Title

Estimating bicycle and pedestrian ridership using the Random Forest algorithm

Permalink

https://escholarship.org/uc/item/1qj4f71g

Author

Kamalapuram, Sravya

Publication Date

2022

Peer reviewed|Thesis/dissertation

Estimating bicycle and pedestrian ridership using the Random Forest algorithm

By

SRAVYA KAMALAPURAM
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Transportation Technology and Policy

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Susan L. Handy, Chair

Dillon T. Fitch

Miguel Jaller

Committee in Charge

2022

ABSTRACT

For reasons related to traffic congestion, emissions, safety, physical activity and health, there has been an increased focus on active transportation modes, including cycling and walking, by transportation planners and policymakers in the United States. In this regard, estimating bicycle and pedestrian volumes is key to evaluating transportation systems, building new infrastructure, safety studies, and understanding the impact of policy changes. Researchers have used various methods to estimate these volumes but most of the studies are limited to a single city or include study locations only in urban areas. This study contributes to the existing literature by including study locations from rural areas and using unique explanatory variables from other tools such as the Strava Fitness app and the Bicycle Network Analysis (BNA) tool from PeopleforBikes'. I built a set of models using the Random Forest algorithm to predict Annual Average Daily Bicycle Traffic (AADBT) at the street level and Annual Average Daily Pedestrian Traffic (AADPT) at the intersection level. The dependent variable in the bicycle models is the AADBT calculated using permanent counts from San Francisco and San Diego and short-term counts from Caltrans District 1 (including Del Norte, Mendocino, Humboldt, and Lake counties). The data from rural locations is limited to four counties in Northern California and thus I built separate models (urban, rural, and generalized: urban + rural) to account for the time and space limitations in the counts. The dependent variable in the pedestrian models is the AADPT calculated using annual average crossing volumes from 1308 intersections in California. Unlike the bicycle count locations, pedestrian count locations are spread across various geographies and thus I developed a generalized pedestrian model that accounts for all neighborhood types.

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Professor Susan Handy. Thank you, Susan, for your continuous support. I am inspired by your passion for bicycling and sustainable transportation. Your attention to detail and critical thinking have helped me improve my research and writing skills immensely.

I am equally thankful to my co-advisor, Dr Dillon Fitch. Dillon, without your consistent support and encouragement, this research would not be possible. I really appreciate the time you took in teaching me statistics and R programming. I am lucky to have you as my mentor. Your guidance and patience have allowed me to grow as a student, researcher, and person.

I would also like to thank Professor Miguel Jaller, who helped me complete this research. Thank you, Miguel, for your time and interest in my work. Your work in bikesharing and sustainable freight has always inspired me, and I look forward to working with you again in the future.

I would also like to thank everyone who helped in my research, directly and indirectly. Thanks to all the students, faculty, and staff in the Institute of Transportation Studies at UC Davis. Many thanks to Annemarie Schaaf for always helping me with the right information and being a pillar of support. Thank you, Matthew Favetti, for designing the amazing web tool and I will miss our Monday meetings.

Finally, thanks to my husband, Yashwanth Kalluru for believing in me and supporting me throughout. A very special thanks to my sister, parents, and parents-in-law. I love you and miss you all.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

Abbreviation	**Description**
AADBT	Annual Average Daily Bicycle Traffic
AADPT	Annual Average Daily Pedestrian Traffic
ACS	American Community Survey
API	Application Programming Interface
BNA	Bicycle Network Analysis
CHP	California Highway Patrol
CART	Classification and Regression Tree
CHTS	California Household Travel Survey
FHWA	Federal Highway Administration
GPS	Global Positioning System
IDW	Inverse Distance Weighting
LASSO	Least Absolute Shrinkage and Selection Operator
LEHD	Longitudinal Employer Household Dynamics
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NHTS	National Household Travel Survey
NOAA	National Oceanic and Atmospheric Administration
OLS	Ordinary Least Squares
OSM	Open Street Maps
RSS	Residual Sum of Squares
RMSE	Root Mean Square Error
SANDAG	San Diego Association of Governments
SFMTA	San Francisco Municipal Transportation Authority
SWITRS	Statewide Integrated Traffic Records System
TAZ	Traffic Analysis Zones

1. Introduction

Transport mode share in the United States is dominated by personal vehicles (car, truck, van). According to the American Community Survey (ACS), in 2019, over 84 percent of workers aged 16 and above commuted to work in personal vehicles. This includes individuals who are driving alone as well as those who are carpooling. Non-motorized/Active transportation modes, such as walking and bicycle, accounted for only 2.65 percent and 0.51 percent, respectively, in terms of mode share to work (US Department of Commerce). But, commuters who walk or bike to work are generally more satisfied with their travel than those who use personal vehicles (St-Louis et al., 2014). Active transportation also has the potential to reduce disease burden and carbon emissions while improving psychological well-being (Lindsay et al., 2011; Martin et al., 2014; Mueller et al., 2015). Many jurisdictions in the US are focusing on increasing walking and biking to reduce crash rates, improve air quality and public health (Hankey et al., 2017).

In this regard, reliable bicycle and pedestrian volume counts can assist in evidence-based infrastructure investments, prioritization of projects based on benefits and assessment of safety by building crash or exposure analysis models (Munira, 2017; Nordback et al., 2019). Many cities now have non-motorized traffic monitoring programs to collect bicycle and pedestrian counts (Turner et al., 2017). But cost is a major hindrance in establishing permanent counts as many communities do not have the resources available to invest in permanent counting infrastructure (Roll & Proulx, 2018). Establishing manual or automated counters across entire networks is also impractical (Miah et al., 2022). As a result, directly measured counts are often collected at a limited set of locations in a single city or a few urban areas with high walking and biking activity.

Researchers have used various approaches to estimate network-wide bicycle and pedestrian volumes such as integration of non-motorized modes into regional travel demand forecasting models based on the four-step modeling process, network simulation models and direct demand models (Aoun et al., 2015). Four step demand models are typically applied to traffic analysis zones (TAZs) which are relatively large and may not accurately capture internal bicycle or pedestrian trips (Schneider et al., 2012; Aoun et al., 2015). Network simulation tools use a representation of a network along with other data such as the street network density, block size and local attractions to determine activity levels. But simulating network wide volumes requires sophisticated software for GIS and spatial analysis (Turner et al., 2017). The most widely used approach to estimate non-motorized volumes is direct demand modeling (Pushkarev and Zupan, 1971; Domencich and McFadden, 1974). Direct demand models are typically regression models that relate demand directly to local conditions such as land-use, demographic, socio-economic, roadway and network design attributes and can be used to predict volumes at similar locations without counts (Ortuzar and Willumsen, 2011). Compared to regional travel demand models, direct demand models generate high-resolution spatial estimates of bike and pedestrian volumes (Le et al., 2018).

Direct demand modeling studies differ among themselves with respect to data collection methods, choice of explanatory variables and modeling approaches. Two key changes can be observed in recent non-motorized demand modeling research: 1) Use of crowdsourced data, 2) Employing new statistical approaches for modeling. With the penetration of Global Positioning System (GPS) enabled mobile devices like smartphones and watches, collection of travel information is no longer limited to travel surveys (Lee & Sener, 2021). Real time travel information from a large pool of the population can be collected through smartphone apps and other online methods known as crowdsourcing (Dadashova & Griffin, 2020). Crowdsourced big data sources like fitness tracking apps (Strava, MapMyRide) or Streetlight data can provide useful information for assessing physical activity (Miah et al., 2022). Crowdsourced data applications have the potential to collect continuous data with broad spatial and temporal resolution and can thus supplement permanent counts data in transportation planning (Hochmair et al., 2019; Roy et al., 2019). Crowdsourced data has been used primarily in bicycling research to estimate existing demand, predict demand at new locations, study risk exposure to crashes and exposure to air pollution (Dadashova et al., 2020; Hochmair et al., 2019; Jestico et al., 2016; Lee & Sener, 2021; Lin & Fan, 2020; Smith, 2015). Besides new data sources, new statistical approaches have been employed in non-motorized direct demand models. Traditionally linear and generalized linear models were used for estimating demand, while recent studies also use algorithms like Least Absolute Shrinkage and selection Operator (LASSO) regression and Random Forest algorithm for variable selection and modeling (Jestico et al., 2016; Dadashova et al., 2018; Nelson et al., 2021). Feature or Variable selection is a statistical method to identify the explanatory variables important in developing predictive models by removing irrelevant and redundant variables (Alsahaf et al., 2022). It is superior to intuition or correlation-based selection, especially useful when dealing with large number of explanatory variables.

This thesis is part of a larger research effort to estimate the quantitative and qualitative benefits from active transportation initiatives for Caltrans Active Transportation Program Benefit – Cost Tool. Estimating the existing bicycle or pedestrian volumes is necessary for this tool to calculate the downstream benefits from new active transportation projects such as increase in physical activity and number of bike or walk miles travelled, reduction in crashes and injuries and vehicle miles travelled (VMT). Therefore, the primary objective of this thesis is to develop direct demand bicycle and pedestrian models which can then be used to predict the annual average daily bicycle traffic (AADBT) and annual average daily pedestrian traffic (AADPT) on the California road network. Considering model performances and the primary goal of this study to predict bicycling and pedestrian volumes in California, which means ability to predict at locations not used in modeling, I chose to use the Random Forest algorithm for all the models. A Random Forest model is a collection of multiple decision trees, and each decision tree is constructed by recursively partitioning the dataset into multiple subsamples. The model can be tuned to improve prediction accuracies by altering the number of trees, depth of each tree, number of variables to be considered before deciding a split in the tree node, etc. The final prediction output is an average of estimates from each decision tree and thus is more robust to overfitting than simple decision tree models.

The current study contributes to the existing literature on bicycle and pedestrian volume estimation by including count locations from urban, suburban, and rural areas. It also discusses the challenges in using short-term counts and how they can affect generalizability of models when long-term counts are not available to build expansion factors. This study is also unique in that it incorporates explanatory variables from other bicycle and pedestrian analysis tools, such as crowdsourced data from the Strava app and network accessibility metrics from PeopleforBikes' Bicycle Network Analysis (BNA) tool (PeopleforBikes Bicycle Network Analysis) and use of a Random Forest approach to modeling. The Strava app is a social network for athletes around the world with a global user base of 95 million in 2021 (Strava, 2021). Despite its drawbacks, such as the disproportionate representation of young adults (25-35 years in age) and male riders, the volume and coverage of Strava data outweigh these disadvantages and can supplement permanent counts data that lack spatial detail (Hochmair et al., 2019; Roy et al., 2019). The use of network accessibility metrics from PeopleforBikes' Bicycle Network Analysis Tool (BNA) is another novel approach. Using stress ratings and accessibility scores of census blocks, these metrics highlight the importance of low stress connected networks on bicycle or pedestrian volumes. They are comparable to the land-use and network related variables used in similar studies but much more robust than methods which use Euclidean distance or network buffers to model non-motorized volumes. Including the network accessibility metrics as explanatory variables is also a crucial step to validate BNA tool outcomes. To understand their importance in improving the model performance, I built two sets of bicycle and pedestrian demand models, one with and one without the use of network accessibility metrics as explanatory variables. Results suggest that these metrics reduce the prediction errors and improve the model performance.

The remainder of this thesis is organized as follows. In Section 2, I present a literature review which helps demonstrate the need for bicycle and pedestrian estimates, research methods used to arrive at these estimates and how they are evolving to incorporate new data sources and statistical approaches. Section 3 discusses the data sources used in this study. Section 4 explains the methods used for modeling and performance metrics for model evaluation and Section 5 contains results and discussion. Section 6 includes conclusion, limitations of this study, and recommendations for future research.

2. Literature Review

In this literature review, I discuss the need to estimate bicycle and pedestrian volumes, the limitations of using regional travel demand models to estimate the volumes and why direct demand models are a popular approach, the common explanatory variables and statistical approaches used to develop the direct demand models and how these have evolved over time.

With the intention to promote physical activity, health and sustainable transportation, transportation planners, health advocates and policy makers are increasingly interested in providing non-motorized infrastructure and promoting walking and cycling activities in communities (Munira, 2017; Tabeshian & Kattan, 2014). As investment in infrastructure increase, so also grows the need to understand the patterns of usage or impact of these facilities (Roll & Proulx, 2018). Bicycle and pedestrian counts are thus a key performance metric for infrastructure planning, funding allocation, crash exposure studies, access, economic activity, and air quality analysis (Roll, 2018). Most methods to obtain non-motorized traffic counts use one of three approaches: direct observation through manual or automated counters, estimation of bicycle and pedestrian activity through regional travel demand models, or estimation using a direct demand model. Data collection through manual or automated counts is expensive and many communities do not have the resources available to invest in permanent counting infrastructure (Roll & Proulx, 2018). Also, installing counters on every street or intersection across the network is expensive and infeasible (J. B. Griswold et al., 2019; Miah et al., 2022), when cities want to compare different possible projects to prioritize infrastructure investments or study the impact of a policy on non-motorized traffic volumes in the entire city.

Researchers have thus tried to estimate existing bicycle and pedestrian volumes or forecast future volumes using various demand modelling approaches. Clifton et.al., (2012) provide a comprehensive overview of the regional travel demand models of metropolitan planning organizations (MPOs) that included non-motorized modes. As of 2012, 63% of the 48 largest MPOs in the US included non-motorized travel in their regional models while 47% of them distinguished between walk and bicycle modes. But traditionally regional travel demand models were more focused on predicting automobile or transit trips and thus are modeled at the level of traffic analysis zones (TAZs) or census tracts (Porter et al., 1999). TAZs or census tracts are relatively large and thus these models fail to capture intra-zonal or short trips by pedestrian or bicycle travel (Aoun et al., 2015, Griswold et. al., 2011). Furthermore, these models only capture utilitarian trips and hence cannot model bicycle or pedestrian trips for leisure or recreational purposes. Building these models also requires robust household travel survey data which several MPOs identified as a barrier to incorporating non-motorized travel.

Direct demand modelling is an alternate approach to regional travel demand models to estimate non-motorized volumes (Kuzmyak et al., 2014; Turner et.al, 2017). These are generally regression models used to estimate volumes based on local characteristics such as land-use or transportation network attributes (Le et al., 2018). They are appealing due to their simplicity in development and application and can be built based on available data. Table 1 shows some studies that built direct demand models to estimate bicycle and pedestrian volumes. Section 2.1 discusses the studies that use crowdsourced data from the Strava app in estimating bicycling volumes.

Table 1: Literature Review - Bicycle and Pedestrian demand models

Study	Location and Data collected	Explanatory Variables used		Modelling Approach	Key takeaways/ Comments
		Pedestrian	Bicycle		
Griswold et al. (2019)	Location: California Data collected: Short-term counts (count duration varying) from more than 1,200 intersections	Number of employees, population, number of street segments, walk commute mode share, number of schools; dummy variables for principal arterial, minor arterial, four-way intersection	Not applicable	Log-linear regression	Variables not selected using feature selection techniques like LASSO regression. High residuals for predicted AADPTs.
Le et al. (2018)	Location: 20 US Metropolitan Statistical Area (MSA). Data collected: 6,342 locations (12,231 bicycle count observations and 10,827 pedestrian count observations).	Household density, number of jobs, multimodal network density, walk commute mode share, density of transit stops.	Water and green space, number of jobs, proximity to university and college campuses, off-street bike facilities, multimodal network density, intersection density, bicycle commute mode share.	Stepwise linear regression	Data from urban and suburban locations included, but not from rural locations. Most bicycle and pedestrian counts focus on fall season (August-November) and methods for counting vary across count locations.
Hankey et al. (2017)	Location: Blackburg, Virginia Data collected: Pedestrian and bicycle count from 4 continuous sites and 97 short-duration sites	Sidewalk length, off-street trail length, household income, count of residential addresses within a buffer, population density, number of bus stops	Household income, population density, on-street facilities, length of major roads	Stepwise linear regression	Stratified location selection by street functional class and used a measure of bicycle trip potential (centrality) to assess weather count locations captured spatial variability of traffic in each road type

Study	Location and Data collected	Explanatory Variables used		Modelling Approach	Key takeaways/ Comments
		Pedestrian	Bicycle		
Tabeshian & Kattan (2014)	Location: City of Calgary, Canada Data collected: Pedestrian and bicycle counts from 34 intersections. 6-hour counts completed during three-time intervals in a day: 7.00-9.00, 11.00 to 13.00 and 16.00 to 18.00	1) Land use variables such as hectares of commercial space, number of schools, 2) Socio-economic variables such as total number of jobs, total number of transit users, 3) Number of bus stops, bus frequency, street length, kilometers of pathways, all calculated within buffer zones	1) Land use variables such as hectares of institutional space, low-density of residential space, commercial space 2) Number of bus stops, all calculated within buffer zones	Multiple linear regression and Poisson regression models	Poisson regression models have slightly better prediction accuracy than the linear regression models
Wang et al. (2013)	Location: Minneapolis, MN Data collected: 6 locations of which 3 are near arterial streets and 3 on trails. Hourly counts aggregated to get 24-hour daily totals	Percentage of African American residents, percentage of residents with college education, percentage of population over 64 or below 6, median household income, population density, recorded high temperature, deviation from the 30-year normal temperature, precipitation, average wind speed, weekend dummy.		Ordinary least squares (OLS) and Negative binomial regression	Negative binomial models perform better than the OLS models. Bicyclists and pedestrians are counted and modelled together. Data are available for only six locations for unequal time periods.
Schneider et al. (2012)	Location: San Francisco, CA Data collected: Manual and automated counts of pedestrians at 50 intersections	Total households, total employment, dummy variables for intersection within 0.25-mile of a university campus, signalized intersection	Not applicable	Log-linear regression	Though the overall correlation was significant, estimated and observed volumes showed notable differences (more than 50%) at most intersections.

