors affect transit ridership. They divide the research into two categories: (1) research that focuses on travelers perceptions, with both travelers and transit managers as the units of analysis, and (2) studies that examine the environmental, system, and behavioral characteristics associated with transit ridership. These types of studies can be categorized as (1) external, or control, factors (nature) and (2) internal, or policy, factors (nurture). Taylor et al. (2009) finds that 26 percent of transit ridership in over 265 urbanized areas can be explained simply by frequency and fare levels.

There has been many studies looking at what factors influence transit ridership regionally (Taylor et al. 2009; Estupiñán and Rodríguez 2008; Chakour and Eluru 2016; Brown and Thompson 2008; Balcombe et al. 2004; Iseki and Ali 2015; Kennedy 2013). Estupiñán and Rodríguez (2008) look at how the built environment around a bus rapid transit stop influences ridership. They find that the built environment does significantly affect ridership, in particular environmental supports for walking and barriers to car use. Chakour and Eluru (2016) study the effect that operational attributes, system infrastructure attributes, and the built environment around a stop has on transit ridership in Montreal, Canada. They find that the most effective way to increase ridership is to increase public transportation service and accessibility, whereas enhancements to land use have a smaller effect on ridership. Iseki and Ali (2015) used panel data models to understand what factors are affecting ridership in 10 U.S. urbanized areas. They find that the short-term increase in gasoline prices have a relatively small effect on bus ridership and marginal effect for rail ridership. Brown and Thompson (2008) use a time-series model to test if the decentralization of population and employment in Atlanta is causing a recent downward trend in transit ridership. They find that the dip in transit patronage is attributable to employment decentralization outside of the transit agencies service area.

Even though the research has ranged in location from Bogota, Columbia to Portland, Oregon there is a general trend to them. They are that ridership is consistently sensitive to fare prices, transit supply, employment density, and population density. Transit supply includes a range of factors that all refer to the amount of service provided, such as frequency of buses or service miles. There are many other factors that are found to be significant when modeling transit ridership, but the most common are the aforementioned.

shows a more exhaustive list of variables that have been included in previous research.

Table 2: Significant Variables in Previous Rail and Bus DRMs

Variables		
Service Attributes	Location of Stop Attributes	Bus Stop or Site Attributes
Primary • Presence of dedicated-lane services (0-1) • Run Speed (Actual) • Reliability • Bus-stop ridership (log) (dependent variable) • Fare • Service Miles (Scheduled) • Dwell Time • Crowded hours	**Primary** • Population density (US census) • Employment density (US census) • Total urban density (persons plus workers) • Distance to the nearest rapid transit stop • Income • Competitive bus-stops (two levels: opposing and same route)	**Primary** • Presence of bus benches (0-1) • Presence of a passenger information system (0-1) • Presence of a bus-stop shelter or canopy (0-1) • Distance to urban center • Bus terminus (indicator) (start/end of stop)
Seconday • Rear Door Boardings • Number of daily Connecting rail-transit lines • Supply demand match index (how well demand matches supply) • Stop frequency (log)	**Secondary** • Percent of population that are elderly • Land Use (residential/agriculture/etc.) • Street connectivity (# of nodes divided by # of links)	**Secondary** • Transer stop (indicator) (can transfer to at least two other routes) •Park-and-ride lot (0-1) •Number of park-and-ride spaces • Presence of BRT-branding or logo at stop (0-1)

Erhardt has looked at San Francisco in particular. (Erhardt et al. 2017) Erhardt estimated a time-series model to look at what factors in recent years influenced transit ridership in San Francisco between 2009 and 2013. Table 3 shows the results of Erhardt's time-series model. Service and fare changes are internal factors that the transit agency has control over. Service cuts and fare increases led to a roughly 9% decrease in MUNI bus ridership. Employment grew significantly causing a roughly 10% increase in MUNI bus and BART ridership. MUNI light rail service increases lead to an increase in MUNI bus ridership. Meaning that the two modes complement each other. Cost of auto travel, price of gas, has a minimal impact. The change in BART ridership was explained well, with only a 3% growth left unexplained. However, MUNI bus ridership still has an 11% unexplained reduction in ridership. Numerous hypotheses attempted to explain the trend, such as decentralized growth and changes in demographics.

Table 3: Factors Contributing to changes in MUNI vs. BART riders: Sep 2009 to Sep 2013

	Change in MUNI Riders		Change in BART Riders	
Associated with change in:	Absolute	Percent	Absolute	Percent
Service and fare changes	-48,425	-8.9%	7,051	2.0%
Employment and employment concentration	61,786	11.3%	43,238	12.2%
Change in MUNI rail service	11,892	2.2%	0	0.0%
Increased cost of auto travel	0	0.0%	1,841	0.5%
Unexplained trends and random error	-60,589	-11.1%	9,994	2.8%
Total Change	-35,337	-6.5%	62,124	17.6%

* This is pulled directly from Erhardt's Thesis

An "Uber Effect" is proposed by Erhardt, but the lack of data on TNC trips prevented further analysis. A Shared Use Mobility Center report refuted Erhardt's claim, but did not look at San Francisco specifically. (Feigon and Murphy 2016) Chapter 4 looks into the effect TNCs have on transit ridership in more detail.

Chapter 3 Discovering Factors Influencing Transit Ridership

The purpose of this chapter is to determine what factors are influencing changes in MUNI bus and light rail ridership. The chapter starts by introducing cross-sectional direct ridership models and previous studies that utilized them. This is followed by explanations of the data used in the model. Then, the model results are provided and interpreted. Next the model is applied and compared to the observed ridership. A sensitivity test is then used to determine how much each factor is contributing to the changes in MUNI light rail and bus ridership. Lastly, the conclusions and limitations of the model are discussed. A version of this chapter has been accepted to be published by the Transportation Research Board.

3.1 Introduction

Transit direct ridership models (DRM), sometimes referred to as direct demand models, are often used for transit planning applications (Cervero 2008). A DRM is a sketch planning tool that reflects the notion that ridership is the sum of riders boarding at each stop. Rather than estimating trips on all modes and then applying a mode choice, as is done in the 4-step modeling paradigm, DRMs estimate transit ridership "directly" as a function of station or stop environments and transit services. Essentially each station becomes a node and a unit of analysis, which provides fine-grained spatial analysis of factors influencing ridership.

A hybrid between a DRM and a 4-step model is the Federal Transit Authority's (FTA) Simplified Trips on Stops (STOPS) model (RSG 2015), which estimates rail and bus rapid transit (BRT) ridership. It is similar to a DRM in that it predicts ridership and is not as extensive as the 4-step model, but it still considers more information than most DRMs. The additional information includes the Census journey-to-work flows, and highway level-of-service skims from a regional travel model.

DRMs have been developed for both rail and bus transit, and used for a variety of applications. The models are used both as descriptive tools to understand the factors that influence ridership, and as predictive tools to forecast the ridership at new stations or in response to changes in service or other attributes.

For example, Cervero, Murakami, and Miller (2010) estimate a DRM of bus rapid transit ridership in the Los Angeles area. They argue for the application of DRMs as a complement to 4-step models, to generate first-cut ridership estimates, and to conduct sensitivity tests of key variables. Pulugurtha and Agurla (2012) estimate DRMs using different spatial modeling methods in Charlotte, North Carolina. They find that using a spatial weight method does not yield better results over a spatial proximity method, a buffer. Kepaptsoglou, Stathopoulos, and Karlaftis (2017) predict ridership for a new light

rail transit system in Cyprus using a DRM. The transit system had not been built yet, so there was not any comparison to observed ridership.

The following papers all explore the effects of various factors on transit ridership. Lee et al. (2015) studies the relationships between heavy and light rail ridership and socio-economic characteristics in Korea. They find that heavy rail ridership is correlated with the density of businesses and households, while light rail ridership is correlated with the density of economically active population and businesses. Liu et al. (2014) estimate a DRM of rail ridership in Maryland to suggest improvements that will boost ridership. Holmgren (2013) determines what factors are affecting transit ridership in Sweden and does this by estimating an econometric first-difference model. Estupiñán and Rodríguez (2008) estimate a DRM of Bogota's BRT ridership to quantify the effect that the built environment has, in particular environmental supports for walking. Dill et al. (2013) describe the effects that transit service characteristics and urban form have on local bus ridership. They find that socio-demographic characteristics seem to have a larger effect on ridership in large urban areas rather than small urban areas.

In nearly all of the above cases, the DRM is estimated from cross-sectional ridership data using ordinary least squares (OLS) regression. An exception is from Kerkman, Martens, and Meurs (2014) who use data from both 2012 and 2013, during which significant service changes occurred. They estimate cross-sectional models and compare those models to a fixed-effects panel data regression model where the dependent variable is the change in ridership from 2012 to 2013. They find notably different sensitivities between the two models, with the panel data models showing the elasticity of ridership with respect to changes in stop frequency as about half of what the cross-sectional models show. They suggest that the panel data models do a better job of mitigating endogeneity where transit agencies tend to run higher frequency service in locations with higher potential demand.

While DRMs are routinely promoted as tools for forecasting, they are rarely evaluated for their ability to forecast (Cervero, Murakami, and Miller 2010; Kepaptsoglou, Stathopoulos, and Karlaftis 2017). In fact, many of the published DRMs evaluate the models exclusively on the goodness of fit against the estimation data set, without any validation against independent data in either location or time (Cervero, Murakami, and Miller 2010; Pulugurtha and Agurla 2012; Kepaptsoglou, Stathopoulos, and Karlaftis 2017; Lee et al. 2015; Liu et al. 2014; Holmgren 2013; Estupiñán and Rodríguez 2008; Dill et al. 2013). An exception to this is Upchurch and Kuby (2013) where a DRM is estimated using data in 9 (other) cities, applied to predict ridership on a new light rail line in Phoenix, and compared to actual ridership. The comparison allowed the authors to understand how well the model could be applied to a new data set, and to understand what caused over/under predictions.

Beyond the efforts of Upchurch and Kuby (2013) there remains a sparsity of evidence on the ability of DRMs to forecast changes in transit ridership. This research contributes to filling that gap by estimating both a rail and bus DRM using 2009 data,

applying that model to forecast 2016 ridership, and comparing the forecasts to actual 2016 conditions. It does this using the example of the MUNI light rail and bus system in San Francisco, California. The analysis goes on to examine the magnitude of factors affecting the change in modeled ridership between these two years. Together, this allows for an evaluation of how well the models perform in a predictive way, what they are able to capture in drivers of transit ridership trends, and what they be missing. Ultimately, this process helps to understand what factors are influencing the diverging ridership trends discussed in chapter 1.

3.2 Data

Direct ridership models estimate boardings or exits or both at a stop for defined periods of time, daily and by time-of-day, as a function of variables related to the stop. This research uses the standards previous research has set for DRMs with slight modifications to improve the models performance in San Francisco. The first modification is that the data are aggregated to the (directional) route-stop level rather than the stop level. That means that if two different transit routes stop at the same location, they are treated as separate observations. While the locational attributes of each record are the same, this allows the service attributes to be tracked separately. Typically quarter-mile buffers are more commonly used, but tenth-mile buffers were selected after a range of buffers were tested. This reflects the high density of stops in San Francisco and avoids excessive overlap between adjacent buffers. The rail stations used larger quarter-mile buffers. Both were tested with a range of buffer sizes.

Previous research was used to determine what other data should be collected and included in the model (Kepaptsoglou, Stathopoulos, and Karlaftis 2017; Lee et al. 2015; Liu et al. 2014; Holmgren 2013; Estupiñán and Rodríguez 2008). These variables are discussed in more detail in chapter 1 and were used as a starting point to guide the data collection process. Table 3 shows a list of the variables considered in this research, along with their calculation method. The remainder of this section describes the data and its processing in further detail.

All the variables with log in parenthesis have undergone a log-transformation. The transformation is the natural log of the data plus one. The plus one is to avoid errors when the data is 0. A sample calculation can be found below.

$$Data_{Log-Transformed} = Ln(1 + Data_{original})$$

Table 3: Variables and Data Sources

Variable	Variable Name	Description	Data Source	Calculation Method
Ridership (Dependent Variable)				
Bus/Rail Route-Stop Ridership (Log) *	LOG_RIDERS	Daily average number of Bus/Rail passengers boarding and alighting a bus at each specific route-stop.	MUNI (Bus Operator)	APC and GTFS Data Fusion
Potential Demand (Independent Variables)				
EDD Employment (Log)	EDD_EMP_LOG	Total employment within walking distance of a stop.	Employment Development Department (SFCTA)	Buffer Aggregation
LEHD Employment Density	EMP_WAC_DEN	Employment per acre within walking distance of a route-stop.	Longitudinal Employer-Household Dynamics	Buffer Aggregation
Income	SHR_INCOME_100P **	Share of households, within various income ranges, within a census tract that a route-stop falls within.	American Community Survey	Spatial Join
Housing Density (Log)	HOUSING_DEN_LOG	Housing units per acre within walking distance of a route-stop.	Census and Housing Inventory (SFCTA)	Buffer Aggregation
Population Density	POP_DEN	Population per acre within walking distance of a route-stop.	Census	Buffer Aggregation
Population	TOTAL_POP	Total population for a census tract that a route-stop falls within.	American Community Survey	Buffer Aggregation
Share of Households with 0 Vehicles	SHR_HH_0VEH	Share of households with 0 vehicles for a census tract that a route-stop falls within.	American Community Survey	Spatial Join
Transit Supply (Independent Variables)				
Frequency (Log)	FREQ_S_LOG	Average number of buses per hour scheduled to serve each route-stop on a typical weekday.	GTFS	Freq = 1/(Headway*60)
Competing Stops	COMP_STOPS	The number of stops within walking distance of a route-stop.	GTFS	Buffer Aggregation
Bus Terminus	EOL_SOL	Whether or not a route-stop is the start or end of a route.	GTFS	N/A
Transbay Terminal (Bus Station)	TRANSBAY	Whether or not a route stop is within walking distance of the Transbay Terminal.	Google Earth	Buffer Aggregation
Reliability	ONTIME5	The on-time share of a route-stop, specified as either 1 minute before or 5 minutes after the scheduled time.	GTFS	N/A
BART Ridership (Log)	AVG_BART_LOG	The average number of BART passengers boarding and alighting within walking distance of a route-stop.	BART OD Matrices	Buffer Aggregation
Route Configuration	LIMITED / EXPRESS	A variable to separate out the limited/rapid and express routes from the rest of the bus fleet.	GTFS	N/A
Bus/Rail Ridership *	MUNI BUS AVG / MUNI_RAIL_AVG	The average number of Bus/Rail passengers boarding and alighting within walking distance of a route-stop.	MUNI (Bus Operator)	Buffer Aggregation

***** For the bus model the bus ridership is the dependent variable with the rail ridership as an independent variable, and vice-versa for the rail model.

** This is one category of income, includes the share of households with an income of over $100k.

3.2.1 Bus Ridership (Dependent Variable)

A portion of the San Francisco Municipal (MUNI) bus fleet is equipped with AVL/APC technology. The AVL equipment records the latitude and longitude along with a timestamp every time a vehicle arrives or departs a stop. Busses are randomly assigned the AVL/APC equipment at the depot each day, but over a number of days all routes have been observed. The data is scaled up to cover the entire system by applying the ratio of total trips over observer trips to the observed data using a previously developed process (Erhardt et al. 2017). In addition to ridership, the data provides other performance metrics, including runtime, speed, on-time performance, and crowding measures. This analysis considers average weekday conditions (excluding holidays) for the fourth quarter of 2009 and 2016.

3.2.2 Rail Ridership (Dependent Variable)

Light rail cars are not equipped with APCs. Instead the San Francisco Municipal Transportation Authority (SFMTA), relies on occasional manual counts. They provide boarding and alighting counts at the station level for average weekday conditions in both 2009 and 2016. Due to the limited ridership data in 2009, some routes were supplemented with 2008 counts. The routes supplemented are route: J, KT, and L.

3.2.3 General Transit Feed Specification (GTFS)

GTFS allows transit operators to publish their schedules in a standard format. Only the current schedule is published, but when a new version is published, the old is archived, so that the differences can be used to systematically identify transit service changes. The GTFS provides service attributes for both the rail and bus models, such as scheduled headways, and is used as trip totals when the APC data is scaled up.

3.2.3 Employment Development Department (EDD)

EDD data for the 4th quarter of 2009 and 2016 was used as the measure of employment. The data are segmented by industry, and are spatially detailed. The data came in the form of point totals, which were aggregated to buffers. Longitudinal Employer-Household Dynamics (LEHD) (US Census Bureau Center for Economic Studies 2010) was considered, but not used because they are currently only available through 2014.

3.2.4 Housing Inventory and Decennial Census

The San Francisco Planning Department provided housing inventory data, which tracks the completion of housing developments in the city with the address, completion date, and net change in housing units. The housing inventory was used to pivot from the 2010 Census to obtain an estimate of the housing stock in each year (US Census Bureau Center for Economic Studies 2010). Census block totals are aggregated to buffers based on the percentage of the census block that is within the buffer. Section 3.3.8 explains this process in more detail.

3.2.5 American Community Survey (ACS)

To test the possible effect of demographics and socio-economic characteristics, we used American Community Survey (ACS) data. The ACS (Bureau 2016) is an annual survey of households conducted by the Census Bureau to collect information such as household income, household composition, number of vehicles per household, etc. For

the 2009 data set, we used the 5-year estimates, from 2005-2009. For 2016, we used the 5-year estimates from 2011-2015, the most recently available. The 5-year estimates are more suitable for our purpose because they provide more spatial detail than the 1-year estimates. In tracking demographic changes, we did not use a stop-buffering approach, but instead tagged each stop with the attributes of the Census tract in which it lies. The income variable for 2015 had to be adjusted to 2009 dollars, to account for inflation. Meaning the share of households above $100k in 2009 dollars is the share above $115k in 2015 dollars. The 2009 income variable is kept at $100k, since it is already in 2009 dollars. The $115k was found by multiplying $100k, the 2009 original amount, by the ratio of the 2016 consumer price index (CPI) over the 2009 CPI.

3.2.6 BART Monthly Entry and Exit Matrices

The Bay Area Rapid Transit (BART) system is a heavy rail system serving four counties in the San Francisco Bay Area. Commuters are a large share of the users because it provides an alternative to the heavily congested Bay Bridge. BART serves as a potential competitor to MUNI for certain trip interchanges within San Francisco, but also serves as a complement because many regional BART trips transfer to MUNI for the "last-mile".

BART uses distance-based fares, so passengers have their tickets read upon both entering and exiting the system. Knowing the number of entries and exits for each station allows BART to publish monthly matrices showing the number of trips to and from each station. (BART 2017) There are nine BART stations in San Francisco that are of interest: Embarcadero, Montgomery, Powell, Civic Center, 16th, 24th, Glen Park, and Balboa Park. Daly City is also included, this is because it falls within the buffer of stops that are within San Francisco. BART ridership is assigned to stops based on which buffer the BART station is located within. Section 3.3.8 explains the stop-buffer process in more detail.

3.2.7 Hourly On-Street Parking Cost

Parking data was provided by SFCTA and came in the form of hourly on street parking prices throughout the day. The prices were averaged to obtain an average hourly price for a meter on a typical weekday. The data was then attached to Micro-Analysis Zones (MAZ) to give the data a spatial component and then assigned to a stop by using a buffer around the stop. The data is only available for 2014, and is not collected annually. The hourly prices are kept in 2014 dollars because that is the year the data is observed. Since the data was only available for one year the same prices are used for 2009 and 2016. Thus the parking prices can only be used to look at the relationship between it and transit ridership, and not how it influences change in transit ridership.

3.2.8 Stop-Buffer Process

The stop-buffer process refers to creating a buffer around each bus, or rail, stop and assigning data to a stop based on its spatial location. The process is perfect for point features, such as the EDD employment data. Data in the form of polygons, such as census blocks, tend to straddle a buffer. Meaning that only a portion of the block falls within a

buffer. In the aforementioned case, the data would be scaled based on the area that fell within the buffer. Meaning that a census block with 100 employees and half of its area within a buffer would have 50 of its employees assigned to the stop associated with the buffer it fell within. Choosing the correct buffer size required trial and error. Previous research have used quarter-mile and half-mile buffers, but ultimately tenth-mile buffers produced the best results (Cervero, Murakami, and Miller 2010), (Kerkman, Martens, and Meurs 2014), (Pulugurtha and Agurla 2012). This is most-likely due to the dense urban network in San Francisco.

3.3 Model Estimation

This section presents the model methodology and estimation results for the bus and rail models.

