Title
A Direct Demand Model for Commuter Rail Ridership in the San Francisco Bay Area

Author
Kwong, Jennifer

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

A Direct Demand Model for Commuter Rail Ridership in the San Francisco Bay Area

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Civil Engineering

by

Jennifer Kwong

Thesis Committee:
Professor Michael G. McNally, Chair
Professor Wenlong Jin
Professor Jean-Daniel Saphores

2018

DEDICATION

To:

阿公,

Whose favorite maxim was “education is the key to success.”

And to my mother and father,

For ensuring both could happen.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis and my time as a master's student at UC Irvine was made possible through a variety of people that I would like to acknowledge and thank.

First, I would like to thank the professors, faculty, and staff at UCI's Institute of Transportation Studies for their time, knowledge, and dedication to providing a space for transportation research. In particular, I would like to thank my professor and advisor, Dr. Michael McNally, for his guidance on this thesis, as well as always having insightful and witty comments about everything in transportation and life. Your classes have cemented my interest in travel demand modeling and transportation, and I've always enjoyed our conversations about transportation and idiosyncrasies – each conversation always sets my curiosity bounding off in another direction.

In addition, I am also grateful to the other two members of my committee – Dr. Jin and Dr. Saphores, for their time and comments on this thesis and their time spent teaching and imparting general knowledge about transportation. I would also like to thank Dr. Houston, Dr. Marantz, and Sarah Catz in the urban planning department – your classes gave an understanding of how transportation is situated in the greater context of the urban environment.

I would also like to dedicate thanks to the friends I've made at ITS over the past two years, and some outside of UCI. In particular - to Monica Ramirez – thank you for always keeping me on track and rescuing me from troubles of my own creation, and for keeping me company in urban planning classes, I don't know how I could have done it without you. To Kat Lee, Lu Xu, Jiaqi Cheng, and Steven Ng – thank you for the laughs, the stress, and everything in between – I'm glad I've met you all and that you've become a consistent part of my life.

Last but not least – to my family - thank you to my parents, who have always supported me, even when it means moving more than five hours away from home and seeing you only three times a year. And to my brother, thank you too, for always lending an ear, sending over good textbooks, helping solve coding snafus, and providing free food.

ABSTRACT OF THE THESIS

A Direct Demand Model of Commuter Rail in the San Francisco Bay Area

By

Jennifer Kwong

Master of Science in Civil Engineering

University of California, Irvine, 2018

Professor Michael McNally, Chair

This thesis documents the development of a direct travel demand model for commuter rail in the San Francisco Bay Area. A direct demand model simultaneously estimates trip generation and attraction, which for this thesis would be trips between an origin-destination pair of stations. In the model, the number of trips assigned to an origin-destination pair of stations is dependent on land use characteristics at the origin and destination stations in combination with travel time on the network during congested peak periods and via transit. The model uses a multiplicative direct demand model to estimate ordinary least square regression parameters for the origin-destination trips. From the model form, the resultant estimated regression parameters are elasticities, and as such, can be used to postulate the effects of the selected land use characteristics and network travel times upon the number of trips made.

At both the origin and destination, the location of the station within the central business districts of the San Francisco Bay region had the largest effect on trip generation and attraction. Higher employment density at the destination and a larger number of workers per household at the origin had a positive effect on trips, while the total number of

industrial workers at the destination and an increased number of two car households had a negative effect on trips. Longer travel times on transit appeared to have a positive effect on trips, yet longer travel times in congested peak periods appeared to have a negative effect on trips.

INTRODUCTION

The continuing increase in population in urban areas has unsurprisingly led to a corresponding increase in driving, and thus congestion. This has led to a variety of ways to attempt to mitigate congestion – from adding additional transportation infrastructure to service the demand, to planning communities that encourage people to drive less, and to also reduce the amount of greenhouse gases emitted due to increased driving.

In recognition of the part that the transportation sector plays in anthropogenic climate change, the government of California passed two bills - Assembly Bill 32 (AB 32) and Senate Bill 375 (SB 375), targeted at reducing greenhouse gas emissions. As a result, the two bills have spurred an increased interest in understanding the relationship between travel behavior and land uses for governmental policy and industry, especially SB 375, which requires regional planning agencies (typically MPOs) to create a Sustainable Communities Strategy (SCS). [1] AB 32 requires 1.2% of emission requirements to come from local governments, and one method in doing so is by changing development patterns to reduce emissions.[2] SB 375 helps reinforce the goals established by AB32, and gives planning agencies more of a direct requirement in the ways they have to implement the GHG emission goals. The Sustainable Communities Strategy is meant to be a complement to the regional planning agencies' Regional Transportation Plan (RTP). The bill specifically states that "changes in land use and transportation policy, based upon established modeling methodology, will provide significant assistance to California's goals to implement the federal and state Clean Air Acts and to reduce its dependence on

[1] https://www.arb.ca.gov/cc/sb375/sb375.htm
[2] http://www.cp-dr.com/articles/node-2140

petroleum." [3] The most powerful method in which SB 375 encourages doing so is in its streamlining of development projects that fit with the SCS, even if they conflict with local plans, especially transit priority projects.

As evidenced by the importance accorded to transit in SB 375 - transit has often been touted as one of the methods that will help alleviate congestion. In addition to new transit projects, the concept of transit-oriented development, such as the changing development pattern policy suggested by AB32, has also been a popular suggestion to improve congestion. This paper aims to evaluate whether there are land use or socioeconomic characteristics, that contribute to differences in transit ridership at stations, in support of this idea. The paper develops a direct demand model for Caltrain, a linear commuter rail corridor in the San Francisco Bay Area, and integrates land use and travel time variables in a log-sum formulation in order to see if changes in land use density and other characteristics typically associated with transit-oriented development strongly influence transit ridership between stations.

Previous research has attempted to evaluate how land use has affected mode choice in an overall setting, with varying results. However, this model focuses specifically on ridership, as it would allow transit agencies to predict expected ridership growth at stations due to land development and adjust operation plans accordingly, or forecast more accurate ridership projections for new capital projects. In addition, it could allow planners to recognize baseline densities for land uses in zoning in order to encourage transit ridership, and supports modeling efforts for the goals that the California government has attempted to put into place with AB 32 and SB 375. It attempts to forecast both origin and

[3] http://www.leginfo.ca.gov/pub/07-08/bill/sen/sb_0351-0400/sb_375_bill_20080930_chaptered.pdf

destinations, whereas most previous studies have chosen solely to focus on origin ridership. In addition, it uses an on-board travel survey and readily available GIS shapefiles to simplify the modeling data collection process, as the data and funds available to transit agencies is typically limited.

Outline

Chapter 1 encompasses a literature review of topics relevant to land use and travel demand interactions, as well as transit ridership modeling. The thesis begins in Chapter 2 with an overview of the study area and travel survey data used for the model. In Chapter 3, the model is formulated and analyzed to see if there are correlations from the specified variables to transit ridership. Chapter 4 makes recommendations for enhancing the research done in this paper and for future efforts and concludes the thesis.

Chapter 1: Literature Review

Travel Demand Modeling

Travel demand modeling primarily began in the 1950s, with the United States' National Defense and Interstate Highway Act of 1956. The construction of the US interstate system was one of the first large-scale infrastructure construction processes in the modern era, and the four-step planning model that was researched, created and implemented during this time period to determine where these highways would go is still used today in transportation planning processes today. Planners of that time needed to understand travel behavior to predict future highway flows, and set a capacity for the highways they would construct. These early attempts to understand travel behavior focused on understanding how many trips would be generated from various land uses, and how they would travel from origin to destination. Requirements for transportation planning were first mandated by the Federal-Aid Highway Act of 1962, including that of models.[4]

As such, transportation models in the United States in practice have primarily focused on highways and car infrastructure, since that was what they were originally developed for. As other modes have increased in usage and popularity, the models have been modified to include techniques that can account for different modes, and to address the issues that arise in specifying increasingly complex models. Specific studies have been undertaken to inform transit ridership modeling, and many were synthesized and reviewed extensively in the 1996 Transit Cooperative Research Program Report, "Transit and Urban Form," and Taylor (2003).

[4] Beimborn (1995), 2

Model Types

Travel demand modeling has primarily been done on a regional scale and using the traditional four-step model. The four step model as typically used in transportation planning is described in McNally (2007). The base activity system (land use characteristics) is used to determine how often trips occur (trip generation). Typical units of analysis are the transportation analysis zones (TAZs), which group together areas with similar characteristics. Trips are estimated in terms of the trips that a land use will produce, and how many it will attract. Then, in the following step (trip distribution), productions are distributed to attractions via calculated travel impedances (e.g. time or cost), reflecting person-trips. These person-trips are then split by mode, based on the understanding of mode-share for the region. Lastly, in route assignment, trips are assigned to segments of the network, and a time-of-day factor is also applied.

Activity based modeling has arisen as an alternate modeling approach to the four-step model that integrates behavioral characteristics and economic utility choices of travelers and treats travel as a byproduct of the activity choices that an individual makes. Recker (1986), Bhat and Koppelman (1999) describe the fundamentals of activity-based modeling. Instead of obtaining trip-based data and outputting individual trips on links for a model, activity-based modeling seeks to replicate the set of choices that travelers make based on the pattern of activities that they typically do on a daily basis. Proponents of activity-based modeling argue for the importance of including the context of a person's daily behavior in determining what trips they will take. Activity based modeling allows for trip chaining and activity rescheduling, as they are options for travelers when faced with travel choices, but trip-based modeling will incorrectly weight the effects of trip chains and cannot account for rescheduling.

The scale of these models is typically on a regional level, and as such, it is more difficult to examine the effects of mesoscopic neighborhood-scale changes for land use on transit line usage by using them to evaluate.

Aggregate, Corridor, and Direct Demand Models

Direct demand models are a model formulation that in a single equation, simultaneously estimate trip generation, distribution, and mode split, and are similar to econometric models of demand. It is an aggregate model that can represent multiple travel choices in a single equation.