Study	Location and Data collected	Explanatory Variables used		Modelling Approach	Key takeaways/ Comments
		Pedestrian	Bicycle		
Hankey et al. (2012)	Location: Minneapolis (MN) Data collected: 230 2-hour (16:00 to 18:00) and 43 12-hour (6.30 to 18.30) manual counts from 2007 – 2010.	1) Socio-demographic variables: percentage of non-white residents, percentage of neighborhood residents over the age of 65 or under the age of 5, percentage of neighborhood residents with a college education, median household income, average number of violent crimes per year 2) Built environment variables: population density, land mix, distance from nearest water body, distance from central business district, number of jobs accessible by transit 3) Weather: Recorded daily high temperature, recorded precipitation		Ordinary least squares (OLS) and negative binomial models	Neighborhood design and urban form play a role in explaining bicycle traffic. Bicycle traffic was higher on streets with bicycle facilities than without. Road classification, proximity to amenities, and activity centers are important explanatory variables for pedestrian traffic.
Miranda-Moreno & Fernandes (2011)	Location: Montreal, Canada Data collected: 8-hour pedestrian counts on weekdays from 1,018 signalized intersections	Population, commercial space, open space, subway, bus stations, schools, percentage of major arterials, street segments four-way intersection dummy, distance to downtown, temperature	Not applicable	Log-linear and negative binomial models	Intersections included in the study were not randomly selected. And Non-signalized intersections were not included.

From the review of bicycling and pedestrian demand modeling research, I conclude that direct demand modeling is a widely used approach to estimate bicycle and pedestrian volumes. Studies show several design differences. Data collection approaches range from short-term manual counts to continuous automated counts and were collected at signalized or unsignalized intersections or at mid-blocks of street segments. Time of day and duration of data collection also vary across studies. Study locations are usually limited to urban or suburban regions. Based on the factors that influence bicycling and pedestrian activity, the common explanatory variables to estimate volumes are related to land-use, transportation network, demographic, socio-economic and weather data.

Land use determines the location of origins of destinations. Residents of neighborhoods with higher levels of urban density, land-use mix, transit accessibility and pedestrian friendliness drive less than residents of neighborhoods with lower levels of these characteristics (Handy et al., 2005). In a study of San Francisco, California, land use diversity at the trip origin and street connectivity were associated with increased likelihood that a trip would be made on bike (Cervero & Duncan, 2003). Various studies have found positive correlations between well-connected streets and walking and bicycling volumes (Le et al., 2018; J. Roll, 2018; Schoner et al., 2014). A more connected network provides shorter routes in addition to a choice of routes (Dill & Voros, 2007). Bicycling and pedestrian volumes also depend on population and density of a neighborhood. Walking among urban residents living in high density regions is far more prevalent than among suburbanites (National Bicycling and Walking Study, 1992). Higher densities of residential and employment land uses define walkability and are associated with higher levels of walking in the neighborhoods (Huang et al., 2019; Le et al., 2018). The more people that are in an area, the greater the likelihood that someone will walk or bike through an intersection on a given day (Sanders et al., 2017).

Bicycle infrastructure is highly correlated with bicycle volumes. Off-street facilities (ex: trails, shared-use paths) had a strong association with bicycle traffic (Le et al., 2018). When controlling for weather, socio-demographics and land use mix, bicycle traffic was higher on streets with bicycle facilities than without (Hankey et al., 2012). Bike trail traffic is also positively correlated with income, neighborhood population density, percentage of neighborhood in commercial user and mean length of street segments (Lindsey et al., 2006).

Demographic factors such as race or ethnicity, income and sex can also influence bicycling and pedestrian volumes. Cycling is more popular among male, younger adults, and transit users (Moudon et al., 2005). Results from Fuller & Winters (2017) show that income inequalities are present in the availability and quality of cycling infrastructure in several Canadian cities and that higher income areas had significantly greater cycling compared to lower income areas. Significant gender and age differences exist in walking for transport and walking for recreation across neighborhoods (Ghani et al., 2016).

The influence of weather on bicycling volumes is studied by Hanson & Hanson (1977), Nankervis (1999), Dill & Voros (2007) and Miranda-Moreno & Nosal (2011). All these studies conclude that precipitation and temperature had significant effects on bicycling ridership especially for recreational rides. Hanson & Hanson (1977) reveals that a larger proportion of commute trips is done by bicycle compared to leisure/recreational trips regardless of weather conditions and according to Thomas et al. (2008), recreational bicycling is more sensitive to weather than utilitarian bicycling. Air temperature and precipitation are highly correlated with pedestrian activity and precipitation, in general lowers pedestrian volumes (Attaset et al. 2010, Aultman-Hall et al., 2009 & Runa, 2020). Temperatures over 80º F and temperatures below 50º F were associated with lower pedestrian volumes in Alameda County, California (Attaset et al., 2010).

2.1 Use of crowdsourced Strava data for bicycling volume estimation

Crowdsourcing refers to real-time data collection from a large pool of the population using internet or smart-phone applications. A popular approach in recent bicycling research is to use crowdsourced Strava data for travel demand estimation (Dadashova et al., 2020; Dadashova & Griffin, 2020; Jestico et al., 2016; Nelson et al., 2021, Kothuri et al., 2022), route choice analysis (Orellana & Guerrero, 2019; Lin & Fan, 2020), risk exposure to crashes (Sanders et al., 2017; Saha et al., 2018; Saad et al., 2019), and exposure to air pollution (Lee & Sener, 2019). In this section, I focused on reviewing the bicycling studies that use Strava data for travel demand estimation.

Jestico et al. (2016) used Strava data to predict categories of ridership as low, medium, and high for all roadways in Victoria, Canada. The Poisson regression model included five explanatory variables: Strava cyclist volumes, street slope, posted speed limit, time of the year (month) and availability of on-street parking. The authors reported that crowdsourced data can be a good proxy for estimating daily, categorical cycling volumes and that the relationship between crowdsourced cyclists and total ridership is strongest when counts are aggregated to morning and evening peak periods.

Roll (2018) developed direct-demand models using negative binomial regression for Central Lane Metropolitan Planning Organization (CLMPO). AADBT was predicted using Strava counts along with other infrastructure, network density and accessibility related variables. The study concludes that including Strava counts significantly improved the pseudo-R squared compared to models which did not include Strava counts.

Using the data from 12 cities across Texas, Dadashova et al. (2020) developed a direct-demand model for estimating AADB. Random Forest algorithm is used for variable selection and final equations to estimate AADB include the annual average daily Strava users, number of households in the census tract with income greater than $200,000, roadway functional classification and number of lanes on the roadway segment. This study notes the importance of having sites from diverse types of bicycle facilities such as urban arterials, trails etc, to improve the prediction accuracy. When AADBs were predicted for Texas using the estimated models, higher fluctuations were expected in rural areas with lower Strava use, as compared to urban areas.

In another study, Dadashova & Griffin (2020) used random parameter models to estimate statewide daily bicycle counts in Texas using Strava data. Random parameter models reduce the model errors to 29% compared to 41% in Dadashova et al. (2020) which used simply scaled Strava data. The authors suggest that using daily bicycle counts can have advantages over aggregated monthly or day of the week count models in capturing the spatial and temporal variation in bicycle count data and that the daily count estimates can be aggregated to monthly or weekly counts for use in transportation planning.

Nelson et al. (2021) built city-specific models and generalized linear models using bicycle and Strava counts from five cities across US and Canada. Acknowledging the selection bias with respect to age and gender (disproportionate representation of young adults and male riders), this study integrates Strava data with multiple data sources such as land-use, topography, demographic and socio-economic characteristics, and roadway design to generate maps of predicted Annual Average Daily Bicycling (AADB) volumes that are more representative of all ages and abilities of bicyclists. The authors conclude that income and safety are key variables critical to bicycling ridership and the models' accuracy is heavily influenced by the official count locations, especially in the city-specific models. Official counts are to be taken over the range of bicycling conditions including low ridership areas and model performance improved when streets with more diverse conditions were included.

Kothuri et al., (2022) used Strava data along with other third-party data sources such as StreetLight and Bikeshare data to employ data fusion methods to estimate annual average daily bicycle volumes. The authors conclude that Strava and StreetLight data give useful insights into bicycling activity, but they reach their full potential in predicting volumes only when combined with other context specific variables such as land-use or network characteristics. Comparing the results from Poisson regression and Random Forest, the authors also suggest that the Random Forest model is more flexible and can outperform conventional count models if sample sizes increased further.

An important concern with Strava data is the presence of sampling bias. Strava users are disproportionately individuals in the age group of 25-35 years of age and male (Jestico et al, 2016; Heesh and Lagdon, 2016; Watkins et al., 2016; Roy et al., 2019) and rode proportionately more for leisure (Garber et al., 2019). Smartphone apps like Strava also tend to under sample lower-income populations and to oversample some minority ethnicity populations (Blanc et al., 2016). Fitness app users ride more frequently and for greater weekly distances compared to non-users (Garber et al., 2019). Despite its drawbacks such as the disproportionate sampling, the volume and coverage of Strava data outweigh these disadvantages and can supplement permanent counts data that lack spatial detail (Hochmair et al., 2019).

Roy et al., (2019) compared Strava data with official bike counts from 104 locations in Maricopa County, Arizona and found that the Ordinary Least Squares Regression can account for 76% variance between the two variables. The authors used LASSO regression to identify geographical covariates such as median household income, percentage of white residents, average segment speed limit, distance to residential areas and distance to green spaces as significant variables to correct the sampling bias in Strava data. Using these variables, a Poisson regression model is built to predict bicyclist volumes in Tempe, Arizona and 86% of the segments were predicted within ±100 AADBT.

Al-Ramini et al. (2022) found a strong positive linear relationship between Strava and ground counter data justifying the use of Strava data to capture cycling trips in Omaha, Nebraska for places with no bike counts. Chen et al. (2020) found that Strava data approximately represents 2% of the total bicyclists in Portland, Eugene – Springfield regions in Oregon and a regression analysis shows positive association with bicycle counter data even controlling seasonal factors.

From the review of bicycling studies using Strava data for travel demand estimation, I conclude that despite its limitations such as the representation bias, Strava data is positively correlated with bicycle counter data and can be an important predictor of bicycling volumes when combined with other variables related to land-use, network connectivity and demographics.

3. Data

The bicycle and pedestrian demand models in this study include data from various data sources. The dependent variables are AADBT (per location per year) and AADPT (per location) calculated from bicycle and pedestrian counts. The explanatory variables are data from the Strava app, network accessibility metrics from PeopleforBikes, roadway characteristics, census tract characteristics and weather data. In this section, I provide a summary of the available data sources, steps followed to calculate AADBT and AADPT from count data and descriptive statistics of the dependent variable. The Appendix shows the detailed steps in data preparation to arrive at the final dataset for modeling by merging the AADBT and AADPT with the explanatory variables.

3.1 Dependent variable

3.1.1 AADBT from bicycle counts

The bicycle count data used in this study consists of permanent counts and short-term counts. A permanent count location in this study is one where data is collected for more than seven days in a given year (>7 days/year), whereas a short-term count location is one where data is collected for less than seven days in a given year (<= 7 days/year). Table 2 shows the description of count data and Figure 1 shows the bicycle counter locations.

The permanent count data come from locations in San Francisco and San Diego and were collected by the San Francisco Municipal Transportation Authority (SFMTA) and San Diego Association of Governments (SANDAG) respectively, using eco-counters. At hourly temporal resolution for 31 locations in San Francisco, counts were available from January 2018 to July 2019 and 6 of these locations contain bi-directional data. The hourly counts in San Francisco are aggregated to the day. This generated full-year counts for 2018 but only January to July counts for 2019.

Table 2: Summary - Bicycle Counter Locations

	San Francisco	**San Diego**	**Caltrans District 1**
Agency	SFMTA	SANDAG	Caltrans
Number of counters	31	9	74
Date range			
Start date	01/01/2018	02/07/2012	06/19/2014
End date	07/29/2019	12/31/2019	09/29/2019
Temporal Resolution	Hourly	Daily	10-15 hours per day

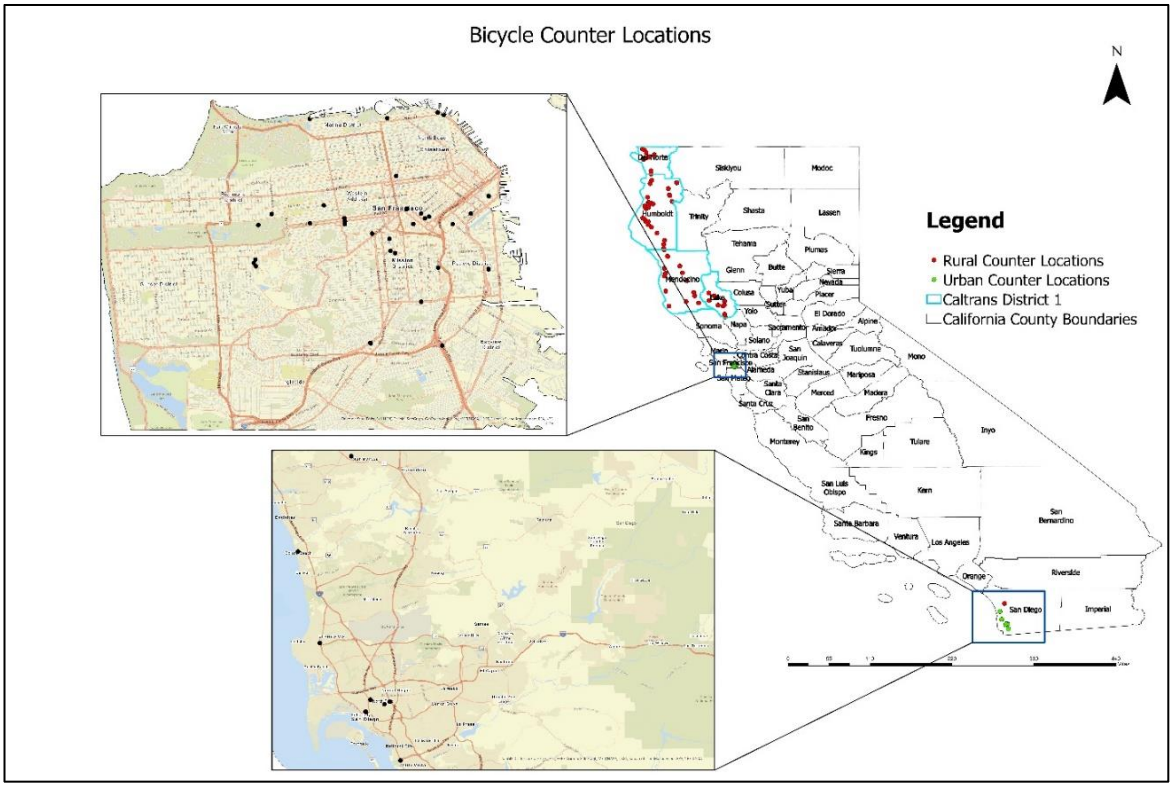


Figure 1: Bicycle Counter Locations

Daily counts were available in San Diego for a total of 9 sites from February 2012 to August 2020 and 7 locations contain bi-directional data. To match the data available in Strava Metro beginning in 2016 and to eliminate ridership disparities in 2020 owing to the COVID -19 pandemic, San Diego data is filtered to retain full-year counts from January 2016 to December 2019.

The number of missing days in full year counts from San Francisco and San Diego is near zero and thus a simple average of daily bicycle volume is used to calculate AADBT. The bicycle counts for San Francisco in 2019 were only available for 7 months (January to July). To reduce the potential error in average daily volume estimation induced by counts with missing days, the Federal Highway Administration (FHWA) recommends using a method developed by the American Association of State Highway and Transportation Officials (AASHTO) as an alternative to a simple average (FHWA, 2014). This method takes advantage of the known periodicity of traffic volumes by both month in a year and day of week. But this calculation requires data from at least one day of each day of week for each month. Given that no counts data were available in San Francisco from August 2019 to December 2019, an average of bicycle volume for all the available days is calculated and taken as AADBT. The results may vary from the value that would have been calculated had all days of data been available.

Short-term counts were obtained from Caltrans District 1 which consists of Del Norte, Humboldt, Lake, and Mendocino counties and collected by the California Department of Transportation (Caltrans). From June 2014 to September 2019, data was collected for 1-7 days at each location. The duration of counts varies by location, from 10-15 hours per day. The facility types of the locations are classified as mid-blocks, junctions, roadways, and roundabouts and have uni-directional or bi-directional counts for roadways and mid-blocks and multi-directional counts for junctions and roundabouts.

Multiple methodologies can be used to estimate AADBT from short- term counts, but most require data from permanent counters to create expansion factors (Miranda-Moreno et al., 2013; El Esawey & Mosa, 2015; El Esawey 2016). Nordback et al. (2019) suggest using week-long (seven-day) short-term counts rather than one-day (24-hour) counts, collected from June to September. If counts must be limited to 24 hours, the authors recommend taking the counts from Tuesday through Thursday to reduce the errors. Roll & Proulx (2018) propose a methodology to estimate AADBT from short-term counts without using permanent counts through a seasonal adjustment regression model. However, this approach requires collecting moderately long counts (for 2 weeks a year) to estimate AADBT with accuracy. As the data collected in Caltrans District 1 is generally limited to 1-7 days per location for 10 – 15 hours per day, extrapolating short-term counts to estimate AADBT using the available methods is not a feasible approach. Thus, short-term counts are assumed to be indicative of the daily counts and averaged. However, short-term counts could be susceptible to changes in weather or special events and may not be an accurate representation of AADBT (FHWA, 2014).

To arrive at the final dataset for modeling, the dependent variable AADBT is combined with all the explanatory variables. Since the nature of counts (permanent vs short-term) vary from San Francisco and San Diego to Caltrans District 1, and the locations in rural areas are limited to only four counties in California, I segmented the data depending on the neighborhood type of the census tract corresponding to the counter's location for estimating three models as follows:

1. Urban model: central city, urban and suburban neighborhood types

2. Rural model: rural neighborhood type
3. Generalized model: all neighborhood types

This segmentation helps in understanding the factors that affect AADBT for each neighborhood type. The results of the segmented models can also be compared with the generalized models. The neighborhood type attribute of a census tract is borrowed from Salon & Handy (2014) and is discussed in the following sections.

Table 3 shows the mean and standard deviation of AADBTs across all neighborhood types.