3.3.1 Modelling Approach

Previous research has used a direct ridership model to determine what factors influence transit ridership. (Liu et al. 2014), (Kepaptsoglou, Stathopoulos, and Karlaftis 2017), (Kerkman, Martens, and Meurs 2014), (Cervero, Murakami, and Miller 2010) Typically DRMs estimate ridership for a bus stop, but more detailed transit data was available. The models estimate ridership for a route-stop and route-stop-TOD. A route-stop refers to a specific route that stops at a specific bus stop. One bus stop typically has multiple routes stopping at it throughout the day. The more detailed data separates out a route's performance and supply. A route-stop aggregates data to a daily total. A route-stop-TOD has transit data aggregated to 7 different times-of-day (TODs). The 7 TODs are: 3am – 5:59am, 6am – 8:59am, 9am – 1:59pm, 2:00pm – 3:59pm, 4:00pm – 6:59pm, 7:00pm – 9:59pm, 10pm – 2:59am. This picks up on changes for variables that change throughout a day, such as frequency.

3.3.2 Model Equation

The models are a hybrid of a log-log and log-level. Some of the explanatory variables have been transformed specifying a log-log relationship. Others are not, keeping them as a log-level relationship. A sample equation for 2009 ridership and the rearranged version that was used to apply the model is shown below. The sample equation only includes one variable for each specification. The full equation can be found in appendix A.

Equation 1: Bus Route-Stop Cross-Sectional DRM Sample Model Equation

$$\ln(Avg\_ride_i) = \alpha + \beta_{emp} * \ln(\mathrm{emp_i}) + \beta_{ontime5} * \mathrm{ontime5_i} + \varepsilon_i$$

Equation 2: Bus Route-Stop Cross-Sectional DRM Sample Applied Equation

$$Avg\_ride_i = e^{\alpha} * \mathrm{emp_i}^{\beta_{emp}} * \mathrm{e}^{\beta_{ontime5} * \mathrm{ontime5_i}} + \varepsilon_i$$

- Avg_ride_i= the average of the boardings and alight for stop i
- α = intercept (adjusts the slope of the fitted line)
- β_{emp} = coefficient for employment density
- $\mathrm{emp_i}$ = Employment density for stop i
- $\beta_{ontime5}$ = coefficient for reliability
- $\mathrm{ontime5_i}$ = reliability for stop i
- ε_i = error term for stop i

3.3.3 Bus Route-Stop Model

A 2009 base year model was estimated for each of the modes. The python statsmodels package ("StatsModels: Statistics in Python — Statsmodels 0.8.0 Documentation" n.d.) was used to run OLS regression on the data. The dependent variable was the log transformed average of passengers boarding and alighting at each route-stop. The log transformation helped to keep the ridership data from skewing towards 0. Table 4 shows the results of the 2009 bus model.

Table 4: 2009 MUNI Bus Direct Ridership Route-Stop-Daily Model Results

Variable	Variable Name	Coefficients	T-Statistics	Observations
Intercept		-0.8151	-6.25	6261
Potential Demand (Independent Variables)				R-squared
EDD Employment (Log+1)	EDD_EMP_LOG	0.1337	15	0.515
Housing Density (Log+1)	HOUSING_DEN_LOG	0.1056	8.45	
High Income Households (2009 $)	SHR_INCOME_100P	-1.2371	-14.71	
On-Street Parking Cost (2014 $)	PARK_HOURLY_AVG_ON_LOG	0.0231	3.44	
BART Ridership (Log+1)	AVG_BART_LOG	0.059	6.87	
Rail Ridership	MUNI_RAIL_AVG	7.40E-05	3.04	
Transbay Terminal (Bus Station)	TRANSBAY	0.8029	3.72	
Transit Supply (Independent Variables)				
Frequency (Log+1)	FREQ_S_LOG	2.8359	57.78	
Reliability	ONTIME5	0.5476	5.96	
Close Route-Stops	CLOSE_STOP	-1.2928	-25.80	
Bus Terminus	EOL_SOL	0.7281	12.75	
Limited Route Configuration	LIMITED	-1.2006	-18.35	
Express Route Configuration	EXPRESS	-1.911	-38.63	

Each of the terms included in the model are significant at the 95% level or better. For those variables that are log-transformed, the coefficients can be interpreted directly as

an elasticity, since the left-hand side is also log transformed. Attributes that are measured based on the area (total employment, housing density, etc.) are calculated within a tenth-mile buffer of the stop. Attributes associated with the service itself (frequency, on-time performance, etc.) are measured at the route-stop itself. Demographic measures from the ACS (income shares) are measured based on the census tract that contains the stop.

The intercept can be interpreted as a scaler. The scaling value is found by taking the exponent of the intercept, which is found to be 0.44. For this model, an intercept is used to adjust the line-of-fit to better fit the data.

The positive employment coefficient is expected, indicating that ridership grows with increasing employment. We also tested employment segmented by different industry categories, including retail, hotel and restaurant, and education and health, but found that these specifications did not result in logical coefficients.

Higher frequency service (measured in average vehicles per hour) is associated with higher ridership. This relationship is logical, but the model appears to be quite sensitive to frequency, with a 1% increase in frequency associated with a 2.8% increase in ridership. For comparison, other studies have found this elasticity to be within the range of 0.1 to 1.04. (Balcombe et al. 2004) This high sensitivity may occur due to endogeneity with the dependent variable, or due to collinearity with other descriptive variables. Either way, the highest frequency service, the highest employment densities and the highest ridership all occur in downtown San Francisco, making it difficult for any cross-sectional model to parse out the differences. Headway, a linear treatment of frequency, and several piece-wise linear specifications were also tested, but with no better results.

Ridership is higher at the start and end of lines, in areas with denser housing, and in areas with higher parking costs, all of which we would expect.

Stops in areas with a higher share of households earning $100,000 or more per year tend to have fewer riders per transit stop. This may be because wealthier households have more options to take other modes. It is also noteworthy because virtually all of the net growth in households over the 2009 through 2016 period has been through an increase in households earning $100,000 or more per year (Erhardt et al. 2017).

We measure reliability as the share of buses on that route arriving at that stop on-time, where on-time is defined as no more than one minute early and no more than 5 minutes late (Kittelson & Associates et al. 2013). This is of particular interest, as reliability has been an important focus of MUNIs operational efforts in recent years.

We measure BART ridership as the log transformed Bart ridership calculated as the average of BART boardings and alightings within the buffer area. The positive coefficient indicates net complementarity between BART and MUNI, with transfers between the two. Boardings and alightings on MUNI light rail stops within the buffer area also have a positive and significant coefficient. Stops near the Transbay Bus

Terminal, which serves AC Transit commuter buses from the East Bay, also have higher ridership, likely due to a similar transfer effect.

CLOSE_STOP is a dummy variable representing whether or not the previous stop on the same route is within 0.2 miles. It is included to account for competition with other stops on the same route, and the negative coefficient indicates that closely spaced stops have lower ridership per stop, as we would expect.

Limited and express routes tend to have fewer riders per stop, which may be associated with relatively worse off-peak service.

The overall R^2 of the model is 0.515 relative to a constants-only model.

3.3.4 Bus Route-Stop-TOD

The bus route-stop-tod model used the same python model estimation package. The dependent variable was the log transformed average of passengers boarding and alighting at each route-stop-tod. The differences from the route-stop model are: picking up on the transit supply variables, such as frequency, changes throughout a day, inclusion of TOD dummy variables, and the changes in rail ridership throughout a day is included. BART ridership and other demand drivers were not available by TOD, and so the same daily ridership is assigned to each TOD. Table 5 shows the results for the route-stop-tod base year model estimation.

Table 5: 2009 MUNI Bus Direct Ridership Route-Stop-TOD Model Results

Variable	Variable Name	Coefficients	T-Statistics	Observations
Intercept		-0.4269	-3.285	31416
Potential Demand (Independent Variables)				R-squared
EDD Employment (Log+1)	EDD_EMP_LOG	0.124	36.682	0.554
Housing Density (Log+1)	HOUSING_DEN_LOG	0.085	18.415	
High Income Households (2009 $)	SHR_INCOME_100P	-1.178	-37.586	
On-Street Parking Cost (2014 $)	PARK_HOURLY_AVG_ ON_LOG	0.053	20.948	
BART Ridership (Log+1)	AVG_BART_LOG	0.054	15.860	
Rail Ridership	MUNI_RAIL_AVG_LOG	0.046	17.381	
Transbay Terminal (Bus Station)	TRANSBAY	0.780	9.439	
Transit Supply (Independent Variables)				
Frequency (Log+1)	FREQ_S_LOG	1.323	85.570	
Reliability	ONTIME5	0.174	6.203	
Close Route-Stops	CLOSE_STOP	-1.000	-8.073	
Bus Terminus	EOL_SOL	0.847	26.772	
Limited Route Configuration	LIMITED	0.102	3.042	
Express Route Configuration	EXPRESS	-0.111	-3.398	
Times-of-Day (TODs)				
3AM-6AM	AM3_6AM	-0.839	-38.062	
6AM-9AM	AM6_9AM	0.502	25.986	
9AM-2PM	AM9_2PM	0.971	50.787	
2PM-4PM	PM2_4PM	0.487	25.469	
4PM-7PM	PM4_7PM	0.708	37.036	
7PM-10PM	PM7_10PM	0.142	7.526	

The route-stop-TOD model can be interpreted in the same fashion as the route-stop. With the log-transformed variables interpreted as an increase of 1 resulting in a β_x percent increase in bus ridership, and the untransformed variables are interpreted as an increase of 1 resulting in a $(e^{\beta_x} - 1) * 100$ percent increase in ridership. When a route-stop-tod is used as the unit of observation the coefficient of variables that vary throughout the day are reduced. This solves the issue of the route-stop model being too sensitive to frequency. The frequency coefficient for the TOD route-stop- model is less than half of the daily route-stop model coefficient. The TOD model has a more reasonable frequency coefficient, based on previous research (Taylor et al. 2009). Reliability follows a similar path. By disaggregating to TOD, the model becomes less sensitive to reliability. The overall R^2 is higher, 0.554, for the route-stop-tod model than the route-stop model. The competing stops (CLOSE_STOP) and terminus stops (EOL_SOL) variables are similar to the daily model results. The highest coefficient, out of the TOD dummy variables, is

9AM-2PM. This is likely because it is the longest TOD, but it does follow trends found in previous research, and suggests that a growing primary use of transit is for recreational purposes during the mid-day peak. (Federal Transit Administration 2016) When comparing to the daily bus DRM results, the muni rail ridership coefficient increased significantly. This is due to the data being log-transformed for this model, which changes the interpretation to a 1% increase in MUNI rail ridership results in 0.046% increase in MUNI bus ridership. The transformation fit better for this model, but the coefficient still shows a complementing relationship between the modes and a rough comparison of coefficients would be multiplying the untransformed coefficient by 100. Limited and Express routes go from a negative coefficient to positive. Those are a fairly new service MUNI is offering named skip-stop routes. The routes only operate during specific TODs, so this disaggregate model picks up better on their effect on bus ridership.

The intercept is again a scaler that is used to fit the data better by adjusting the slope. After taking the exponent, the scaler is found to be 0.66.

3.3.5 Rail Route-Stop Model

The rail model used the same unit of observation, route-stop. The dependent variable was the average of MUNI light-rail passengers boarding and alighting at each route-stop. There are fewer observations because there are less rail stops than bus stops within San Francisco. Table 6 shows the estimation results for this model.

Table 6 MUNI Rail Direct Ridership Route-Stop-Daily Model Results

Variable	Variable Name	Coefficients	T-Statistics	Observations
Intercept		2.1822	-6.245	315
Potential Demand (Independent Variables)				R-squared
EDD Employment (Log)	EDD_EMP_LOG	0.1607	4.29	0.482
Housing Density (Log)	HOUSING_DEN_LOG	0.2755	4.349	
BART Ridership (Log)	AVG_BART_LOG	-0.0652	-2.935	
Bus Ridership	MUNI_BUS_AVG	6.15E-05	4.452	
Transit Supply (Independent Variables)				
Frequency (Log)	FREQ_S_LOG	0.3042	1.686	
Close Stop	CLOSE_STOP	-0.3706	-3.18	
Bus Terminus	EOL_SOL	0.3129	1.89	
Route J Route-Stops	J	-0.6166	-5.158	

The rail model specification broadly follows that of the bus model, although it is simplified both as some insignificant variables drop out of the model, and because we also lack some of the operational measures, such as on time performance, that are available through the APC data on the buses. Station-area attributes for the rail model are tabulated using a 0.25 mile buffer, reflecting a higher willingness to walk to rail and longer distances between the stations. 0.1 and 0.33 mile buffers were also tested and rejected. All variables included in the model are significant at the 90% level.

Since the model is multiplicative, the intercept is interpreted as a scaler, and it is used to fit the data better. The scaler is found to be 8.87.

The coefficient on the log transformed employment is 0.16, slightly higher than what is found in the bus model. Specifications that segmented the employment by industry were unsatisfactory. Housing density is also positive and significant.

The log of transit frequency in the rail model is 0.3, and significant at the 90% level. This is much lower than is found in the bus model, and may reflect the fact that with fewer rail routes, there is less variation in frequency between stops.

Unlike with the bus ridership, BART boardings in the buffer area are negatively correlated with MUNI rail ridership. This may reflect stronger competition and less complementarity between the two rail systems. The model does still show

complementarity with MUNI bus boardings and alightings within the buffer area. That symmetry is encouraging.

As with the bus model, a nearby rail stop on the same route draws down the ridership at the stop in question. Both end of line and start of line stations tend to have higher ridership as expected.

A constant on the J-line is negative and significant. Constants tested on each of the other lines were insignificant and left out of the model. Many of the J-line stops are also BART stations. The negative coefficient could mean that the J-line is competing more directly with BART than the other routes.

The overall R^2 of the model is 0.482 relative to a constants-only model.

3.3.6 Rail Route-Stop-TOD

The rail route-stop-tod model used the same python model estimation package. The dependent variable was the log transformed average of MUNI light-rail passengers boarding and alighting at each route-stop-tod. The differences from the route-stop model are: picking up on the transit supply variables, such as frequency, changes throughout a day, inclusion of TOD dummy variables, and the changes in bus ridership throughout a day is included. BART ridership was not available by TOD, and so the same daily ridership is assigned to each TOD. Table 7 shows the results for the route-stop-tod base year model estimation.

Table 7 MUNI Rail Direct Ridership Route-Stop-TOD Model Results

Variable	Variable Name	Coefficients	T-Statistics	Observations
Intercept		-0.557	-2.514	2176
Potential Demand (Independent Variables)				R-Squared
EDD Employment (Log)	EDD_EMP_LOG	0.1703	10.201	0.553
Housing Density (Log)	HOUSING_DEN_LOG	0.1214	4.539	
BART Ridership	AVG_BART_LOG	-0.049	-5.077	
Bus Ridership	MUNI_BUS_AVG	5.11E-05	8.457	
Transit Supply (Independent Variables)				
Frequency (Log)	FREQ_S_LOG	0.409	7.022	
Route J Route-Stops	J	-0.602	-11.411	
Times-of-Day (TODs) (Independent Variables)				
3AM - 6AM	AM3_6AM	0.617	8.977	
6AM - 9AM	AM6_9AM	1.442	18.700	
9AM - 2PM	AM9_2PM	1.956	25.661	
2PM - 4PM	PM2_4PM	1.401	19.595	
4PM - 7PM	PM4_7PM	1.791	23.762	
7PM - 10PM	PM7_10PM	0.763	10.754	

The route-stop-tod model can be interpreted in the same fashion as the route-stop. With the log-transformed variables interpreted as an increase of 1 resulting in a β_x percent increase in rail ridership, and the untransformed variables are interpreted as an increase of 1 resulting in a $(e^{\beta_x} - 1) * 100$ percent increase in ridership. When a route-stop-tod is used as the unit of observation the coefficient of variables that change throughout a day are reduced. The reduction is not as significant as it is in the bus models, but the variables were more reasonable to begin with. The coefficient for housing density is cut in half, meaning that housing density may have been too sensitive in the route-stop model. This is similar to the bus TOD model properly reducing its sensitivity to frequency. The highest coefficient, out of the TOD dummies, is the same as the bus TOD model, 9AM-2PM. This is most likely due to the 9am-2pm variable being the longest TOD. When comparing to the daily rail DRM, the model's sensitivity to housing density is more than halved. Both coefficients are within a reasonable range, but most likely the daily rail model was over-sensitive to housing density. Similar to the dynamic the bus models have with frequency, but at a lesser magnitude. The overall R^2 is higher, 0.553, for the route-stop-tod model than the route-stop model. The competing stops (CLOSE_STOP) and terminus stops (EOL_SOL) variables were insignificant in the route-stop-tod model.

3.3.7 Other Variables Tested

There were numerous variables included in the estimation file and not included in the final model. A few of importance are the share of households with 0 vehicles and fares. The share of households with 0 vehicles is potentially explained with other variables, such as housing density. However, fares only change 25 cents between 2009 and 2016. After accounting for inflation the change is only 6 cents, which was considered to be negligible.

3.4 Model Assumptions

The direct ridership models in this thesis use multiple-linear regression. Multiple-linear regression models have four assumptions that have to be valid, before a model can be valid. The four assumptions are (Greene 2003):

1) There must be a linear relationship between the outcome variable and the independent variables (not a curvilinear).
2) Multivariate Normality: Residuals must be normally distributed.
3) No Multi-collinearity: The independent variables are not highly correlated to each other.
4) Homoscedasticity: The variance of error terms are similar across the values of the independent variables. In other words, a scatter plot of the residuals does not have a noticeable pattern.

There are techniques in python to check the validity of each of the assumptions. The rest of this section is divided up explaining the bus TOD model assumptions followed by the same discussion for the rail TOD model. The assumptions for the daily models were checked, but are not shown. This is because the TOD models were concluded to be better than the daily.

3.4.1 Route-Stop-TOD Bus Model Assumptions

The outcome variable and independent variables having a linear, rather than curvilinear, relationship was checked first. Figure 5 shows the scatter plots of the independent variables versus the predicted response variable are made to check the assumption. The far left plot is predicted ridership versus observed and the far right is the distribution of log-transformed ridership. The plots in-between are all of the explanatory variables. The dummy variables are not included to simplify the figure. The BART and rail ridership variables have many zeros, because many of the route-stops do not have a rail station near them. The EOL_SOL variable ranges from 0 to 1. Values in-between 0 and 1 are route-stops that are a terminus stop for either one or two of the three months included. The spread is fairly high for many of the plots, but there is not any curvilinear relationships.

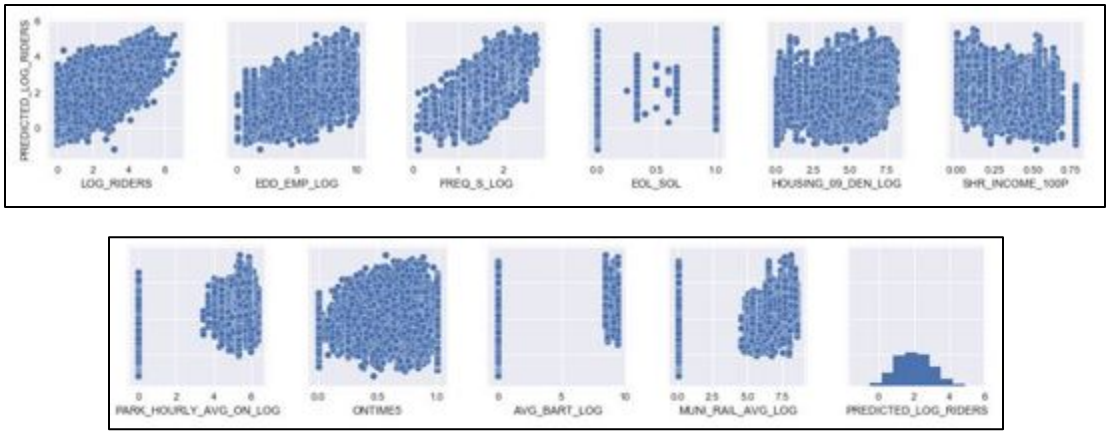

Figure 5: Response Variable vs. Explanatory Variables Scatter Plots

Multivariate normality assumes that the residuals are normally distributed (Greene 2003). Figure 6 and Figure 7 show the distribution of residuals and QQ plot for the route-stop-TOD bus model respectively. If the residuals had a perfect normal distribution, then the QQ plot would be a straight diagonal line. The distribution is not perfectly normal, but is adequate for the multivariate normality assumption to be valid.