Ortuzar and Willumsen (2002) give a brief overview of direct demand modeling – the models were first developed in the 1960s and fell out of favor in comparison to the standard travel models seen today. They also note that corridor and direct demand models are typically used in practice to substitute for full scale modeling, due to data limitations. A large amount of data is typically needed for conventional modeling, which typically is difficult to obtain in practice due to the inability [5] to gather that amount of data on an as-needed basis. Corridor models typically remove assignment problems, which simplifies the network structure and creates savings in data collection and coding.

One of the initial formulations for direct demand models was the Kraft-SARC model, documented in Kraft (1963). The model form as formulated for estimating the demand in the Boston-Washington corridor is as follows:

[5] Ortuzar and Williumsen (2002), 421

$$T_{ijm} = \beta_{mo} \prod_{r}^{R} A_{ijr}^{\beta_{mr}} \prod_{n}^{M} \prod_{s}^{S} C_{ijms}^{\alpha_{mns}} (m = 1,2, \dots M)$$

In this formulation, T_{ijm} is the travel demand between an origin i and destination j by mode m, A_{ijr} describes the socioeconomic activity variables between i and j, and C_{ijns} describes impedance variables (cost, travel time, etc) on mode m. R stands for the r-th variable within the set of socioeconomic activity variables, and s for the s-th variable within the set of transportation impedance variables.

Talvitie (1973) was also one of the first to conceptualize a direct travel demand model for different modes, looking primarily at bus and rail modes to the downtown area. He based the estimated work trips on travel times and costs for each mode, and jobs in each zone, using ordinary least squares (OLS) and constrained least squares (CLS) to estimate the elasticities. Cost of travel appeared to be a minor factor, in comparison to access time to the downtown, supporting that travel demand is typically more responsive to travel time than pricing.

Direct demand modeling of the effects of service variation was done by Lago, Mayworm, and McEnroe. They examined elasticities of ridership response to service changes, and which areas were more responsive to service changes, finding that the response is similar acros modes, as are headway and vehicle-mile elasticities. However, they also found ridership is more responsive to changes in lower-service areas, as well as off-peak ridership.

De Cea et al. (1986) describe a formulation of a marginal demand model that simplifies the modeling approach for a particular problem by concentrating on the part of transport demand likely to be affected by the project or policy. They used this modeling approach to examine an extension of the Santiago, Chile metro network, but note that it is a pragmatic methodology whose virtues and limitations can only be seen through practice.

Preston (1991) models rail demand in Britain with three different models, including a direct demand model which he calls an "aggregate simultaneous model." He attempted to forecast flow between stations based on population, residents in two social classes, and number of workplaces within 800 meters of rail stations, along with an index of competition variable that was defined as the travel time for rail divided by the sum of the travel time for rail, bus, and auto. Wardman (1997) also uses direct demand modeling to examine the interaction between modes in an inter-urban area in Britain. Using both the constant elasticity form and exponential form, and solely estimating rail demand, he develops a generalized time variable that combines station to station journey time, headways, and transfers, and subject to fare pricing, but removing the generation and attraction variables.

Cervero (2006) analyzes the effects of smart growth using direct demand models, as traditional models typically are not fine-grained enough to examine the effects of land use patterns and are meant to be applied over a regional area, not at a neighborhood level. He describes three direct demand modeling efforts for transit-oriented development scenarios, including Charlotte, the San Francisco Bay Area, and St. Louis, concluding that though direct demand modeling cannot supplant typical modeling procedures, they can supplement the models in developing order-of-magnitude effects of planning decisions.

The "D's": Land Use Variables

Research has shifted towards how specific land use policies can affect travel forecasting. Cervero and Kockelman (1997) is one of the first articles to attempt to evaluate how the primary dimensions (typically referred to in literature afterwards as the 3Ds) of density, diversity, and design influence travel demand. They noted that density, land-use diversity, and pedestrian-oriented design seem to affect travel behavior, though the influence is minimal.

Later articles explore how destination accessibility and distance to transit affect travel demand. Ewing and Cervero (2001) summarize how changes in the built environment can affect trip frequency, trip length, mode choice, vehicle miles traveled, and vehicle hours traveled, as well as review past literature. They assess how previous studies have defined different types of neighborhoods, and what those studies concluded about mode choice, trip frequency, or trip lengths. They note that dense, mixed-used development that is not connected to the greater regional network only provides minor improvements in travel times and lengths, which is why transit-oriented development has become more popular, as it combines local land use changes with high capacity transit, rather than relying on isolated land use changes in affecting mode choices.

Additional Factors of Interest

Cao, Mokhtarian, and Handy (2009) evaluate the role that residential self-selection plays upon travel behavior, and in contrast to Mokhtarian's early research, concludes that with structural equation modeling, the built environment does have a strong influence on travel demand, more so than residential self-selection.

Others, mainly Donald Shoup, have evaluated how parking plays a role in travel demand. Shoup (1997) argues that unfairly priced parking (by making it free) has caused an excessive amount of driving, and Willson and Shoup (1990) notes that employer-paid parking subsidies greatly increases the occurrence of driving. After those subsidies are removed, solo drivers tend to shift towards either carpooling or transit as commuting options. Another major discussion has also been the importance of regional variables rather than local accessibility in determining travel demand.

Handy (1992) notes the importance of evaluating how the existing structural form (regional accessibility) of the area ties into local accessibility measures in determining how much people will drive, based on the options that they have available. Further research by Spears, Boarnet and Handy (2010) for the California Air Resources Board evaluate how regional accessibility via transit, distance to the central business district, and number of destinations available nearby play a role in determining VMT.

Mathematical Models and Variable Specification

Earlier models tend to use ordinary least squares regression, negative binomial regression, and other forms of linear regression models to estimate their travel behaviors. Vickerman (1972) is one of the first articles to explore how non-work travel is generated and attracted by various origins and destinations, and uses regression models to explore travel behavior in England. Washington, Karlaftis and Mannering (2010) describe in greater detail the process for regression estimation in transportation modeling. They outline assumptions that need to be fulfilled for regression modeling, and the differences between the commonly used regression methods. In particular, regression models require the data to consist of exogenous variables, which means that the variables have no observable influence on each other.

However, in social and behavior research, it is more likely the case that variables are endogenous and affect each other (e.g. population density will be related to transportation network density, as noted in various papers.) [6] As a result, more recent research has focused on structural equation modeling, which better captures the social and behavioral aspects of travel behavior and can account for endogenous variables, as well as machine learning to better replicate observed travel behavior from large data sets. Bagley and Mokhtarian (2002) use structural equation modeling to evaluate endogenous variables for residential neighborhoods in the San Francisco Bay Area affect travel demand. They conclude that land use configuration and travel patterns have no direct causal relationship, but are primarily due to correlations of those variables with each other, and the primary influencers of travel demand are travel attitudes, lifestyle, and sociodemographic variables.

Many of the studies in the 1996 TCRP addressed what factors have been observed to raise overall transit ridership in the specific total service areas of the analyzed transit systems, without attempting to make predictions on the station level. The primary conclusions of papers in the study noted that increasing jobs in the CBD tended to have a corresponding increase in ridership on transit systems, while moving jobs from the CBD to the suburbs tended to remove riders from the system, even if they had been previous transit users.

Cardozo, Palomares, and Gutierrez (2012) attempt to use geographically weighted regression and ordinary least squares to directly forecast transit ridership at the station level. However, they predict only ridership origins, and do not attempt to predict where riders will be attracted to based on travel costs. In addition, two out of the four selected variables that they used as predictors are the number of lines running to a station, as well

[6] Cervero and Murakami (2010), Crane and Crepeau (1998), etc.

as suburban bus lines running to a station, which in reality, have strong feedback effects to ridership, even if the calculated multicollinearity is low.

Travel Survey Data Review

A discussion of travel behavior analysis and transit demand forecasting should also reference the typical data used to inform the models, and also the limitations for analysis based on the data collection methods. Travel behavior data is typically taken from information sources provided by the government. On the federal level, the National Household Travel Survey (NHTS) is typically performed every eight years. The last dataset available for use is from 2009, and the 2017 survey is currently in the process of assembling the collected data for publication in 2018. [7] Information included in the national survey includes household demographic data, vehicle data, trip purposes, mode choice, travel length (in distance and time), and regional transportation characteristics. It does not include travel information costs, route choices, changes in travel patterns over time, location-identifying characteristics, and rationale for mode choice. As a result, travel cannot be tied directly to individual traffic analysis zones (TAZs) or census block groups, which hinders land use and travel behavior interaction modeling efforts.

For research that requires data at that fine of a level, typically state departments of transportation or metropolitan planning organizations (MPOs) will conduct additional travel surveys every ten years. That data will typically include TAZ locations of the households that are surveyed, which allows for the households to be associated with specific census block groups and then land use characteristics. This allows for the specific

[7] http://nhts.ornl.gov/

effects of land use variables and locations to be modeled against vehicle miles traveled or trip length, whereas the national survey primarily serves to model travel behavior against socioeconomic and demographic characteristics. These travel surveys are also typically for intra-regional trips and daily travel, but since 2001, long-distance trips have also been included as part of the NHTS' record.

The NHTS states that it is conducted via a random digit dialing, computer-assisted telephone interviewing survey conducted over an entire year, from the civilian population of the United States. The Southern California Association of Governments (SCAG)[8] and other MPOs[9,10,11] use similar methods to acquire data for their regional travel surveys. Stopher and Greaves (2007) reviews the limitations of the type of data that these surveys procure, as well as how they are conducted. The majority of travel behavior surveys are collected this way, but they note that most surveys underreport trips by approximately 20-30%. The overall response rate for these surveys (when accounting for recruitment rate and completion rate) is also fairly low, at approximately 35%, and non-respondents are households that tend to travel more or have more members. The first set of households tend to be underrecruited, due to their travel schedules, and the latter tend to not complete the surveys once recruited, since additional household members makes completing the surveys more arduous. Both of these limitations are significant, especially in terms of modeling VMT accurately. Those who travel significantly more often than the general population are a key group for modelers to be able to accurately represent. Underreporting

---

[8] http://www.scag.ca.gov/Documents/Final%20Survey%20Methods%20Report.pdf
[9] http://www.baltometro.org/our-work/regional-data-forecasting/household-travel-survey
[10] http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/chts.html
[11] http://www.cmap.illinois.gov/data/transportation/travel-tracker-survey

trips (especially if emissions models are being based off of VMT) will correspond to emission rates that are higher in actuality than those that the model predicts. Though later models, once accounting for historical trends, will likely be able to capture this irregularity, this will only occur after the additional particulate matter has been emitted and damage has already been incurred, and only if the models are validated against previous data and models. Do-not-call registries and screening for telemarketers has also reduced the number of households willing to be recruited for the surveys, and likewise, the survey data's quality has decreased as the sampling has become less accurate.