Table 3: Descriptive Statistics - AADBT

	Mean	**Standard Deviation**
Urban	544.96	647.92
Rural	8.53	12.79
Generalized	232.6	494.86

The standard deviation in all the three neighborhood types is much higher than the mean indicating substantial spread in the data and the presence of extreme values. This can be seen in the frequency plots in Figure 2, Figure 3, and Figure 4. San Francisco counts are available only for 2018 and 2019, which explains the missing high-volume counts in 2016 and 2017 for the urban and generalized plots. Most of the high-volume counts in 2018 and 2019 are from locations close to transit stops in downtown San Francisco. Rural data is mostly from low-volume locations. The only site with relatively high AADBT is the San Macros Trail (bi-directional) in San Diego County.

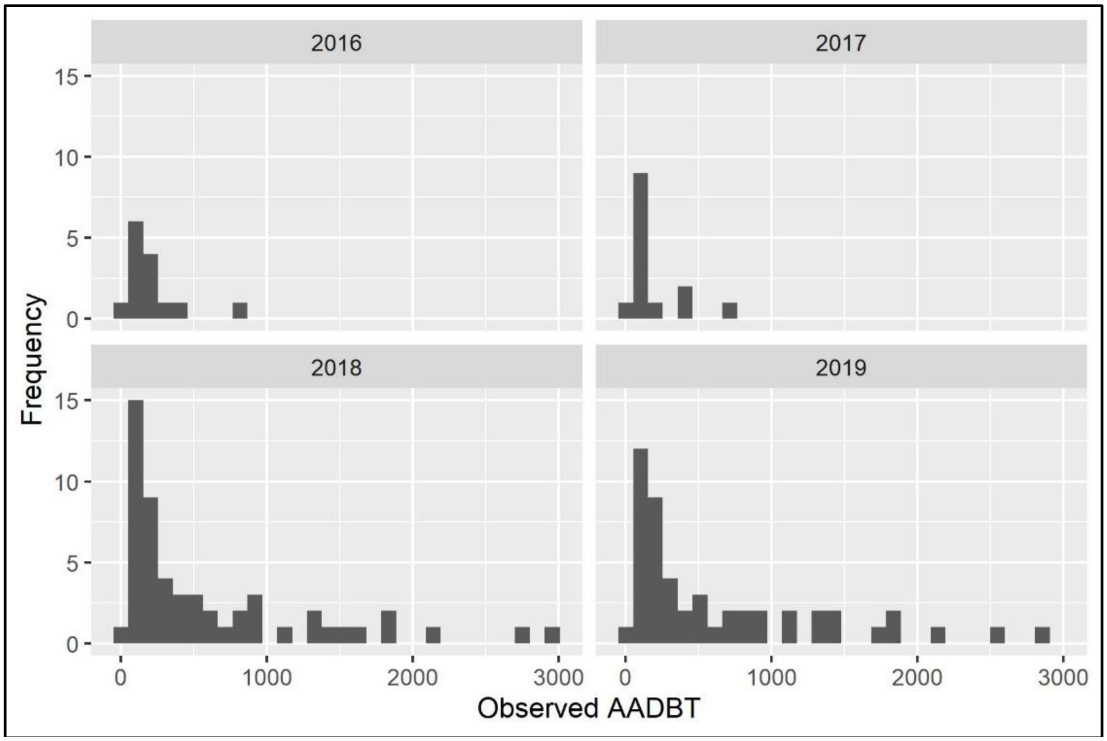


Figure 2: Observed AADBT – Urban Locations

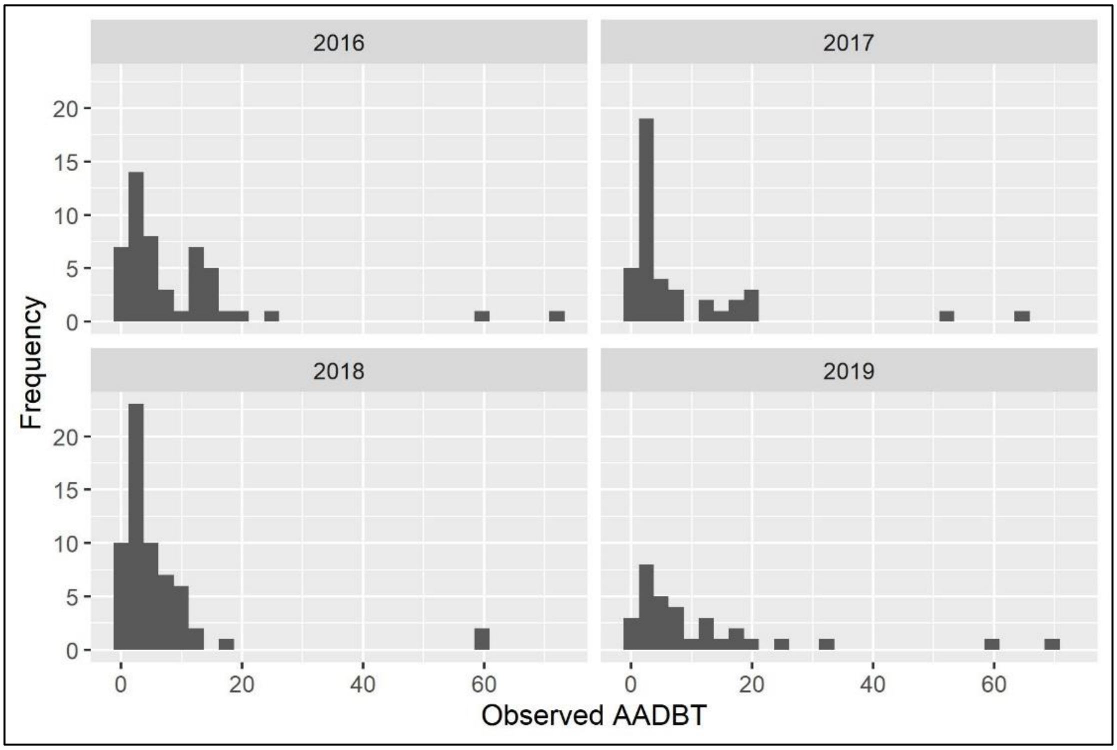


Figure 3: Observed AADBT – Rural Locations

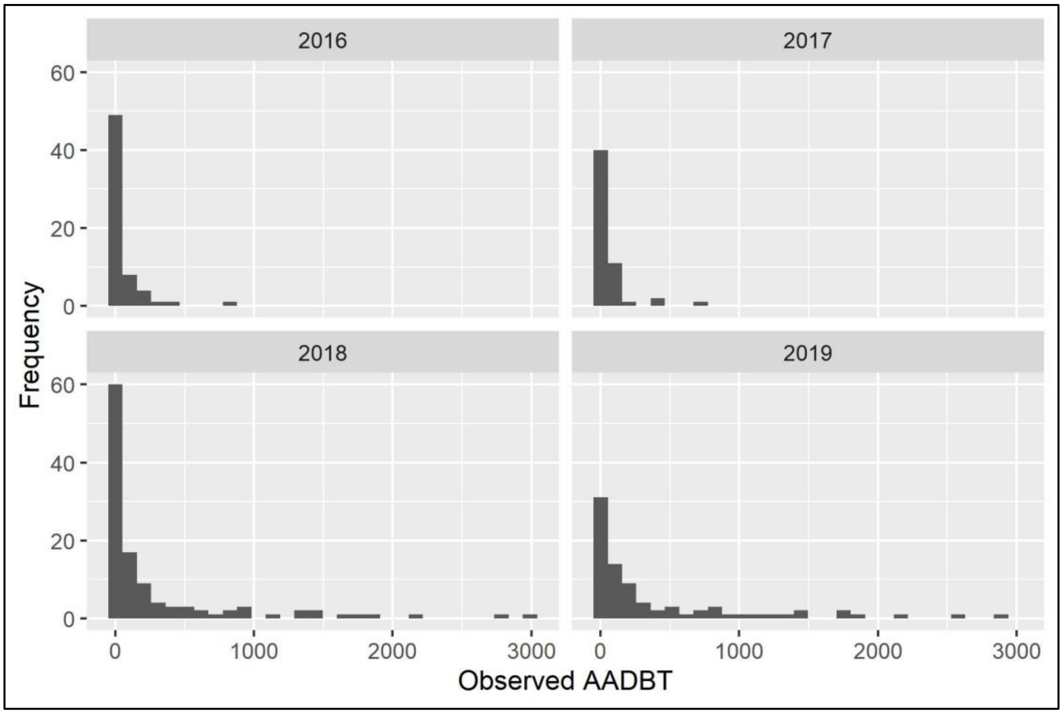


Figure 4: Observed AADBT – Generalized (All locations)

3.1.2 AADPT from pedestrian counts

The pedestrian counts are estimates of intersection crossing volumes in California by Griswold et al., (2018). The authors collected short-term counts for 1308 intersections, with count durations ranging from 1 to 86 hours and built expansion factors (hour-to-weekday, day-to-week and week-to-year) using long-term counts to estimate annual intersection pedestrain crossing volumes. These estimates were normalized to 2016 population using adjustment factors developed from American Community Survey (ACS) estimates for California.

For the current study, I borrowed the outputs from Griswold et al., (2018) which are the annual intersection pedestrian crossing volume estimates normalized to 2016 population and divided these estimates by 365 to get the annual average daily pedestrian traffic (AADPT), which is the dependent variable for pedestrian demand models. A limitation with this dependent variable is that it is an estimate of annual intersection crossing volumes from Griswold et al., (2018) and is subject to unknown error in the estimates.

Figure 5 shows the locations of pedestrian counters. In contrast to the bicycle counter locations, pedestrian counter locations are spread across the state in urban, suburban and rural count locations. Thus a generalized pedestrian model is developed using all the study locations.

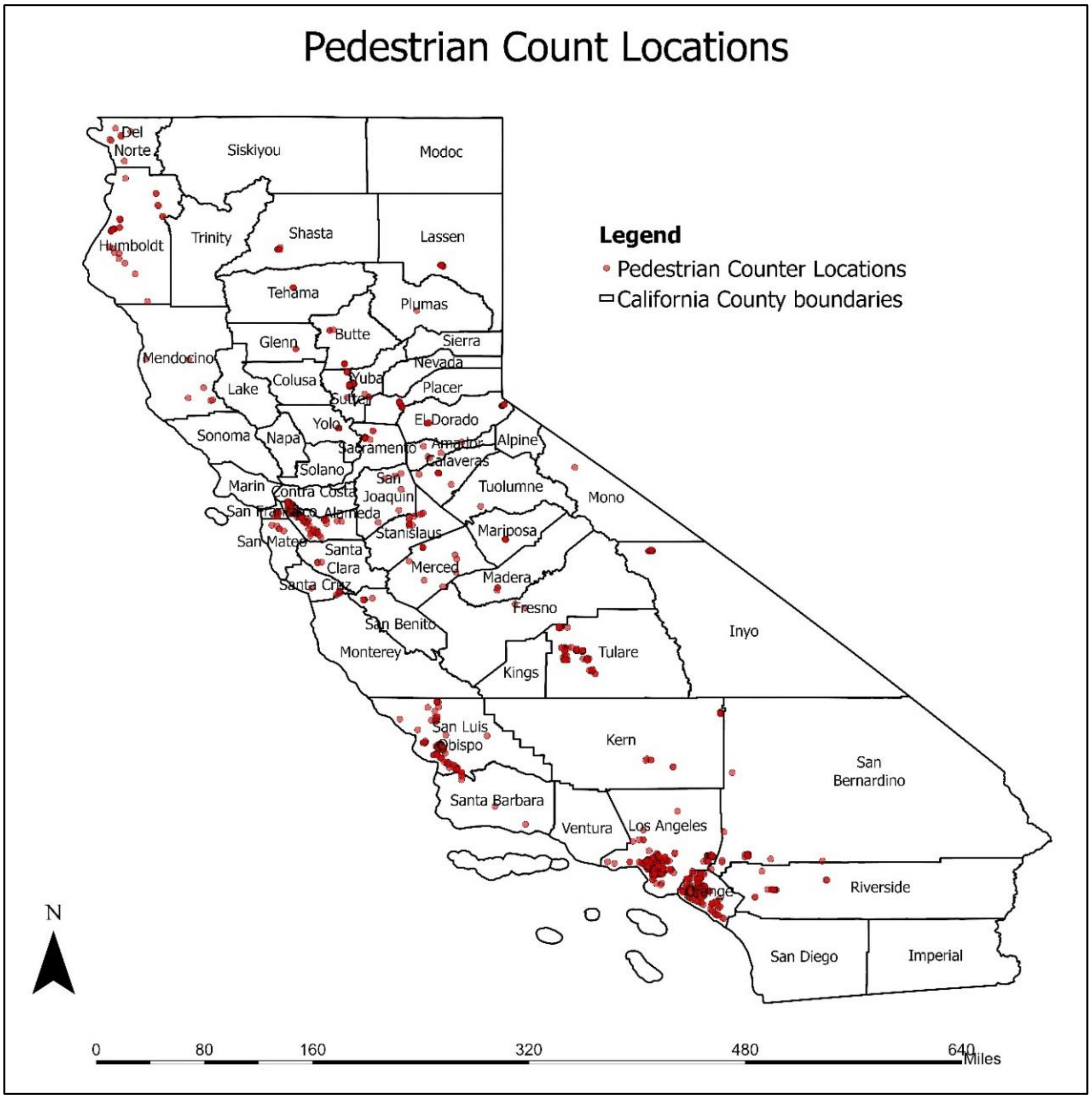


Figure 5: Pedestrian Counter Locations

Of the total counts obtained for 1308 intersections in California, 38 intersections contained null values and are omitted from the analysis. Outliers with pedestrian crossing volumes greater than 20,000 per day are removed. The resultant dataset contains 1238 intersections with a mean and standard deviation of AADPT of 1380.6 and 2623.12 respectively. Figure 6 shows the frequency distribution of AADPTs. Like the bicycle counts, the pedestrian counts are positively skewed with a long tail to the right and the standard deviation is much greater than the mean.

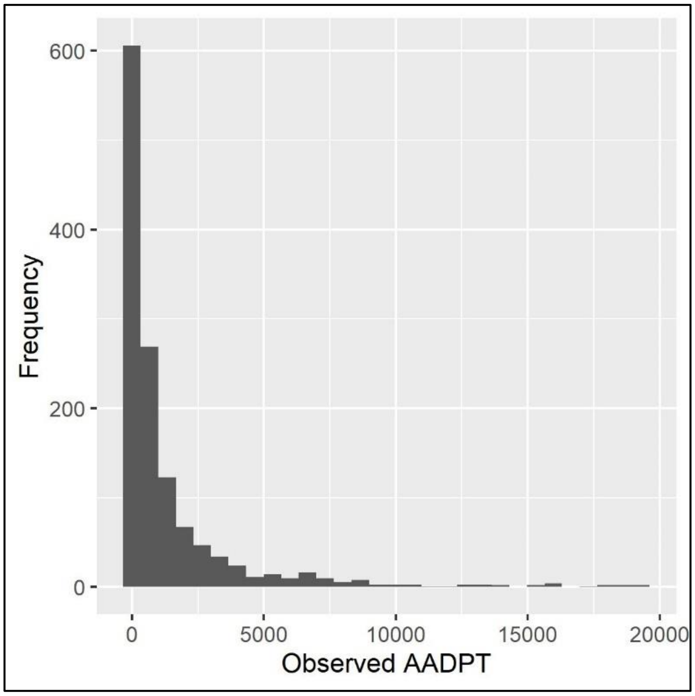


Figure 6: Observed AADPT – Frequency Distribution

3.2 Explanatory Variables

In this section, I describe the explanatory variables considered for modeling. Table 4 shows the comprehensive list of all the explanatory variables, their data sources, and data types. Exploratory data analysis and descriptive statistics for these variables are shown in the Appendix.

3.2.1 Strava Metro

Strava Metro is a cloud-based, aggregated, and anonymized data platform of the Strava Fitness app that tracks and reports crowdsourced location data for physical activities like biking, running, walking, and hiking. Data is available from 2016 to the present (the current version contains data from 2018 – 2022 only) and can be acquired for the entire road network, a specified area, or an individual road link. The activities are grouped into two categories: 1) “bike” and 2) “run, hike, walk,” and counts can be aggregated to the desired temporal resolution: hour, day, month, or year. Strava output data consists of:

- **Network shapefile**: A road link on the Strava network is termed as an “edge” with “Edge ID” being the unique identifier. The Edge IDs are unique to the specific base map used at the time of creation. The Strava base map is also different from the Open Street Map (OSM) network. Each OSM link is separated at decision points into edges. Thus, each OSM ID could have multiple Edge IDs
- **Edge volumes**: For each edge on the network, Strava provides aggregated counts related to trip activity, time of day, user gender and age group. To protect user privacy, individual demographics and details of discrete trips are not provided (Lee & Sener, 2021). Counts can be summarized to hour/day/month/year. The outputs are also bi-directional for any selected edge on the network, including “forward” and “reverse” directional counts for each variable.