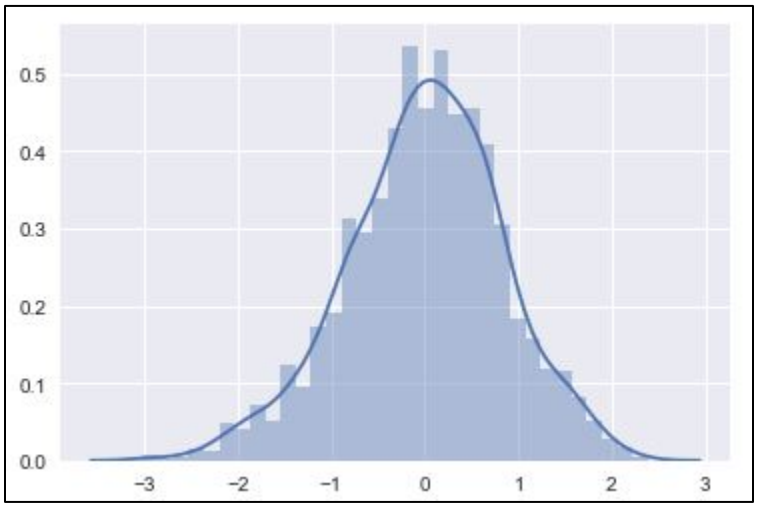

Figure 6: Bus Route-Stop-TOD Model Distribution of Residuals

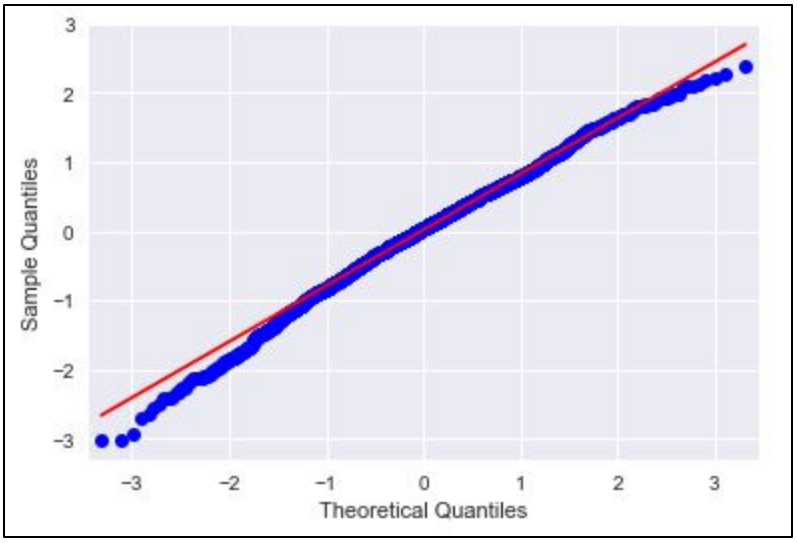

Figure 7: Bus Route-Stop-TOD Model QQ Plot

The no multi-collinearity assumption is to prevent two variables from explaining the same variance. This is an issue for DRMs. Variables such as employment density and frequency are going to be spatially correlated. Meaning that areas with high frequency typically have high employment density. There are a couple of ways to verify if your model has multi-collinearity. One is to create a correlation matrix and the other is to calculate variance inflation factors (VIF) (Murray et al. 2012; O'brien 2007). VIFs measure how much the variance of the estimated regression coefficients are inflated, compared to when the explanatory variables are not linearly related. Table 8 shows the correlation matrix and Table 9 shows the VIFs, both are for the bus route-stop-TOD model. To simplify Table 8, all dummy variables are left out of the correlation matrix. The column and row names have been abbreviated to help the table fit better. Appendix B is the full correlation matrix. The correlation of two variables ranges from -1 to 1, meaning perfectly uncorrelated and perfectly correlated respectively. Table 9 includes the same variables shown in Table 8. A rule-of-thumb is any variance factor above 10 is considered to be a result of multi-collinearity. Both of the tables show that employment density, frequency, housing density, and reliability variables invalidate the no multi-collinearity assumption. This is a problem that spatial models, such as cross-sectional DRMs, face (Blainey and Preston 2010; Cardozo, García-Palomares, and Gutiérrez 2012). In section 3.7 this problem is discussed further and a solution is discussed.

Table 8: Bus Route-Stop-TOD Model Correlation Matrix

	EDD EMP	FREQ_S	EOL SOL	HOUSING DEN_LOG	INCOME 100P	PARK HOURLY	ONTIME5	BART	MUNI RAIL
EDD_EMP	1.00	0.16	0.07	0.00	-0.15	0.66	0.11	0.21	0.15
FREQ_S	0.16	1.00	-0.01	0.07	-0.08	0.14	0.02	0.02	-0.02
EOL_SOL	0.07	-0.01	1.00	-0.10	0.06	0.06	-0.01	0.04	0.12
HOUSING_09_DEN	0.00	0.07	-0.10	1.00	-0.23	0.19	0.04	-0.12	-0.12
SHR_INCOME_100P	-0.15	-0.08	0.06	-0.23	1.00	-0.20	0.01	-0.07	0.03
PARK_HOURLY_AVG_ON	0.66	0.14	0.06	0.19	-0.20	1.00	0.09	0.17	0.08
ONTIME5	0.11	0.02	-0.01	0.04	0.01	0.09	1.00	0.01	0.00
AVG_BART	0.21	0.02	0.04	-0.12	-0.07	0.17	0.01	1.00	0.37
MUNI_RAIL_AVG	0.15	-0.02	0.12	-0.12	0.03	0.08	0.00	0.37	1.00

Table 9: Bus Route-Stop-TOD Model Variance Inflation Factors

VIF Factor	Features
13.9	EDD_EMP_LOG
25.4	FREQ_S_LOG
1.1	EOL_SOL
27.3	HOUSING_09_DEN_LOG
6.3	SHR_INCOME_100P
3.1	PARK_HOURLY_AVG_ON_LOG
14.6	ONTIME5
1.3	AVG_BART_LOG
68.6	CLOSE_STOP
1.3	MUNI_RAIL_AVG_LOG

Heteroscedasticity assumes that the residuals of the model are not correlated with each other. This can be verified by plotting the residuals and visually checking for any patterns. Figure 8 shows the scatter plot of the residuals for the route-stop-TOD bus model. There is not a clear visual pattern, so the heteroscedasticity assumption is valid.

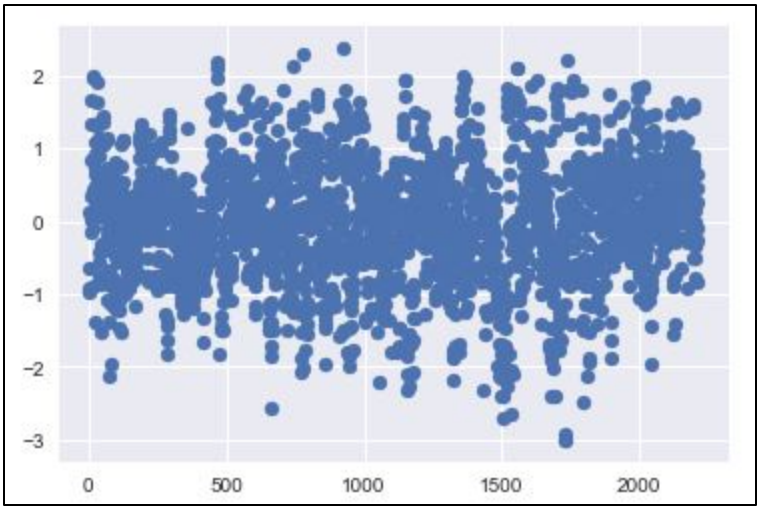

Figure 8: Bus Route-Stop-TOD Scatter Plot of Residuals

3.4.2 Route-Stop-TOD Rail Model Assumptions

The rail route-stop-TOD model requires the same four assumptions to be valid. The first being that the outcome variable and explanatory variables have a linear relationship. Figure 9 is scatter plots of the response variable, predicted log-transformed ridership, versus the explanatory variables. The far left is predicted ridership versus observed ridership and the far right is the distribution of the predicted ridership. Similar to the bus model, the spread is large, but there is not any evidence of a curvilinear relationship. Thus, the linear relationship assumption is valid.

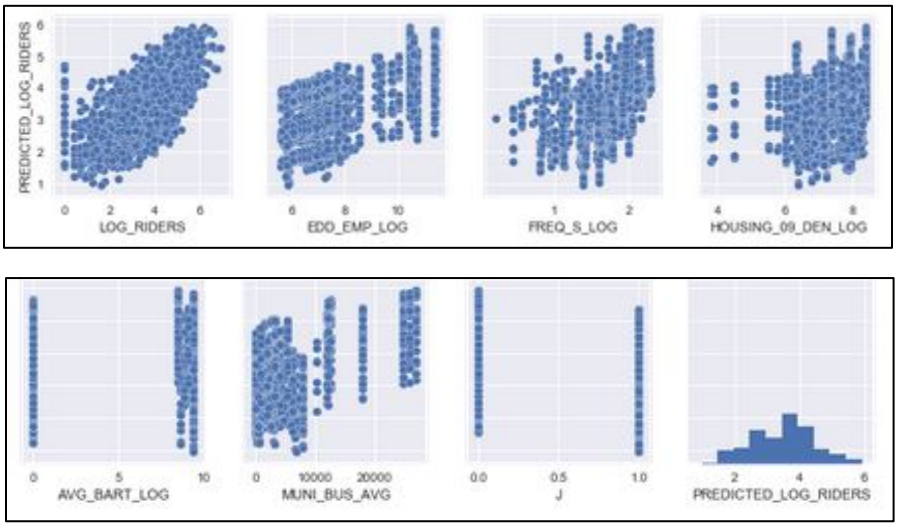

Figure 9: Response Variable vs. Explanatory Variables Scatter Plots

Multivariate normality assumes that the residuals are normally distributed. Figure 10 and Figure 11 show the distribution of residuals and QQ plot for the route-stop-TOD rail model respectively. If the residuals had a perfect normal distribution, then the QQ plot would be a straight diagonal line. The distribution is not perfectly normal, but is adequate for the multivariate normality assumption to be valid.

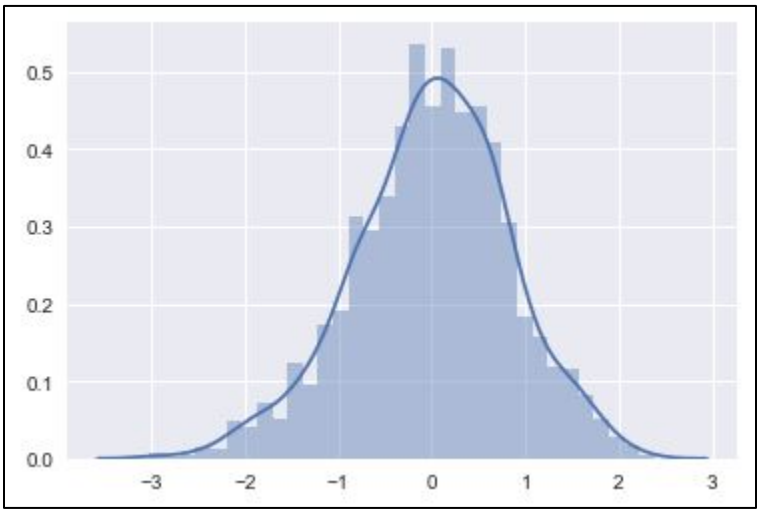

Figure 10: Rail Route-Stop-TOD Distribution of Residuals

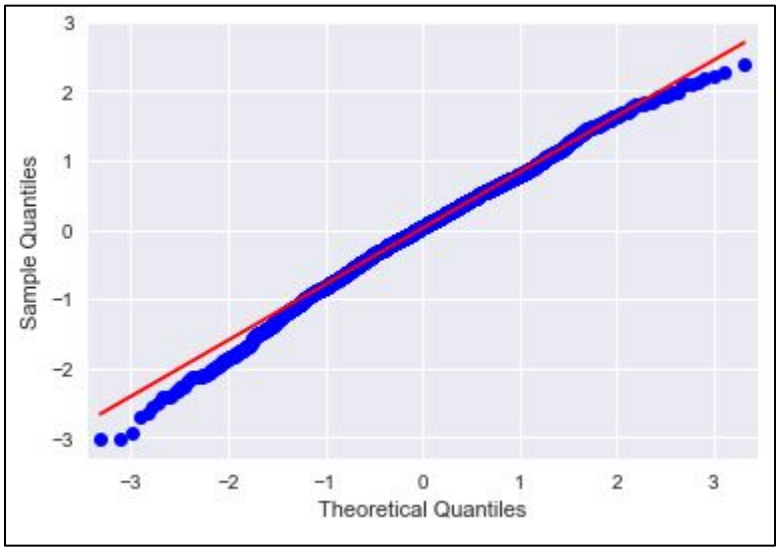

Figure 11: Rail Route-Stop-TOD QQ Plot

The no multi-collinearity assumption, for the rail model, is verified the same way the bus model is. Table 10 and Table 11 are the correlation matrix and variance inflation factors, respectively, for the rail route-stop-TOD model. The rail and bus DRMs have the same issue, having certain explanatory variables correlated with each other. Both tables show that frequency, employment density, and housing density are correlated for the rail

model. The same three variables were found to be correlated in the bus model, section 3.7 discusses the issue in more detail, and they are addressed in chapter 4. In spite of these limitations, we continue with the exercise of applying and validating these models, because the purpose is understanding what factors are influencing ridership. Chapter 4 quantifies the effect that the influencing variables have on ridership.

Table 10: Rail Route-Stop-TOD Model Correlation Matrix

	EDD EMP	FREQ_S	HOUSING_DEN	AVG BART	MUNI BUS	AM3 6AM	AM6 9AM	AM9 2PM	PM2 4PM	PM4 7PM	PM7 10PM	J
EDD_EMP_LOG	1.00	0.03	0.26	0.52	0.75	0.00	0.00	0.00	0.00	0.00	0.00	-0.02
FREQ_S_LOG	0.03	1.00	0.03	0.02	0.03	-0.23	0.26	0.23	0.00	0.19	-0.05	-0.03
HOUSING_09_DEN_LOG	0.26	0.03	1.00	0.08	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.23
AVG_BART_LOG	0.52	0.02	0.08	1.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.13
MUNI_BUS_AVG	0.75	0.03	0.24	0.82	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
AM3_6AM	0.00	-0.23	0.00	0.00	0.00	1.00	-0.17	-0.17	-0.17	-0.17	-0.17	0.00
AM6_9AM	0.00	0.26	0.00	0.00	0.00	-0.17	1.00	-0.17	-0.17	-0.17	-0.17	0.00
AM9_2PM	0.00	0.23	0.00	0.00	0.00	-0.17	-0.17	1.00	-0.17	-0.17	-0.17	0.00
PM2_4PM	0.00	0.00	0.00	0.00	0.00	-0.17	-0.17	-0.17	1.00	-0.17	-0.17	0.00
PM4_7PM	0.00	0.19	0.00	0.00	0.00	-0.17	-0.17	-0.17	-0.17	1.00	-0.17	0.00
PM7_10PM	0.00	-0.05	0.00	0.00	0.00	-0.17	-0.17	-0.17	-0.17	-0.17	1.00	0.00
J	-0.02	-0.03	0.23	0.13	0.09	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 11: Rail Route-Stop-TOD Model Variance Inflation Factors

VIF Factor	Variables
44.2	EDD_EMP_LOG
28.9	FREQ_S_LOG
48.4	HOUSING_09_DEN_LOG
3.7	AVG_BART_LOG
6.0	MUNI_BUS_AVG
2.0	AM3_6AM
2.6	AM6_9AM
2.6	AM9_2PM
2.2	PM2_4PM
2.5	PM4_7PM
2.2	PM7_10PM
1.3	J

Heteroscedasticity assumes that the residuals of the model are not correlated with each other (Greene 2003). This can be verified by plotting the residuals and visually checking for any patterns. Figure 12 shows the scatter plot of the residuals for the route-stop-TOD rail model. There is not a clear visual pattern, so the homoscedasticity assumption is valid.

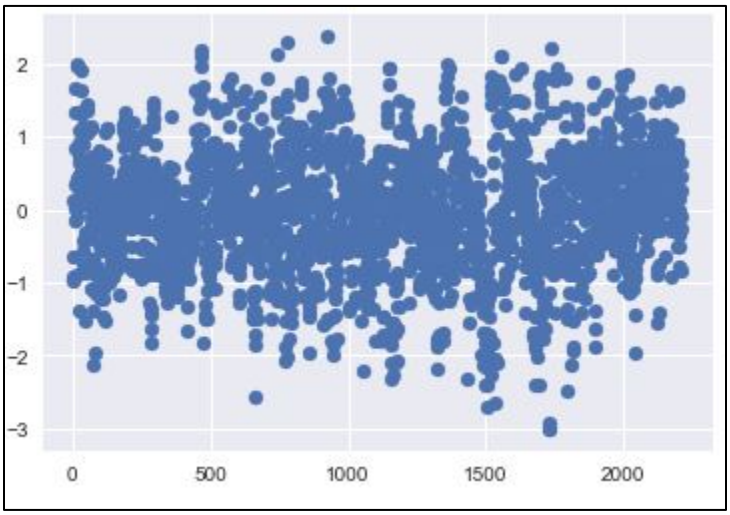

Figure 12: Rail Route-Stop-TOD Model Scatter Plot of Residuals

3.5 Model Application

The model was applied with the 2009 data to evaluate the models accuracy. To evaluate the forecasting ability of the models, the 2009 bus and rail models were applied using an independent data set, 2016 data that was processed in the same manner as the 2009 data. In contrast to the model estimation, which limited the data set to route-stops with observed data for all variables, the models were applied to the full set of route-stops in San Francisco (excluding a few outside the county line). The focus of the examination was on how well the models predict change, rather than their cross-sectional fit.

3.5.1 Route-Stop-Daily Model Results

Table 12 shows the system-level results. The 2009 bus model under-predicts total ridership by 11%, and the rail model under-predicts total ridership by 13%. Between 2009 and 2016, the data show that bus ridership decreases by 13%, whereas the model predicts a 2% increase. The data show rail ridership increasing by 17% over this period, while the model only predicts a 3% increase. For the rail model, the percent root mean square error (RMSE) is slightly better for the 2016 application than the 2009 application. The opposite result is found for the bus model. The percent RMSE are fairly large due to the disaggregate nature of a route-stop. The percent RMSE is given at two aggregated

levels, stop and route. The general trend is that a better percent RMSE is found when more aggregated data is used. The average observed ridership for a bus route-stop is 70, a stop is 130, and a route is 5,500. For rail, the average observed ridership for a route-stop is 410, a stop is 530, and a route is 25,600.

Table 12: System Level Route-Stop-Daily Model Application Results

		Observed Ridership	Modeled Ridership	Difference	Percent Difference	Route-Stop % RMSE	Stop % RMSE	Route % RMSE
Bus	2009	515,059	459,602	-55,456	-11%	409%	369%	150%
	2016	449,819	467,522	17,704	4%	424%	369%	72%
	Change	-65,240	7,920					
	P Change	-13%	2%					
Rail	2009	147,470	128,042	-19,428	-13%	67%	82%	20%
	2016	171,442	131,721	-29,666	-18%	60%	79%	21%
	Change	23,972	3,679					
	P Change	16%	3%					

3.5.2 Route-Stop-TOD Model Results

Table 13 and Table 14 show the system-level results for the bus TOD model. The ridership totals for each year are slightly different than the daily totals. The process of aggregating the initial GTFS data from a time-of-day to a daily total causes the slight change. The 2009 bus model under-predicts total ridership by 30%. Between 2009 and 2016, the data show that bus ridership decreases by 14%, whereas the model only predicts a 6% decline. Generally the time period that has the largest miss-prediction is during the mid-day period, 9AM-2PM. Followed up by the typical AM and PM peak periods, 6AM-9AM and 4PM-7PM. The model results in Table 5 show that 9am-2pm had the highest coefficient. The large over-prediction could be due to the 9am-2pm being the longest TOD. The predicted totals are off by a good amount, but the focus of applying this model is to predict the change in ridership correctly. The TOD model predicts the change in ridership significantly better than the previous daily model.

Table 13: Bus System Level Route-Stop-TOD Predicted Totals Results

			Observed Ridership	Modeled Ridership	Difference	Percent Difference	Route-Stop % RMSE	Stop % RMSE	Route % RMSE
Bus .	2009 .	3AM - 6AM	5,244	4,600	-644	-12%			
		6AM - 9AM	101,219	74,180	-27,039	-27%			
		9AM - 2PM	146,600	97,182	-49,418	-34%			
		2PM - 4PM	80,002	61,643	-18,359	-23%			
		4PM - 7PM	114,692	86,702	-27,990	-24%			
		7PM - 10PM	33,824	25,304	-8,520	-25%			
		10PM - 3AM	17,768	14,375	-3,393	-19%			
		Total	522,531	363,985	-158,546	-30%	167%	184%	87%
	2016	3AM - 6AM	4,225	3,887	-338	-8%			
		6AM - 9AM	93,816	67,092	-26,724	-28%			
		9AM - 2PM	136,013	99,951	-36,062	-27%			
		2PM - 4PM	73,553	58,457	-15,096	-21%			
		4PM - 7PM	98,835	77,922	-20,913	-21%			
		7PM - 10PM	30,208	24,492	-5,716	-19%			
		10PM - 3AM	14,911	12,155	-2,756	-18%			
		Total	451,562	343,956	-107,606	-24%	165%	200%	69%

Table 14: Bus System Level Route-Stop-TOD Predicted Change Results

TOD	Observed Change	Modeled Change	Observed % Change	Modeled % Change
3AM - 6AM	-1,019	-713	-19%	-16%
6AM - 9AM	-7,403	-7,088	-7%	-10%
9AM - 2PM	-10,587	2,769	-7%	3%
2PM - 4PM	-6,449	-3,186	-8%	-5%
4PM - 7PM	-15,857	-8,780	-14%	-10%
7PM - 10PM	-3,616	-812	-11%	-3%
10PM - 3AM	-2,857	-2,220	-16%	-15%
Total	-70,969	-20,030	-14%	-6%

Table 15 and Table 16 show the system-level results for the rail TOD model. Similar to the bus models, the observed TOD ridership totals are slightly different than the daily totals. The slight change is due to the same aggregating process. The 2009 rail model under-predicts total ridership by 26%. Between 2009 and 2016, the data show that rail ridership increases by 17%, whereas the model under-predicts a 6% increase. The

time-period constants do not follow the same trend as the bus model. The only noticeable trend for the rail model is that the overnight periods are generally under-predicted. The model results in Table 7 do show that the overnight TODs have the lowest coefficients. With that being said, it is intuitive for them to be lower, because overnight ridership during weekdays should be lower than during the day. The estimated ridership totals are off by a good amount for the rail model too, but the focus was on how well the model predicted change. When comparing the TOD and daily rail models, the accuracy of the model to predict change increases.