There are specific limitations in using household travel survey data to estimate transit behavior, and transit demand analysis is typically done by supplanting typical household travel survey data with on-board travel surveys. Typical household travel surveys tend to undersample transit riders. For example, only 3% of the trips observed in the California Household Travel Survey (CHTS) are made by transit, which covers both bus, commuter rail, light rail, and metro rail systems across California over multiple metropolitan areas, whereas statewide in a similar year (2013)[12], the state saw roughly 5.3% of work trips made by transit.[13] This data is split over multiple transit systems and multiple geographic areas, which may behave differently due to system design and overall transit coverage in the urban area, making it hard to evaluate variations in user behavior for a single system.

The proliferation of smartphones has also added another method for which travel behavior can be recorded and measured. Haghani (2010) describes how bluetooth sensors

[12] http://www.energy.ca.gov/almanac/transportation_data/transit.html
[13] https://www.bts.gov/sites/bts.dot.gov/files/legacy/california.pdf

use the Machine Access Control (MAC) addresses of smart phones, cameras, and other connected devices passing through the sensors' area to determine accurate travel times on roads. Travel time data via smartphones is also recorded and used commercially by corporations such as Microsoft and Google in order to estimate real-time traffic for their proprietary mapping software. Ridesharing companies also have a treasure trove of travel data in their databases, and have begun partnering in limited capacity with governments to analyze and use said data.[14]

Smartphone data has also been used to record travel path and routing choice data that previously was difficult to acquire. Hood, Sall and Charlton (2013) use smartphones to acquire bicycle route choice data in San Francisco, and then pair that data with a network model to better understand which routes bicyclists prefer. They analyze road segment variables such as the presence of bicycle lanes, length, turns, grade of the segment, traffic volume, speeds, number of lanes, crime rates, and time of day. Nitsche (2012) and (2014) also describes how smartphones can be used to supplement and recreate the data in traditional travel surveys, and the relative accuracy of the models based on different types of modes, e.g. walking, biking, train, and automobile. Gong, Chen, Bialostozky, and Lawson (2011) designed a GIS algorithm to process data from GPS units to determine mode and route choice and other travel behaviors in New York City using the city's multi-modal transportation network and validate their data against the self-reported travel diaries for fairly high accuracy in reporting (at approximately 80%). Cottrill et. al (2013) discuss how a smartphone-based travel survey can be developed, with a test case in Singapore for the Land Transport Authority (the city-state's urban planning agency.)

[14] https://movement.uber.com/cities

It is worth noting that smartphone data may tend to over represent certain economic classes, based on how data is collected. The initial cost of purchasing a smart phone may be a factor, as well as how recruitment efforts for these surveys are targeted. The Hood (2013) survey contacted local bicycle coalitions and university groups and asked them to send out a link for downloading their survey application, which gave out incentives for response – but the recruited are more likely to be highly educated and with greater purchasing power. However, smartphone ownership covers a huge swath of the American population nowadays – the Pew Research Center estimates that 77% of all Americans now own smartphones, as well as 64% of individuals in lower-income households earning less than $30,000. Thus, the possible skewing of data based on smartphone owner demographics may be less of a factor than what they were previously.[15]

Current Relevance

Climate change issues have come to be one of the driving motivators in transportation research, especially with the awareness that the transportation sector is responsible for 27% of the greenhouse gas (GHG) emissions within the United States. [16] Vehicle miles traveled (VMT) per capita has a high correlation with declines in local environmental quality (via particulate matter and road runoff) and increases in energy resource consumption. As more miles are driven, more particulate matter is emitted, more acreage is set aside for transportation infrastructure, and more fossil fuels are consumed. Cervero and Murakami (2010) note that GHG emissions can be roughly determined by multiplying miles per gallon (fuel consumption) versus the fuel's carbon content, and then

[15] http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/
[16] https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

by a person's vehicle miles traveled. The first are determined by the technology advances available for transportation, and the latter by the activities that the person chooses to partake in as well as their origin (which is typically their residence.) They refer to these two concepts as sustainability mobility – for advances in transportation technology, and sustainable urbanism – for improved land-use planning measures that reduce total vehicle miles traveled. They also note that the majority of research has focused on the sustainable mobility portion, as the benefits in improving transportation technology are more easily measurable. At present, the available literature tying changes in land use to VMT and GHG reductions has been inconsistent in determining the benefits that can be attributed to differing land uses, though a number have concluded that there are measurable, if small, benefits to doing so.

Chapter 2: Study Area and Background

Study Area Overview

The study focuses on the Caltrain commuter rail network, which is a commuter rail line with only one line and 32 stations that serves the cities and counties in between the two cities of San Francisco and San Jose (Figure 1.) It averages approximately 62,000 riders on the weekdays, and 12,000 on the weekends. Two highways, US-101 and I-280, parallel the Caltrain system for its length, and are marked on the figure below. In the year the study data was acquired, Caltrain was notable for being one of the few American transit systems that saw ridership gains rather than declines, due to the strong economic growth of the Bay Area. Ridership gains have also been driven by increased congestion on the two highways paralleling the lines (US-101 and I-280) for the majority of the peak periods during the same time period.

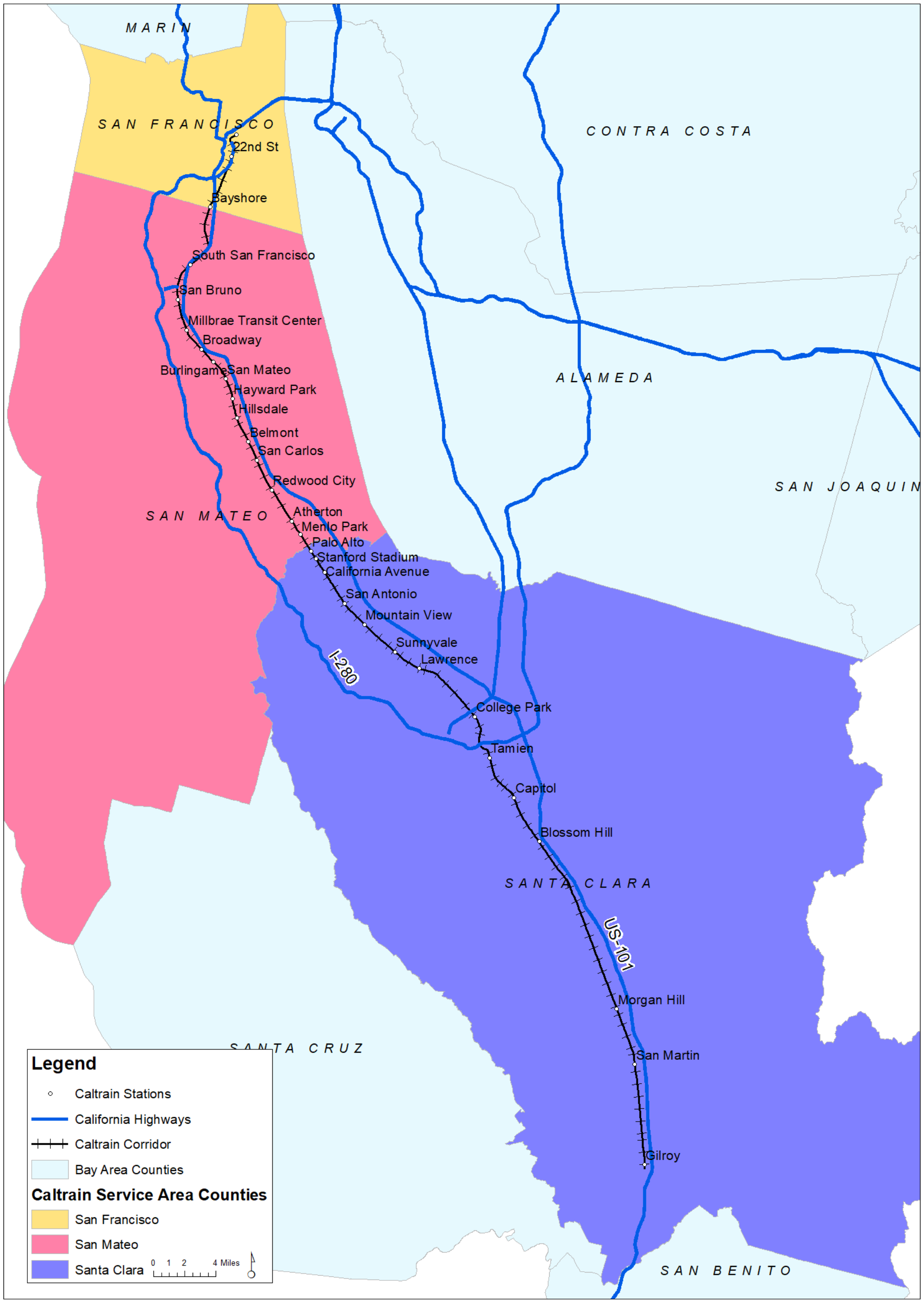

Figure 1: Caltrain Stations and Service Area

Data Overview

The data used for this study came from five primary sources – the EPA's Smart Location Database, the 2014 Caltrain On-Board Travel Survey, Google Maps, and Caltrans' GIS and Performance Measurement System (PeMS) Databases.

Land use data was obtained through the EPA's Smart Location Database. The EPA first published the SLD in 2012, with a complete update in 2014. It is intended to be a nationwide geographic data resource for measuring location efficiency and is available for public use in a GIS format. Sample variables included within the dataset include (on a block group level) – residential density, population density, employment density, employment and housing entropy, street intersections per square mile, distance to nearest transit stop, destination accessibility via car and transit, and more. Land use and accessibility measures were computed for the SLD based on typical measures used previously in transportation research. The data year for the SLD data matches that of the MTC Caltrain data, so it does not need to be scaled for population/job changes. Google Maps and PeMS were used to obtain free flow and congested travel times. GIS data obtained from the Caltrans system included the Caltrain rail network, highway network in the Bay Area, and stations on the network. The Caltrain webpage also provided the total number of parking spaces at each Caltrain station.