According to Strava (Strava, 2020), “commuting” refers to all non-leisure trips and is derived from one of the two methods: 1) a commute tag by the Strava member, 2) an automated process that detects trip purpose by latitude and longitude of trip start and end locations within duration and distance constraints. Commute tags by Strava users are used to validate and

Table 4: List of explanatory variables considered

Category	Data Source	Variable	Data Type
Strava counts	Strava Metro	Total number of trips	Numeric
		Number of commute trips	
		Number of leisure trips	
		Number of morning trips	
		Number of evening trips	
		Total number of riders	
		Number of male riders	
		Number of female riders	
		Number of riders aged 13 - 19 years	
		Number of riders aged 20 - 34 years	
		Number of riders aged 35 - 54 years	
		Number of riders aged 55 - 64 years	
		Number of riders aged 65 and above	
Network Accessibility Metrics	Bicycle Network Analysis (BNA) tool, PeopleforBikes	Population in other census blocks that can be accessed via low stress connections from this block	Numeric
		Population in other census blocks that can be accessed via all connections from this block	
		Jobs in other census blocks that can be accessed via low stress connections from this block	
		Jobs in other census blocks that can be accessed via all connections from this block	
		K-12 schools in other census blocks that can be accessed via low stress connections from this block	
		K-12 schools in other census blocks that can be accessed via all connections from this block	
		Universities in other census blocks that can be accessed via low stress connections from this block	

Category	Data Source	Variable	Data Type
Network Accessibility Metrics	Bicycle Network Analysis (BNA) tool, PeopleforBikes	Universities in other census blocks that can be accessed via all connections from this block	Numeric
		Tech/vocational colleges in other census blocks that can be accessed via low stress connections from this block	
		Tech/vocational colleges in other census blocks that can be accessed via all connections from this block	
		Doctor offices in other census blocks that can be accessed via low stress connections from this block	
		Doctor offices in other census blocks that can be accessed via all connections from this block	
		Dentist offices in other census blocks that can be accessed via low stress connections from this block	
		Dentist offices in other census blocks that can be accessed via all connections from this block	
		Hospitals in other census blocks that can be accessed via low stress connections from this block	
		Hospitals in other census blocks that can be accessed via all connections from this block	
		Pharmacies in other census blocks that can be accessed via low stress connections from this block	
		Pharmacies in other census blocks that can be accessed via all connections from this block	
		Retail centers in other census blocks that can be accessed via low stress connections from this block	
		Retail centers in other census blocks that can be accessed via all connections from this block	
		Supermarkets and groceries in other census blocks that can be accessed via low stress connections from this block	
		Supermarkets and groceries in other census blocks that can be accessed via all connections from this block	

Category	Data Source	Variable	Data Type
Network Accessibility Metrics	Bicycle Network Analysis (BNA) tool, PeopleforBikes	Social services in other census blocks that can be accessed via low stress connections from this block	Numeric
		Social services in other census blocks that can be accessed via all connections from this block	
		Parks in other census blocks that can be accessed via low stress connections from this block	
		Parks in other census blocks that can be accessed via all connections from this block	
		Trails in other census blocks that can be accessed via low stress connections from this block	
		Trails in other census blocks that can be accessed via all connections from this block	
		Community centers in other census blocks that can be accessed via low stress connections from this block	
		Community centers in other census blocks that can be accessed via all connections from this block	
		Transit hubs in other census blocks that can be accessed via low stress connections from this block	
		Transit hubs in other census blocks that can be accessed via all connections from this block	
Roadway characteristics	Bicycle Network Analysis (BNA) tool, PeopleforBikes	Open Street Map Functional Class	Categorical Levels: path, residential, tertiary, tertiary_link, secondary, secondary_link, primary, primary_link, trunk, trunk_link, motorway, motorway_link

Category	Data Source	Variable	Data Type
Roadway characteristics	Bicycle Network Analysis (BNA) tool, PeopleforBikes	Speed limit	Numeric
		One way for car traffic	Categorical: Levels: Yes, No
		One way for bike traffic	Categorical: Levels: Yes, No
		Presence of bike infrastructure	Categorical: Levels: Yes, No
Census tract characteristics	Salon & Handy (2014)	Neighborhood type	Categorical Levels: Suburb, Urban, Rural, Central City
	2015 ACS 5 - Year Estimates	Percent of commuters using transit	Numeric
		Percent of commuters using walk	
		Percent of commuters using bike	
		Percent of male population	
		Percent of female population	
		Median household income	
		Percent of White alone population	
		Percent of Black or African American alone population	
		Percent of Asian alone population	
		Percent of American Indians alone population	
		Percent of Hispanic or Latino population	
	SWITRS dataset	Number of severe injuries	Numeric
		Number of visible injuries	
		Number of complaints of pain	
		Number of pedestrians killed	
		Number of pedestrians injured	
		Number of bicyclists killed	
		Number of bicyclists injured	

Category	Data Source	Variable	Data Type
Weather data	National Oceanic and Atmospheric Administration (NOAA)	Precipitation (mm)	Numeric
		Minimum temperature (ºC)	
		Maximum temperature (ºC)	

improve trip purpose predictions over time (Sunde, 2019). Strava counts are also rounded in an unusual way (this applies to all time intervals: hourly, daily, monthly, and yearly) (Kothuri et al., 2022).

- Trip counts <3 are not included and are set to zero.
- Trip counts >=3 are rounded to the nearest 5.

For the edges that correspond to the latitude and longitude of the counter locations, I obtained the Strava trip counts from the Strava Metro website. These counts are then combined with the bicycle counter data based on location, direction, and date. Strava Metro data for the purposes of this research is available for streets or edges only, not the intersections. Therefore, I used Strava counts only in the bicycle demand model and not in the pedestrian demand model. Table 5 shows the characteristics of Strava counts for urban, rural, and generalized study locations.

Table 5: Strava count characteristics – Bicycle data

Variable	Urban	Rural	Generalized
Commute trips	40.47%	6.76%	39.7%
Leisure trips	59.32%	93.23%	60.1%
Morning trips	47.22%	51.21%	47.32%
Evening trips	28.77%	3.9%	28.20%
Gender			
% Male riders	84.69%	92.87%	84.81%
% Female riders	13.18%	7.12%	13.09%
Age category			
13- 19 years	0.02%	0	0.02%
20- 34 years	27.62%	17.45%	27.47%
35- 54 years	46.64%	37.39%	45.51%
55- 64 years	10.61%	8.82%	10.58%
65 years and above	3.2%	3.73%	3.21%

Nearly 40% of the total Strava trips in urban locations are commute trips. This is in line with the counter placement in San Francisco and San Diego where most counters are placed in downtown areas or closer to transit stations with commute trip purposes. In rural locations however, trip purpose is highly skewed towards leisure. Looking at the gender and age statistics, more than 80% of the riders are male across all study regions, and around 40% of the riders belong to the 35-54 years age group.

To understand the correlation between the total number of Strava trips and bicycle counter data, I used Spearman's rank method which deals with skewed data (Neter et al., 1985). Spearman's correlation coefficient does not assume a linear relationship between two variables but determines the strength and direction of a monotonic relationship which is less restrictive than the linear relationship. Table 6 shows the values of Spearman correlation coefficient for all the three study areas. The values suggest a strong positive correlation in urban study areas. The correlation in rural areas is positive but shows a weak relationship.

Table 6: Spearman Correlation for Strava and bicycle counter data

Study Area	Spearman Coefficient
Urban	0.65
Rural	0.37

3.2.2 Network Accessibility metrics

The network accessibility metrics used in this study are the low-stress and high-stress connections of census blocks from PeopleForBikes' Bicycle Network Analysis (BNA) tool. The BNA tool uses data from 2010 decennial census, Open Street Map (OSM), and the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) dataset and establishes street segment stress ratings based on the Level of Traffic Stress (LTS) scale originally developed by Furth et al. (2012). The LTS 1 rating corresponds to little traffic stress demanding little attention from cyclists, which is suitable for almost all cyclists including children. The LTS 2 rating demands a little more attention than might be expected from children and therefore is suitable to most adult cyclists. The LTS 3 rating is on longer or higher speeds roads than allowed by LTS 2 but is still considered acceptably safe to most adults. LTS 4 is a level of stress beyond LTS 3. The BNA tool differentiates LTS 1 and LTS 2 from LTS 3 and LTS 4 and the low-stress bicycling in the BNA tool corresponds to LTS 1 and LTS 2. After establishing stress ratings for every street segment, the tool evaluates each census block to determine which other census blocks are within biking or walking distance from the given census block and can be accessed via the low-stress network. A low-stress route is only assumed if it does not require a detour of more than 25% when compared to a car trip.

The BNA tool then summarizes the number and types of destinations in every census block such as retail, doctors, pharmacies, trails, and transit stations. It uses this information to calculate the total number of destinations in other census blocks accessible via the low stress network from a given census block and the total number of destinations in other census blocks that are within a biking or walking distance regardless of the low stress network. These results are termed as “accessible via low stress network” and “accessible via high stress network” in the tool outputs. For each destination type, points are allocated on a scale of 0 to 100 based on the number of low-stress destinations available as well as the ratio of low-stress destinations to all destinations within biking/walking distance.

For the current study, I assumed the biking and walking distances to be 1.67 miles (2608 meters), and 0.25 miles (402 meters) respectively. According to the National Institute of Health, the mean speed of cyclists aged 14 and older was 9.7 miles per hour. PeopleforBikes BNA methodology assumes a biking distance of 1.67 miles as measured along streets or paths, the distance an average rider would travel in ten minutes biking at ten miles per hour. To determine the optimal walking distance, I evaluated the modelling results with 0.25-mile, 0.5-mile, 0.75-mile and 1-mile buffers and found that all performed equally well as predictors. Thus, I chose the shortest walking distance of 0.25-mile to reduce the computational time.

The accessibility scores for each census block are calculated using the BNA methodology described above. Then, the scores are normalized using the Inverse Distance Weighting (IDW) technique. The steps followed to calculate IDW normalized accessibility scores are:

- As streets/edges correlate to bicycle counter locations and intersections/nodes to pedestrian counter locations, I generated the latitude and longitude of intersections as the points where two streets/edges intersect on the Strava network shapefile.
- For edge or node on the Strava network that corresponds to a bicycle or pedestrian counter location, I listed the census blocks that are within the biking or walking distance respectively and calculated the distance from the counter location to these census blocks.
- Then the weighted average of the accessibility scores is computed using the formula below.

$$z = \frac{\sum_{i=1}^{n} \frac{z}{d^2}}{\sum_{i=1}^{n} \frac{1}{d^2}}$$

where,

z = *Inverse Distance Weighted (IDW) accessibility score*

n = *Number of census blocks within biking/walking distance from a census block* *d* = *Distance from Strava edge to a census block*

The final network accessibility metrics are IDW normalized accessibility scores for the census block corresponding to the counter location. These scores are used as explanatory variables in the model. Using IDW means that more weight is given to the destinations that are closer to the census block of a counter location.

3.2.3 Roadway characteristics

Roadway characteristics for the bicycle demand model are extracted from the Bicycle Network Analysis (BNA) tool outputs from PeopleforBikes. The BNA tool is sourced from the Open Street Maps (OSM) database and the required parameters such as the roadway functional classification, speed limit, presence of bike lanes is extracted using OSM tags. For the pedestrian demand model, roadway characteristics such as the binary variables for principal arterial, minor arterial, and four way intersections are borrowed from Griswold et al. (2019) .

3.2.4 Census tract characteristics

In this study, census tract characteristics are included from three data sources: 1) Salon & Handy (2014) and 2) the US Census Bureau 3) California Highway Patrol Statewide Integrated Traffic Records System (SWITRS) dataset.

For the current study, I borrowed the neighborhood type classification of a census tract from Salon & Handy (2014). Walk and bike miles travelled estimated using data from the 2009 National Household Travel Survey (NHTS) and the 2010 – 2012 California Household Travel Survey (CHTS) survey respondents were averaged for gender-age-neighborhood type categories and these averages were used to estimate distances walked and biked in the State of California based on census tract population from the 2010 decennial Census. Salon & Handy (2014) used a k-means clustering algorithm to classify each census tract into one of the four neighborhood types: central city, urban, suburban, or rural. This algorithm takes characteristics of each tract (population density, road density, local job access, regional job access, restaurants within 10-minute walk, percentage of walk/bike commuters, percent of single family detached housing, percent of old housing, percent of new housing and median house value) and organizes the tracts into groups that are internally similar. In general, downtown census tracts were classified as Central City, tracts in relatively dense areas of the city as Urban, tracts in less dense parts of the metropolitan areas as Suburb, and tracts on the outskirts of the cities as Rural. However, this neighborhood type classification is based on the 2010 decennial Census data and might be not reflective of the changing demographic and transportation scenarios of a census tract.

I incorporated other census tract characteristics such as the percentage of White alone, Asian alone, Black or African American alone, American Indian alone, Hispanic or Latino population, total male or female population, median household income, percentage of commuters who use transit or bike or walk as a means of transportation to work statistics of a census tract from 2015 American Community Survey 5-Year Estimates obtained through the US Census Bureau website. I obtained the bicycle and pedestrian crash data from the SWITRS dataset. SWITRS is a database from the California Highway Patrol that serves as a means to collect and process data gathered from a collision scene. To correspond with the date ranges of bicycle and pedestrian counts, I downloaded the bicycle crash data from 2016-2019 and pedestrian crash data for 2016 and spatially joined the latitude and longitude of crashes with the California census tract shapefile and aggregated the number of crashes in each census tract to the corresponding year.

3.2.5 Weather data

Both the bicycle and pedestrian demand models include annual average of precipitation, minimum and maximum temperature obtained from the National Oceanic and Atmospheric

Administration (NOAA) website. To get this data, I first retrieved the closest weather station to every counter location. Station IDs vary based on the required parameter (precipitation or temperature). For example, the closest station to Rose Canyon Bike Path bicycle counter in San Diego County for precipitation data is “US1CASD0060” and for the temperature data is “USC00047741”. If the data at the closest station is not available for the required year, I considered next nearby station subject to a 50km (31 mile) radius of the counter location, checked for data availability, and obtained the precipitation and temperature data.

4. Modeling

I developed a set of models to estimate the AADBT or AADPT using the Random Forest algorithm. Before deciding on using the Random Forest algorithm, I experimented with the following methods:

- Random Forest or Gradient Boosting Machines (GBM) for variable selection followed by Poisson or negative binomial regression models for parameter estimation
- Gradient Boosting Machines for variable selection and parameter estimation
- Random Forest for variable selection and parameter estimation
- Mixed effects or multi-level models

Feature or variable selection refers to techniques that identify the explanatory variables important in developing predictive models by removing irrelevant and redundant variables (Alsahaf et al., 2022). Effective feature selection can improve prediction performance by reducing over-fitting, computation and data acquisition costs (Guyon & Elisseeff, 2003; Kou et al., 2020). With fewer variables, models also become easier to interpret and work better when used with new data (Laib & Kanevski, 2019; Otchere et al., 2022). Given the large number of explanatory variables used in this study, variable selection is critical for determining a subset of variables to employ in modeling. Previous studies to predict ridership using crowdsourced data (Roy et al., 2019; Nelson et al., 2021) employ LASSO regression (Tibshirani, 1996) for variable selection. Although widely employed in variable selection, LASSO regression only models' linear correlations between variables and cannot detect non-linear dependencies (Xu et al., 2014). And regression-based methods for variable selection like LASSO or stepwise regression techniques perform better for smaller datasets. Tree-based methods like the Random Forest algorithm or GBM can discover non-linear connections between variables and perform better with larger datasets (Sanchez-Pinto et al., 2018). Using Random Forest or GBM for variable selection, I developed Poisson and Negative binomial models for parameter estimation. Compared to ordinary least squares regression (OLS), Poisson regression models are useful for count data especially when the data is highly positively skewed and non-normally distributed (Hutchinson & Holtman, 2005; Coxe et al., 2009, Hankey et al., 2012). The Poisson regression models converged when estimated with the subset of variables after variable selection, but the negative binomial models did not. As a result, I eliminated the negative binomial models and retained the Poisson regression models. I then compared the prediction errors from the Poisson regression models with models that used GBM and Random Forest for both variable selection and parameter estimation. The Random Forest model performed the best when comparing the Root Mean Squared error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

In bicycle and pedestrian modeling, study locations can be grouped into clusters based on spatial or temporal factors such as neighborhood/pathway/city or time of day/day of the week/year of data collection. Traditional linear or generalized linear regression models assume the counts at study locations to be independent of each other and ignore the relationships between counts from locations that belong to a cluster. Mixed effects /multi-level models account for this unobserved heterogeneity in the data from locations within a cluster. Types of multi-level models include random intercept models where the intercept term in the regression line is allowed to vary across clusters and random coefficients models where the intercept and the slope of the regression line are allowed to vary across clusters. Mixed effects models were used in (Dadashova & Griffin, 2020) to estimate bicycle volumes and the prediction error reduced with mixed effects models compared to traditional models. Further understanding of mixed effects models with regard to computational costs is required, especially considering the end goal of this thesis to deploy the models developed to predict AADBT or AADPT for the Caltrans Active Transportation Benefit-Cost Tool. Thus, taking into consideration the prediction errors and computational challenges, I chose the Random Forest algorithm to build models in this thesis.

4.1 Random Forest

Random Forest is an algorithm originally proposed by Breiman (2001). A Random Forest model is a collection of multiple decision trees. A decision tree is constructed by recursively partitioning the dataset into multiple subsamples. This procedure is called bootstrapping. A root node (often called first parent) is split into left and right child nodes and these nodes can be further split into left and right nodes. The bottom of the decision trees which are the terminal nodes are called leaf nodes or leaves. Figure 7 shows the graphical representation of decision trees.

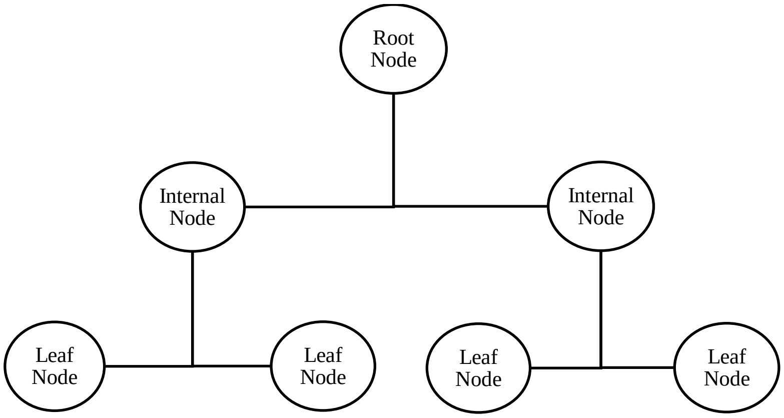


Figure 7: Graphical Representation of a Decision Tree

A drawback of decision trees is that they are prone to overfitting, which means the model fits too well to the training data and performs poorly on a new dataset. Compared to decision trees, Random Forests use a two-stage randomization procedure (X. Chen & Ishwaran, 2012). In addition to bootstrapping the dataset, the procedure introduces a second layer of randomization at the node level when growing the tree. Rather than splitting a node using all variables, Random Forest selects a random set of variables at each node of each tree and only uses these variables to find an optimal split for the node. Using information gain theory, it detects the features that provide maximum information about the regression outputs and thus important to the forest (Kothuri et al., 2020). Gain indicates a variable's relative contribution to the model. A higher value of variable importance implies it is a better feature for prediction. For Random Forest regression, gain is determined by the decrease in variance (detailed explanation is shown in the Appendix). The final Random Forest prediction is the average of predictions from individual decision trees and thus reduces overfitting. Compared to individual decision trees, Random Forests are difficult to interpret since they aggregate several decision trees. On the other hand, they typically outperform decision trees in terms of prediction (Schonlau & Zou, 2020). In comparison to decision trees, they also predict error rates more accurately. Figure 8 shows the graphical representation of a Random Forest model.