Table 15: Rail System Level Route-Stop-TOD Predicted Ridership Totals Results

			Observed Ridership	Modeled Ridership	Difference	Percent Difference	Route-Stop % RMSE	Stop % RMSE	Route % RMSE
Rail.	2009.	3AM - 6AM	9,701	6,422	(3,279)	-34%			
		6AM - 9AM	25,870	18,076	(7,793)	-30%			
		9AM - 2PM	40,064	30,620	(9,444)	-24%			
		2PM - 4PM	19,883	16,364	(3,519)	-18%			
		4PM - 7PM	35,176	25,584	(9,592)	-27%			
		7PM - 10PM	11,158	7,958	(3,200)	-29%			
		10PM - 3AM	4,273	2,999	(1,273)	-30%			
		Total	146,124	108,024	(38,100)	-26%	103%	124%	38%
	2016	3AM - 6AM	2,925	6,640	3,715	127%			
		6AM - 9AM	38,768	19,892	(18,876)	-49%			
		9AM - 2PM	43,070	31,602	(11,468)	-27%			
		2PM - 4PM	24,734	18,581	(6,153)	-25%			
		4PM - 7PM	44,049	27,369	(16,680)	-38%			
		7PM - 10PM	14,638	8,199	(6,439)	-44%			
		10PM - 3AM	3,258	2,701	(557)	-17%			
		Total	171,442	114,985	(56,457)	-33%	111%	165%	53%

Table 16: Rail System Level Route-Stop-TOD Predicted Ridership Change Results

TOD	Observed Change	Modeled Change	Observed % Change	Modeled % Change
3AM - 6AM	(6,776)	218	-70%	3%
6AM - 9AM	12,898	1,816	50%	10%
9AM - 2PM	3,006	982	8%	3%
2PM - 4PM	4,851	2,217	24%	14%
4PM - 7PM	8,873	1,785	25%	7%
7PM - 10PM	3,480	241	31%	3%
10PM - 3AM	(1,015)	(298)	-24%	-10%
Total	25,318	6,961	17%	6%

3.5.3 Model Accuracy

Both the rail and bus models miss-predict ridership. One way to combat miss-predictions is to calibrate the TOD dummies to match the TOD observed totals. Calibrating the TOD dummies was tested, but adjusting the TOD dummies to match 2009 conditions then applying them to 2016 significantly altered the predicted change. The purpose of this chapter is to apply the model to understand what factors are influencing changes in transit ridership, so the model was not calibrated.

Figure 13 and Figure 14 show maps comparing the growth in observed ridership versus the growth in route-stop modeled ridership. Green circles indicate stops where the model predicts too much ridership growth (or too little decline), while red circles indicate stops where the model predicts too little ridership growth (or too much decline). Larger circles indicate a larger absolute difference between the observed and modeled change. The model was applied at an aggregated stop-level for mapping purposes. The route-stops map to the same location and overlap each other on the map.

The bus map does not show an obvious pattern, beyond a general under-estimation of the decline in bus ridership, focused in particular along the Market Street corridor in the Northeast portion of San Francisco. The rail map shows that the model under-estimates the growth of rail ridership at all but a few stops. The stops with the largest radii, meaning largest absolute difference, are underground stations that are shared with Bay Area Rapid Transit (BART). These are high-traffic stops with many riders boarding and alighting.

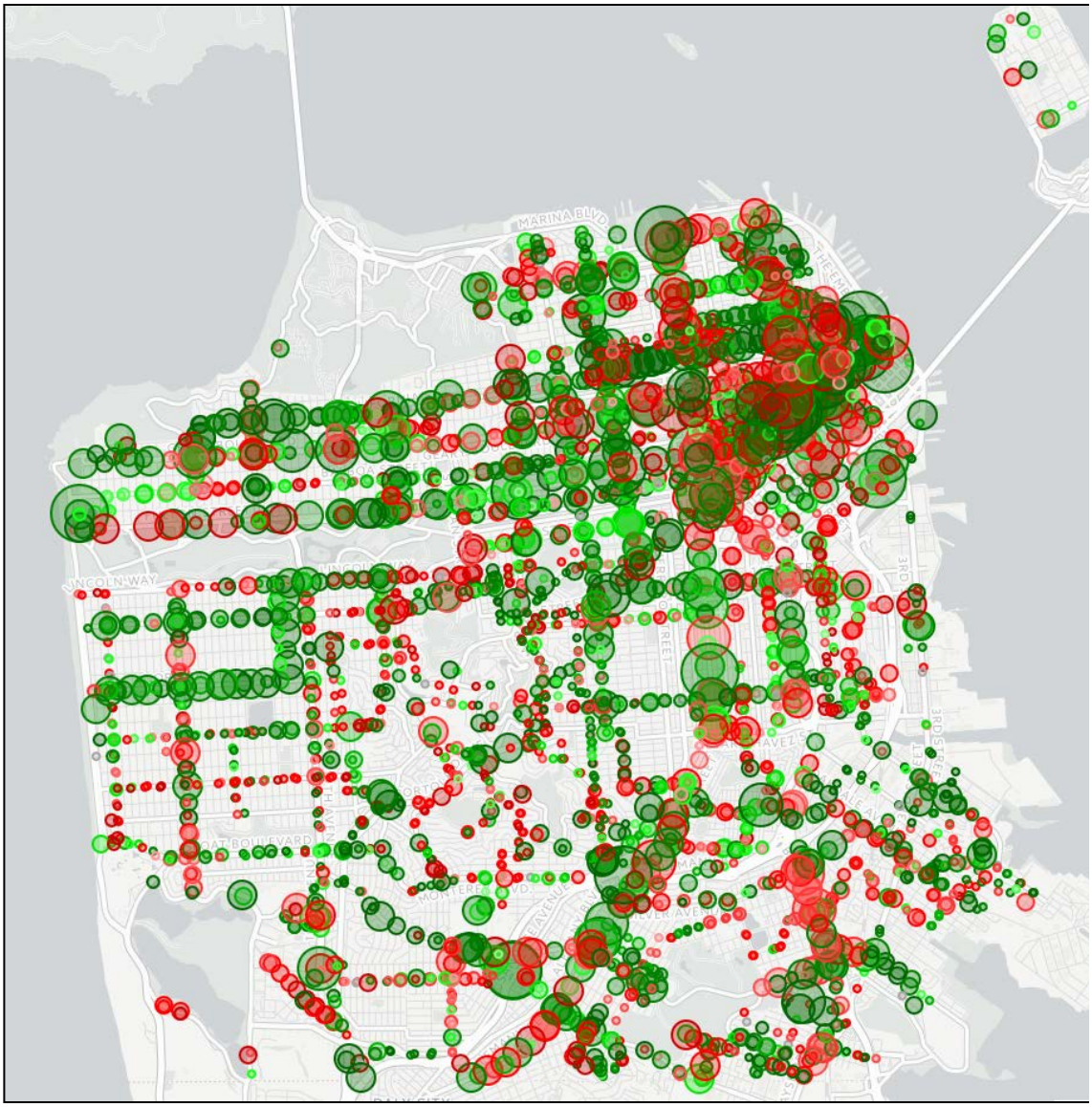

Figure 13: Snapshot of Bus Model Accuracy

Each marker is a bus stop, with the radius being based on the absolute difference between modeled growth and observed growth and the color based on percent difference between modeled growth and observed growth, dark green being when the model greatly overestimates and the opposite for dark red. 1 inch is approximately 3 inches.

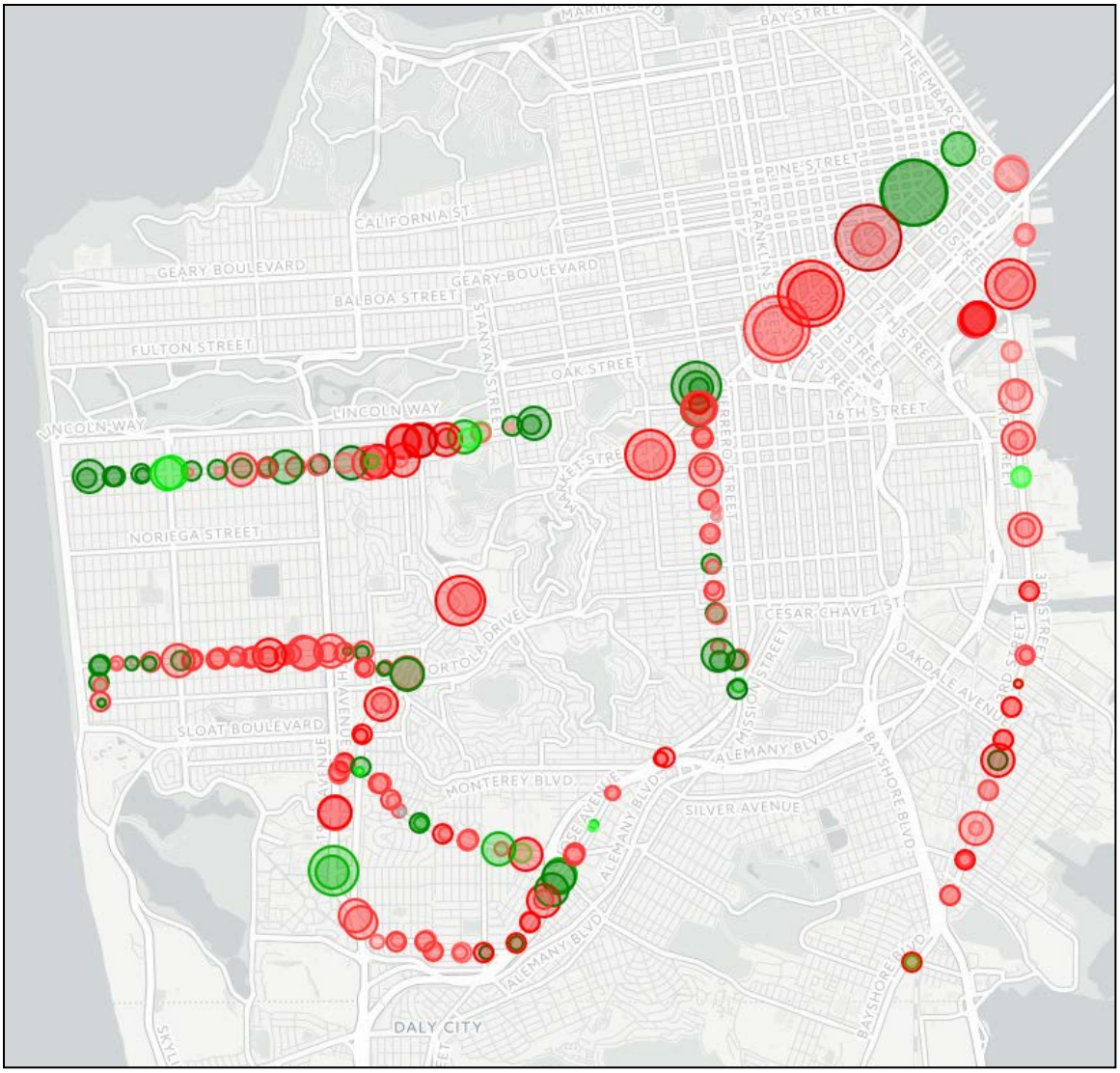

Figure 14: Snapshot of Rail Model Accuracy

Each marker is a bus stop, with the radius being based on the absolute difference between modeled growth and observed growth and the color based on percent difference between modeled growth and observed growth, dark green being when the model greatly overestimates and the opposite for dark red. 1 inch is approximately 3 miles.

3.6 Factors Affecting Change

To better understand what is driving the change in ridership between the two years, a series of sensitivity tests were conducted with the models. The tests focused on the subset of route-stops present in both 2009 and 2016. A baseline total in ridership is estimated by applying the model to route-stops present in both years, using 2009 data only. Then one isolation variable is chosen, for which 2016 data is substituted. The total ridership found after substituting for the isolation variable is then compared to the 2009

baseline total. This follows past work to understand the factors that drive changes in transit ridership, (Upchurch and Kuby 2013), and provides a means for understanding the magnitude of change that can be attributed to that variable.

3.6.1 Route-Stop-Daily Factors Affecting Change Results

Table 17 shows the results of factors affecting change (FAC) exercise using the bus model. The coefficient for frequency was mentioned before to be high, and it shows up here again with a 3 percent increase in frequency resulting in ridership increasing by 21 percent. San Francisco has experienced favorable economic growth in recent years and that is picked up by the high income variable increasing 6 percent, resulting in a 4 percent decrease in bus ridership. MUNI rail ridership has grown significantly and is found to complement bus ridership leading to an increase of 8 percent. Close stops is picking up recent changes in route configurations where the MUNI forward program is reducing the number of stops along a route. Employment has grown resulting in bus ridership increasing 4 percent. The bus system overall has become more unreliable resulting in a decrease of 4 percent in ridership. The other variables do not experience much change and thus have a minimal effect on bus ridership. There is a decline in ridership associated with route-stops in 2009 but not 2016, and an increase associated with route-stops in 2016 but not 2009. The removed and added route-stops dominate the change in ridership, contributing to a 36% decline and 15% incline in bus ridership respectively. This is due to issues when linking the route-stops between years. It covers cases where a stop has been moved, eliminated or added, or where a route does not have a direct correspondence between the two years. Given the schedule changes associated with Muni Forward, this is common. Parking data was only available for 2014, so there was not any observed change in the data. This resulted in the on-street parking cost variable not contributing to the change in ridership.

Table 17: FAC in Route-Stop-Daily Modeled Bus Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	Elasticity _
Intercept	-0.8151						
Potential Demand							
EDD Employment (Log)	0.1337	1197	1434	20%	10,403	4%	0.19
Housing Density (Log)	0.1056	384	396	3%	2,236	1%	0.26
On-Street Parking Cost (Log) (2014 $)	0.0231	$0.94	$0.93	-1%	(173)	0%	0.10
Share of High Income Households (2009 $)	-1.2371	0.36	0.38	6%	(11,245)	-4%	-0.75
BART Ridership (Log)	0.059	282	294	4%	601	0%	0.05
MUNI Rail Ridership	7.40E-05	162	480	197%	11,959	8%	0.04
Transbay Terminal (Bus Station)	0.8029	0.002	0.008	278%	5,076	2%	0.01
Transit Supply							
Frequency (Log)	2.8359	3.93	4.05	3%	37,224	21%	6.89
Reliability	0.5476	0.63	0.58	-8%	(11,091)	-4%	0.49
Close Route-Stops	-1.2928	0.90	0.90	0%	18,458	7%	-0.72
Bus Terminus	0.7281	0.05	0.05	1%	573	0%	0.16
Limited Route Configuration	-1.2006	0.02	0.03	18%	(589)	0%	-0.01
Express Route Configuration	-1.911	0.10	0.10	-2%	1,120	0%	-0.18
Total for Route Stops Present in Both Years--each term applied separately *Percent change uses ridership for route-stops present in both years					64,552	35%	
Total for Route Stops Present in Both Years - all terms applied together *Percent change uses ridership for all route-stops					64,081	12%	
Total for Route Stops Dropped					(187,043)	-36%	
Total for Route Stops Added					130,882	25%	
System Total					7,920	2%	

Table 18 shows the factors affecting change (FAC) sensitivity analysis for the rail model. The changes appear reasonable given the trends in the data, and the frequency contribution is more modest than in the bus model. For the rail model, the bus ridership variable leads to a decrease in rail ridership, whereas the opposite effect is found for rail ridership in the bus model. This is because the rail system complements the bus system. Rail ridership has increased leading to an increase in bus ridership, while bus ridership has experienced a decline resulting in a decrease in bus ridership. Rail ridership being more sensitive to housing density than buses is shown by the contribution to rail ridership being 4 times larger than bus. MUNI rail has not made any skip-stop service change, so the close route stop variable is picking up on the route-stops competing with each other.

Table 18: FAC in Route-Stop-Daily Modeled Rail Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		Elasticity
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	—
Intercept	2.1822						
Potential Demand							
EDD Employment (Log)	0.1607	11926	14933	25%	4,119	3%	0.13
Housing Density (Log)	0.2755	1789	1985	11%	5,116	4%	0.37
BART Ridership (Log)	-0.0652	1473	1348	-8%	-21	0%	-0.07
Bus Ridership	6.15E-05	4297	3977	-7%	-3,959	-3%	0.42
Transit Supply							
Frequency (Log)	0.3042	4.5	4.83	7%	3,181	3%	0.34
Close Route-Stop	-0.3706	0.83	0.86	3%	-3,290	-3%	-0.83
Rail Terminus	0.3129	0.08	0.08	-6%	-415	0%	0.05
Route J Route-Stops	-0.6166	0.16	0.16	0%	-	0%	-0.62
Totals							
Total for Route Stops Present in Both Years--each term applied separately *Percent change uses ridership for route-stops present in both years					4,733	4%	
Total for Route Stops Present in Both Years - all terms applied together *Percent change uses ridership for all route-stops					4,115	3%	
Total for Route Stops Dropped					-2,067	-1%	
Total for Route Stops Added					1,631	1%	
System Total					3,679	3%	

3.6.2 Route-Stop-TOD Factors Affecting Change Results

Table 19 shows the results of this exercise using the bus route-stop-TOD model. The coefficient for frequency is reduced to a more reasonable effect, and it shows up here again with increases in frequency resulting in ridership increasing by 2 percent. The high income variable shows up again with an increase in high income households resulting in a 4 percent decrease of bus ridership. MUNI light-rail ridership is still found to complement bus ridership. The bus system overall has become more unreliable resulting in a decrease of 1 percent in ridership. The Transbay terminal was moved during this period to accommodate the construction of a new terminal. Most likely the 2 percent increase in ridership is associated to the terminal moving to an area of higher ridership. The other variables do not experience much change and thus have a minimal effect on bus ridership. There is a decline in ridership associated with route-stops that were removed some-time between 2009 and 2016, and an increase associated with route-stops added some-time between 2009 and 2016. Again the change in bus ridership is dominated

by the removed and added stops. The time-of-day variables are excluded, because the variables are time-invariant and to simplify the table. Even after accounting for intuitive variables, such as employment and frequency, the model is still under-predicting the change in ridership by 8%. In chapter 4 a transportation network company (TNC) variable is introduced, which may be able to explain a large portion of the residual.

Table 19: FAC in Route-Stop-TOD Modeled Bus Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	Elasticity_
Intercept							
Potential Demand							
EDD Employment (Log)	0.124	935	1136	21%	7,929	4%	0.12
Housing Density (Log)	0.085	395	406	3%	1,308	1%	0.08
On-Street Parking Cost (Log) (2014 $)	0.053	$0.84	$0.83	-1%	(119)	0%	0.05
High Income Households (2009 $)	-1.178	0.37	0.39	5%	(9,344)	-4%	-0.69
BART Ridership (Log)	0.054	236	248	5%	433	0%	0.05
MUNI Rail Ridership (Log)	0.046	122	342	180%	1,894	1%	0.05
Transbay Terminal (Bus Station)	0.78	0.00139	0.00796	473%	5,168	2%	1.18
Transit Supply							
Frequency (Log)	1.323	4.19	4.22	1%	5,267	2%	1.32
Reliability	0.174	0.64	0.60	-6%	(2,622)	-1%	0.19
Close Route-Stops	-1.000	0.98	0.97	-1%	8	0%	-0.63
Bus Terminus	0.847	0.0356	0.0359	1%	0	0%	1.33
Limited Route Configuration	0.102	0.0139	0.0158	14%	-	0%	0.11
Express Route Configuration	-0.111	0.00841	0.00787	-6%	-	0%	-0.11
Total for Route Stops Present in Both Years--each term applied separately *Percent change uses ridership for route-stops present in both years					9,923	7%	
Total for Route Stops Present in Both Years - all terms applied together *Percent change uses ridership for all route-stops					9,784	2%	
Total for Route Stops Dropped					-171,529	-33%	
Total for Route Stops Added					131,636	25%	
System Total					-30,108	-6%	

Table 20 shows the sensitivity analysis for the rail model. The changes appear reasonable given the trends in the data. Since bus ridership is decreasing, the bus ridership variable leads to a decrease in rail ridership. The rail model finds bus ridership complementing rail ridership. The increase in employment contributes the most change in ridership, with a 4 percent increase. Similar to the daily model the decline in bus ridership is decreasing rail ridership, due to the two modes complementing each other.