MTC's Caltrain On-Board Travel Survey data included information about traveler socioeconomic data, such as workers in household, vehicles in household, household income; trip data, including location (boarding and alighting stations, trip departure hour and predicted return hour), as well as information about the home TAZ, destination TAZ, school TAZ, and workplace TAZ. In addition, a GIS shapefile of the TAZs was provided. The survey consisted of two major portions – the on-to-off element, in order to identify boarding and alighting patterns of transit riders, and to expand the results of the main survey. The main survey was meant to create a detailed profile of typical riders on Caltrain, and gathered socioeconomic data. Over 19,000 on-to-off surveys were completed, as well as approximately 5,000 main surveys.[17] A table of the origin and destination trips by station pairs follows.

---

[17] http://www.caltrain.com/Assets/_MarketDevelopment/pdf/Caltrain+Origin+$!26+Destination+Survey+2014.pdf

Table 1a: O\D Pairs for Caltrain Daily Boardings and Alightings

O \ D	San Francisco	22nd Street	Bay-shore	So. San Francisco	San Bruno	Mill-brae	Burlin-game	San Mateo	Hayward Park	Hillsdale	Belmont	San Carlos	Redwood City	Menlo Park	Palo Alto
San Francisco	0	3	0	45	85	138	81	456	75	507	78	276	834	598	2308
22nd Street	145	0	1	4	7	23	2	47	24	98	8	39	127	90	854
Bayshore	27	0	0	3	2	2	13	5	8	4	0	25	21	0	143
So. San Francisco	154	1	0	0	13	0	10	2	0	52	5	33	39	55	93
San Bruno	221	7	0	3	0	0	2	8	6	27	0	5	59	6	154
Millbrae	446	0	3	0	0	0	32	65	14	45	47	117	265	149	442
Burlin-game	401	15	0	0	7	11	0	9	0	133	14	74	72	166	352
San Mateo	722	23	4	11	20	86	29	0	3	40	33	53	167	102	435
Hayward Park	78	7	0	0	11	16	0	0	0	11	0	20	36	10	75
Hillsdale	1444	23	7	30	18	112	45	33	0	0	0	68	27	81	423
Belmont	208	0	0	7	12	37	19	3	0	1	0	11	37	11	133
San Carlos	297	70	0	23	29	40	15	55	13	27	56	0	18	16	337
Redwood City	1270	69	20	113	31	167	33	100	7	140	6	30	0	17	238
Menlo Park	663	64	0	19	9	157	10	20	13	32	0	0	54	0	113
Palo Alto	2158	22	0	31	12	134	51	43	0	86	7	19	199	14	0
California Ave	607	48	0	12	0	93	4	24	0	21	7	17	21	14	17
San Antonio	349	13	0	8	15	40	0	13	0	67	5	42	60	27	67
Mountain View	2819	56	0	90	8	288	72	64	15	127	13	47	124	128	311
Sunnyvale	2752	14	0	33	44	191	20	156	31	68	40	56	167	103	415
Lawrence	568	0	0	59	3	97	1	15	0	23	0	10	74	23	106
Santa Clara	597	5	0	5	0	105	0	39	3	15	21	40	41	64	211
College Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
San Jose	3499	99	0	26	178	406	104	258	5	325	24	191	431	178	1235
Tamien	473	1	0	7	0	37	4	10	13	10	0	15	143	45	299
Capitol	15	0	0	0	0	0	0	0	0	2	0	13	0	0	1
Blossom Hill	16	0	0	0	0	0	0	0	0	7	0	0	9	6	28
Morgan Hill	29	0	0	0	0	0	0	0	0	0	0	7	0	15	21
San Martin	8	0	0	0	0	8	0	0	0	7	0	0	13	2	15
Gilroy	18	0	0	0	0	1	0	0	0	2	0	0	0	2	31
Subtotal	19985	538	34	529	504	2192	546	1425	230	1876	363	1208	3040	1923	8860

Table 1b: O\D Pairs for Caltrain Daily Boardings and Alightings

O \ D	California Ave	San Antonio	Mountain View	Sunnyvale	Lawrence	Santa Clara	College Park	San Jose	Tamien	Capitol	Blossom Hill
San Francisco	342	158	1156	379	138	315	0	1297	24	13	3
22nd Street	52	21	480	95	23	29	0	140	0	0	0
Bayshore	7	0	2	0	7	15	0	45	0	0	0
So. San Francisco	0	0	18	7	0	61	0	41	0	0	0
San Bruno	5	0	76	59	7	18	0	34	0	0	0
Millbrae	86	6	430	27	13	111	7	278	7	0	0
Burlin-game	107	0	183	32	38	35	0	104	3	0	0
San Mateo	88	10	176	25	32	19	0	184	7	0	0
Hayward Park	13	1	0	7	0	0	0	66	3	0	0
Hillsdale	41	15	228	68	47	41	0	96	16	0	0
Belmont	11	18	14	0	0	38	0	138	0	0	0
San Carlos	35	4	199	59	53	94	21	171	18	0	0
Redwood City	46	3	88	79	11	17	15	195	7	0	0
Menlo Park	9	5	60	11	19	82	38	183	0	0	0
Palo Alto	13	33	139	29	38	14	23	284	0	0	0
California Ave	0	0	46	13	27	29	0	78	12	0	0
San Antonio	0	0	14	15	13	18	11	23	0	0	0
Mountain View	109	34	0	0	54	18	50	205	0	0	0
Sunnyvale	150	18	177	0	0	36	11	122	7	0	0
Lawrence	58	12	0	0	0	0	11	39	0	0	0
Santa Clara	78	0	15	20	0	0	13	42	5	0	0
College Park	0	0	0	0	0	0	0	0	0	0	0
San Jose	265	63	148	129	27	88	13	0	0	0	0
Tamien	99	12	47	30	22	7	0	6	0	0	0
Capitol	0	0	6	0	0	0	0	0	0	0	0
Blossom Hill	7	0	27	13	30	0	10	0	0	0	0
Morgan Hill	22	2	47	26	31	0	0	0	0	0	0
San Martin	33	0	0	0	12	0	0	0	0	0	0
Gilroy	21	0	17	17	13	7	0	0	7	0	0
Subtotal	1695	414	3795	1140	656	1090	222	3770	116	13	3

Chapter 3: Model Methodology, Formulation, and Results

Methodology

In order to calculate the land use variables, the service area of the station had to be defined. A distance threshold for the catchment area of the stations was defined by taking a Euclidean distance of 0.5 miles around the station, which is the typical walking distance used for rail station service areas in most research, and has been extensively studied (Parsons Brinckerhoff 1996, Walters and Cervero 2003, O'Neill Ramsey and Chou 1992). The data from the census block groups that had portions which fell within the 0.5 mile buffer were aggregated. The counts for variables that were hypothesized to correlate with ridership were created for each station, and then divided by the unprotected acreage to create density measures for the station catchment areas. Counts and unprotected acreage for each census block group was taken from the EPA's Smart Location Database, which in turn obtained its data from the US Census's TIGER shapefiles. Land use variables that were created in this step are listed below (Table 2), and were taken from the variables of interest previously reviewed and established in the literature review. A table of values by station for the selected variables of interest (Table 3) follows the hypothesized variables.

Table 2: Hypothesized Ridership Variables

Variable	Description
Population	Population of census block groups that have at least some portion falling within the 0.5 mile station catchment area, abbreviated as 0.5 mile station catchment area CBGs hereafter
Population Density	Calculated population divided by the sum of unprotected acreage of in the 0.5 mile station catchment area CBGs
Housing Units	Housing units in the as 0.5 mile station catchment area CBGs
Housing Density	Calculated housing units divided by the sum of unprotected acreage of in the 0.5 mile station catchment area CBGs
Workers	Total workers in station catchment area
No-car households	Non-car households in the 0.5 mile station catchment area CBGs
One-car households	One-car households in the 0.5 mile station catchment area CBGs
Two-car households	Two-car households in the 0.5 mile station catchment area CBGs
Employment	Total jobs in the as 0.5 mile station catchment area CBGs
Office Employment	Total office jobs in the as 0.5 mile station catchment area CBGs
Industrial Employment	Total industrial jobs in the as 0.5 mile station catchment area CBGs
Entertainment Employment	Total entertainment jobs in the as 0.5 mile station catchment area CBGs
Retail Employment	Total retail jobs in the as 0.5 mile station catchment area CBGs
Employment Density	Calculated total jobs divided by the sum of unprotected acreage of in the 0.5 mile station catchment area CBGs
Parking spaces	Parking spaces at the station
Land-use mix	Entropy mix calculated via employment and household entropy
Network density	Road network density for the 0.5 mile station catchment area CBGs
CBD	Generated-variable that flagged the two central business districts (San Jose and San Francisco)