A Random Forest model is tuned using hyperparameters, to get the most accurate predictions. Hyperparameters are external to the model and their values control the learning process of the algorithm. Some hyperparameters for a Random Forest model include number of trees, maximum depth of a tree, maximum number of variables to consider before splitting each node, etc. In this study, I employed a grid search to get the number of trees and maximum number of variables to consider before splitting each node. The best combination has the lowest Root Mean Square Error and is chosen for the next steps in the modeling process.

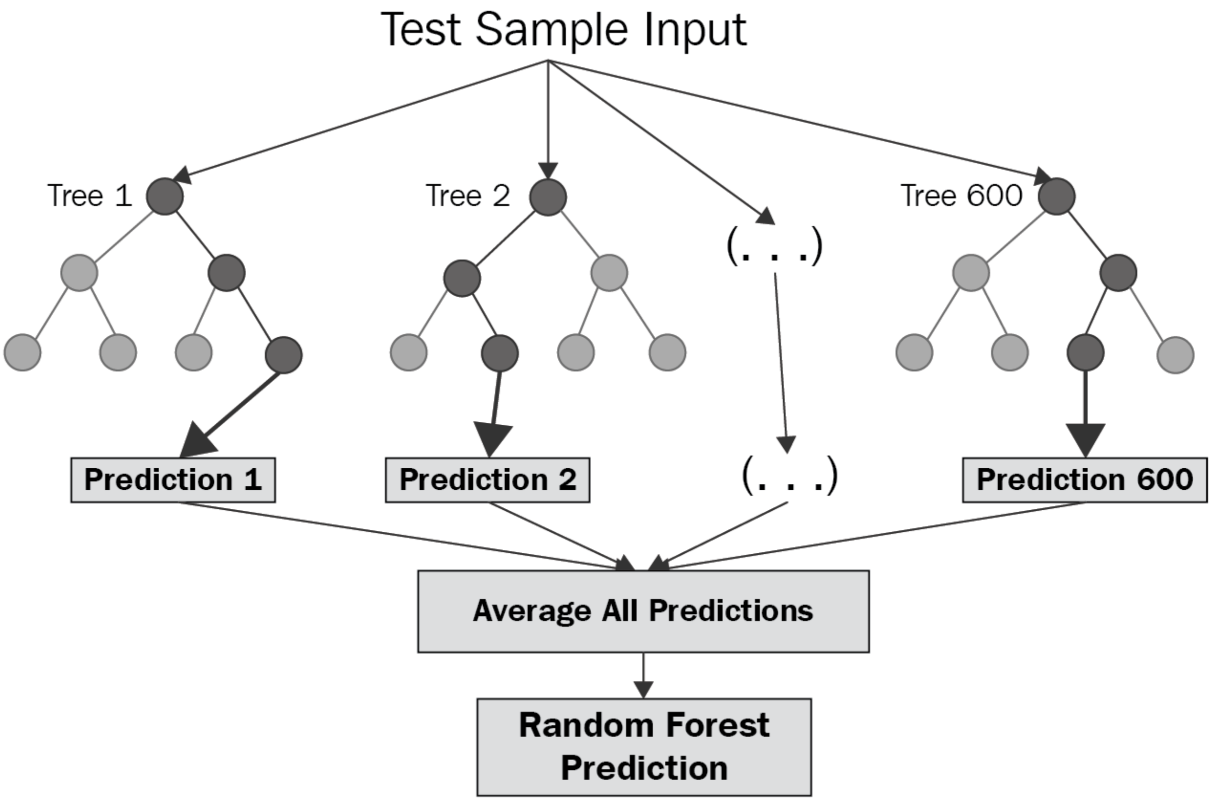


Figure 8: Graphical Representation of a Random Forest model

4.2 Model development and analysis

As discussed earlier, I segmented the bicycle demand models into Urban, Rural, or Generalized (combining urban and rural) models while pedestrian demand models are limited to generalized models estimated across all study locations. Bicycle and pedestrian demand models are further divided into two sets based on the explanatory variables used for modeling as shown in Table 7 and Table 8.

Table 7: Bicycle demand models

Study Area	Without network accessibility metrics	With network accessibility metrics
Urban	B_UM1	B_UM2
Rural	B_RM1	B_RM2
Generalized	B_GM1	B_GM2

Table 8: Pedestrian demand models

	Without network accessibility metrics	**With network accessibility metrics**
Pedestrian Demand Models	PM1	PM2

The first of bicycle and pedestrian models use roadway characteristics, census data, bicycle or pedestrian crash data from SWITRS and weather data. These data sources are free, easily accessible, and interpretable for use by transportation agencies or researchers to validate or repeat this work in different settings. The bicycle demand models also use Strava data. Strava works with urban planners, city governments and safe infrastructure advocates to understand mobility patterns, identify opportunities for active transportation investments and evaluate the impact of infrastructure changes. Though a license is required to access the Strava Metro dataset, Strava recently made it available free of charge (Strava, 2020).

The accessibility scores from the BNA tool include a complex calculation procedure and IDW normalization could be computationally challenging with large datasets. To separate data sources that require little processing like in the first set with the network accessibility metrics, I developed a second set of models that use all the explanatory variables used in the first set along with Inverse Distance Weighted network accessibility metrics from the BNA tool. Separating the models gives an opportunity to compare the prediction accuracies, make an informed decision in model selection based on the available data and computational capabilities and also validate the BNA tool outputs.

I assess the model performance to determine the best model by comparing three performance metrics: Root Mean Squared error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). RMSE is the standard deviation of the residuals (prediction errors). It measures the spread of the residuals around the best fit line. MAE is the mean of absolute errors.

It tells the average error to expect from the predicted values. MAPE expresses mean absolute percent error of each prediction. The equations used in calculation are shown below:

$RMSE = \sqrt{\frac{\sum(y_i - y_p)^2}{n}}$ …………………………………….. *(Equation 4.1)*

$MAE = \frac{|(y_i - y_p)|}{n}$………………………………………….. *(Equation 4.2)*

$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - y_p}{y_i}\right|$…………………………… *(Equation 4.3)*

where,

y_i = *actual value*

y_p = *predicted value*

n = *number of observations*

All the available count data is used for training and testing the models. Each model is estimated using a 10-fold cross validation technique by splitting the entire dataset into 10 random folds. I also calculated the prediction errors as the absolute difference between the observed and predicted values and reported the prediction errors for 25%, 50%, 75% and 99% for data points in each model.

5. Results and Discussion

In this section, I report the results of bicycle and pedestrian demand models, including the performance metrics (RMSE, MAE, MAPE), error margins of the predicted AADBT or AADPT and variable importance plots from the Random Forest algorithm.

5.1 Bicycle demand model

Table 9 shows the performance metrics from two sets of bike demand models. The RMSE, MAE and MAPE in general improve when network accessibility metrics are included as explanatory variables. Comparing all the performance metrics, the Generalized model with network accessibility metrics (B_GM2) is the best performing bicycle demand model.

Table 9: Random Forest model performance – Bicycle demand models

	Without network accessibility metrics			**With network accessibility metrics**		
	Urban (B_UM1)	**Rural (B_RM1)**	**Generalized (B_GM1)**	**Urban (B_UM2)**	**Rural (B_RM2)**	**Generalized (B_GM2)**
RMSE	322.51	5.16	203.84	265	4.96	173.04
MAE	192.26	3.84	82.99	160	3.75	71.78
MAPE	0.551	1.06	0.95	0.39	1.05	.978

For the two sets of bicycle demand models, Table 10 shows how well bicycling volumes were predicted by showing an absolute difference between the observed and predicted AADBT. For example, for urban areas without using network accessibility metrics, 25% of the data points were predicted within ±27 of the observed AADBT, 50% of the data points were predicted within ±61 of the observed AADBT, 75% of the data points were predicted within ±246 of the observed AADBT and 99% of the data points were predicted within ±1600 of the observed AADBT. Likewise, comparing other results from Table 10, predicted errors decline with the use of network accessibility metrics for all the three study areas.

Table 10: Error margins of predicted AADBT

Study Area	**Number of data points***	**% of data points predicted**	**Error margins of predicted AADBT**	
			Without network accessibility metrics	**With network accessibility metrics**
Urban	132	25%	±27	±21
		50%	±61	±57
		75%	±246	±190
		99%	±1600	±1000
Rural	184	25%	±2	±1
		50%	±5	±3
		75%	±7	±5
		99%	±16	±16
Generalized	316	25%	±2	±2
		50%	±7	±7
		75%	±46	±42
		99%	±957	±783

**Number of data points denotes AADBTs per location per year.*

Figure 9 to Figure 11 show a comparison of variable importance plots for the two sets of bike Demand models from the Random Forest algorithm. These plots show the top 25 variables of importance. The variable importance plots are only indicative of the variables selected for the Random Forest model and the variable importance values or order cannot be used to define the relationship of explanatory variables with the dependent variable or for calculation of odds like in traditional models. Figure 12 shows the observed vs predicted AADBTs for Urban, Rural and Generalized bicycle demand models with the network accessibility metrics.

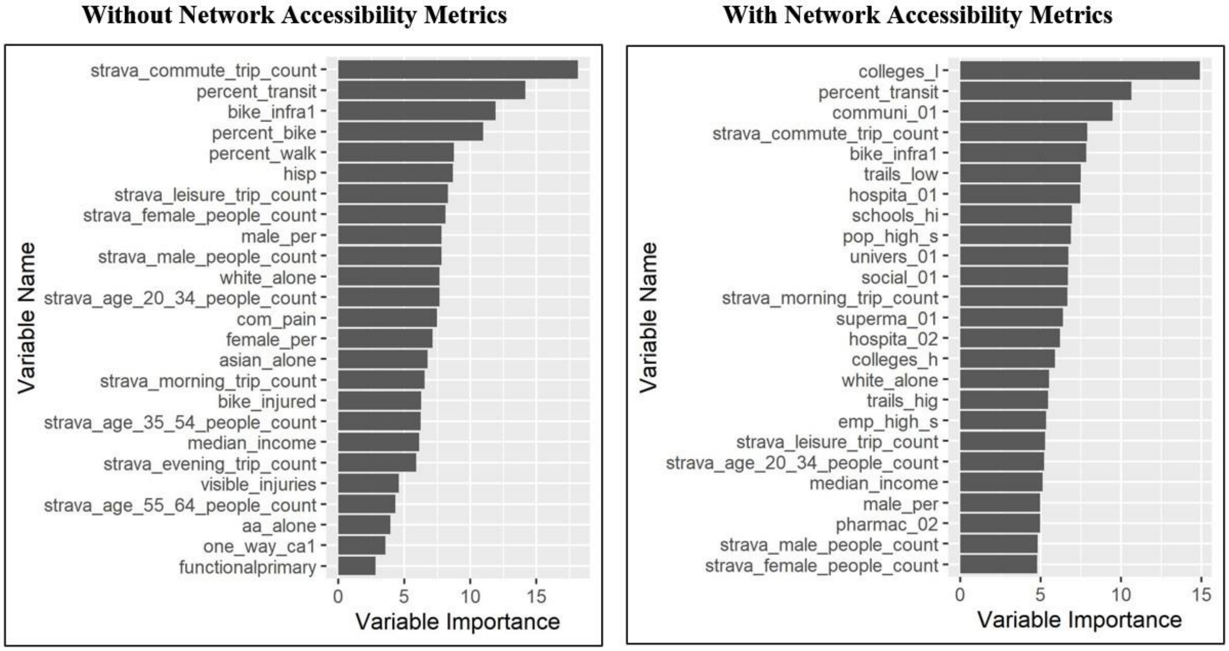


Figure 9: Variable Importance – Urban bicycle demand models

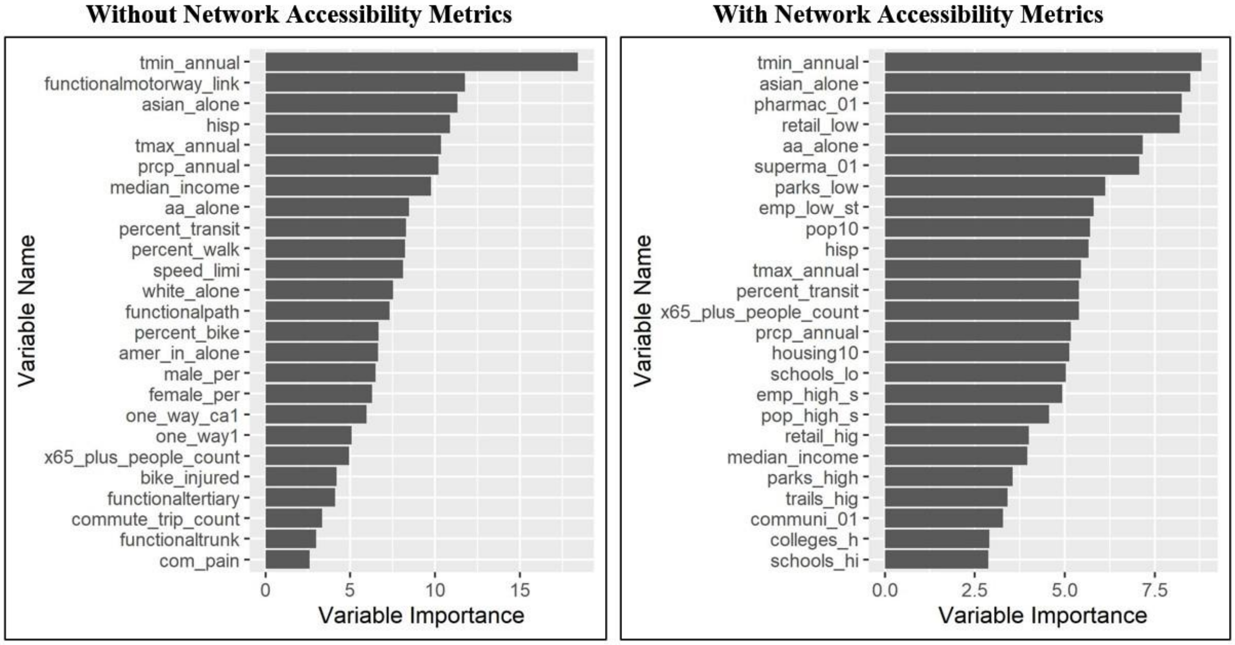
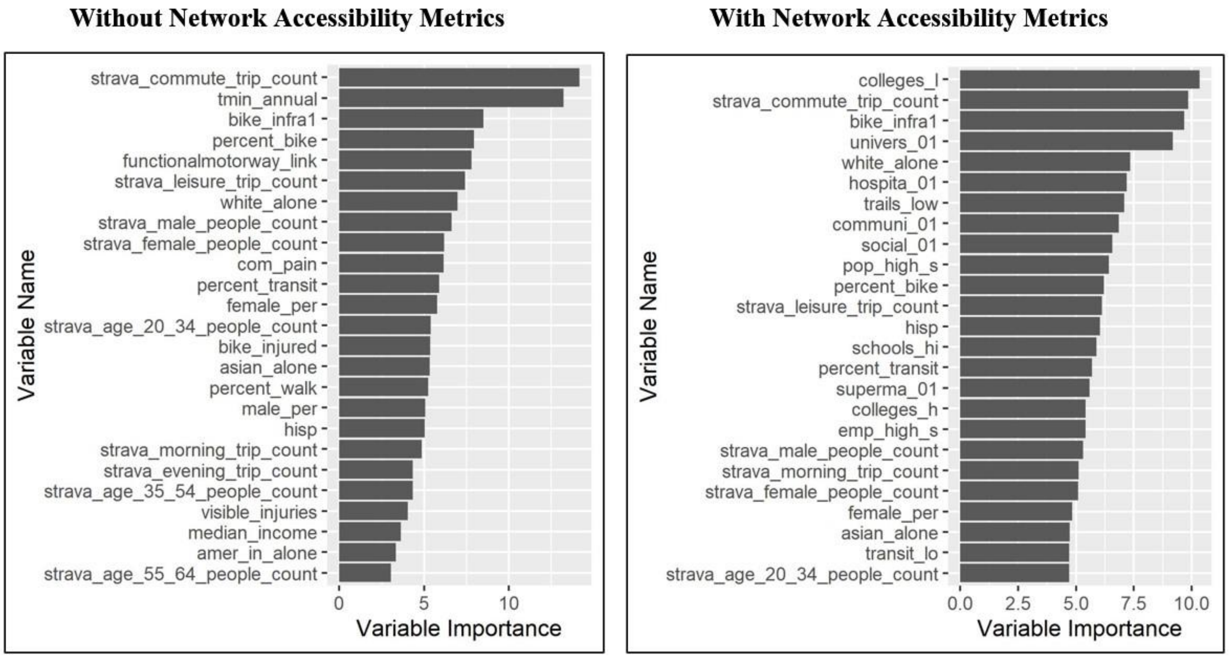