Table 20: FAC in Route-Stop-TOD Rail Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	Elasticity
Intercept	-0.575						
Potential Demand							
EDD Employment (Log)	0.170	11618	15167	31%	3,893	4%	0.17
Housing Density (Log)	0.121	1774	2019	14%	1,613	2%	0.12
BART Ridership (Log)	-0.0491	1542	1384	-10%	38	0%	-0.05
Bus Ridership	5.11E-0.5	4288	4067	-5%	(2,451)	-2%	5.11E-0.5
Transit Supply							
Frequency (Log)	0.409	4.93	5.32	8%	2,039	2%	0.41
Route J Route-Stops	-0.6019	0.15	0.16	7%	-	0%	-0.45
Totals							
Total for Route Stops Present in Both Years--each term applied separately *Percent change uses ridership for route-stops present in both years					5,227	6%	
Total for Route Stops Present in Both Years - all terms applied together *Percent change uses ridership for all route-stops					5,230	6%	
Total for Route Stops Dropped					-2,067	-1%	
Total for Route Stops Added					1,631	1%	
System Total					4,669	6%	

Overall the time-of-day models predict the change in MUNI bus and light rail ridership better than the daily model. The time-of-day bus model has a more reasonable sensitivity to frequency than the daily, which provided better results. These results suggest variables that change throughout a day, such as transit performance and supply, should be estimated by time-of-day. Having the data at a time-of-day aggregation allows for the model to see how a variable varies throughout the day and provides more reasonable results.

3.7 Conclusions

This chapter demonstrated the application of a rail and cross-sectional bus direct ridership model (DRM) to understand what factors are affecting transit ridership. This was for the MUNI transit system in San Francisco. The model is effective in predicting ridership cross-sectionally for a given year, and the TOD models even predicted the direction of change correctly for both modes. The purpose of this chapter was to explore the relationships between the MUNI rail and bus ridership with various factors. The result is that bus ridership is most-sensitive to frequency and employment, whereas rail ridership is most-sensitive to employment and housing density.

Cross-sectional DRMs were estimated at two temporal resolutions, daily and by time-of-day (TOD). The daily bus model was found to be overly-sensitive to frequency, relative to previous research (Erhardt et al. 2017; Taylor et al. 2009). Once frequency is disaggregated to a TOD, then the model's sensitivity becomes more reasonable. This suggests that transit performance measures, such as reliability and frequency, should be estimated at TOD temporal resolution to understand their effect on ridership.

There are limitations to cross-sectional ridership models. In particular issues with multi-collinearity. The no multi-collinearity assumption is discussed in Section 3.5, and a few of the explanatory variables are found to be correlated with each other. Meaning that the area of San Francisco with the highest frequency is also the area with the highest employment density. This makes it difficult for the estimation process to separate the effect of each variable. This issue is difficult to avoid when using cross-sectional models, but chapter 4 uses a fixed effects panel model to overcome the multi-collinearity issue.

To better understand what factors drive the changes in transit ridership, a series of sensitivity tests were conducted to isolate the effects of each variable in the model. This analysis showed modest changes across a number of variables in the rail model. While these sensitivities appear reasonable, they do not capture all of the observed increase in rail ridership. The bus ridership decline was under-predicted by 8% and the rail incline was under-predicted by 11%.

It appears that there are real-world trends, over this period, that the models are not able to capture. The divergent rail and bus ridership trajectories are noteworthy, it would be valuable to understand what is driving that divergence. One possibility is that travelers are making an increased use of transportation network companies (TNCs) as a substitute for bus trips, while increasing accessibility for rail trips, as hypothesized by Erhardt (2016). TNCs have experienced noteworthy growth with the inception of the mode in 2009, and with TNCs now comprising 15% of intra-San Francisco vehicle trips (Castiglione et al. 2017). Current research has various results on TNCs relationship with transit, but the lack of TNC trip data rather than geographic location is most likely the cause (Murphy et al. 2016; Clewlow and Mishra 2017). Chapter 4 estimates fixed-effects panel models with TNC trip data, as a TNC variable, to understand the impact that TNCs have on transit ridership.

Chapter 4 Evaluating the Effects of TNCs on Transit Ridership

The previous chapter used cross-sectional direct ridership models (DRMs) to understand what factors are influencing MUNI light rail and bus ridership. The significant variables were used as a starting point for the fixed-effects panel models estimated in this chapter. The purpose of this chapter is to use fixed-effects panel models, along with a transportation network company (TNC) variable, to understand what is contributing to the diverging rail and bus ridership trends. First, background on panel models and TNCs is provided. Then the data is discussed, including a brief discussion of the TNC data processing. Next, the model results are presented and interpreted. The model is then applied and compared to observed ridership. Sensitivity tests are used to determine how much each variable contributes to the change in ridership. Finally, conclusions and next steps are discussed.

4.1 Introduction

DRMs are typically estimated from cross-sectional data for a specific time period, typically a year. This helps to find correlations between transit ridership and variables (Kerkman, Martens, and Meurs 2014). Cross-sectional DRMs use the spatial location of a bus stop to assign demand driver data, such as employment. This can lead to explanatory variables being correlated with each other. For example, in chapter 2 we found highly dense employment areas to be where high frequency transit operates. The correlation voids the no multi-collinearity assumption that all linear models must assume. This chapter overcomes the limitation of multi-collinearity and lack of a TNC variable, by introducing a TNC variable in a fixed-effects panel model.

Panel models have data for the same entities for multiple points in time, and they look at how much each explanatory variable contributes to the dependent variable's change. This helps to determine causation for the factors found to be correlated with transit ridership. For example, we might interpret employment increases to cause transit ridership increases. Fixed-effects panel model adds an intercept to each entity, in this case a bus or rail stop, to help avoid multi-collinearity.

Panel models are an econometric tool and are more commonly used for research in economics (Greene 2003). There are instances where this tool has been applied for research involving transit ridership (Iseki and Ali 2015; Kennedy 2013; Kitamura 1990). Iseki and Ali (2015) estimate panel models for 10 US urbanized areas. They find that transit ridership is influenced greater by internal factors, meaning that factors controlled by the transit agency, such as reliability, have a greater influence than external factors, such as gas prices. Kennedy (2013) performs a corridor level analysis for selected New Zealand cities. He looks at service changes to specific corridors, and ultimately helps to optimize the New Zealand transit service changes. Kitamura (1990) provides an overview of the proper utilization of panel models to understand travel behavior. He discusses how panel models help to eliminate multi-collinearity in detail.

The term transportation network company was coined by the California legislation when Uber, formerly known as UberCab, first began in 2009. The term refers to companies that are considered ride-sourcing, such as Uber and Lyft. TNCs are properly defined as companies who use online platforms to connect passengers with drivers who use personal, non-commercial, vehicles. New features of Uber and Lyft, UberPool and Lyft Line respectively, fall under ride-splitting rather than ride-sourcing. These services allow multiple passengers to share the same vehicle and the passengers split the cost of the trip. The latest change announced, which is not being implemented currently, is fixed stop locations. The change is only for UberPool and encourages users to walk a block or two before being picked up. The location is more efficient for the overall trip. The structure resembles fixed route transit, which is the same way MUNI buses operate.

Shared-Use Mobility Center (SUMC) published a report that summarizes results from a survey given to transit agencies in 7 major cities (Murphy et al. 2016). The results concluded that recreational/social trips were over half of the ride-sourcing trip purposes. The report concluded with the statement that shared modes, in general, reduce automobile ownership and thus complement transit. San Francisco was included in the study and Figure 15 shows a couple of graphs from the report. The graphs compare weekday transit capacity to weekday TNC demand. Units for transit capacity are seat stops per hour, and units for TNC demand are mean surge multiplier.

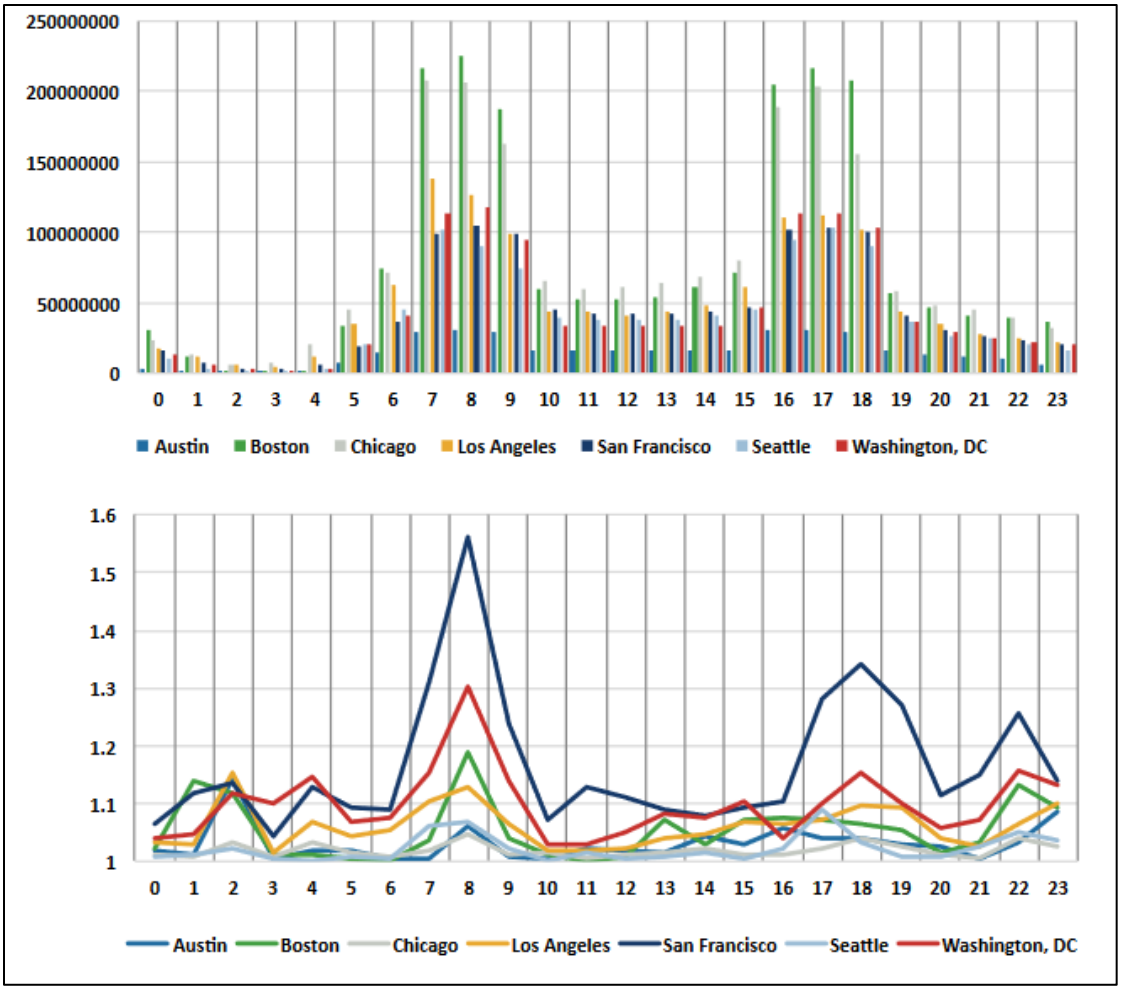

Figure 15: ***Weekday Transit Capacity (Top) vs. Average Ride-sourcing Demand (Bottom)***

*This figure is taken directly from (Murphy et al. 2016)

San Francisco follows intuitive expectations regarding transit capacity, with it having an AM/PM peak periods and a peak of around 100,000,000 seat stops per hour. However, San Francisco does stand out when looking at the TNC demand graph. For many of the cities there is an AM and PM peak in demand, but San Francisco has more pronounced peaks. This could be attributed to Uber starting in San Francisco and having a more mature user base. Surge multiplier isn't directly demand. However, the graph in Figure 16, from the SFCTA report, has a similar distribution of trips throughout an average weekday. SUMC concludes that TNCs fill in the gaps that transit service coverage leaves, typically late at night. Figure 15 and Figure 16 contradict the conclusion and show that throughout an average weekday TNC use follows a similar pattern that transit use has.

Uber and Lyft have a long history with San Francisco. Both companies were founded and started there. There are over 170,000 TNC trips made on an average

weekday within San Francisco, in comparison to 11,000 transit vehicle trips and 940,000 automobile trips. (Castiglione et al. 2017). They find that TNC ridership is distributed diurnally throughout the day, meaning that there is two distinct peaks. Figure 16 shows the distribution for a typical Wednesday. The PM peak is much larger than the AM and decreases at a slower rate. This suggests that, in San Francisco, TNCs are used for daily commutes.

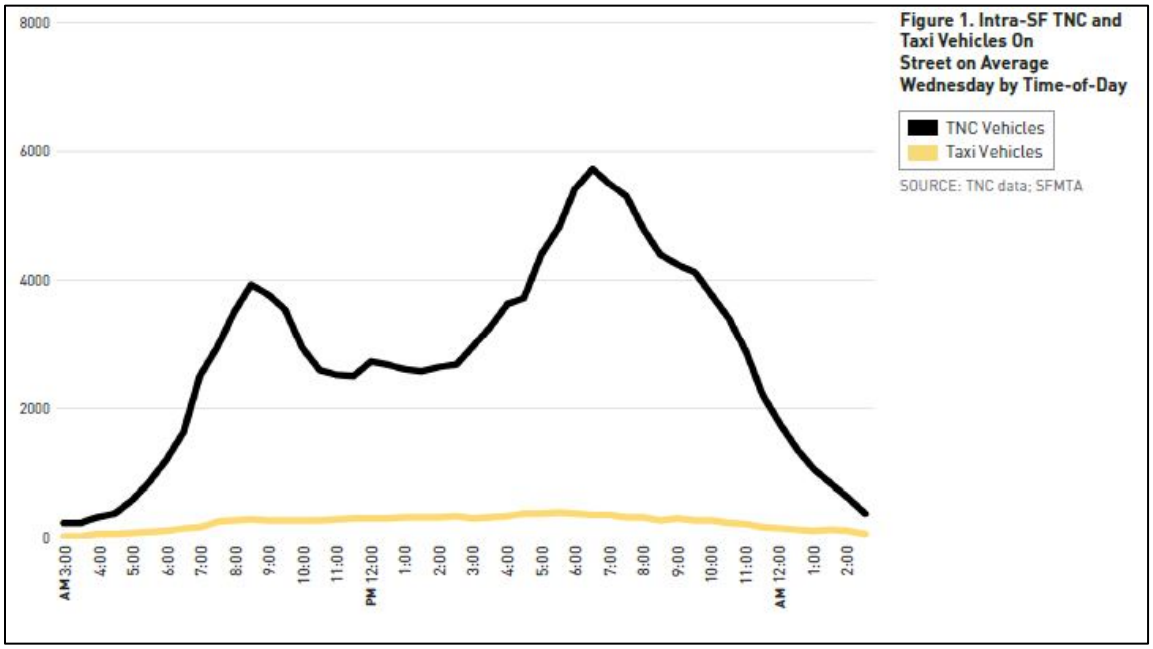

Figure 16: Distribution of TNC Ridership vs. Taxi Ridership (Average Wednesday)

*This figure is taken directly from (Castiglione et al. 2017)

A related question is that of the effect on auto ownership. A report on households joining car-sharing (such as Zipcar) shows that households reduce the number of private vehicles owned after joining a car-sharing program (Martin, Shaheen, and Lidicker 2010). The effect of TNCs on car ownership remains unknown.

A survey of TNC users in San Francisco found that most TNC trips were made for recreational purposes by wealthy and young individuals. (Rayle et al. 2014) The survey helped to provide insight on who is using TNCs. Other research has found that TNCs complement transit ridership, but only Bay Area Rapid Transit (BART) riders are sampled for San Francisco. (Murphy et al. 2016) The paper focuses on TNCs improving accessibility to transit stations, but the relationship seems more intuitive for commuter rail ridership and not bus ridership. A report by the Institute of Transportation Studies states that the effect TNCs have on transit ridership is complex and varies by mode. (Clewlow and Mishra 2017) The report uses survey data to find that after using a TNC a person uses buses 6% less, light rail 3% less, and commuter rail 3% more.

Previous research is not unified on whether or not TNCs complement or compete with transit. The diverging ridership trends discussed in chapter 1 promotes the hypothesis that TNCs complement MUNI light-rail, but compete with MUNI buses. The hypothesis is tested in this chapter and ultimately the effect TNCs have on MUNI light-rail and bus ridership is quantified. It does so using a novel trip data set that provides the most complete inventory to date of TNC usage.

4.2 Data

The purpose of this section is to discuss the data used for the panel models. The new TNC variable and nuances of panel model data is discussed.

The significant variables found in chapter 3 are used as a starting point for the panel models. The data used in the panel models is the same that is used in the cross-sectional DRMs. Refer to section 3.2 for descriptions of the data. For this section only the new TNC variable is described. All variables with log in parenthesis have been log-transformed the same way they are for the cross-sectional DRMs.

The models in chapter 3 estimate ridership for a route-stop and route-stop-TOD, but the to estimate the panel models properly the route-stop-TODs have to be present in both years. The route-stop-TOD service changes in the bus data caused a large portion of the route-stop-TODs to not be present in both years. Thus, the bus data was aggregated to a stop-TOD level. A bus stop-TOD has bus data aggregated, across all routes stopping at a bus stop, to the same 7 times-of-day (TODs). The MUNI light rail system did not experience as many service changes, and have significantly fewer route-stops. Meaning that the MUNI light rail data did not have to be aggregated to a stop level.

Panel Models drop out time-invariant variables, meaning variables that do not change over the study period are dropped out. These may be variables that are correlated with ridership, but do not cause a change in ridership. An example would be terminus stops. In the previous chapter end-of-line and start-of-line stops were found to be positively correlated with transit ridership. The total number of terminus stops do not change much through the years, so they do not cause ridership to change.

4.2.1 Transportation Network Company (TNC) Trip Data

TNCs, such as Uber, operate through an application programming interface (API). APIs are typically used in the background of apps to provide a convenient interface for users. A TNC driver “clocks in” and sends his location to the API, which is then sent to the TNC app for riders to be linked with. The API works as a middle man between a driver and user. Researchers in San Francisco at Northeastern University developed a methodology to “scrape” the Uber and Lyft APIs for archived trip data. This provided TNC trip data for a 6-week period in the fall of 2016. The TNC trip data includes all Uber and Lyft services, such as UberPool and Lyft Line. A TNC vehicle’s location is pinged every five seconds, and the user’s pickup and drop off locations are reported. The pickup location is defined as the location a TNC vehicle goes “offline”, meaning that the

driver has accepted a ride. This is not exactly where the user is being picked up, but it is likely to be close, when driving in a dense city, like San Francisco. The drop off location is when a TNC vehicle goes back “online”. This is the moment a user reaches their destination, and the driver can accept a new ride. For this research, additional processing was performed by SFCTA, and a detailed description can be found in their report (Castiglione et al. 2017). The same stop-buffer process discussed in chapter 3 is used to get the number of TNC pickups and drop offs within walking distance of a stop by time-of-day. The years considered in this chapter are the same as chapter 3, 2009 and 2016. Since there are only two years it is identical to a first-difference panel model. The TNC trip data is used to create a TNC variable for 2016. Uber and Lyft launched in San Francisco after 2009, so ridership is considered to be 0 in 2009.

4.3 Model Estimation

This section presents the model methodology and estimation results for the MUNI bus and light rail panel models.

4.3.1 Modeling Approach

A random effects panel model requires the assumption that the independent variables are not correlated between stations. Variables, such as employment and frequent transit, cannot be clustered in the same part of a city. This is found to be true in chapter 3, so a random effects model cannot be used.

Pooled ordinary least squares (OLS) models are another type of panel model, but they are typically used to determine correlation. Correlated variables have already been found using cross-sectional DRMs.

A fixed effects model adds an intercept to each entity, bus/rail stop. The intercept removes any cross-sectional correlation within a specific year. Instead the model only considers the change in ridership for each stop. Essentially the model eliminates the effect of where the bus is located. This mitigates or solves the issue of multi-collinearity that cross-sectional DRMs can have.

A limitation of panel models is that the entities, bus/rail stops, must link between years. Meaning a route-stop-tod in 2009 must also be present in 2016. This was an issue for the bus route-stop-TODs, and is why the data had to be aggregated from a route-stop-tod to a stop-tod.