Table 3a: Hypothesized Ridership Variables Values by Station

STATION	Parking	Housing Units	Housing Density	Population	Population Density	0 Car HH	1 Car HH	2+ Car HH
SAN FRANCISCO	0	6288	11.459	11629	21.192	1616	3659	1007
22ND ST	0	2408	5.561	4577	10.570	367	1167	869
BAYSHORE	38	3428	1.598	11344	5.288	427	1032	1962
SOUTH SAN FRANCISCO	81	2787	1.070	9453	3.630	323	964	1494
SAN BRUNO	178	2395	5.154	7619	16.398	130	828	1432
MILLBRAE TRANSIT CENTER	175	4544	3.105	10651	7.278	415	2010	2109
BURLINGAME	68	4147	9.161	7626	16.847	414	2145	1582
SAN MATEO	42	4206	9.893	10797	25.395	587	2221	1388
HAYWARD PARK	213	4521	4.094	10699	9.690	240	1608	2664
HILLSDALE	518	4169	4.882	9984	11.690	340	1457	2366
BELMONT	375	4700	3.490	11088	8.232	263	1507	2919
SAN CARLOS	219	3195	2.840	6277	5.579	251	1192	1748
REDWOOD CITY	557	3854	4.811	9934	12.401	412	1799	1634
MENLO PARK	150	4337	2.866	8836	5.839	385	1898	2043
PALO ALTO	389	7044	4.007	14483	8.238	1069	4011	1956
CALIFORNIA AVENUE	185	3068	4.504	6993	10.265	191	863	2007
SAN ANTONIO	199	4385	6.495	9347	13.844	431	1853	2093
MOUNTAIN VIEW	340	3264	6.611	5980	12.113	267	1580	1411
SUNNYVALE	367	4498	2.393	9992	5.315	340	2091	2060
LAWRENCE	122	3357	2.094	8098	5.052	92	1362	1898
SANTA CLARA	289	1981	0.804	6332	2.570	153	794	1029
COLLEGE PARK	0	2160	3.544	4642	7.616	252	1006	899
SAN JOSE	581	12026	4.176	26099	9.064	976	5732	5300
TAMIEN	275	3081	5.374	8196	14.295	271	945	1860
CAPITOL	379	3686	3.360	9488	8.650	83	1332	2268
BLOSSOM HILL	425	1555	1.804	5054	5.862	27	210	1313
MORGAN HILL	486	3217	1.537	9537	4.557	69	601	2543
SAN MARTIN	167	924	0.134	2971	0.429	4	137	782
GILROY	471	2247	2.968	8333	11.006	182	702	1355

Table 3b: Hypothesized Ridership Variables Values by Station

STATION	Office	Retail	Industrial	Entertainment	Total Workers	Avg. Network Density
SAN FRANCISCO	4402	824	3516	3298	3047	33.258
22ND ST	227	213	1786	578	2146	22.171
BAYSHORE	1795	442	3791	144	5174	7.456
SOUTH SAN FRANCISCO	2564	2300	19949	2234	3941	14.853
SAN BRUNO	63	225	69	540	3197	25.029
MILLBRAE TRANSIT CENTER	1756	1074	12620	3983	4828	19.052
BURLINGAME	805	1037	259	933	3363	21.386
SAN MATEO	720	805	412	1383	3903	28.131
HAYWARD PARK	2956	1089	772	1232	5419	21.481
HILLSDALE	1227	2221	342	1133	5171	21.320
BELMONT	771	584	2257	870	5110	19.780
SAN CARLOS	863	1542	4933	846	2946	19.242
REDWOOD CITY	5549	1950	1102	1019	3992	24.130
MENLO PARK	1979	1040	426	1175	3720	15.383
PALO ALTO	3559	3008	822	3473	4743	27.580
CALIFORNIA AVENUE	1085	902	579	770	2665	26.719
SAN ANTONIO	497	1779	175	636	4167	21.680
MOUNTAIN VIEW	1088	175	145	812	2961	26.998
SUNNYVALE	2109	1077	16991	1069	5016	16.106
LAWRENCE	1054	1286	15550	546	3515	13.802
SANTA CLARA	2584	1236	19435	1278	1903	15.927
COLLEGE PARK	1772	620	1266	862	2010	22.996
SAN JOSE	6445	2793	5028	3694	11083	25.471
TAMIEN	78	90	956	76	2967	24.139
CAPITOL	175	78	338	148	3870	17.628
BLOSSOM HILL	92	446	3143	201	2261	22.134
MORGAN HILL	475	212	3120	595	3669	14.593
SAN MARTIN	20	83	281	258	1102	3.940
GILROY	417	511	550	324	2960	20.840

Travel costs were calculated using congested travel time on highways against travel time on the Caltrain network. Free flow travel time was estimated using the average time at 3 am from station to station using Google Maps. Then, peak period travel time on the highway was estimated using the average travel time index from PeMS during peak periods for the stretch of highway corresponding to the Caltrain link. The free flow travel speed was multiplied by the travel time index to get a time estimate, as the travel time index is defined by PeMS as the ratio of the average travel time for all travelers in a certain area to the free-flow travel time. The PeMS data used was taken from two weeks in October 2014 (during the same time period as the on-board travel survey) for Tuesday, Wednesday, and Thursday each week and averaged to best replicate typical conditions. Travel time on the Caltrain network was calculated using the published Caltrain timetables.

The provided Caltrain data was synthesized to create origin and destination pairs from the survey data (Table 4.) The weighting factor attributed to each trip was summed based on the home station and destination station provided by the survey respondent. The accuracy of the home and destination station for the respondents was checked by matching the stations to the provided home TAZ code and destination TAZ code. The home TAZ was considered the origin TAZ, and the employment/school/other TAZ would be considered the destination TAZ. In addition, a table of origin/destination pairs by station was also created and follows below.

Table 4: Caltrain Daily Ridership Boardings and Alightings by O\D

STATION	***ORIGIN***	***DESTINATION***
SAN FRANCISCO	9311	19985
22ND ST	2309	538
BAYSHORE	331	34
SOUTH SAN FRANCISCO	584	529
SAN BRUNO	696	504
MILLBRAE TRANSIT CENTER	2588	2192
BROADWAY	0	0
BURLINGAME	1758	546
SAN MATEO	2267	1425
HAYWARD PARK	354	230
HILLSDALE	2864	1876
BELMONT	698	363
SAN CARLOS	1654	1208
REDWOOD CITY	2699	3040
ATHERTON	0	0
MENLO PARK	1561	1923
PALO ALTO	3349	8860
STANFORD STADIUM	0	0
CALIFORNIA AVENUE	1088	1695
SAN ANTONIO	799	414
MOUNTAIN VIEW	4630	3795
SUNNYVALE	4613	1140
LAWRENCE	1099	656
SANTA CLARA	1318	1090
COLLEGE PARK	3	222
SAN JOSE	7692	3770
TAMIEN	1279	116
CAPITOL	38	13
BLOSSOM HILL	151	3
MORGAN HILL	199	0
SAN MARTIN	99	0
GILROY	135	0
TOTAL	56166	56167

Formulation

The aggregate ridership model as tested consisted of two formulations. One replicated typical methods of estimating ridership via ordinary least squares regression, in order to validate hypothesis about variables' effects on ridership.

The first was formulated as

$T_i = f(origin\ land\ use\ charecteristics)$ and $T_j = f(destination\ land\ use\ charecteristics)$

Where

$$T_i = number\ of\ trips\ originating\ at\ station\ i$$

$$T_j = number\ of\ trips\ originating\ at\ station\ j$$

The second tested a log-sum formulation that related the land use characteristics at each end to the number of riders on its origin/destination station pair, and included travel-time costs between stations for both transit and congested peak speeds. A hypothesized relationship for ridership between stations, travel time between stations, and land use variables at origin and destinations was proposed in the form of:

$$T_{ij} = f(travel\ time, origin\ land\ use\ characteristics, destination\ land\ use\ characteristics)$$

This posits that the interaction between two locations (via trips between the two locations) is related to characteristics of urban development at the origin and destination, as well as taking into account the travel time between destination and origin. The formulation was tested as:

$$T_{ij} = T_i * T_j * t_{ij}$$

Where:

$$T_{ij} = trips\ between\ station\ i\ to\ station\ j$$
$$T_i = trips\ originating\ at\ station\ i$$
$$T_j = trips\ destined\ to\ station\ j$$

$$t_{ij} = travel\ time\ between\ i\ and\ j$$

As the model seeks to estimate the effects of the independent variables upon the dependent variables, a log-log transformation of the variables was taken in order to produce an equation that could estimate the percent change in the dependent variable (trips between a station pair) with a percent change in the independent variables. The result could then be transformed back to determine the predicted number of trips for that station pair. This leads to a form:

$$\log(T_{ij}) = \log(T_o) + \log(T_j) + \log(t_{ij})$$

Direct forecasting of ridership at a station level has been done primarily with ordinary least squares regression or multiple regression analysis, which is seen in Parsons Brinckerhoff 1996, Cardozo et. al 2012, Cervero 2006. In addition, this replicates the estimation methods of the four-step model (McNally 2007), which tends to use regressions to estimate trips. This method was used to estimate origins and destinations from the available land-use variables. The model would ideally be able to predict the variation in ridership from land-use, before using the variables in the log-sum formulation.

Model Estimation, OLS

The final model for origins incorporated three independent variables, which were households owning 2-cars or more, total workers in the catchment area, and the central business district variable. These three variables were all statistically significant at the 0.05 level, and explained approximately 68% of the variation found within the destination data. None of the variables appeared to show multicollinearity, based on their variance inflation factors (VIF). The variables also appeared to be correlated with the data in the hypothesized ways – an increase of households owning two or more cars in the station catchment area saw a decrease in transit trip origins, and increasing the number of workers in the catchment area also seems to correlate with increased transit trip origins. The two stations in the CBDs of San Jose and San Francisco also were associated with an increased number of origins.

Table 5: Final OLS Model, Origins

Variable	**Coefficient Estimate**	**Standard Error**	***t*-Statistic**	**Probability**	**VIF**
Intercept	1432.9774	603.9327	2.373	0.0257	-
Households with two or more cars	-1.6215	0.7019	-2.310	0.0294	6.54
Total workers	0.8043	0.3550	2.266	0.0324	7.24
CBD	6499.4585	1066.3058	6.095	.00000227	1.36
# of observations	29				
# of variables	3				
Adjusted R^2	0.6893				
R^2	0.7225				
Residual standard error	1246				
Degrees of freedom	25				
F-statistic	21.7	*p-value (F-statistic)*	.0000003881		

In addition to the final model, other models for origins were tested, which included the two variables for single-car households and no car household with the other tested variables, as well as including population and parking spaces at the station. However, the correlations given by the coefficients and their significance levels were less significant, and also had problems with redundancy as indicated by their variance inflation factors. The model coefficients for the variables of interests are presented below.