Figure 10: Variable Importance –Rural bicycle demand models

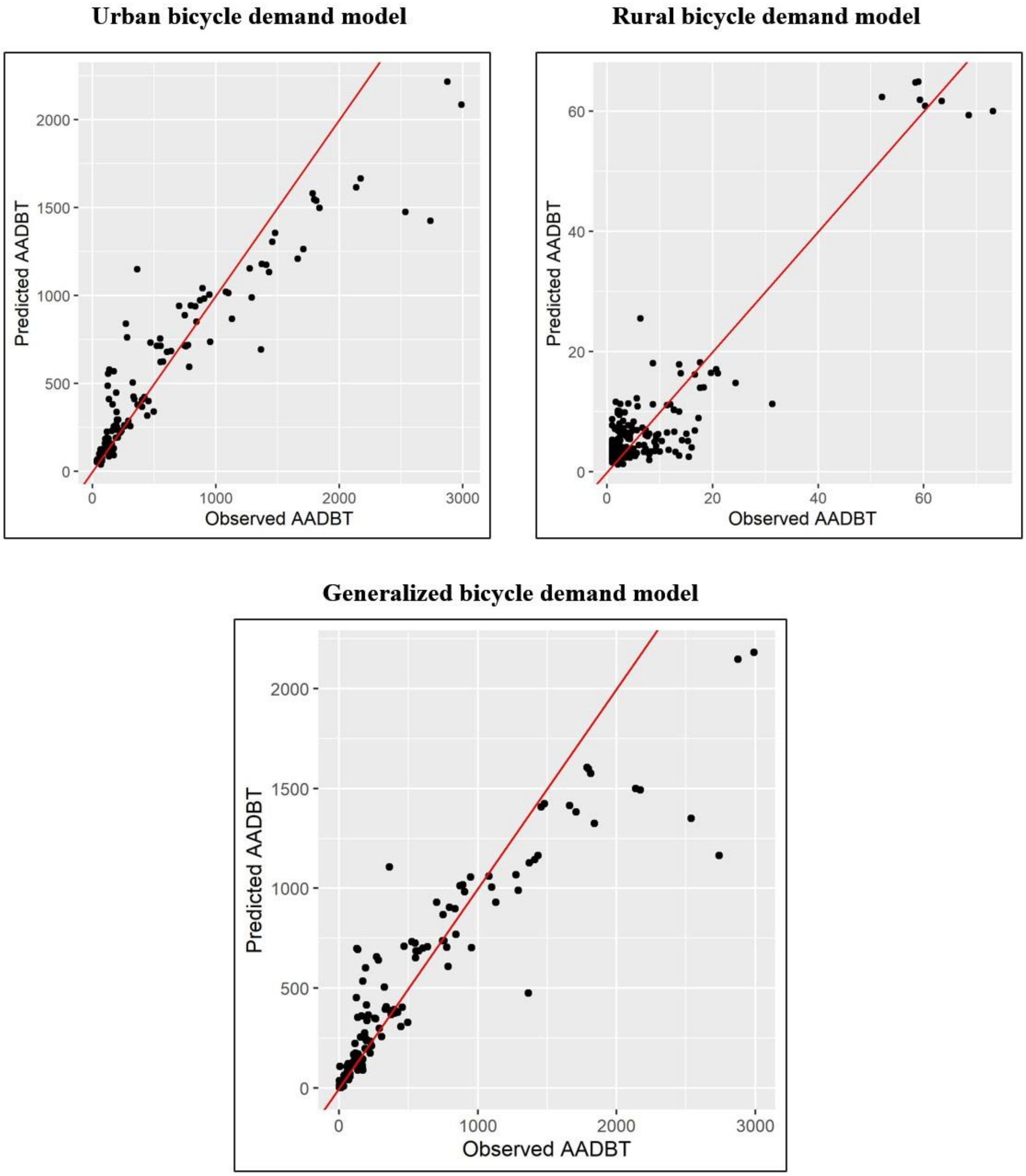

Without Network Accessibility Metrics

With Network Accessibility Metrics

Figure 11: Variable Importance – Generalized bicycle demand models

Compared to the bicycle demand models without the network accessibility metrics, the models with network accessibility metrics in general perform better. The performance metrics (RMSE, MAE and MAPE) improve across all study areas when network accessibility metrics are added to the set of explanatory variables. Similarly, the absolute difference between observed and predicted AADBT decreases with network accessibility metrics.

The variable importance plots for models with and without the network accessibility metrics show a similar pattern for all three study areas. Without network accessibility metrics, roadway functional classification and bicycle crash related variables feature among the top 25 variables of importance. When network accessibility metrics are included, the most important predictors are low-stress connections to destinations such as colleges, universities, pharmacies, and retail. Strava counts are strong predictors when counts from urban areas are included i.e., in Urban only and Generalized bicycle models, with or without the network accessibility metrics.



Urban bicycle demand model

Rural bicycle demand model

Generalized bicycle demand model

Figure 12: Observed vs Predicted AADBT for bicycle demand models with the network accessibility metrics

The San Francisco and San Diego bicycle counters are mostly placed in downtown areas or near transit stations and have bike infrastructure in at least one direction. Presence of bike infrastructure and percentage of commuters in a census tract who use transit or bike as a means of transportation to work are the most important predictor variables in the Urban bicycle models. This suggests an association between infrastructure and bicycling in urban areas and that bicycling could be complementing transit trips. This observation is in line with findings from Dill & Carr (2003) and Winters et al. (2010), who suggest an association between infrastructure and bicycling at the city level. Presence of bike infrastructure is not a strong predictor variable in rural areas because of the 74 count locations in rural areas, only 3 locations had presence of some bike infrastructure including one location from the San Marcos Inland Rail Trail in San Diego County. Bicyclists prefer bicycle facilities when they are available, which are not generally present in rural areas (Griffin & Jiao, 2015).

Number of commute trips from the Strava app is another key predictor in both sets of Urban and Generalized models. Strava commute trips were used in similar studies that estimate bicycling volumes using crowdsourced data (Kothuri et al., 2022.; Nelson et al., 2021).The presence of Strava commute trips as one of the top explanatory variables suggests that Strava counts are not limited to recreational riders and are associated with bicycling trip purpose at the counter location. Although Strava represents only a small proportion of total bicyclists, it can be used an important predictor of bicycling volumes, especially in urban areas. Other Strava variables such as the number of male and female riders, number of leisure trips and number of riders in the age group of 25-34 years are also good predictors in urban and generalized contexts with or without the network accessibility metrics.

Strava counts rarely feature among the top 25 variables in the Rural bicycle demand models. Apart from the San Marcos Inland Rail Trail in San Diego County, less than 10 locations have recorded Strava counts. Bicycle counter data for rural areas used in this study is limited to short-term counts and the average of the counts for all the available days is assumed to be AADBT. To match the bicycle counter data, Strava data is retrieved on days only when bicycle counter data is available. These counts are not a true representation of annual average Strava ridership in rural areas. Taking annually aggregated Strava counts when bicycle counter data is available for only 1-7 days in a year for a location, would be an over representation of Strava counts for the available bicycle counter data and often Strava counts exceeded the AADBT of a location. Hence the influence of Strava counts on estimating bicycling volumes in rural areas is inconclusive from the current results. The results highlight the importance of official bicycle count availability and completeness to use crowdsourced Strava data or other data sources with high temporal coverage to estimate bicycling volumes. Nelson (2021) also discussed the importance of official count data to build a generalized bike demand model. The models' accuracy was heavily influenced by the number of official count locations and how representative the official counts were of the full range of conditions on the bicycling network.

Weather has a higher influence on counts in rural areas compared to urban areas. Minimum temperature is the most predictor variable in both sets of Rural bike demand models. Maximum temperature and precipitation also feature among the top 25 variables. According to the FHWA (2020), only 7 percent of non-motorized trips in rural and small-town regions are work related, compared to 9 percent in urban regions. For the count locations in this study, however, Strava trip purpose statistics shown in Table 5 suggest that nearly 90% of the trips in rural areas were leisure/recreational. This explains the higher influence of temperature and precipitation in Rural bicycle demand models compared to Urban/Generalized bicycle demand models.

Network accessibility metrics especially low-stress connections to destinations such as colleges, universities, pharmacies, retail, supermarkets, and parks feature among the top variables of importance and improve model performance when included. This highlights the importance of low stress connected network and access to destinations for bicycling volumes. Most studies that estimate bicycling volumes use network characteristics as explanatory variables. Roll (2018) found a strong positive correlation of bicycle counts with intersection density and student population accessibility and a negative correlation with local streets density. Le et.al (2018) also found a positive relationship between multi-modal network density and bicycling volumes in 20 US Metropolitan Areas. Nelson et.al (2021) use distance to green spaces, residential areas, bike parking, educational institutions, and seashore in estimating bicycling volumes. These distances were found to be significant predictors in only three of the five city-specific models and do not feature in the generalized bike demand model which pools data from all the cities. On the contrary, Kothuri et.al (2022) found that distance metrics (distance to water body, industrial areas, park, Central Business Districts, and forest) and area weighted metrics calculated by using half mile and one-mile buffers such as number of jobs, number of students, length of roadway segments around a count station as important predictors to estimate AADBT when pooled data from many cities across the US was used. The findings of the current study with respect to network accessibility metrics are thus in line with previous research and indicate the importance of well-connected grid networks for bicycling volumes (Schoner et al., 2014). The use of network accessibility metrics, particularly low-stress access to destinations as explanatory variables to estimate bicycling volumes is a novel approach used in this study. This methodology is an improvement over previous methodologies that use network connectivity or land use metrics as a proxy to access to destinations. The results validate the BNA tool outputs for bicycling mode. Use of BNA tool outputs also automates the process of accessibility metrics calculation for different types of destinations and saves computational time and burden compared to network and land-use data acquisition using GIS tools or Open Street Maps as in other studies.

Among the race or ethnicity related variables, the percentage of White alone and Hispanic or Latino population of a census tract are important predictors in the Urban and Generalized models. According to Buehler (2012), Whites were associated with 3.43 times greater odds to cycle to work than non-whites. Non-whites were considerably less likely to use non-motorized modes for travel to work (Plaut, 2005). This result could also be because of disparities in presence of bike infrastructure. A case study in Chicago by Prelog (2015) shows that the presence of bike infrastructure was correlated with wealthier, whiter neighborhoods. However, a nationwide survey of 9,616 people ages 16 years and older found that Hispanics were most likely to have cycled within the past 30 days, followed by non-Hispanic whites, individuals of other races and Blacks (National Survey of Bicyclist and Pedestrian Attitudes and Behavior, 2002).

A surprising result is the high variable importance of percentage of Asian alone, Hispanic or Latino, and Black or African American alone population on bicycling volumes in rural areas. The mean values of percentage of Asian alone, Hispanic or Latino, and Black or African American alone population at the census tract level in rural areas are 1.9%, 16.2% and 0.88% respectively, compared to the mean of percentage of White alone population which is 79.2% (shown in Appendix). To cross-verify these results, I checked the correlation of these variables with the observed AADBT in rural areas. Percentage of Asian alone, Hispanic or Latino and Black or African American alone population have a correlation of 0.84, 0.659 and 0.27

respectively while the percentage of White alone population has a correlation of -0.17. Further investigation is required into these results as previous research on race or ethnicity factors influencing bicycling in rural areas is limited. Official counts from rural areas used in this study are short-term counts from only four counties in Northern California affecting the generalizability of the results.

Roadway characteristics and bicycle crash data feature in the top 25 variables of importance only in models without the network accessibility metrics. Motorway, path (cycleways) and tertiary roads in rural areas and speed limit in urban and rural areas are key to bicycling volumes. According to the OSM functional classification, a motorway is a highway and motorway links are link roads/ramps to/from a motorway. The high influence of motorway links on bicycling volumes in rural areas could be because of the following reasons: 1) Counter placement: Counters in Caltrans District 1 are placed near on or off ramps to US Highway 101 and other State Routes such as 20, 36, 197, 200, etc. in Northern California. Hence bicycling counts could be biased towards motorway links than those on other roadway functional classifications, 2) Bike infrastructure in rural areas is meagre as discussed earlier and bicyclists are forced to use motorways or motorway links and 3) Leisure or recreational bicyclists seek roadways that are uninterrupted by stop lights (Griffin & Jiao, 2015). Number of bicyclists killed, number of complaints of pain and visible injuries are other important variables in urban and rural models without network accessibility metrics. Safety metrics are surrogates for bicycling activity. When the network accessibility metrics are included, they better account for bicycling activity in general.

5.2 Pedestrian demand model

Table 11 shows the performance metrics for two sets of pedestrian demand models.

Table 11: Random Forest model performance - Pedestrian demand models

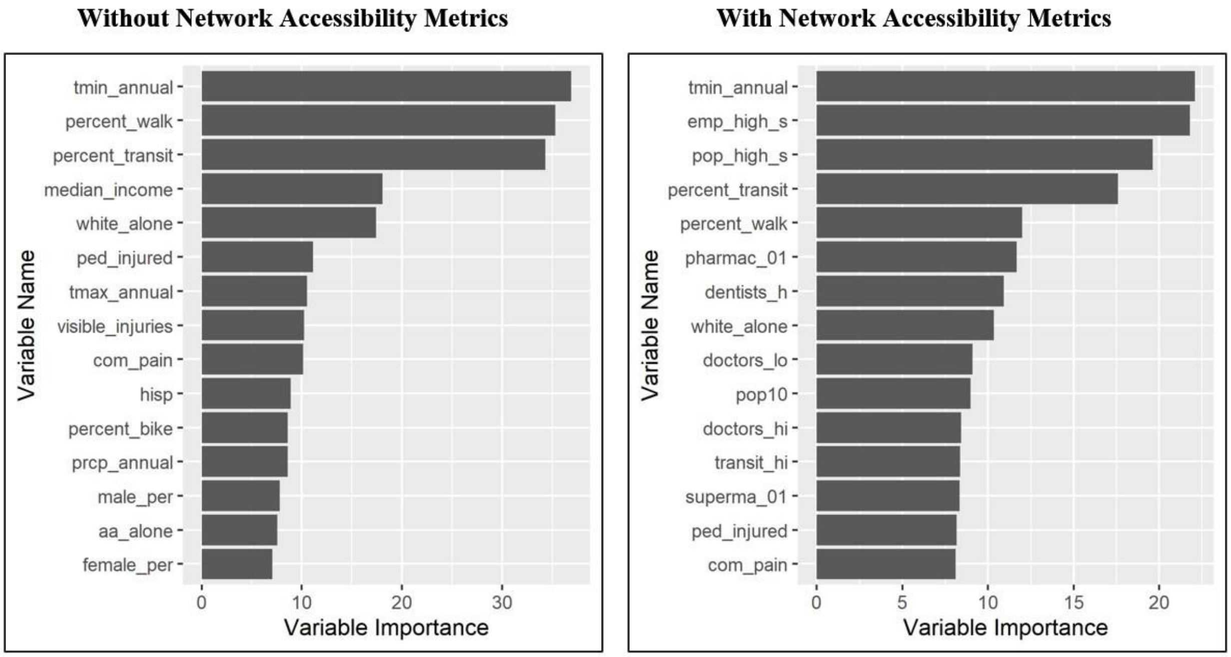
	Without network accessibility metrics (PM_1)	With network accessibility metrics (PM_2)
RMSE	1881.13	1647.58
MAE	973.03	884.11
MAPE	8.68	7.69

PM_2 (with the network accessibility metrics) is the best performing pedestrian demand model when comparing all the model performance metrics. Table 12 shows the error margins of predicted AADPT when compared to the observed values for both the pedestrian demand models. The error margins decrease when the network accessibility metrics are included.

Table 12: Error margins of predicted AADPT

Model	Number of locations	% of locations predicted	Error margins of predicted AADPT
PM_1	1238	25%	±141
		50%	±386
		75%	±1140
		99%	±7910
PM_2	1238	25%	±137
		50%	±367
		75%	±1000
		99%	±7380

Figure 13 shows the variable importance plots for the pedestrian demand models and Figure 14 shows the observed vs predicted plot for AADPT when using the network accessibility metrics (PM_2).



Without Network Accessibility Metrics

With Network Accessibility Metrics

Figure 13: Variable Importance plots for Pedestrian demand models

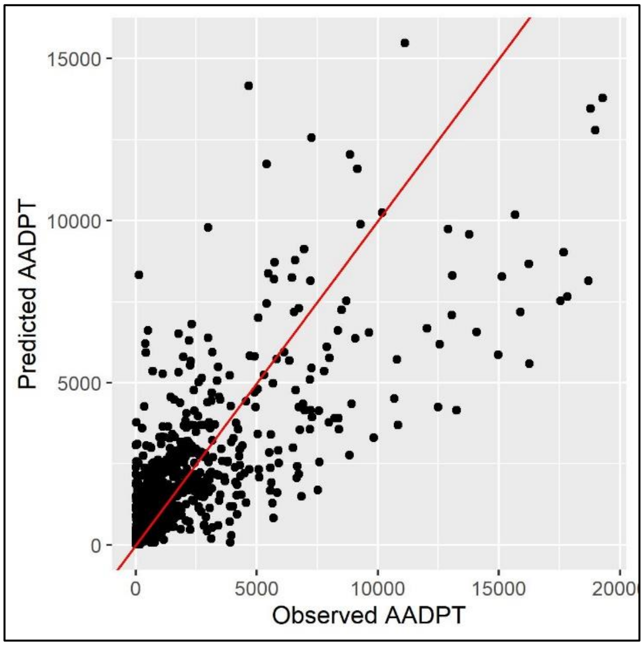


Figure 14: Observed vs Predicted AADPT for pedestrian demand model with network accessibility metrics

From the results, it can be understood that the minimum temperature is an important predictor of pedestrian volumes. Maximum temperature and precipitation also feature among the top 15 variables of importance, when network accessibility metrics are not included. Miranda-Moreno & Fernandes (2011) also included precipitation and temperature in their final models to estimate pedestrian activity at signalized intersections. Other top variables of importance in both the models, with or without the network accessibility metrics are the percentage of commuters who use transit and walk as a means of transportation to work. The results indicate that walk could be an important mode for first and last mile connections to transit.