4.3.2 Model Equation

The models are a hybrid of a log-log and log-level. Some of the explanatory variables have been transformed specifying a log-log relationship. Others are not, keeping them as a log-level relationship. A sample equation for the bus fixed-effects panel model and the rearranged version that was used to apply the model is shown below.

The sample equation only includes one variable for each specification. The full equation can be found in appendix C.

Equation 3: Bus Stop Fixed-Effects Panel Sample Model Equation

$$\ln\left(Avg\_ride_{16,i}\right) - \ln\left(Avg_{Ride_{09,i}}\right) = \alpha_i + \beta_{emp} * \left[\ln\left(\mathrm{emp}_{16,i}\right) - \ln\left(\mathrm{emp}_{09,i}\right)\right] + \beta_{ontime5} * \left(\mathrm{ontime5}_{16,i} - \mathrm{ontime5}_{09,i}\right) + \varepsilon_{t,i}$$

Equation 4: Bus Stop Fixed-Effects Panel Sample Applied Equation

$$Avg\_ride16_i = \left[e^{\alpha_i} * \left(\frac{\mathrm{emp}_{16,i}}{\mathrm{emp}_{09,i}}\right)^{\beta_{emp}} * \mathrm{e}^{\beta_{ontime5} * (\mathrm{ontime5}_{16,i} - \mathrm{ontime5}_{09,i})}\right] * Avg\_ride_{09,i} + \varepsilon_{t,i}$$

- $Avg\_ride_{16,i}$ = The average of the boardings and alightings at stop i for 2016
- $Avg\_ride_{09,i}$ = the average of the boardings and alight at stop i for 2009
- α_i = an intercept added to stop i (eliminates spatial correlation)
- β_{emp} = coefficient for employment density
- $Emp_{16,i}$= Employment density at stop i for 2016
- $Emp_{09,i}$ = Employment density at stop i for 2009
- $\beta_{ontime5}$ = coefficient for reliability
- $Ontime5_{16,i}$ = reliability at stop i for 2016
- $Ontime5_{09,i}$ = reliability at stop i for 2009
- $\varepsilon_{t,i}$ = error term for each stop in each year

4.3.3 Bus Stop-TOD Fixed Effects Panel Model Results

The python linear models package (Python n.d.) was used to estimate the fixed effect panel models. The dependent variable is the difference of 2016 and 2009 log transformed ridership. Ridership is defined as the average of passengers boarding and alighting at each stop-tod. The log transformation is made due to the ridership data being skewed towards 0. Table 21 shows the results of the model estimation.

Table 21: Bus Stop-TOD Fixed Effects Panel Model Results

Variable	Variable Name	Coefficients	T-Statistics	
Potential Demand (Independent Variables)				R-squared Within
EDD Employment (Log)	EDD_EMP_LOG	0.035	7.15	0.039
Housing Density (Log)	HOUSING_DEN_LOG	0.208	5.47	
Share of High Income Households (2009 $)	SHR_INCOME_100P	-0.108	-2.42	
TNC Ridership (Log)	AVG_TNC_LOG	-0.033	-15.41	
Transit Supply (Independent Variables)				R-Squared Between
Frequency (Log)	FREQ_S_LOG	0.229	13.75	0.724
Reliability	ONTIME5	0.192	8.89	
TOD Trends				
10pm - 3am Trend	PM10_3AM	-0.080	-7.51	

Each of the terms included in the model are significant at the 95% level, or better. For those variables that are log-transformed, the coefficients can be interpreted directly as an elasticity, since the left-hand side is also log transformed. The untransformed coefficients can be interpreted by an increase of 1 results in a $(e^{\beta_x} - 1) * 100$ percent increase in ridership. For a rough comparison, untransformed variable coefficients can be multiplied by 100 to compare to log-transformed coefficients. When an untransformed variable's coefficient is more than 0.1 away from zero, the equation above must be used (Greene 2003). Attributes that are measured based on the area (total employment, housing density, etc.) are calculated within a quarter-mile buffer of the stop. Attributes associated with the service itself (frequency, on-time performance, etc.) are measured at the stop itself. Demographic measures from the ACS (income shares) are measured based on the census tract that contains the stop.

The positive employment coefficient is expected, indicating that ridership grows with increasing employment. The coefficient is relatively small compared to other demand drivers, such as housing density. The data assigned based on a quarter-mile buffer was used instead of a tenth-mile buffer to help solve this issue, but compared to elasticities found in Erhardt (2016) and Taylor et al. (2009) it is still small.

Housing density has a highly-positive impact on bus ridership. Meaning that when you increase housing density around a bus stop the ridership increases.

Increasing frequency results in an increase in bus ridership, and the elasticity is within reason. The model is most-sensitive to transit performance and supply, such as frequency and reliability.

The reliability coefficient can be interpreted as a bus that becomes on-time all the time boosts ridership by 21 percent.

The only time-of-day variable that showed up as significant was 10pm-3am. This could be attributed to overnight service cuts, or a trend in ridership, with people using MUNI buses less late at night.

TNC ridership was found to have a negative impact on bus ridership. Meaning that TNC trips are substituting for bus trips and that the modes compete with each other. This relationship is discussed in more detail in section 4.6.1, where the factors affecting change tables quantify the effect TNCs have on bus ridership.

There are two R-squared measures for panel models. The within R-squared measures how well the model estimates the change in rail ridership. While the between R-squared measures how well the model estimates ridership for each year. The between R-squared is similar to the R-squared reported for the cross-sectional DRMs in chapter 3. The within R-squared is 0.039. This is most likely due to the difficulty in predicting the generally small changes for roughly 22,000 stop-TODs. The between R-squared is 0.724, and is higher than the cross-sectional DRM in chapter 3. This may be due to the inclusion of a TNC variable.

4.3.4 Rail Route-Stop-TOD Fixed Effects Panel Model Results

The rail route-stop-tod model used the same python model estimation package. The dependent variable was the log transformed average of MUNI light-rail passengers boarding and alighting at each route-stop-tod. Table 22 shows the model results. Frequency does not change much between 2009 and 2016, which causes it to be insignificant. Changes in frequency is something that should be accounted for. To do this, the dependent variable is adjusted using the frequency coefficient found with the rail route-stop-tod DRM.

Table 22: Rail Route-Stop-TOD Fixed Effects Panel Model Results

Variable	Variable Name	Coefficients	T-Statistics	R-Squared Within
Potential Demand (Independent Variables)				
EDD Employment (Log)	EDD_EMP_LOG	0.044	1.78	0.325
Housing Density (Log)	HOUSING_DEN_LOG	0.245	3.89	
TNC Ridership	AVG_TNC	0.0001	3.46	
Frequency (Log)	FREQ_S_LOG	0.347	Fixed*	
Times-of-Day (TODs) (Independent Variables)				R-Squared Between 0.812
3AM - 6AM	AM3_6AM	-1.125	-27.79	
6AM - 9AM	AM6_9AM	0.271	7.04	
9AM - 2PM	AM9_2PM	0.065	1.61	
2PM - 4PM	PM2_4PM	0.104	2.73	
4PM - 7PM	PM4_7PM	0.085	2.17	
7PM - 10PM	PM7_10PM	0.105	2.64	
10PM - 3AM	PM10_3AM	-0.190	-4.59	

*Frequency is manually included in the model by fixing the coefficient

The rail model can be interpreted in the same manner as the bus panel model. The employment coefficient is showing up small again. Housing density and the growth in the am peak period are the variables that the model is most-sensitive to.

Multiple specifications of transit supply were tested, such as frequency and headway, but they all were highly insignificant. Frequency does not change much over the study period, and it could be caused by the variable's lack of variance. The panel model estimated coefficients for frequency were negative, so the coefficient from the rail route-stop-TOD model is used. The dependent variable was adjusted accordingly, so that the changes in supply were accounted for.

The 9am-2pm TOD and employment variables are significant with 90% confidence. The rest of the variables are significant with 95% confidence. The dummies are picking up on the overall increasing trend in MUNI rail ridership. The negative overnight coefficients could be picking up overnight service cuts, or a trend in rail ridership. The trend possibly being that late at night people are using MUNI light rail less.

The TNC variable is untransformed in this model. The untransformed variable fit the data better for the rail model. The log-transformed variable was tested in an attempt to

be consistent with the bus model, but the results were unreasonable. The coefficient is positive, meaning that TNC ridership complements MUNI light rail ridership. This relationship is discussed in more detail in section 4.6.2. The section includes factors affecting change tables that quantify the effect TNCs have on MUNI light rail ridership.

There are two R-squared measures for panel models. The within R-squared measures how well the model estimates the change in rail ridership. While the between R-Squared measures how well the model estimates ridership for each year. The between R-squared is similar to the R-squared reported for the DRMs in chapter 3, and is higher at 0.812. The rail model's within R-squared is significantly higher than the bus, at 0.282. This is most likely due to the roughly 2,000 entities the rail model is predicting change for, rather than 22,000.

When comparing the rail to the bus model results, light rail ridership is most-sensitive to changes in housing density and bus ridership is most-sensitive to changes in frequency and reliability. TNC ridership complements MUNI rail ridership and competes with MUNI bus ridership. The rail model explains more of the variance, and this is most likely due to the significantly fewer entities.

4.4 Model Assumptions

The panel models in this thesis use multiple-linear regression. Multiple-linear regression models have four assumptions (Greene 2003). Before a model can be valid, the assumptions must be valid. The four assumptions are:

1) There must be a linear relationship between the outcome variable and the independent variables (not a curvilinear).
2) Multivariate Normality: Residuals must be normally distributed.
3) No Multi-collinearity: The independent variables are not highly correlated to each other.
4) Homoscedasticity: The variance of error terms are similar across the values of the independent variables. In other words, a scatter plot of the residuals does not have a noticeable pattern.

There are techniques in python to check the validity of each of the assumptions. The rest of this section is divided up explaining the bus panel model assumptions followed by the same discussion for the rail panel model.

4.4.1 Stop-TOD Bus Panel Model Assumptions

The outcome variable and explanatory variables having a linear, rather than curvilinear, relationship was checked first. Scatter plots of the independent variables versus the predicted response variable are made to check the assumption. Figure 17 shows the scatter plots. The bottom far right is the distribution of log-transformed ridership. The other plots are all of the explanatory variables. The spread is fairly high for many of the plots, but there is not any curvilinear relationships.

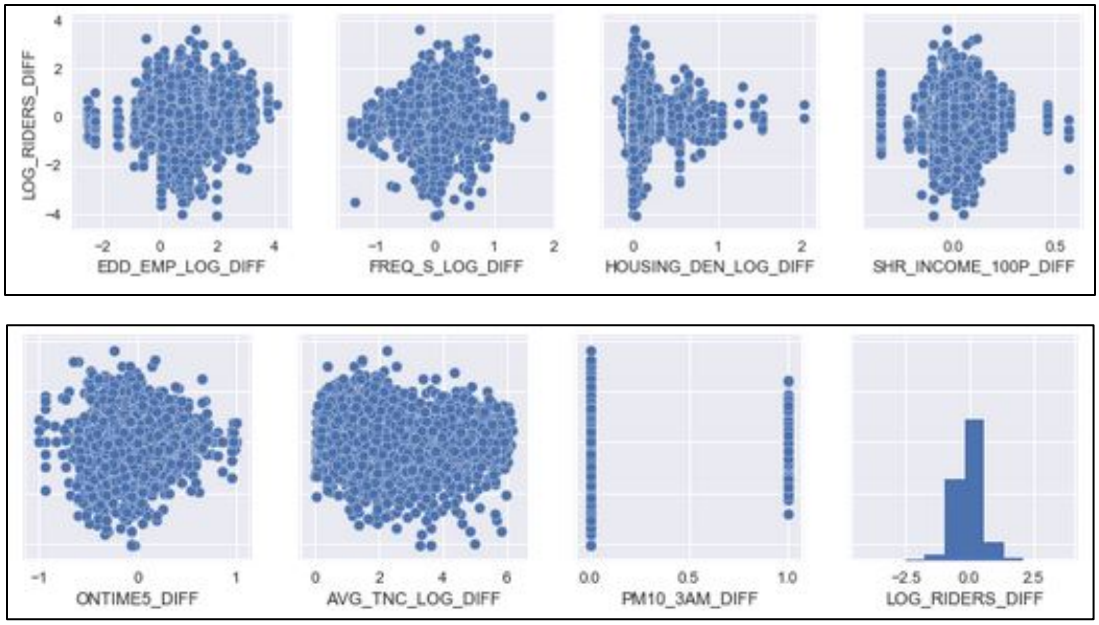

Figure 17: Response Variable vs. Explanatory Variables Scatter Plots

Multivariate normality assumes that the residuals are normally distributed. Figure 18 and Figure 19 show the distribution of residuals and QQ plot for the route-stop-TOD bus model respectively. If the residuals had a perfect normal distribution, then the QQ plot would be a straight diagonal line. The distribution is not perfectly normal, but is adequate for the multivariate normality assumption to be valid. The distribution of residuals for the cross-sectional DRMs look closer to a normal distribution, but this is because the residuals for the panel models are concentrated around zero.

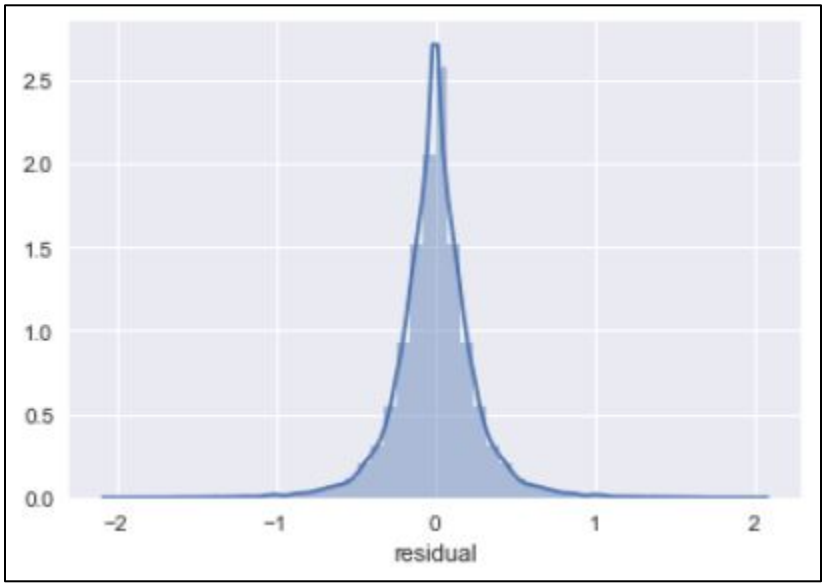

Figure 18: Bus Stop-TOD Model Distribution of Residuals

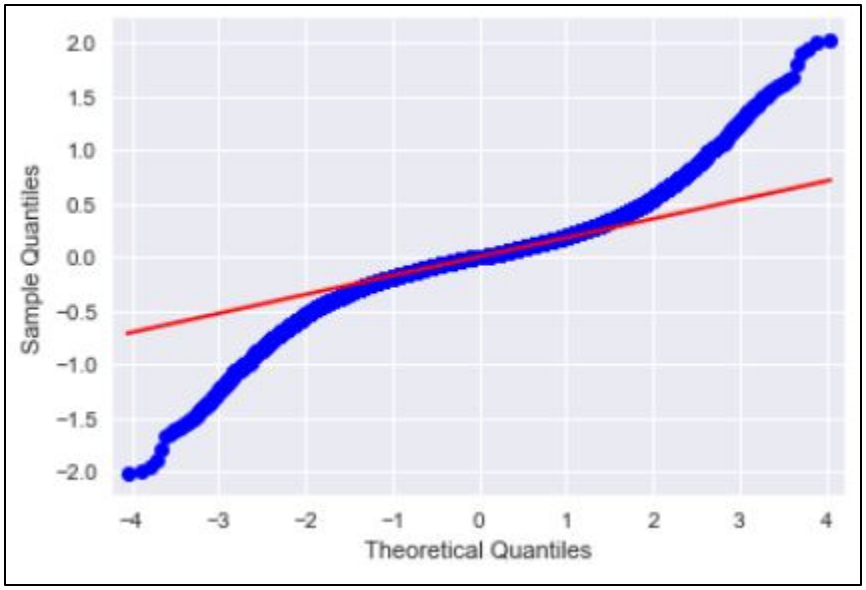

Figure 19: Bus Stop-TOD Model QQ Plot

The no multi-collinearity assumption is to prevent two variables from explaining the same variance. There are a couple of ways to verify if your model has multi-collinearity. One is to create a correlation matrix and the other is to calculate variance inflation factors (VIF) (Murray et al. 2012). VIFs measure how much the variance of the estimated regression coefficients are inflated, compared to when the explanatory variables are not linearly related. Table 23 and Table 24 shows the correlation matrix and VIF, respectively, for the bus stop-TOD model. The column and row names have been abbreviated to help the table fit better. The correlation of two variables ranges from -1 to 1, meaning perfectly uncorrelated and perfectly correlated respectively. A rule-of-thumb is any variance inflation factor above 10 is considered to be multi-collinear (O'brien 2007). Both tables show that none of the variables void the no multi-collinearity assumption. Although the assumption is not invalid, the correlation matrix does show a bit of the story in San Francisco. The change in TNC ridership is somewhat correlated with housing density and high income households. During the study period new condos have been built in SOMA, a now wealthier part of town. The correlation suggests that the increase in dense high income housing would lead to an increase in TNC use, and ultimately a decline bus ridership.

Table 23: Bus Stop-TOD Model Correlation Matrix

	EDD_EMP	FREQ_S	HOUSING_DEN	SHR_INCOME_100P	ONTIME5	AVG_TNC	PM10_3AM
EDD_EMP	1.00	-0.01	-0.07	-0.03	0.01	-0.12	0.00
FREQ_S	-0.01	1.00	0.00	-0.01	-0.02	0.00	-0.09
HOUSING_DEN	-0.07	0.00	1.00	0.06	-0.02	0.23	0.01
SHR_INCOME_100P	-0.03	-0.01	0.06	1.00	0.00	0.25	-0.01
ONTIME5	0.01	-0.02	-0.02	0.00	1.00	-0.03	0.08
AVG_TNC_LOG	-0.12	0.00	0.23	0.25	-0.03	1.00	0.02
PM10_3AM	0.00	-0.09	0.01	-0.01	0.08	0.02	1.00

Table 24: Bus Stop-TOD Model Variance Inflation Factors

VIF Factor	Features
2.1	EDD_EMP_LOG_DIFF
1.0	FREQ_S_LOG_DIFF
t1.1	HOUSING_DEN_LOG_DIFF
1.1	SHR_INCOME_100P_DIFF
1.0	ONTIME5_DIFF
2.5	AVG_TNC_LOG_DIFF
1.2	PM10_3AM_DIFF

The heteroscedasticity assumption states that the residuals of the model are not correlated with each other (Greene 2003). This can be verified by plotting the residuals and visually checking for any patterns. Figure 20 shows the scatter plot of the residuals for the stop-TOD bus model. There is not a clear visual pattern, so the homoscedasticity assumption is valid.

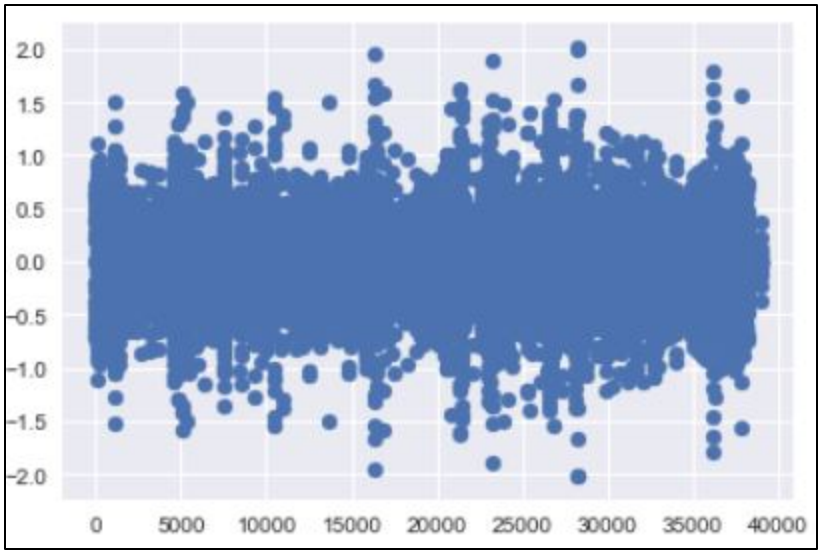

Figure 20: Bus Stop-TOD Model Scatter Plot of Residuals

4.4.2 Route-Stop-TOD Rail Model Assumptions

The rail route-stop-TOD model requires the same four assumptions to be valid. The first being that the outcome variable and explanatory variables have a linear relationship. Figure 21 is scatter plots of the response variable, log-transformed ridership, versus the explanatory variables. The bottom far right is the distribution of log-transformed ridership. Similar to the bus model, the spread is large, but there is not any evidence of a curvilinear relationship. The TNC ridership variable shows a bit of a pattern, but when it is log-transformed the contribution to light rail ridership is unreasonable. Thus, the linear relationship assumption is valid.