Table 6: Preliminary OLS Model, Origins – Single Car, Zero Car Households

Variable	Coefficient Estimate	Standard Error	*t*-Statistic	Probability	VIF
Intercept	1138.0675	706.9632	1.610	0.12108	-
Zero-car households	-0.1372	1.7877	-0.077	0.93949	8.57
One-car households	1.240	0.5516	2.038	0.05322	9.64
Households with two or more cars	-0.9679	0.7808	-1.240	0.22763	10.47
Total workers	0.1421	0.3871	0.367	0.71696	11.13
CBD	4311.9165	1408.1324	3.062	0.00552	3.07
# of observations	29				
# of variables	3				
Adjusted R^2	0.7596				
R^2	0.8025				
Residual standard error	1096				
Degrees of freedom	23				
F-statistic	18.7	*p-value (F-statistic)*	.0000002005		

A model consisting of the zero-car and one-car households, with CBD, and population, and parking spaces at stations variables was also tested, but likewise, had problems with redundancy, even if it explained additional variation in the model and the correlation coefficients passed typical levels of significance.

Table 7: Preliminary OLS Model, Origins – Additional Variables

Variable	Coefficient Estimate	Standard Error	*t*-Statistic	Probability	VIF
Intercept	1004.0757	547.4947	1.834	0.080226	
Zero-car households	3.3838	1.3604	2.487	0.020936	7.567
One-car households	1.0459	0.4411	2.371	0.026901	9.4041
Parking	3.2027	1.1874	2.697	0.0131	1.56
Population	-0.6859	0.1908	-3.594	0.001613	21.93
Total workers	0.8389	0.3692	2.272	0.033197	15.44
CBD	4282.9282	1082.5133	3.956	0.000671	2.76
# of observations	29				
# of variables	6				
Adjusted R^2	0.8425				
R^2	0.8762				
Residual standard error	887.7				
Degrees of freedom	25				
F-statistic	25.94	*p-value (F-statistic)*	.000000000644		

The final model for destinations incorporated three independent variables, one of which was the central business district indicator variable, and the other two related to employment – specifically, employment density and total industrial employment. These three variables were all statistically significant at the 0.001 level, and explained approximately 75% of the variation found within the destination data. None of the variables appeared to show multicollinearity, based on their variance inflation factors. The variables also followed the expected correlation – stations with a greater employment density in the catchment area were likely to have more trip ends. The two stations located in the two central business districts that anchored the census statistical area also had more trip ends. A higher number of industrial workers seemed to be associated with less trip ends.

Total employment was also tested instead of employment density, and also saw a statistically significant correlation at a probability of 0.06 and a t-value of 1.940. However, the total employment model had a greater standard error of 2743, lower adjusted R^2 and R^2 values (0.5624, 0.5099), and a lower F-statistic, and so the employment density model was chosen for inclusion instead.

Table 8: Final OLS Model, Destinations

Variable	Coefficient Estimate	Standard Error	*t*-Statistic	Probability	VIF
Intercept	-1477	671.0	-2.202	0.037119	-
Employment Density	352.9	62.51	5.645	0.0000071	1.54
CBD	6164	1638	3.762	0.000909	1.31
Total Industrial Workers	-.1918	.06587	-2.912	0.007448	1.22
# of observations	29				
# of variables	3				
Adjusted R^2	0.7787				
R^2	0.7521				
Residual standard error	1951				
Degrees of freedom	25				
F-statistic	29.31	*p-value (F-statistic)*	.00000002383		

As the census business district stations appear in both formulations as a significant factor, It could be hypothesized that stations in these areas tend to behave significantly differently than other stations – they both produce and attract a significant amount of trips, irrespective of other land-use factors. However, this is also because central business districts tend to be some of the densest, concentrated employment centers of their geographic region, and San Francisco and San Jose also hold a significant amount of residents for the region. A table (Table 9) of the selected variables and their values for each station follows on the next page.

Table 9: Station Land Use Variable Values

STATION	ORIGIN	DESTINATION	CBD	2 Car + Ownership Households	Total Industrial Workers	Total Workers	Employment Density
SAN FRANCISCO	9311	19985	1	1007	3516	3047	34.576
22ND ST	2309	538	0	869	1786	2146	9.143
BAYSHORE	331	34	0	1962	3791	5174	3.848
SOUTH SAN FRANCISCO	584	529	0	1494	19949	3941	14.115
SAN BRUNO	696	504	0	1432	69	3197	2.849
MILLBRAE TRANSIT CENTER	2588	2192	0	2109	12620	4828	19.017
BURLINGAME	1758	546	0	1582	259	3363	9.530
SAN MATEO	2267	1425	0	1388	412	3903	17.177
HAYWARD PARK	354	230	0	2664	772	5419	8.959
HILLSDALE	2864	1876	0	2366	342	5171	6.951
BELMONT	698	363	0	2919	2257	5110	5.232
SAN CARLOS	1654	1208	0	1748	4933	2946	10.516
REDWOOD CITY	2699	3040	0	1634	1102	3992	20.382
MENLO PARK	1561	1923	0	2043	426	3720	6.477
PALO ALTO	3349	8860	0	1956	822	4743	16.869
CALIFORNIA AVENUE	1088	1695	0	2007	579	2665	9.706
SAN ANTONIO	799	414	0	2093	175	4167	8.390
MOUNTAIN VIEW	4630	3795	0	1411	145	2961	11.809
SUNNYVALE	4613	1140	0	2060	16991	5016	17.123
LAWRENCE	1099	656	0	1898	15550	3515	17.761
SANTA CLARA	1318	1090	0	1029	19435	1903	14.910
COLLEGE PARK	3	222	0	899	1266	2010	13.421
SAN JOSE	7692	3770	1	5300	5028	11083	10.825
TAMIEN	1279	116	0	1860	956	2967	3.105
CAPITOL	38	13	0	2268	338	3870	1.036
BLOSSOM HILL	151	3	0	1313	3143	2261	10.942
MORGAN HILL	199	0	0	2543	3120	3669	3.186
SAN MARTIN	99	0	0	782	281	1102	0.124
GILROY	135	0	0	1355	550	2960	3.245

Model Estimation, Log-sum

The direct demand model is estimated using the form In the previous sections, an assessment of the land use variables hypothesized to have an effect on creation of trips originating and trips destined to stations was established. The equations for those are as follows:

$$T_o = 1432 - 1.625HH_{2car} + 0.8043W_{total} + 6499CBD$$
$$T_j = -1147 + 3529E_{density} - 0.1918W_{industrial} + 6164CBD$$

Where:

$$HH_{1car} = households\ owning\ 2\ or\ more\ cars$$
$$W_{total} = total\ workers$$
$$E_{density} = employment\ density$$
$$W_{industrial} = total\ industrial\ workers$$
$$CBD = identifier\ for\ central\ business\ district\ location$$

Congested network travel times and travel times between stations via transit were also obtained in order to see if there was a relation between an increase in network travel time and transit time to ridership.

The general multiplicative form of a direct demand model can be formulated as:

$$T_{ij} = a^{X^b}Y^cZ^d$$

Where a, y, and z are some hypothetical parameters (in this case, land use and travel time.)

As $e^{T_{ij}} = \left(\frac{dT}{dX}\right)\left(\frac{X}{T}\right) = b$, the equation can also be written as a log-log formulation of:

$$\log(T_{ij}) = \log(a) + b\ log(X) + c\ log(Y) + d\ \ log(Z)$$

Using the log log formulation, the following model was developed:

Table 10: Final Direct Demand Model Estimation

Variable	**Coefficient Estimate**	**Standard Error**	***t*-Statistic**	**Probability**
Intercept	-0.65243	1.287	-0.507	0.6122
CBD_j	2.2263	0.234	9.519	< 2e-16
E_density_j	0.81911	0.061	13.406	< 2e-16
E_industrial_j	-0.1628	0.0409	-3.982	7.44e-05
HH_2car_i	-0.99111	0.30504	-3.249	0.00121
CBD_i	1.821	0.2462	7.396	3.53e-13
HH_workers	1.311	0.3057	4.29	2.01e-05
Peak AM TT	-3.98051	0.61664	-6.455	1.87e-10
Transit TT	3.6452	0.65509	5.564	3.58e-08
# of observations	29			
# of variables	3			
Adjusted R^2	0.3984			
R^2	0.4043			
Residual standard error	1.613			
Degrees of freedom	803			
F-statistic	68.12		*p-value (F-statistic)*	< 2.2e-16

The model then becomes:

$$\begin{aligned} log(T_{ij}) = {} & 2.263\log(CBD_j) + 0.819 log\left(E_{density_j}\right) - 0.1628\log\left(E_{industrial_j}\right) \\ & - 0.9911 log(HH_{2car_i}) + 1.821\log(CBD_i) + 1.311 log(HH_{workers_i}) \\ & - 3.98\log(TT_{am}) + 3.6452\log(TT_{transit}) \end{aligned}$$

In terms of the land use variables, the relative magnitude and direction of the land use variables effect in the log-log formulation on trips produced at origin and attracted to destination stays similar to that observed in the linear formulations. The station being located either in a CBD for its origin or destination appears to be one of the biggest factors in whether a transit trip will occur at it. Congested travel times between stations and transit travel times between stations were considered to see if there was an effect upon ridership. A table of the predicted values of the trips by origin and destination follows below in Table 11.