When network accessibility metrics are included, access to employment and population also feature among the top variables of importance. Schneider et al. (2009) used total population within a 0.5-mile radius, number of jobs within a 0.25-mile radius, number of commercial retail properties within a 0.25-mile radius and presence of regional transit stations within a 0.1-mile radius of an intersection to estimate pedestrian intersection crossing volumes in Alameda County, California. Griswold et al. (2019) included the number of employees, population and walk commute mode share calculated using 0.25-mile and 0.5-mile buffers as significant explanatory variables in their final pedestrian exposure model. Sanders et al. (2017) included the number of households and commercial properties within 0.25 miles of an intersection and presence of a university within 0.25 miles of an intersection as predictor variables in the pedestrian volume estimation model for Seattle, Washington. The current study, however, includes the inverse distance weighted network accessibility metrics which highlight the importance of a well-connected network on pedestrian and is an improvement over the previous studies which use counts of destinations from network buffers.

Similar to the current study, Griswold et al. (2019) built a pedestrian exposure model using the annual pedestrian crossing volume estimates from Griswold et al. (2018). However, the dependent variable is the annual pedestrian crossing volume estimate as opposed to AADPT in this study. To compare the results of Griswold et al. (2019) with the pedestrian demand models in this study, I calculated the coefficient of variation of the predicted AADPT from the best performing pedestrian demand model (PM_2). The coefficient of variation for PM_2 is 1.32 which indicates 32% error relative to the mean estimate, while the corresponding value in Griswold et al. (2019) is 1.5. This indicates a decline in error relative to the mean in the current study compared to Griswold et al. (2019) using the same estimates from Griswold et al. (2018). Use of network accessibility metrics, census tract characteristics, weather data and pedestrian crash data from SWITRS is likely to have improved the model performance in the current study. The key difference of this study from Griswold et al (2019) is the inclusion of roadway characteristics (principal arterial, minor arterial and four-way intersection categorial variables) in the final pedestrian exposure model, which are assigned lower variable importance from the Random Forest algorithm (as shown in Figure 13). Griswold et al (2019) used collinearity or correlation based variable selection as opposed to Random Forest variable selection in the current study. It is not practical to test all the possible combinations when the number of variables is huge, and this could cause exclusion of variables which are good predictors of the dependent variable. The Random Forest algorithm considers non-linear relationships between variables and automates the process of variable selection, which is practical in use and avoids exclusion of important predictors in the models.

Compared to the bicycle demand models, two key differences can be observed in the pedestrian demand models. Bicycle demand models have low stress connections from network accessibility metrics as important predictors while the pedestrian demand models have high stress connections as the top predictors and crashes have a higher influence on pedestrian volumes compared to bicycle volumes. Greater variable importance to high-stress connections in pedestrian demand models is counter intuitive considering walking trips are in general shorter than the bicycling trips. This is probably because the Level of Stress ratings by Furth et al. (2012) and the BNA tool were designed for bicycling comfort and not walking. The high stress connections in the BNA tool outputs include all the connections or destinations that can be accessed from a given census block which means that walking or pedestrian activity is more influenced by network access rather than the bicycling comfort level.

6. Conclusion and Limitations

In this study, I developed two sets of demand models to estimate bicycle and pedestrian volumes using the Random Forest algorithm. Through the results, it can be concluded that network accessibility metrics from the BNA tool are strong predictors of bicycle and pedestrian volumes and when included, they improve the overall model performance. Crowdsourced data from the Strava app is also a good predictor of bicycling volumes in urban areas and the spatial and temporal coverage of crowdsourced data can be leveraged to predict bicycling volumes in locations where counter data is sparse. Strava data has also evolved over the years to reduce the sampling bias in trip purposes and the percentage of commute or leisure trips on Strava can change based on the location. The number of commute trips from Strava is a top predictor variable when estimating bicycling volumes in urban locations. The influence of weather on bicycling is higher in rural areas compared to urban areas. In the pedestrian demand model, connections to employment, population and pharmacies are top predictors of intersection volumes, along with percentage of commuters in a census tract who use transit and walk as a means of transportation to work. Crash related variables such as number of visible injuries and number of complaints of pain have greater influence on pedestrian volumes when compared to bicycle volumes.

There are certain limitations in this study which need to be addressed in the future. A major limitation is the availability of official count data. Urban bicycle counters are located on facilities where bicycle traffic is expected to be higher such as downtown streets or bike trails or near transit stations in San Francisco and San Diego. However, most of the street segments in the US have low bicycle volumes (Kothuri et.al., 2022) and count sites are not randomly selected (Le et al., 2018). Placing permanent counters in low bicycle volume streets is critical for data completeness and variability when studying network-wide volumes and using crowdsourced data to estimate ridership. Site selection guidelines in FHWA (2014) to include streets with various functional classification and volume groups can be followed by local government agencies while setting up counter locations to capture varying cycling patterns across the city. The spatial and temporal coverage of crowdsourced data also provides an opportunity to stratify streets as low, medium, and high volumes and can be leveraged to select counter locations (Brum-Bastos et al., 2019). Rural bicycle counts in this study are limited to short-term counts. Due to lack of availability of permanent counts, expansion factors could not be calculated with existing methodologies and an average of the short-term counts is used in modeling. Short-term counts of less than seven-day 24-hours are highly susceptible to errors. Given California's geographic variance, rural counts from only four counties may possibly be insufficient to generalize the model. Results can be improved by using reliable short-term or permanent count data from more rural locations. Future research can also estimate the bicycle models at the daily level. Modeling at the daily level provides an opportunity to leverage the temporal resolution of crowdsourced data from the Strava app and analyze how the effect of various factors or variables changes when compared to modeling with the annual averages as in this study. Unlike the bicycle demand model, the dependent variable in the pedestrian demand model is an estimate of AADPT from short-term counts. These estimates have unknown errors and need to be validated before using them for further research. Strava data is also not incorporated in the pedestrian demand models. Future studies might look for opportunities to obtain intersection or node level pedestrian counts from the Strava Metro dataset to understand the relationship of crowdsourced data with pedestrian volumes. Though all the models are calibrated using a cross validation procedure, future research should evaluate the model accuracy using on-ground traffic counts from additional counter locations as they become available.

Despite these limitations, this work contributes to the existing research by incorporating different data sources such as the Strava data and BNA network accessibility metrics and demonstrating their importance and influence for non-motorized volume estimation, highlighting the limitations with short-term counts in rural locations, and the use of Random Forest algorithm in non-motorized transportation research. Apart the direct application of this study in the Caltrans Active Transportation Program Benefit – Cost Tool, the findings are relevant to researchers, transportation planners, local government agencies and policy makers in prioritizing active transportation infrastructure, crash analysis studies, designing traffic monitoring programs, exploring new data sources and statistical methods for bicycle or pedestrian volume estimation.

7. Appendix

7.1 Data Preparation

In this section, I discuss the detailed data preparation steps for bicycle and pedestrian data, which includes combining all the exploratory variable datasets with the dependent variable.

7.1.1 Bicycle data

Figure 15 outlines the different datasets used in the bicycle demand model, their relationships with one another and the data preparation pipeline. The relationships can be classified as one to one or one to many. One to one implies that one record in the source table matches only one record in the target table. One to many implies that one record in the source table matches many records in the target table. For ease of understanding, data is divided into:

- *Static data*, which does not vary by year. This includes Network accessibility metrics, roadway, and census tract characteristics. (Roadway and census tract characteristics can change over time. For example, a new bike lane can be added to a roadway or female population of a census tract can change from year to year. But these datasets are assumed to be static for the purposes of modeling)
- *Dynamic data*, which varies by year. This includes bicycle counts, Strava counts and weather data.

AADBTs across all the counter locations are summarized into the "Counts" table. Location IDs in the "Counts" table are random numbers and often not unique with respect to type of pathway and direction. Unique location IDs are thus generated based on pathway and direction, increasing the number of locations from 114 to 310 (San Francisco – 38, San Diego – 20, Caltrans District 1 – 252). These are referred to as "Location ID" here on and is the unique identifier in the "All Locations" table. A Location ID is therefore specific to a combination of location description, direction, and pathway. The naming convention is to begin with an abbreviation for a city/county (San Francisco- SF, San Diego- SD, Del Norte- DN, Lake – LAK, Humboldt – HUM and Mendocino – MEN), followed by a number. The “Counts” table is updated with new Location IDs.

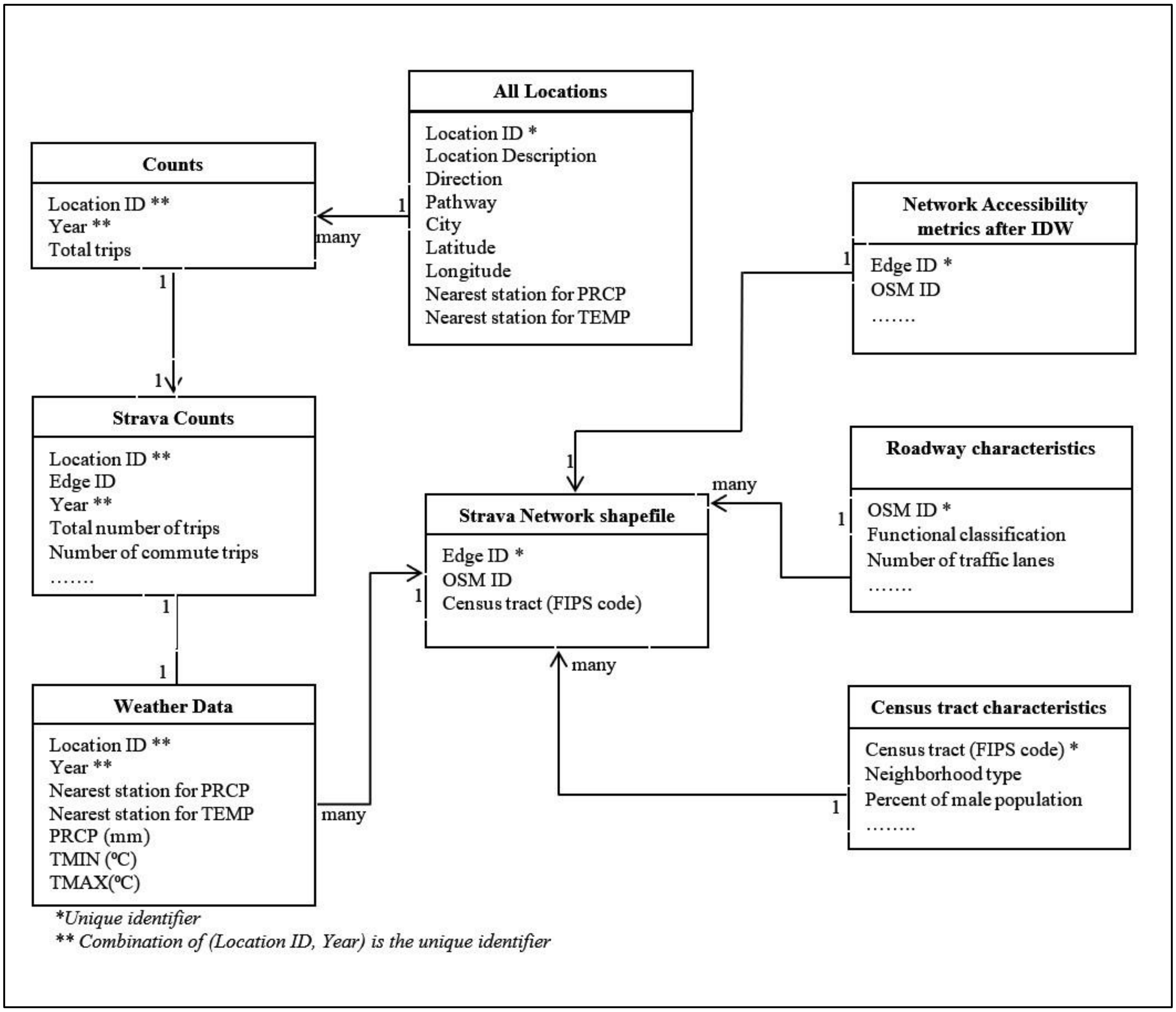


Figure 15: Data preparation pipeline- Bicycle data

The “Weather data” table consists of annual summaries of precipitation, minimum and maximum temperatures from NOAA website queried using the “rnoaa” package in R. meteo_nearby_stations()” function in this package returns stations within a specified radius from a given latitude and longitude. The stations vary based on the given variable (prcp –

precipitation, tmin – minimum temperature and tmax – maximum temperature) and date range. The closest station to a counter location is checked for completeness of the data. If the data is not available for the required dates, other nearby stations within a 50km (31mile) radius are considered. The list of stations thus generated, is taken as input in the “meteo_pull_monitors()” function which retrieves annual summaries from the NOAA website. The units of measurement are “tenths of a millimeter (mm)” and “tenths of a degree Celsius (ºC)”, respectively. With a division of 10, these are transformed to “(mm)” and “(ºC)”. The “Weather data” is then merged to the “Counts” data by Location ID and year.

Daily bicycle volume on a network link corresponding to latitude and longitude of counter locations in “All Locations” table is obtained from Strava Metro. Strava Metro outputs consist of bi-directional data for every link, shown as “forward” or “reverse”. When the Counts data includes bi-directional data at a location, the “forward” and “reverse” columns are separated and a Location ID from the “All Locations” table is matched based on location description and direction. On the Strava network, the “forward” direction correlates to the direction of the same link on OSM. Thus, OSM was used to determine the “forward” direction and assign Location IDs accordingly. Figure 16 shows this processing of Strava outputs by taking two sample variables, number of trips (trip count) and number of males (male people count). Location IDs are manually added to the processed outputs. Because Location IDs are unique to a location in a specific direction, even if the edge ID remains the same, the output tables will have distinct Location IDs.

Daily counts are averaged for the days for which counts are available to get annual average Strava counts. Such output tables for all study area locations are combined to a “Strava Counts” table, shown in Figure 16. "Strava Counts" is then merged with the "Counts" based on Location ID and Year.

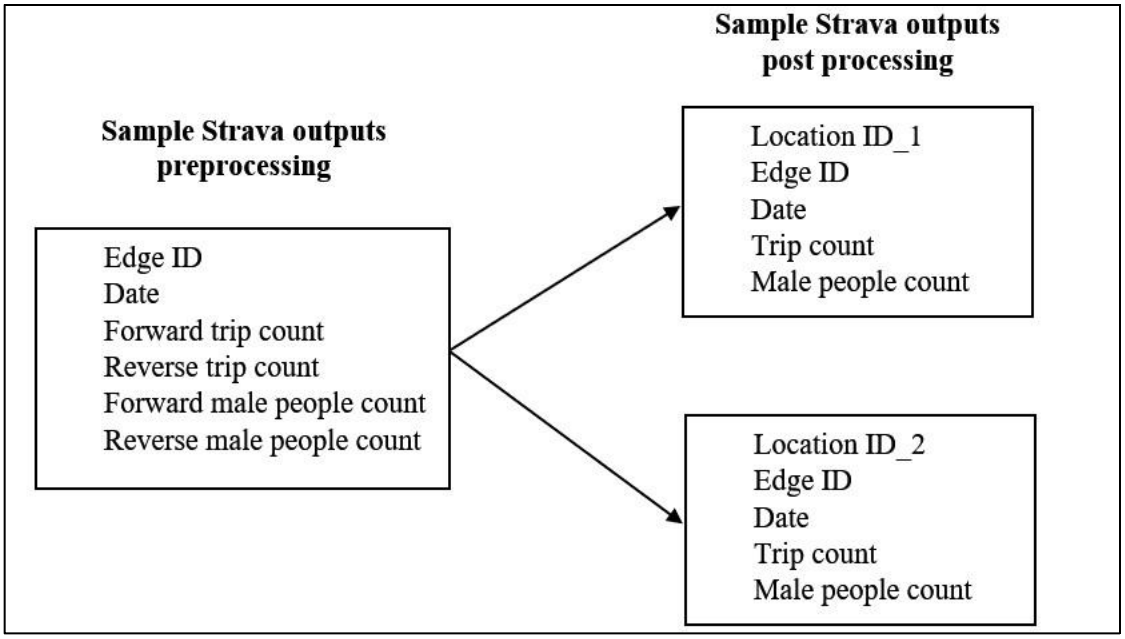


Figure 16: Strava data processing pipeline

The other output from Strava Metro is the Network shapefile for the State of California. The Strava network shapefile with "Edge ID" as the unique identifier is a bridge to join the static and dynamic datasets. The network consists of two columns "Edge ID" and "OSM ID". The Strava base map is different from the Open Street Map (OSM) network. Each OSM line is separated at decision points into edges. Thus, each OSM ID could have multiple edge IDs. The "Edge ID"s in the shapefile are filtered to retain only the "Edge ID" s that match the study area locations from the "Strava Counts" table. The resulting network is then spatially joined with the California Census tract shapefile to extract the Census tract (FIPS) code within which each an edge lies. The final network shapefile as shown in Figure 1 consists of "Edge ID", "OSM ID" and "Census tract (FIPS) code".

The static data tables “Network Accessibility metrics after IDW”, “Roadway characteristics” and “Census tract characteristics” have “Edge ID”, “OSM ID”, and “Census tract (FIPS) code” as unique identifiers respectively, which are merged to the Strava network shapefile. This output is then merged to the dynamic data by matching the “Edge ID” s from “Strava Counts” table.

After filtering the rows with zero bicycle counts and segmenting the data based on neighborhood type of the counter location, there are 54 unique urban and 158 unique rural locations. As one location can have multiple data points depending on the year for which AADBT is calculated, the final dataset for modeling contains 132 urban and 184 rural data points.

7.1.2 Pedestrian data

The pedestrian data preparation pipeline is similar to that of the bicycle data pipeline, but with fewer datasets as shown in Figure 17. The “Intersection Counts” table contains the dataset used for pedestrian exposure model in Griswold et.al. (2019), with Location ID as the unique identifier of an intersection. This table is spatially joined to the California Census tract shapefile to obtain the Census tract FIPS code for each location. Then the table is filtered to retain AADPT, Census tract ID and other necessary roadway characteristics data in the final “Intersection Counts” table. The “Network Accessibility metrics after IDW” table contains IDW weighted BNA tool outputs and is joined to the “Intersection Counts” table based on Location ID. “Census tract characteristics” are joined to counts based on the Census tract FIPS code to get the final dataset.