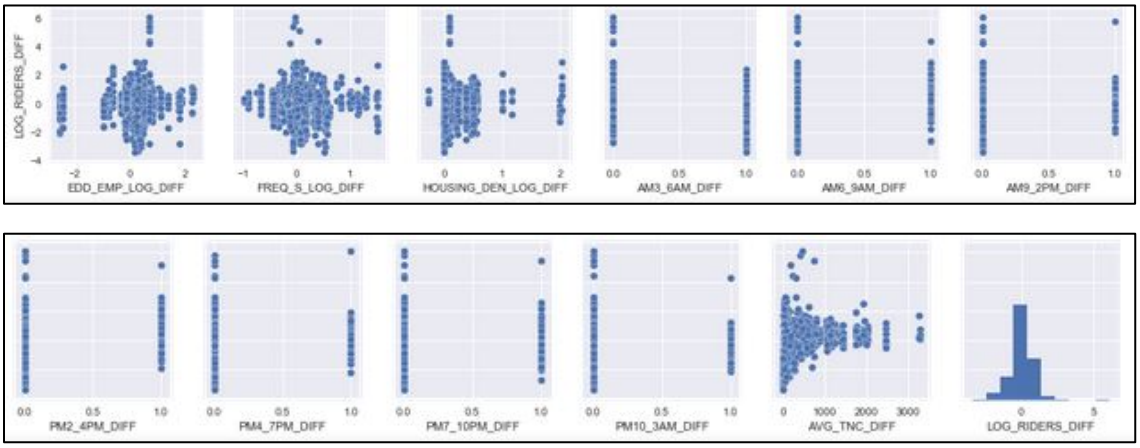

Figure 21: Response Variable vs. Explanatory Variables Scatter Plots

Multivariate normality assumes that the residuals are normally distributed (Greene 2003). Figure 22 and Figure 23 show the distribution of residuals and QQ plot for the rail route-stop-TOD bus model respectively. If the residuals had a perfect normal distribution, then the QQ plot would be a straight diagonal line. The distribution is not perfectly normal, but is adequate for the multivariate normality assumption to be valid.

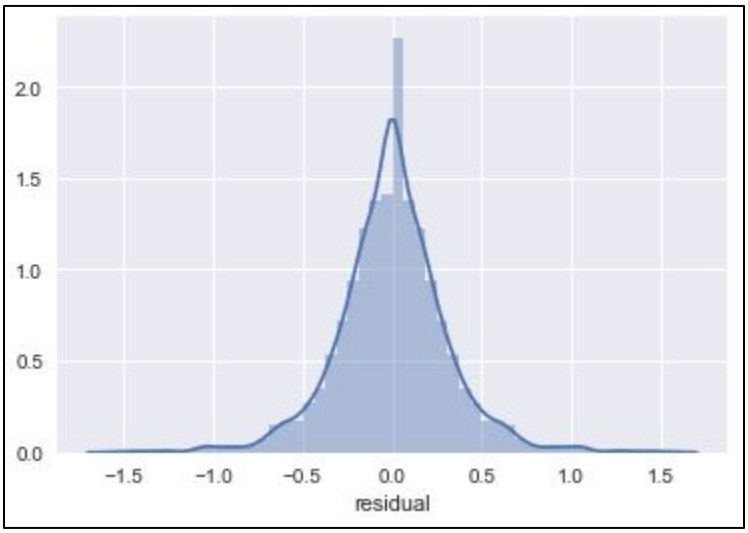

Figure 22: Rail Route-Stop-TOD Model Distribution of Residuals

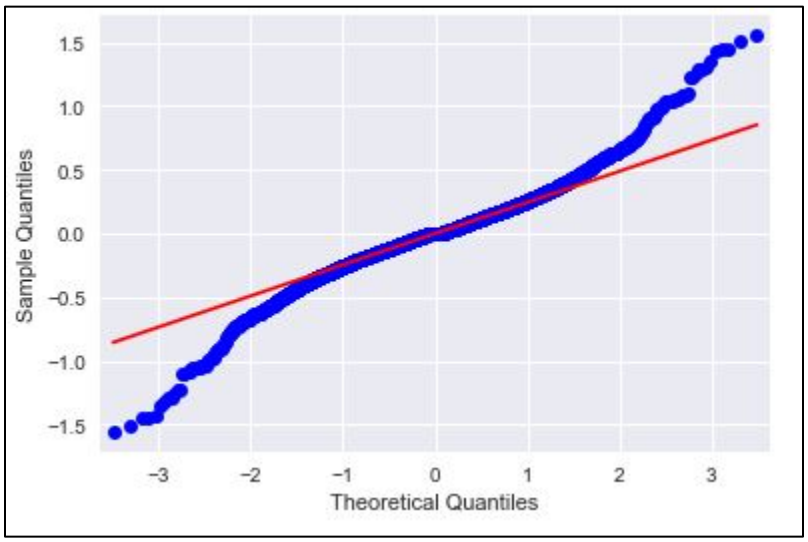

Figure 23: Rail Route-Stop-TOD Model QQ Plot

The no multi-collinearity assumption is to prevent two variables from explaining the same variance (Greene 2003). There are a couple of ways to verify if your model has multi-collinearity. One is to create a correlation matrix and the other is to calculate variance inflation factors (VIF) (Murray et al. 2012). VIFs measure how much the variance of the estimated regression coefficients are inflated, compared to when the

explanatory variables are not linearly related. Table 25 and Table 26 shows the correlation matrix and VIF, respectively, for the rail route-stop-TOD model. The column names have been abbreviated to help the table fit better. The correlation of two variables ranges from -1 to 1, meaning perfectly uncorrelated and perfectly correlated respectively. A rule-of-thumb is any variance inflation factor above 10 is considered to be multi-collinear (O'brien 2007). Both tables show that none of the variables void the no multi-collinearity assumption. Although the no multi-collinearity assumption is not invalid, similar to the bus data, the growth of TNC use is somewhat correlated with the growth of housing density.

Table 25: Rail Route-Stop-TOD Model Correlation Matrix

	EDD_EMP	FREQ_S	HOUSING_DEN	AM3_6 AM	AM6_9 AM	AM9_ 2PM	PM2_4 PM	PM4_7 PM	PM7_10PM	PM10_ 3AM	AVG_TNC
EDD_EMP_LOG_DIFF	1.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	-0.01	0.04
FREQ_S_LOG_DIFF	0.00	1.00	-0.01	0.11	0.06	-0.07	0.23	0.01	-0.15	-0.19	-0.08
HOUSING_DEN_LOG_ DIFF	0.00	-0.01	1.00	0.02	0.00	0.00	0.00	0.00	0.00	-0.01	0.24
AM3_6AM_DIFF	-0.01	0.11	0.02	1.00	-0.16	-0.16	-0.16	-0.16	-0.16	-0.15	-0.19
AM6_9AM_DIFF	0.00	0.06	0.00	-0.16	1.00	-0.18	-0.18	-0.18	-0.17	-0.16	-0.01
AM9_2PM_DIFF	0.00	-0.07	0.00	-0.16	-0.18	1.00	-0.18	-0.18	-0.17	-0.16	0.17
PM2_4PM_DIFF	0.00	0.23	0.00	-0.16	-0.18	-0.18	1.00	-0.18	-0.17	-0.16	-0.09
PM4_7PM_DIFF	0.00	0.01	0.00	-0.16	-0.18	-0.18	-0.18	1.00	-0.17	-0.16	0.09
PM7_10PM_DIFF	0.00	-0.15	0.00	-0.16	-0.17	-0.17	-0.17	-0.17	1.00	-0.16	0.11
PM10_3AM_DIFF	-0.01	-0.19	-0.01	-0.15	-0.16	-0.16	-0.16	-0.16	-0.16	1.00	-0.10
AVG_TNC_DIFF	0.04	-0.08	0.24	-0.19	-0.01	0.17	-0.09	0.09	0.11	-0.10	1.00

Table 26: Rail Route-Stop-TOD Model Variance Inflation Factors

VIF Factor	Features
1.0	EDD_EMP_LOG_DIFF
1.1	HOUSING_DEN_LOG_DIFF
1.0	AM3_6AM_DIFF
1.1	AM6_9AM_DIFF
1.2	AM9_2PM_DIFF
1.1	PM2_4PM_DIFF
1.1	PM4_7PM_DIFF
1.1	PM7_10PM_DIFF
1.0	PM10_3AM_DIFF
1.2	AVG_TNC_DIFF

The heteroscedasticity assumption states that the residuals of the model are not correlated with each other (Greene 2003). This can be verified by plotting the residuals

and visually checking for any patterns. Figure 24 shows the scatter plot of the residuals for the route-stop-TOD rail model. There is not a clear visual pattern, so the homoscedasticity assumption is valid.

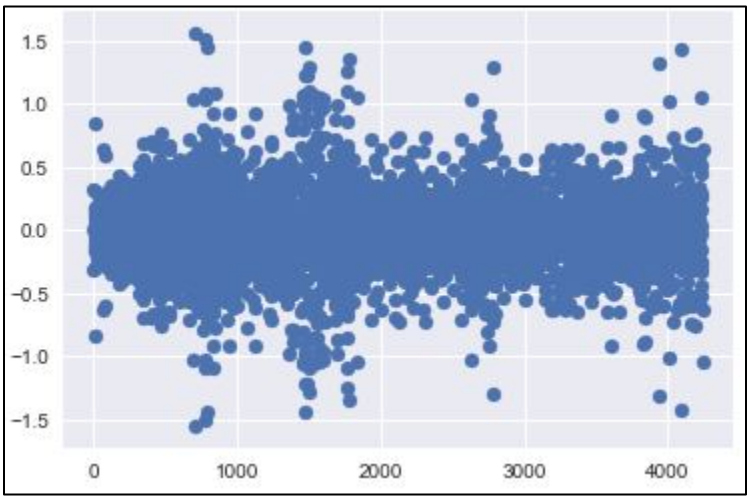

Figure 24: Rail Route-Stop-TOD Model Residual Scatter Plot

4.5 Model Application

The model was applied to evaluate the models accuracy. To evaluate the ability of the models to predict change. In contrast to the model estimation, which limited the data set to route-stops with observed data for all models, the models were applied to the full set of route-stops, or stops, that are present in both years in San Francisco (excluding a few outside the county line). The focus of the examination was on how well the models predict change, rather than their cross-sectional fit.

Table 27 is the aggregated bus and rail system level model application results. Both panel models predict the change in ridership better than the daily and TOD DRMs. The rail model predicts within 1,100 riders of the observed change. However, the bus model predicts within 25,000 riders of the observed change. The total ridership is different than the cross-sectional DRM totals. Panel models can only be applied to "entities", route-stops or stops, which are observed in both years. The totals shown for each year are for "entities" that are present in both years. The percent root mean squared error (RMSE) is calculated for various aggregations for the rail model. The data for the bus model is already aggregated to a stop level. Thus, the stop level percent RMSE is the only measure that can be calculated. Overall the percent RMSE decreases when the data is more aggregated, this is because the average ridership for an observation increases. Average 2016 observed ridership for the rail route-stop, stop, and route is 41, 607, and 28,547 riders respectively. The average 2016 bus stop ridership is 21.

Table 27: System Level Panel Model Results

		Observed Ridership	Modeled Ridership	Difference	Percent Difference	Route-Stop % RMSE	Stop % RMSE	Route % RMSE
Bus	2009	500,866	500,866*					
	2016	456,885	433,859	-23,027	-5%		126%	
	Change	-43,981	-67,007					
	P Change	-9%	-13%					
Rail	2009	142,733	142,733*					
	2016	169,520	170,437	917	1%	85%	43%	13%
	Change	26,787	27,704					
	P Change	19%	19%					

* The observed 2009 ridership is used for 2009 modeled ridership. This provided a starting point for 2016 ridership.

4.6 Factors Affecting Change

To better understand what is driving the change in ridership between the two years, a series of sensitivity tests were conducted with the models. The tests focused on the subset of route-stops (rail) or stops (bus) present in both 2009 and 2016. The process is similar to what was calculated for the cross-sectional DRMs in chapter 3. After manipulating the model equation, the dependent variable becomes the ratio of 2016 ridership over 2009 ridership. Zeros are used as the baseline data. The exponential of zero is 1, meaning that the observed 2009 ridership is the baseline. The change in one variable is added, while the others are kept at zero, to find the variables contribution to change. This provides a means for understanding the magnitude of change that can be attributed to a variable.

4.6.1 Factors Affecting Change Bus Stop Ridership Results

Table 28 shows the results of the factors affecting change (FAC) exercise using the bus model. San Francisco has experienced favorable economic growth in recent years and that is shown by the large increase in employment, resulting in a 3% increase in bus ridership. The employment contribution seems to be low, considering how much it has increased. Section 4.7.2 explains the next steps to potentially solve this issue. Housing density is similar with the correct direction of change but the magnitude seeming small. If housing density and employment were to increase, then the variables with negative coefficients would have larger impacts on bus ridership too. The bus system overall has become more unreliable resulting in a 1 percent decrease in ridership. The TNC ridership variable contributes the most, with a 10% reduction in bus ridership. The overnight TOD

variable was significant, but had a negligible effect on bus ridership. Overall the model over-predicts the change in ridership by 4 percent.

In comparison to the bus DRM FAC table, high income households is the only variable that is noticeably different. The DRM did not have a TNC variable and TNC ridership typically services wealthier clientele. The effect found in the DRM FAC table for high income households is now explained by the TNC variable.

Table 28: FAC in Modeled Bus Stop Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		Elasticity
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	
Potential Demand							
EDD Employment (Log)	0.035	1197	1434	20%	13,796	3%	0.035
Housing Density (Log)	0.208	384	396	3%	5,211	1%	0.208
High Income Households (2009 $)	-0.108	0.36	0.38	6%	(1,690)	0%	-0.102
TNC Ridership (Log)	-0.033	0	294		(49,734)	-10%	-0.107
Transit Supply							
Frequency (Log)	0.229	3.93	4.05	3%	7,709	2%	0.229
Reliability	0.192	0.63	0.58	-8%	(5,948)	-1%	0.192
TOD Trends							
10pm-3am	-0.243				(1,514)	0%	-0.215
Total for Route Stops Present in Both Years--each term applied separately *Percent change uses ridership for stops present in both years					(32,170)	-6%	
Total for Route Stops Present in Both Years--all terms applied together *Percent change uses ridership for all stops					(33,054)	-7%	
Total for Route Stops Dropped					(50,581)	-11%	
Total for Route Stops Added					21,636	5%	
System Total					(61,999)	-13%	

4.6.2 Factors Affecting Change Rail Route-Stop Ridership Results

Table 29 shows the results of the sensitivity analysis using the rail model. Employment and housing density have grown significantly over the study period, but have a relatively small impact on ridership, similar to the bus model. If these effects were to be a more reasonable percent change, then the variables with a negative coefficient would become larger, and positive coefficients would become smaller.

TNCs have a positive impact on ridership, contributing a 7% increase in rail ridership. Rail ridership is complemented by housing density, and a 3% growth in rail ridership can be contributed to it. The 3am-6am variable had a significant coefficient, -1.13, and causes a 4% decrease in rail ridership. Overall the model predicts the change in

rail ridership within 1,000 riders, with the percent change being the same as the observed change.

When comparing the panel model FAC table to the DRM FAC table, many variables drop out due to their time-invariant nature. The results are similar for the variables present in both. Employment is less impactful in the panel model, but both are seemingly low contributions.

Table 29: FAC in Modeled Rail Route-Stop Ridership between 2009 and 2016

		Average Values (without log)			Associated Ridership Change		
Variable	Coefficient	2009	2016	% Diff	Absolute	Percent%	Elasticity _
Potential Demand							
EDD Employment (Log)	0.044	11926	14933	25%	1,217	1%	0.04%
Housing Density (Log)	0.245	1789	1985	11%	4,800	3%	0.25%
TNC Ridership	0.0001	0	140		10,566	7%	0.0001%
Frequency (Log)	0.347	4.5	4.83	7%	2,759	2%	0.35%
TOD Trends							
3AM_6AM	-1.125	316	268	-15%	(6,250)	-4%	-0.68%
6AM_9AM	0.271	316	308	-3%	2,879	2%	0.31%
9AM_2PM	0.065	316	308	-3%	1,691	1%	0.07%
2PM_4PM	0.104	315	308	-2%	4,274	3%	0.11%
4PM_7PM	0.085	315	308	-2%	1,731	1%	0.09%
7PM_10PM	0.105	313	307	-2%	3,832	3%	0.11%
10PM_3AM	-0.190	315	259	-18%	(1,899)	-1%	-0.17%
Totals							
Total for Route Stops Present in Both Years--each term applied separately *Compared to the ridership for route-stops present in both years					25,601	18%	
Total for Route Stops Present in Both Years--all terms applied together					25,500	18%	
Total for Route Stops Dropped					-1,923	-1%	
Total for Route Stops Added					3,395	2%	
System Total					26,972	19%	

Overall TNCs have a diverging effect on MUNI rail and bus ridership. They contribute to a 7% increase in rail ridership, and a 10% decrease in bus ridership. This helps to explain the diverging ridership trends that San Francisco has experienced. Clewlow and Mishra have similar results from their survey (Clewlow and Mishra 2017). They found that TNC use decreased bus ridership 6%, complemented commuter rail ridership, and decreased light rail ridership 3%.

4.6.3 TNC Variable Sensitivity Test

Depending on the model specification, the TNC variable can contribute a 5% to 21% growth in rail ridership, and a 4% to 17% decline in bus ridership. The final models have the highest goodness of fit measurements, while also providing reasonable results. Specifying TNC and transit ridership to have a log-log relationship resulted in TNCs having the highest contribution of change in ridership. Removing certain variables increases the TNC variables contribution as well. The increased contribution is inflated and does not represent the true effect TNCs have. The final models include all of the significant cross-sectional DRM variables that were also significant in the panel models.

4.7 Conclusions and Next Steps

This section concludes the chapter and discusses possible next steps for the research.

4.7.1 Conclusions

The cross-sectional DRMs in chapter 3 predicted the change of MUNI bus and light rail ridership within 10% of the observed change. The DRMs had issues with multi-collinearity, and arguably voided the no multi-collinearity assumption. Panel data models are used to tackle the multi-collinearity issue, and they predict changes in ridership within 4% of the observed change. Panel models are found to predict change better than cross-sectional DRMs.

In chapter 1 research on the diverging ridership trends are discussed, and Erhardt concluded with a -11% unexplained trend in MUNI bus ridership (G..D. Erhardt et al. 2017). The TNC variable helped to explain the diverging trend, by contributing a 10% decrease in bus ridership and a 7% increase in MUNI rail ridership. The 10% decline in ridership predicts somewhat bigger change than Clewlow's and Mishra's findings (Clewlow and Mishra 2017). They find that TNC use results in a 6% decline in bus ridership.

The percent changes calculate to 50,000 less riders using MUNI buses and 10,500 riders using MUNI light rail also use TNCs to access the mode. This means that nearly a third of the 170,000 average weekday TNC trips made are substituting for bus trips.

Other than TNC ridership, MUNI rail ridership is found to be most-sensitive to changes in demand drivers, such as housing density. Other than TNC ridership, MUNI bus ridership is most-sensitive to changes in transit performance and supply, such as reliability and frequency. However, data was not available on reliability for MUNI rail, so it could not be included in the model.

4.7.2 Next Steps

Employment was previously mentioned to contribute a less than expected growth in ridership. This may be due to the buffering process not capturing the full utility of a transit stop. The buffer only accounts for the employment around the bus stop. That does

not account for what a bus stop allows a rider to access though. An accessibility calculation of how many jobs can a bus stop reach within 30 minutes would provide the full effect employment has on transit ridership. Tools such as Transport Analyst should be used to do the calculation (Conveyal 2017).

A limitation of panel models is that the entities, stops, must be represented in both years. The rail stops do not change much, so it is not a problem. However, many bus route-stops have changed route names or locations. This lead to the data being aggregate to a stop level, but there is still a 6% decrease in bus ridership being contributed to stops that have been removed. Assigning bus performance data to road segments, and building buffers around the segments would eliminate the linking problem. This could be another alternative to solving the under-estimated employment coefficient too.

San Francisco is one of many cities facing diverging transit ridership trends. The results have not been confirmed by an independent data set, so they may be regionally specific. A natural extension of this research would be to perform a similar analysis on a different city. A sprawling Phoenix, Arizona would be a prime candidate to complement the results found in a dense San Francisco.

The result of this research is that TNCs do have an effect on transit ridership and it is more complex than lumping all transit modes into one category. Other modes, such as commuter rail, in San Francisco should be analyzed. Once the effect is understand, something similar to a DRM should be developed to understand what drives TNC demand. Then policies can be optimized to incentivize TNC demand to areas that benefit the city as a whole.