Table 11a: Model Predicted Trips by O\D Pair

O\D	SAN FRANCISCO	22ND ST	BAYSHORE	SOUTH SAN FRANCISCO	SAN BRUNO	MILLBRAE TRANSIT CENTER	BURLIN GAME	SAN MATEO	HAYWARD PARK	HILLSDALE	BELMONT
SAN FRANCISCO		71	24	59	45	99	104	146	78	73	43
22ND ST	207		3	7	6	13	13	18	10	9	5
BAYSHORE	228	9		15	12	25	24	31	16	15	9
SOUTH SAN FRANCISCO	234	10	6		19	34	27	31	16	15	8
SAN BRUNO	208	9	6	22		30	20	23	12	11	6
MILLBRAE TRANSIT CENTER	263	12	7	23	17		20	22	13	12	7
BURLINGAME	214	9	5	14	9	16		15	10	10	6
SAN MATEO	275	12	6	15	9	16	14		23	19	11
HAYWARD PARK	225	10	5	12	8	14	15	35		16	9
HILLSDALE	238	10	5	12	8	15	16	33	19		11
BELMONT	192	8	4	10	6	12	13	25	14	15	
SAN CARLOS	147	6	3	7	5	9	9	16	8	7	3
REDWOOD CITY	233	10	5	11	7	14	15	25	13	12	6
MENLO PARK	173	7	3	8	5	10	11	18	10	9	5
PALO ALTO	254	11	5	12	8	15	17	28	15	14	8
CALIFORNIA AVENUE	119	5	2	6	4	7	8	13	7	7	4
SAN ANTONIO	206	9	4	9	6	12	14	22	12	11	7
MOUNTAIN VIEW	192	8	4	9	6	11	13	20	11	10	6
SUNNYVALE	248	10	5	11	7	14	16	25	13	12	7
LAWRENCE	158	7	3	7	5	9	10	15	8	7	4
SANTA CLARA	115	5	2	5	3	6	7	10	5	5	3
COLLEGE PARK	134	5	2	5	4	7	7	12	6	5	3
SAN JOSE	1207	49	22	48	32	63	66	100	52	47	27
TAMIEN	95	4	2	4	2	5	5	8	4	4	2
CAPITOL	101	4	2	4	3	5	5	8	4	4	2
BLOSSOM HILL	88	4	2	3	2	5	5	7	4	3	2
MORGAN HILL	89	4	2	4	2	5	5	7	4	4	2
SAN MARTIN	61	2	1	2	2	3	3	5	3	2	1
GILROY	135	6	2	5	4	7	8	12	6	6	3
Grand Total	6038	316	142	349	244	480	488	730	396	364	211

Table 11b: Model Predicted Trips by O\D Pair

O\D	SAN CARLOS	REDWOOD CITY	MENLO PARK	PALO ALTO	CALIFORNIA AVENUE	SAN ANTONIO	MOUNTAIN VIEW	SUNNYVALE	LAWRENCE
SAN FRANCISCO	63	138	64	129	89	96	129	76	75
22ND ST	8	17	8	16	11	12	16	9	9
BAYSHORE	12	26	12	25	17	18	24	14	14
SOUTH SAN FRANCISCO	12	25	12	23	16	17	23	13	13
SAN BRUNO	9	19	9	18	13	13	18	10	10
MILLBRAE TRANSIT CENTER	10	21	10	20	14	15	20	12	11
BURLINGAME	8	18	8	17	12	13	17	10	9
SAN MATEO	13	28	13	27	18	20	26	14	13
HAYWARD PARK	10	22	10	22	15	16	21	12	11
HILLSDALE	10	23	11	23	16	17	23	12	11
BELMONT	6	17	9	19	13	14	18	10	9
SAN CARLOS		19	9	19	13	13	17	9	8
REDWOOD CITY	14		18	37	25	24	29	14	12
MENLO PARK	10	29		36	22	20	22	10	9
PALO ALTO	16	43	27		35	28	31	14	12
CALIFORNIA AVENUE	7	20	11	24		12	13	6	5
SAN ANTONIO	12	31	16	30	19		22	9	7
MOUNTAIN VIEW	11	25	12	23	14	15		7	6
SUNNYVALE	12	27	12	23	14	14	15		9
LAWRENCE	7	15	7	12	8	7	9	6	
SANTA CLARA	4	10	4	7	5	4	5	3	3
COLLEGE PARK	5	11	5	8	5	5	6	3	3
SAN JOSE	42	90	39	71	44	44	54	32	32
TAMIEN	3	7	3	6	4	4	4	3	3
CAPITOL	3	7	3	6	4	4	5	3	3
BLOSSOM HILL	3	7	3	5	3	4	5	3	3
MORGAN HILL	3	7	3	6	4	4	5	3	3
SAN MARTIN	2	5	2	4	3	3	4	2	2
GILROY	5	11	5	10	6	7	9	6	6
Grand Total	323	717	344	667	462	463	589	324	309

Table 11c: Model Predicted Trips by O\D Pair

O\D	SANTA CLARA	COLLEGE PARK	SAN JOSE	TAMIEN	CAPITOL	BLOSSOM HILL	MORGAN HILL	SAN MARTIN	GILROY
SAN FRANCISCO	55	75	419	21	9	45	17	2	24
22ND ST	7	9	50	2	1	5	2	0	3
BAYSHORE	10	13	73	4	2	8	3	0	4
SOUTH SAN FRANCISCO	9	12	66	3	1	7	3	0	4
SAN BRUNO	7	9	51	2	1	5	2	0	3
MILLBRAE TRANSIT CENTER	8	10	58	3	1	6	2	0	3
BURLINGAME	6	9	47	2	1	5	2	0	3
SAN MATEO	9	12	66	3	1	7	3	0	4
HAYWARD PARK	7	10	52	3	1	5	2	0	3
HILLSDALE	8	10	54	3	1	6	2	0	3
BELMONT	6	8	42	2	1	4	2	0	3
SAN CARLOS	5	6	34	2	1	4	1	0	2
REDWOOD CITY	8	10	53	3	1	6	2	0	3
MENLO PARK	5	7	37	2	1	4	2	0	2
PALO ALTO	7	9	49	2	1	5	2	0	3
CALIFORNIA AVENUE	3	4	21	1	0	2	1	0	2
SAN ANTONIO	5	6	33	2	1	4	2	0	3
MOUNTAIN VIEW	4	5	28	1	1	3	1	0	2
SUNNYVALE	5	6	36	2	1	5	2	0	3
LAWRENCE	3	4	24	1	1	3	1	0	2
SANTA CLARA		4	22	1	0	3	1	0	2
COLLEGE PARK	3		31	2	1	4	2	0	3
SAN JOSE	31	50		22	7	44	20	2	35
TAMIEN	3	4	35		0	4	2	0	3
CAPITOL	3	4	26	1		14	4	0	6
BLOSSOM HILL	3	4	30	2	2		3	0	5
MORGAN HILL	3	5	37	2	2	9		1	6
SAN MARTIN	2	4	27	2	1	6	4		4
GILROY	6	9	67	4	3	14	7	1	
Grand Total	229	316	1566	99	45	236	96	11	146

The model appears to account for approximately 40% of the variation in trips between OD pairs, and a comparison follows below. The model appears to underpredict trips between station pairs that have higher ridership, and a comparison of the estimated origins and estimated destinations against the observed origins and destinations follows in Table 12. This is less apparent in the log-log formulation, as seen in the tabularized summary statistics for the model and the graph below.

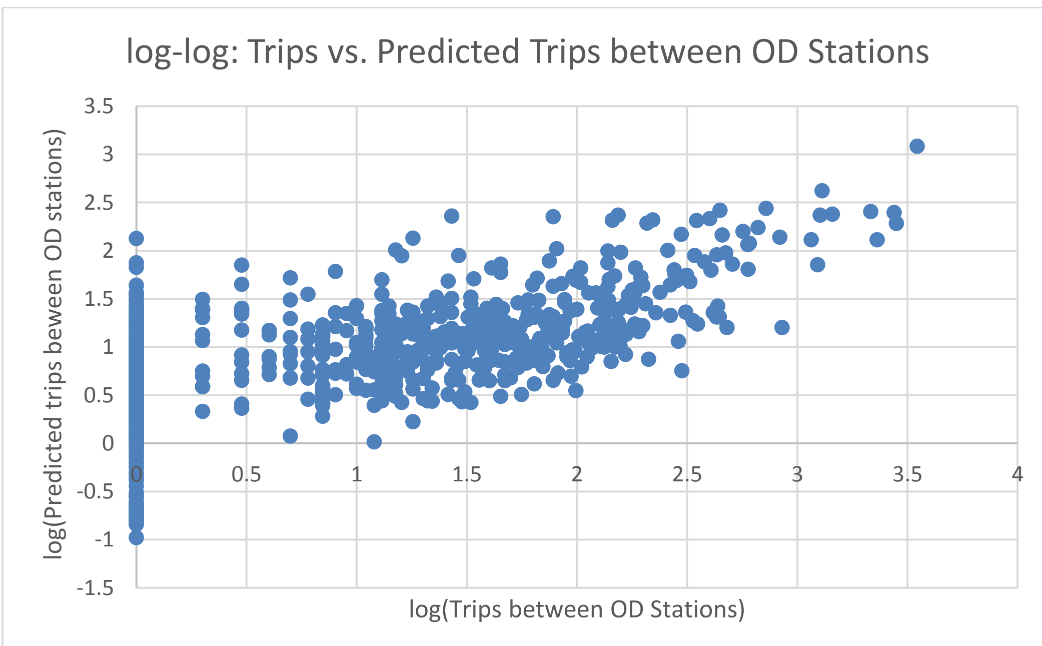

Figure 2: Log of Trips vs Log of Predicted Trips for OD Station Pairs

However, when examining the log-log residuals, it is more apparent that the model has difficulty making an accurate prediction for the station pairs with higher ridership, and that a pattern exists for the residual .

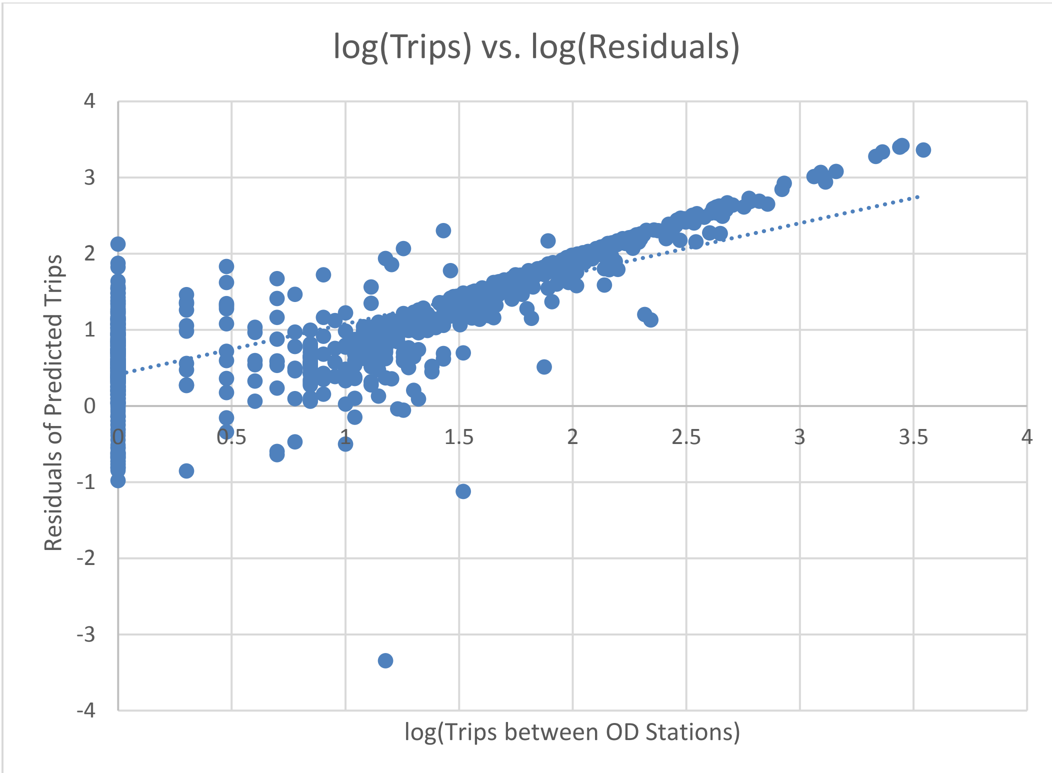

Figure 3: Residual Plot of the log of Trips vs. Predicted Trips

Thus, improvements on the model form could likely be made in order to account for the variation in ridership a higher OD-pair stations.

Table 12: Observed and Estimated Origins and Destinations by Station

	Observed Origins	Estimated Origins	Difference	Observed Destinations	Estimated Destinations	Difference
San Francisco	19985	6038	-13947	9311	2269	7043
22nd Street	538	316	-222	2309	477	1832
Bayshore	34	142	109	331	661	-330
So. San Francisco	529	349	-180	584	659	-76
San Bruno	504	244	-260	696	546	150
Millbrae	2192	480	-1711	2588	622	1966
Burlin-game	546	488	-58	1758	497	1261
San Mateo	1425	730	-695	2267	676	1590
Hayward Park	230	396	166	354	569	-215
Hillsdale	1876	364	-1512	2864	600	2264
Belmont	363	211	-152	698	481	216
San Carlos	1208	323	-885	1654	381	1272
Redwood City	3040	717	-2323	2699	608	2091
Menlo Park	1923	344	-1580	1561	478	1083
Palo Alto	8860	667	-8194	3349	671	2679
California Ave	1695	462	-1234	1088	312	776
San Antonio	414	463	49	799	513	287
Mountain View	3795	589	-3205	4630	449	4181
Sunnyvale	1140	324	-816	4613	555	4059
Lawrence	656	309	-347	1099	341	758
Santa Clara	1090	229	-861	1318	245	1073
College Park	222	316	94	3	286	-283
San Jose	3770	1566	-2205	7692	2373	5319
Tamien	116	99	-17	1279	222	1057
Capitol	13	45	32	38	238	-200
Blossom Hill	3	236	233	151	209	-57
Morgan Hill	0	96	96	199	229	-30
San Martin	0	11	11	99	164	-65
Gilroy	0	146	146	135	367	-232

Chapter 4: Conclusion, Future Work and Recommendations

In conclusion, the model sought to evaluate the creation of a direct demand model for transit ridership that would be able to integrate origin and destination land use characteristics, in addition to network travel time information. The model used a multiplicative direct demand model to estimate elasticities between land use variables and network characteristics for the origin-destination trips.

At both the origin and destination, the location of the station within the central business districts of the San Francisco Bay region had the largest effect on trip generation and attraction. Higher employment density at the destination and a larger number of workers per household at the origin had a positive effect on trips, while the total number of industrial workers at the destination and an increased number of two car households had a negative effect on trips. Longer travel times on transit appeared to have a positive effect on trips, yet longer travel times in congested peak periods appeared to have a negative effect on trips.

Future work would include the examination of additional variables to see if any explanatory variables could account for the variation at stations that generate and attract more riders. High ridership stations are usually the points of interests for agencies, in order to evaluate peak load and peak load capacities. As such, a model that could accurately evaluate the effects of changing land use characteristics and travel times as a result of land development in a transit corridor for high ridership station would be beneficial. A likely explanatory variable could be subcenters for employment. Adding subcenters as a variable would require a regional analysis for subcenters to be determined. In addition, future work could examine the effects of aggregation at a TAZ level instead of a station level, which would likely be useful for evaluating land development projects. Future work could also evaluate if this model form works on a non-linear corridor system with multiple branch lines, which would likely add to the complexity.

More models similar to this could also be easily constructed if more transit agencies had readily available data, either from obtaining the data from transit card usage or performing similar on-board travel surveys.

References

Bagley, Michael N., and Patricia L. Mokhtarian. "The impact of residential neighborhood type on travel behavior: A structural equations modeling approach." *The Annals of regional science* 36.2 (2002): 279-297.

Beimborn, Edward A. "A transportation modeling primer." *Center for Urban Transportation Studies, University of Wisconsin—Milwaukee. Available http://www. uwm. edu/Dept/CUTS/primer. htm* (1995).

Bhat, Chandra R., and Frank S. Koppelman. "Activity-based modeling of travel demand." *Handbook of transportation Science*. Springer, Boston, MA, 1999. 35-61.
Cardozo, Osvaldo Daniel, Juan Carlos García-Palomares, and Javier Gutiérrez. "Application of geographically weighted regression to the direct forecasting of transit ridership at station-level." *Applied Geography* 34 (2012): 548-558.

Cervero, Robert, and Kara Kockelman. "Travel demand and the 3Ds: density, diversity, and design." *Transportation Research Part D: Transport and Environment* 2.3 (1997): 199-219.

Cervero, Robert. "Alternative approaches to modeling the travel-demand impacts of smart growth." *Journal of the American Planning Association* 72.3 (2006): 285-295.

Cervero, Robert, and Jin Murakami. "Effects of built environments on vehicle miles traveled: evidence from 370 US urbanized areas." *Environment and planning A* 42.2 (2010): 400-418.

Cottrill, Caitlin, et al. "Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore." *Transportation Research Record: Journal of the Transportation Research Board* 2354 (2013): 59-67.

De Cea, J., J. De D. Ortúzar, and L. G. Willumsen. "Evaluating marginal improvements to a transport network: An application to the Santiago underground." *Transportation* 13.3 (1986): 211-233.

Domencich, Thomas A., Gerald Kraft, and Jean Paul Valette. *Estimation of urban passenger travel behavior: An economic demand model*. Charles River Associates, 1968.

Ellis, Raymond H., Paul R. Rassam, and John C. Bennett. "Consideration of Intermodal Competition in the Forecasting of National Intercity Travel." *Highway Research Record* 369 (1971).

Ewing, Reid, and Robert Cervero. "Travel and the built environment: a synthesis." *Transportation Research Record: Journal of the Transportation Research Board* 1780 (2001): 87-114.

Gong, Hongmian, et al. "A GPS/GIS method for travel mode detection in New York City." *Computers, Environment and Urban Systems* 36.2 (2012): 131-139.

Haghani, Ali, et al. "Data collection of freeway travel time ground truth with bluetooth sensors." *Transportation Research Record: Journal of the Transportation Research Board* 2160 (2010): 60-68.

Handy, Susan L. "Regional versus local accessibility: neo-traditional development and its implications for non-work travel." *Built Environment (1978-)* (1992): 253-267.

Handy, Susan, Xinyu Cao, and Patricia Mokhtarian. "Correlation or causality between the built environment and travel behavior? Evidence from Northern California." *Transportation Research Part D: Transport and Environment*10.6 (2005): 427-444.

Hood, Jeffrey, Elizabeth Sall, and Billy Charlton. "A GPS-based bicycle route choice model for San Francisco, California." *Transportation letters* 3.1 (2011): 63-75.

Kraft, G (1963) "Demand for intercity passenger travel in the Washington-Boston corridor." *Part V, Northeast Corridor Project, US Department of Transportation* (1963).

Manheim, Marvin L. "Practical implications of some fundamental properties of travel-demand models." *Highway Research Record* 422 (1973).

McNally, Michael G. "The four-step model." *Handbook of Transport Modelling: 2nd Edition*. Emerald Group Publishing Limited, 2007. 35-53.

Widhalm, Peter, Philippe Nitsche, and Norbert Brändie. "Transport mode detection with realistic smartphone sensor data." *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012.

Parsons, Brinckerhoff, Quade & Douglas, United States. Federal Transit Administration, & Transit Development Corporation. *Transit and Urban Form: Pt. III, A guidebook for practioners; Pt. IV, Public policy and transit-oriented development: six international case studies*. Vol. 16. National Academy Press, 1996.

Preston, Jonathan. "Demand forecasting for new local rail stations and services." *Journal of Transport Economics and Policy* (1991): 183-202.

Recker, Wilfred W., Michael G. McNally, and Gregory S. Root. "A model of complex travel behavior: Part II—An operational model." *Transportation Research Part A: General* 20.4 (1986): 319-330.

Shoup, Donald C. "The high cost of free parking." *Journal of planning education and research* 17.1 (1997): 3-20.

Spears, Steven, Marlon G. Boarnet, and Susan Handy. "Draft policy brief on the impacts of road user pricing based on a review of the empirical literature, for research on impacts of transportation and land use-related policies, California Air Resources Board." *California Air Resources Board (http://arb. ca. gov/cc/sb375/policies/policies. htm)* (2010).

Stopher, Peter R., and Stephen P. Greaves. "Household travel surveys: Where are we going?." *Transportation Research Part A: Policy and Practice* 41.5 (2007): 367-381.

Talvitie, Antti. "A direct demand model for downtown work trips." *Transportation* 2.2 (1973): 121-152.

Taylor, Brian D., and Camille NY Fink. "The factors influencing transit ridership: A review and analysis of the ridership literature." (2003).

Vickerman, Roger W. "Accessibility, attraction, and potential: a review of some concepts and their use in determining mobility." *Environment and Planning A* 6.6 (1974): 675-691.

Washington, Simon P., Matthew G. Karlaftis, and Fred Mannering. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.

Willson, Richard W., and Donald C. Shoup. "Parking subsidies and travel choices: assessing the evidence." *Transportation*17.2 (1990): 141-157.