Appendix A

Bus Direct Ridership Model Equation:

$$\ln(Avg\_ride_i) = \alpha + \beta_{Emp} * \ln(\text{Emp}_\text{i}) + \beta_{Housing\ Density} * \ln(\text{Housing Density}_\text{i}) + \beta_{High\ Income} * \text{Share of High Income Households}_\text{i} + \beta_{On-Street\ Park} * \ln(On-Street\ Parking\ Price_\text{i}) + \beta_{BART} * \ln(\text{BART}_\text{i}) + \beta_{MUNI\ Rail} * \ln(\text{MUNI Rail}_\text{i}) + \beta_{Transbay} * \text{Transbay}_\text{i} + \beta_{Freq} * \ln(\text{Freq}_\text{i}) + \beta_{ontime5} * \text{ontime5}_\text{i} + \beta_{Limited} * \text{Limited}_\text{i} + \beta_{Express} * \text{Express}_\text{i} + \beta_{3AM-6AM} * 3\text{AM} - 6\text{AM}_\text{i} + \beta_{6AM-9AM} * 6\text{AM} - 9\text{AM}_\text{i} + \beta_{9AM-2PM} * 9AM - 2PM_\text{i} + \beta_{2PM-4PM} * 2\text{PM} - 4\text{PM}_\text{i} + \beta_{4PM-7PM} * 4\text{PM} - 7\text{PM}_\text{i} + \beta_{7PM-10PM} * 7PM - 10PM_\text{i}$$

Rail Direct Ridership Model Equation:

$$\ln(Avg\_ride_i) = \alpha + \beta_{Emp} * \ln(\text{Emp}_\text{i}) + \beta_{Housing\ Density} * \ln(\text{Housing Density}_\text{i}) + \beta_{BART} * \ln(\text{BART}_\text{i}) + \beta_{MUNI\ Bus} * \ln(\text{MUNI Bus}_\text{i}) + \beta_{Freq} * \ln(\text{Freq}_\text{i}) + \beta_{3AM-6AM} * 3\text{AM} - 6\text{AM}_\text{i} + \beta_{6AM-9AM} * 6\text{AM} - 9\text{AM}_\text{i} + \beta_{9AM-2PM} * 9AM - 2PM_\text{i} + \beta_{2PM-4PM} * 2\text{PM} - 4\text{PM}_\text{i} + \beta_{4PM-7PM} * 4\text{PM} - 7\text{PM}_\text{i} + \beta_{7PM-10PM} * 7PM - 10PM_\text{i}$$

Appendix B

Full Bus Route-Stop-TOD DRM Correlation Matrix:

	EDD_EMP_LOG	FREQ_S_LOG	EOL_SOL	HOUSING_09_DEN_LOG	SHR_INCOME_100P	PARK_HOURLY_AVG_ON_LOG	ONTIME5	AVG_BART_LOG	CLOSE_STOP	LIMITED	EXPRESS	AM3_6AM	AM6_9AM	AM9_2PM	PM2_4PM	PM4_7PM	PM7_10PM	TRANSBAY	MUNI_RAIL_AVG_LOG
EDD_EMP_LOG	1	0.16	0.07	0	-0.15	0.66	0.11	0.21	0.01	0.09	0.04	0.01	0.01	0.01	0	0	-0.01	0.12	0.15
FREQ_S_LOG	0.16	1	-0.01	0.07	-0.08	0.14	0.02	0.02	-0.02	0.15	0.22	-0.05	0.19	0.07	0.1	0.12	-0.12	0.03	-0.02
EOL_SOL	0.07	-0.01	1	-0.1	0.06	0.06	-0.01	0.04	0	0.01	0.01	0	0.01	0	0	0	0	0.15	0.12
HOUSING_09_DEN_LOG	0	0.07	-0.1	1	-0.23	0.19	0.04	-0.12	0.05	-0.02	0.01	0	0	0	0	0.01	0.01	-0.23	-0.12
SHR_INCOME_100P	-0.15	-0.08	0.06	-0.23	1	-0.2	0.01	-0.07	-0.03	-0.02	-0.12	0	0	0.01	0	-0.01	0	0.07	0.03

PARK_HOURLY_AVG_ON_LOG	0.66	0.14	0.06	0.19	-0.2	1	0.09	0.17	0.01	0.09	0.05	0	0.01	0	0	0	-0.01	0.09	0.08
ONTIME5	0.11	0.02	-0.01	0.04	0.01	0.09	1	0.01	0	0.05	-0.05	0.15	0.08	0.02	-0.04	-0.16	-0.03	0.03	0
AVG_BART_LOG	0.21	0.02	0.04	-0.12	-0.07	0.17	0.01	1	0	0.05	0	-0.01	0	0	0	0	-0.01	-0.01	0.37
CLOSE_STOP	0.01	-0.02	0	0.05	-0.03	0.01	0	0	1	-0.03	-0.03	-0.32	0.06	0.07	0.01	0.07	0.07	-0.02	-0.03
LIMITED	0.09	0.15	0.01	-0.02	-0.02	0.09	0.05	0.05	-0.03	1	-0.04	-0.06	0.05	0.01	0.06	0.03	-0.03	0.07	0.02
EXPRESS	0.04	0.22	0.01	0.01	-0.12	0.05	-0.05	0	-0.03	-0.04	1	-0.04	0.08	-0.03	0.01	0.08	-0.05	-0.01	0.01
AM3_6AM	0.01	-0.05	0	0	0	0	0.15	-0.01	-0.32	-0.06	-0.04	1	-0.15	-0.14	-0.15	-0.15	-0.14	0	-0.01
AM6_9AM	0.01	0.19	0.01	0	0	0.01	0.08	0	0.06	0.05	0.08	-0.15	1	-0.18	-0.19	-0.19	-0.17	0	0

AM9_2PM	0.01	0.07	0	0	0.01	0	0.02	0	0.07	0.01	-0.03	-0.14	-0.18	1	-0.18	-0.18	-0.16	0	-0.01
PM2_4PM	0	0.1	0	0	0	0	-0.04	0	0.01	0.06	0.01	-0.15	-0.19	-0.18	1	-0.18	-0.17	0	0
PM4_7PM	0	0.12	0	0.01	-0.01	0	-0.16	0	0.07	0.03	0.08	-0.15	-0.19	-0.18	-0.18	1	-0.17	0	-0.01
PM7_10PM	-0.01	-0.12	0	0.01	0	-0.01	-0.03	-0.01	0.07	-0.03	-0.05	-0.14	-0.17	-0.16	-0.17	-0.17	1	0	-0.01
TRANSBAY	0.12	0.03	0.15	-0.23	0.07	0.09	0.03	-0.01	-0.02	0.07	-0.01	0	0	0	0	0	0	1	-0.02
MUNI_RAIL_AVG_LOG	0.15	-0.02	0.12	-0.12	0.03	0.08	0	0.37	-0.03	0.02	0.01	-0.01	0	-0.01	0	-0.01	-0.01	-0.02	1

Appendix C

Bus Fixed-Effects Panel Model Equation:

$$\begin{aligned}\ln(Avg\_ride_{16,i}) - \ln(Avg\_ride_{09,i}) &= \alpha_i + \beta_{emp} * [\ln(\text{Emp}_{16,\text{i}}) - \ln(\text{Emp}_{09,\text{i}})] \\ &+ \beta_{Housing} * [\ln(\text{Housing}_{16,\text{i}}) - \ln(\text{Housing}_{09,\text{i}})] \\ &+ \beta_{High\ Income} * (\text{Share High Income Households}_{16,\text{i}} - \text{Share High Income Households}_{09,\text{i}}) + \beta_{Freq} * [\ln(\text{Freq}_{16,\text{i}}) - \ln(\text{Freq}_{09,\text{i}})] + \beta_{ontime5} \\ &* (\text{ontime5}_{16,\text{i}} - \text{ontime5}_{09,\text{i}}) + \beta_{10PM-3AM} * (10\text{PM} - 3\text{AM}_{16,\text{i}} - 10\text{PM} - 3\text{AM}_{09,\text{i}}) + \varepsilon_{t,i}\end{aligned}$$

Rail Fixed-Effects Panel Model Equation:

$$\begin{aligned}\ln(Avg\_ride_{16,i}) - \ln(Avg\_ride_{09,i}) &= \alpha_i + \beta_{emp} * [\ln(\text{Emp}_{16,\text{i}}) - \ln(\text{Emp}_{09,\text{i}})] \\ &+ \beta_{Housing} * [\ln(\text{Housing}_{16,\text{i}}) - \ln(\text{Housing}_{09,\text{i}})] \\ &+ \beta_{High\ Income} * (\text{Share High Income Households}_{16,\text{i}} - \text{Share High Income Households}_{09,\text{i}}) + \beta_{Freq} * [\ln(\text{Freq}_{16,\text{i}}) - \ln(\text{Freq}_{09,\text{i}})] + \beta_{ontime5} \\ &* (\text{ontime5}_{16,\text{i}} - \text{ontime5}_{09,\text{i}}) + \beta_{3AM-6AM} * (3\text{AM} - 6\text{AM}_{16,\text{i}} - 3\text{AM} - 6\text{AM}_{09,\text{i}}) + \beta_{6AM-9AM} * (6\text{AM} - 9\text{AM}_{16,\text{i}} - 6\text{AM} - 9\text{AM}_{09,\text{i}}) \\ &+ \beta_{9AM-2PM} * (9\text{AM} - 2\text{PM}_{16,\text{i}} - 9\text{AM} - 2\text{PM}_{09,\text{i}}) + \beta_{2PM-4PM} * (2\text{PM} - 4\text{PM}_{16,\text{i}} - 2\text{PM} - 4\text{PM}_{09,\text{i}}) + \beta_{4PM-7PM} * (4\text{PM} - 7\text{PM}_{16,\text{i}} - 4\text{PM} \\ &- 7\text{PM}_{09,\text{i}}) + \beta_{7PM-10PM} * (7\text{PM} - 10\text{PM}_{16,\text{i}} - 7\text{PM} - 10\text{PM}_{09,\text{i}}) + \beta_{10PM-3AM} * (10\text{PM} - 3\text{AM}_{16,\text{i}} - 10\text{PM} - 3\text{AM}_{09,\text{i}}\end{aligned}$$

References

Balcombe, R, R Mackett, J Preston, J Shires, H Titheridge, M Wardman, and P White. 2004. "The Demand for Public Transport: A Practical Guide." http://www.trpa.org/documents/rseis/New%20References%20for%20Final%20EIS/Balcombe%20et%20al%202004.pdf.

BART. 2017. "Ridership Reports | Bart.Gov." Ridership Reports. 2017. https://www.bart.gov/about/reports/ridership.

Blainey, Simon, and John Preston. 2010. *A Geographically Weighted Regression Based Analysis of Rail Commuting around Cardiff, South Wales*.

Brown, Jeffrey R., and Gregory L. Thompson. 2008. "The Relationship between Transit Ridership and Urban Decentralisation: Insights from Atlanta." *Urban Studies* 45 (5–6):1119–1139.

Bureau, U. S. Census. 2016. "American FactFinder - Download Center." American Fact Finder. 2016. https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml.

Caltrain. 2017. "Caltrain 2017 Annual Passenger Count and Key Findings." http://www.caltrain.com/Assets/_Marketing/caltrain/pdf/2016/2017+Annual+Count+Key+Findings+Report.pdf.

Cardozo, Osvaldo Daniel, Juan Carlos García-Palomares, and Javier Gutiérrez. 2012. "Application of Geographically Weighted Regression to the Direct Forecasting of Transit Ridership at Station-Level." *Applied Geography* 34 (Supplement C):548–58. https://doi.org/10.1016/j.apgeog.2012.01.005.

Castiglione, Joe, Tilly Chang, Drew Cooper, Jeff Hobson, Warren Logan, Eric Young, Billy Charlton, et al. 2017. "TNCs Today A Profile of San Francisco Transportation Network Company Activity." San Francisco Transportation Authority.

Cervero, Robert. 2008. "Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth." *University of California Transportation Center*, January. http://escholarship.org/uc/item/1861b3db.

Cervero, Robert, Jin Murakami, and Mark Miller. 2010. "Direct Ridership Model of Bus Rapid Transit in Los Angeles County, California." *Transportation Research Record* 2145:1–7. https://doi.org/10.3141/2145-01.

Chakour, Vincent, and Naveen Eluru. 2016. "Examining the Influence of Stop Level Infrastructure and Built Environment on Bus Ridership in Montreal." *Journal of Transport Geography* 51 (February):205–17. https://doi.org/10.1016/j.jtrangeo.2016.01.007.

Clark, Hugh, Karen Basinger, Katey Maloney, and Callie Whiteman. 2017. "Who Rides Public Transportation." American Public Transportation Association. https://www.apta.com/resources/reportsandpublications/Documents/APTA-Who-Rides-Public-Transportation-2017.pdf.

Clewlow, Regina, and Gouri Mishra. 2017. "Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States."

Conveyal. 2017. "Transport Analyst." 2017. https://github.com/conveyal/analysis-backend.

Dill, Jennifer, Marc Schlossberg, Liang Ma, and Cody Meyer. 2013. "Predicting Transit Ridership at the Stop Level: The Role of Service and Urban Form." In *92nd Annual Meeting of the Transportation Research Board, Washington, DC*. http://docs.trb.org/prp/13-4693.pdf.

Erhardt, Gregory D., Oliver Lock, Elsa Arcaute, and Michael Batty. 2017. "A Big Data Mashing Tool for Measuring Transit System Performance." In *Seeing Cities Through Big Data*, edited by Piyushimita (Vonu) Thakuriah, Nebiyou Tilahun, and Moira Zellner, 257–78. Springer Geography. Springer International Publishing. https://doi.org/10.1007/978-3-319-40902-3_15.

Estupiñán, Nicolás, and Daniel A. Rodríguez. 2008. "The Relationship between Urban Form and Station Boardings for Bogotá's BRT." *Transportation Research Part A: Policy and Practice* 42 (2):296–306. https://doi.org/10.1016/j.tra.2007.10.006.

Federal Transit Administration. 2016. "2015 National Transit Summary and Trends." U.S. Department of Transportation.

Feigon, Sharon, and Colin Murphy. 2016. "Shared Mobility and the Transformation of Public Transit." Transit Cooperative Research Program TCRP 188. Washington, D.C.: Transportation Research Board. https://doi.org/10.17226/23578.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, N.J: Prentice Hall.

Holmgren, Johan. 2013. "An Analysis of the Determinants of Local Public Transport Demand Focusing the Effects of Income Changes." *European Transport Research Review* 5 (2):101–7. https://doi.org/10.1007/s12544-013-0094-0.

Iseki, Hiroyuki, and Rubaba Ali. 2015. "Fixed-Effects Panel Data Analysis of Gasoline Prices, Fare, Service Supply, and Service Frequency on Transit Ridership in 10 U.S. Urbanized Areas." *Transportation Research Record: Journal of the Transportation Research Board* 2537 (January):71–80. https://doi.org/10.3141/2537-08.

Kennedy, David. 2013. "Panel Data Analysis of Public Transport Patronage Growth–an Innovative Econometric Approach." In *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia*. http://atrf.info/papers/2013/2013_kennedy.pdf.

Kepaptsoglou, Konstantinos, Antony Stathopoulos, and Matthew G. Karlaftis. 2017. "Ridership Estimation of a New LRT System: Direct Demand Model Approach." *Journal of Transport Geography* 58 (January):146–56. https://doi.org/10.1016/j.jtrangeo.2016.12.004.

Kerkman, Kasper, Karel Martens, and Henk Meurs. 2014. "Factors Influencing Bus-Stop Level Ridership in the Arnhem Nijmegen City Region." http://rstrail.nl/new/wp-content/uploads/2014/12/kerkman_kasper.pdf.

Kitamura, Ryuichi. 1990. "Panel Analysis in Transportation Planning: An Overview." *Transportation Research Part A: General* 24 (6):401–15. https://doi.org/10.1016/0191-2607(90)90032-2.

Kittelson & Associates, National Research Council (U.S.), Transit Cooperative Research Program, United States, and Transit Development Corporation, eds. 2013. *Transit Capacity and Quality of Service Manual*. 3rd ed. TCRP Report 165. Washington, D.C: Transportation Research Board.

Lee, Kwang Sub, Jin Ki Eom, So Young You, Jae Hong Min, and Keun Yul Yang. 2015. "An Empirical Study on the Relationship between Urban Railway Ridership and Socio-Economic Characteristics." *Procedia Computer Science* 52:106–12. https://doi.org/10.1016/j.procs.2015.05.037.

Liu, Chao, Sevgi Erdogan, Ting Ma, and Frederick W. Ducca. 2014. "How to Increase Rail Ridership in Maryland? Direct Ridership Models (DRM) for Policy Guidance." In *TRB 93rd Annual Meeting Compendium of Papers*. Vol. 2. http://docs.trb.org/prp/14-3417.pdf.

Martin, Elliot, Susan Shaheen, and Jeffrey Lidicker. 2010. "Impact of Carsharing on Household Vehicle Holdings: Results from North American Shared-Use Vehicle Survey." *Transportation Research Record: Journal of the Transportation Research Board* 2143 (December):150–58. https://doi.org/10.3141/2143-19.

Murphy, Sharon Feigon and Colin, Transit Cooperative Research Program, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine. 2016. *Shared Mobility and the Transformation of Public Transit*. Washington, D.C.: Transportation Research Board. https://doi.org/10.17226/23578.

Murray, Leigh, Hien Nguyen, Yu-Feng Lee, Marta D. Remmenga, and David W. Smith. 2012. "VARIANCE INFLATION FACTORS IN REGRESSION MODELS WITH DUMMY VARIABLES."

O'brien, Robert M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality & Quantity* 41 (5):673–90. https://doi.org/10.1007/s11135-006-9018-6.

Pulugurtha, Srinivas S., and Mahesh Agurla. 2012. "Assessment of Models to Estimate Bus-Stop Level Transit Ridership Using Spatial Modeling Methods." *Journal of Public Transportation* 15 (1):3.

Python. n.d. "Linearmodels 2.0." Linearmodels 2.0. Accessed October 23, 2017. https://pypi.python.org/pypi/linearmodels/2.0.

Rayle, Lisa, Susan Shaheen, Nelson Chan, Danielle Dai, and Robert Cervero. 2014. "App-Based, on-Demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco University of California Transportation Center (Uctc)." *University of California, Berkeley, United States Rogers, B.(2015) The Social Costs of Uber. James E. Beasley School of Law, Temple University, Philadelphia, United States*. http://www.its.dot.gov/itspac/dec2014/ridesourcingwhitepaper_nov2014.pdf.

RSG. 2015. "User Guide: STOPS Simplified Trips-on-Project Software." Federal Transit Administration. https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/STOPS_1.50_user_documentation_v5.pdf.

San Francisco Municipal Transportation Agency. 2014. "MuniForward." Muni Forward - Get On Board. 2014. http://muniforward.com.

"StatsModels: Statistics in Python — Statsmodels 0.8.0 Documentation." n.d. Statsmodels's Documentation. Accessed August 1, 2017. http://www.statsmodels.org/stable/index.html.

Taylor, Brian D., Douglas Miller, Hiroyuki Iseki, and Camille Fink. 2009. "Nature and/or Nurture? Analyzing the Determinants of Transit Ridership across US Urbanized

Areas." *Transportation Research Part A: Policy and Practice* 43 (1):60–77. https://doi.org/10.1016/j.tra.2008.06.007.

Upchurch, Christopher, and Michael Kuby. 2013. "Evaluating Light Rail Sketch Planning: Actual versus Predicted Station Boardings in Phoenix." *Transportation* 41 (1):173–92. https://doi.org/10.1007/s11116-013-9499-9.

US Census Bureau Center for Economic Studies. n.d. "US Census Bureau Center for Economic Studies Publications and Reports Page." Longitudinal Employer-Household Dynamics. Accessed June 20, 2017. https://lehd.ces.census.gov/data/.

Vita

Richard Alexander Mucci

Education:

Bachelor's Degree - University of Kentucky 2016

Professional Positions:

- Graduate Assistant (2016-Present)
- Kentucky Transportation Cabinet Intern (Sumer of 2013, 2014, 2015, 2016)

Scholastic and Professional Honors:

- Kentucky Transportation Cabinet Scholarship Student (2012-Present)
- Kentucky Institute of Transportation Annual Scholarship Student (2017)
- Passed the Fundamentals of Engineering Exam (2015)
- Passed the Principles of Engineering Exam (2017)

Professional Publications

- Transportation Research Record by Transportation Research Board
    - Title: Evaluating the ability of transit direct ridership models to forecast medium-term ridership changes: Evidence from San Francisco
    - Date: 1/7/2017
    - Authors: Alex Mucci and Greg Erhardt

- World Transportation Convention
    - Title: Speed Estimation in a Data Poor Environment, A Case Study Site Along Interstate-75 in Kentucky
    - Date: 6/6/17
    - Authors: Alex Mucci, Jacob Brashear, Teng Wang