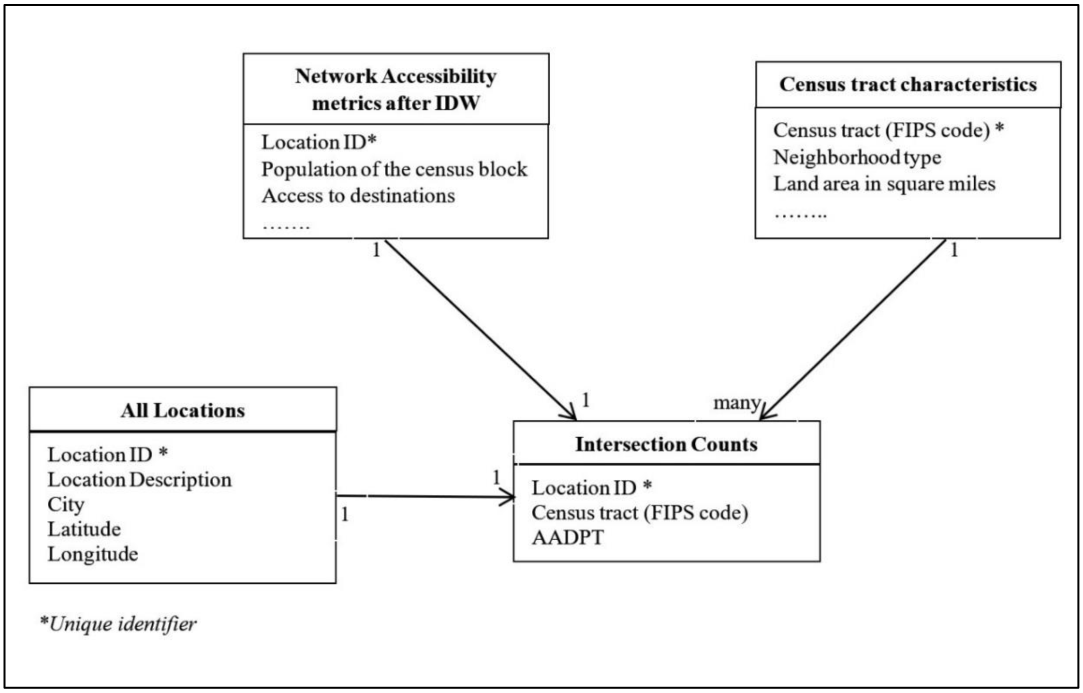


Figure 17: Data preparation pipeline- Pedestrian data

7.2 Exploratory Data Analysis and Descriptive Statistics

In this section, I present the results of exploratory data analysis and descriptive statistics for bicycle and pedestrian datasets to better understand the nature of variables and their relationships with one another.

Table 13 shows the minimum, mean, maximum and standard deviation of continuous/numerical variables in the urban and rural bicycle study locations. The dataset for the Generalized model is the summation of urban and rural datasets. To avoid redundancy, the descriptive statistics of the Generalized model dataset is not presented here. Table 14 shows the descriptive statistics for the pedestrian data. The network accessibility metrics used as explanatory variables are IDW normalized BNA tool outputs and their absolute values are hard to interpret. Hence their descriptive statistics are not shown here.

Table 13: Descriptive statistics for explanatory variables in bicycle demand models

	Urban				Rural			
	Min	**Max**	**Mean**	**Std.Dev**	**Min**	**Max**	**Mean**	**Std.Dev**
Percent of commuters using transit	0	57	21	17	0	8	0.9	1.8
Percent of commuters using walk	0	40	10	10	0	33	5	6.5
Percent of commuters using bike	0	12	3	3	0	7	1.1	2.1
Percent of male population	42	70	52.9	7	42	59	50.5	3.7
Percent of female population	30	58	47	7	41	58	49.4	3.7
Median household income	23757	176875	72177	33975	19538	62587	39192	10788
Percent of White alone population	12	94	68	17.2	12	94	79.2	13.5
Percent of Black or African American alone population	0	16	5	4.8	0	4	0.8	0.88
Percent of Asian alone population	0	72	15	14.2	0	6	1.9	2.51
Percent of American Indians alone population	0	1	0.2	0.4	0	12	2	11.9
Percent of Hispanic or Latino population	0	47	19	12	3	61	16.2	12.6
Number of severe injuries	0	7	0.75	1.31	0	1	0.375	0.48
Number of visible injuries	0	21	3.4	4.1	0	1	0.942	1.05
Number of complaints of pain	0	32	3.9	5.69	0	3	0.51	0.85
Number of bicyclists killed	0	1	0.03	0.18	0	1	0.04	0.21
Number of bicyclists injured	0	45	7.1	9.56	0	8	1.23	1.88
Precipitation (mm)	0.49	2.19	1.27	0.49	0	8.211	3.25	1.76
Minimum Temperature (ºC)	3.97	14.11	8.077	1.87	11.04	12.76	15.7	1.94
Maximum Temperature (ºC)	12.98	25.61	18.8	2.06	17.74	22.87	19.63	3.37

Table 14: Descriptive statistics for explanatory variables in pedestrian demand model

	Pedestrian Data			
	Min	**Max**	**Mean**	**Std.Dev**
Percent of commuters using transit	0	57	6.25	8.36
Percent of commuters using walk	0	50	5.3	7.2
Percent of commuters using bike	0	16	2.15	3.13
Percent of male population	20	90	50.9	6.61
Percent of female population	10	80	49.3	6.61
Median household income	4541	233026	54993	27120
Percent of White alone population	2	97	64	22.8
Percent of Black or African American alone population	0	86	8.22	14.8
Percent of Asian alone population	0	85	11.8	14.55
Percent of American Indians alone population	0	80	10	4.06
Percent of Hispanic or Latino population	10	98	30.57	24.25
Number of severe injuries	0	6	0.63	0.937
Number of visible injuries	0	20	2.058	2.57
Number of complaints of pain	0	36	2.818	3.87
Number of pedestrians killed	0	3	0.25	0.553
Number of pedestrains injured	0	60	5.241	6.177
Precipitation (mm)	0	7.47	1.32	0.99
Minimum Temperature (^{0}C)	0	31.76	23	3.23
Maximum Temperature (^{0}C)	-3.8	17.3	11.74	2.93

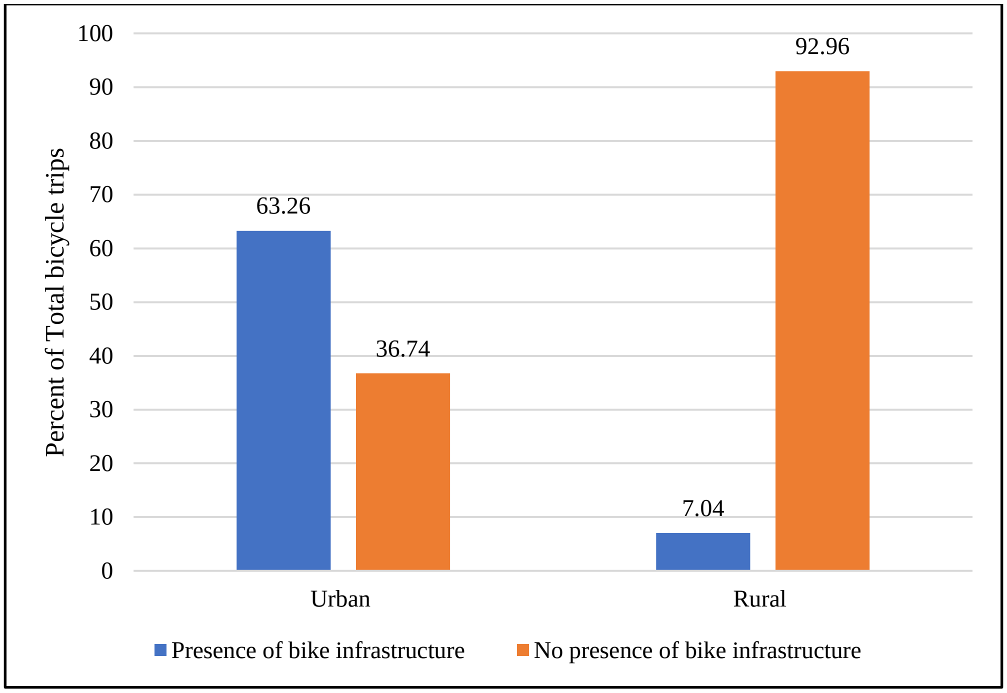


Figure 18: Total bicycle trips based on presence of bike infrastructure

Presence of bike infrastructure implies that the counter location has bike lane or bike path or sharrows. 40 of the 54 urban locations have bike infrastructure, while only 6 of the 158 rural locations have bike infrastructure. Nearly 60% of the total bicycle trips in urban areas are on street segments which have some bike infrastructure (Figure 18).

7.3 Random Forest for Regression

Random Forest algorithm was proposed by Breiman (2001) and the Random Forest for regression algorithm includes the following steps (Hastie et al., 2009) :

1. For *b = 1 to B*, *where B* is the total number of trees in a Random Forest:
    - Draw a bootstrap sample Z^* of size *N* from the training data
    - Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size is reached.

- Select *m* variables at random from the *p* variables.
- Pick the best variable/split-point among the *m*.
- Split the node into two daughter nodes.

2. Output the ensemble of trees $\{Tb\}_1^B$

Prediction at a new point x is calculated by: $\frac{1}{B}\sum_{b=1}^{B} T_b(x)$

The decision trees in a Random Forest algorithm are constructed using the methodology of CART (Classification and Regression Tree) (Breiman et al., 1984). Let the data consist of p explanatory variables denoted by $x_i = (x_1, x_2, \ldots \ldots \ldots \ldots \ldots \ldots .., x_p)\ for\ i = 1,2, \ldots p.$ Let the response variable be denoted by y. In the current study, x_i can be understood as the explanatory variables such as the Strava counts, network accessibility metrics, census tract characteristics, roadway characteristics and weather data. Response variable, y, is the AADBT and AADPT for bicycle and pedestrian demand models respectively. Let the data contain *N* observations. Then $x_k = (x_{k1}, x_{k2}, \ldots \ldots \ldots \ldots \ldots \ldots .., x_{kN})\ for\ k = 1,2, \ldots p, and\ y = (y_1, y_2, \ldots \ldots \ldots \ldots \ldots \ldots .., y_N).$

For each variable $x_k\ where\ 1 \leq k \leq p$, the algorithm determines the optimal split point *s* by the formula:

$$\min_s [MSE(y_i|X_{ki} \leq s) + MSE(y_i|X_{ki} > s)]$$

MSE is the mean squared error calculated using the predicted value of y_i, $\widehat{y_i}$ using the formula:

$$\text{MSE} = \frac{\sum(y_i - \widehat{y_i})^2}{n}$$

Finally, using the combination of every variable x_k with its optimal split point *s*, $\widehat{y_i}$ is predicted and MSE is calculated. The variable x_k that yields the lowest MSE is chosen as a tree node.

And this procedure is repeated until the MSE-gain becomes too small.

7.3.1 Calculating variable importance in Random Forest regression

For a single decision tree *T*, when performing split of a region R into two subregions R_1 and R_2 , the squared improvement of that split is defined as the difference of the residual sum of squares (RSS), before and after the split. The squared improvement of the split for a variable x_k $where$ $1 \ \leq k \leq p$ is calculated by the formula:

$$\tau^2 \coloneqq \sum_{i:x_{ki} \in R} (y_i - \hat{y}_R)^2 - (\sum_{i:x_{ki} \in R_1} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_{ki} \in R_2} \left(y_i - \hat{y}_{R_2}\right)^2)$$

The variable importance of a predictor variable x_k is then calculated by:

$$I_k^2(T) = \sum_{t=1}^{J-1} \tau_t^2 \ 1_{\{s_t = k\}}$$

Where:

- *t* corresponds to the split performed (total of J-1 splits in the tree)
- τ_t^2 is the squared improvement obtained from split *t*
- s_t is the predictor variable over which split *t* was done

The squared relative importance of variable x_k is thus the sum of squared improvements over all the internal nodes for which it was chosen as the splitting variable. This importance measure is then averaged over all the decision trees to get the relative importance of a variable x_k in the Random Forest. The higher the variable importance of x_k, the more useful it is to make accurate predictions.

References

Al-Ramini, A., Takallou, M. A., Piatkowski, D. P., & Alsaleem, F. (2022). Quantifying changes in bicycle volumes using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/23998083211066103

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, *187*. https://doi.org/10.1016/j.eswa.2021.115895

Aoun, A., Bjornstad, J., Dubose, B., Mitman, M., Pelon, M., & Peers, F. &. (2015). *White Paper Series Bicycle and Pedestrian Forecasting Tools: State of the Practice*. www.pedbikeinfo.org

Attaset, V., Schneider, R. J., & Arnold, L. S. (2010). *Effects of Weather Variables on Pedestrian Volumes in Alameda County, California*. https://escholarship.org/uc/item/3zn9f4cr

Aultman-Hall, L., Lane, D., & Lambert, R. R. (2009). *Assessing the Impact of Weather and Season on Pedestrian Traffic Volumes*.

Blanc, B, Figliozzi, M, Clifton, K (2016) How representative of bicycling populations are smartphone application surveys of travel behavior? *Transportation Research Record: Journal of the Transportation Research Board 2587*: 78–89.

Breiman, L. (2001). *Random Forests. Machine Learning* (Vol. 45). http://dx.doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Brum-Bastos, V., Ferster, C. J., Nelson, T., & Winters, M. (2019). Where to put bike counters? Stratifying bicycling patterns in the city using crowdsourced data. *Transport Findings*. https://doi.org/10.32866/10828

Buehler, R. (2012). Determinants of bicycle commuting in the Washington, DC region: The role of bicycle parking, cyclist showers, and free car parking at work. *Transportation Research Part D: Transport and Environment*, *17*(7), 525–531. https://doi.org/10.1016/j.trd.2012.06.003

Chen, C., Wang, H., Roll, J., Nordback, K., & Wang, Y. (2020). Using bicycle app data to develop Safety Performance Functions (SPFs) for bicyclists at intersections: A generic framework. *Transportation Research Part A: Policy and Practice*, *132*, 1034–1052. https://doi.org/10.1016/j.tra.2019.12.034

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. In *Genomics* (Vol. 99, Issue 6, pp. 323–329). https://doi.org/10.1016/j.ygeno.2012.04.003

Clifton, K. J., & Patrick A. Singleton. (2012). *Pedestrians in Regional Travel Demand Forecasting Models: State-of-the-Practice*. https://nacto.org/docs/usdg/13-4857.pdf

Dadashova, B., Griffin, G., Das, S., Turner, S., & Graham, M. (2018). *Guide for Seasonal Adjustment and Crowdsourced data scaling*. https://static.tti.tamu.edu/tti.tamu.edu/documents/0-6927-P6.pdf

Dadashova, B., & Griffin, G. P. (2020). Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. *Transportation Research Part D: Transport and Environment*, *84*. https://doi.org/10.1016/j.trd.2020.102368

Dadashova, B., Griffin, G. P., Das, S., Turner, S., & Sherman, B. (2020). Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data. *Transportation Research Record*, *2674*(11), 390–402. https://doi.org/10.1177/0361198120946016

Dawn Royal, & Darby Miller-Steiger. (2002). *National Survey of Bicyclist and Pedestrian Attitudes and Behavior Volume II Findings Report*. www.nhtsa.dot.gov.

Dill, J., & Carr, T. (2003). Bicycle Commuting and Facilities in Major U.S. Cities If You Build Them, Commuters Will Use Them. *N Transportation Research Record, TRB, National Research Council, Washington, D.C.*, *1828*(1), 116–123. https://doi.org/10.3141/1828-14

Dill, J., & Voros, K. (2007). Factors affecting bicycling demand: Initial survey findings from the Portland, Oregon, region. *Transportation Research Record*, *2031*, 9–17. https://doi.org/10.3141/2031-02

Federal Highway Administration. (2014). *Assessing Roadway Traffic Count Duration and Frequency Impacts on Annual Average Daily Traffic (AADT)*

Federal Highway Administration. (2020). *NON-MOTORIZED TRAVEL 2017 National Household Travel Survey Who Reports Walk/Bike Trips?* http://nhts.ornl.gov

Fuller, D., & Winters, M. (2017). Income inequalities in Bike Score and bicycling to work in Canada. *Journal of Transport and Health*, *7*, 264–268. https://doi.org/10.1016/j.jth.2017.09.005

Garber, M. D., Watkins, K. E., & Kramer, M. R. (2019). Comparing bicyclists who use smartphone apps to record rides with those who do not: Implications for representativeness and selection bias. *Journal of Transport and Health*, *15*. https://doi.org/10.1016/j.jth.2019.100661

Ghani, F., Rachele, J. N., Washington, S., & Turrell, G. (2016). Gender and age differences in walking for transport and recreation: Are the relationships the same in all neighborhoods? *Preventive Medicine Reports*, *4*, 75–80. https://doi.org/10.1016/j.pmedr.2016.05.001

Griffin, G. P., & Jiao, J. (2015). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport and Health*, *2*(2), 238–247. https://doi.org/10.1016/j.jth.2014.12.001

Griswold, J. B., Medury, A., Schneider, R. J., Amos, D., Li, A., & Grembek, O. (2019). A Pedestrian Exposure Model for the California State Highway System. *Transportation Research Record*, *2673*(4), 941–950. https://doi.org/10.1177/0361198119837235

Griswold, J., Medury, A., Huang, L., Amos, D., Lu, J., Schneider, R., & Grembek, O. (2018). *CA18-2452 2. GOVERNMENT ASSOCIATION NUMBER 3. RECIPIENT'S CATALOG NUMBER 4. TITLE AND SUBTITLE Pedestrian Safety Improvement Program: Phase 2*.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. In *Journal of Machine Learning Research* (Vol. 3).

Handy, S., Cao, X., & Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. *Transportation Research Part D: Transport and Environment*, *10*(6), 427–444. https://doi.org/10.1016/j.trd.2005.05.002

Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., & Xu, Z. (2012). Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. *Landscape and Urban Planning*, *107*(3), 307–316. https://doi.org/10.1016/j.landurbplan.2012.06.005

Hankey, S., Lu, T., Mondschein, A., & Buehler, R. (2017). Spatial models of active travel in small communities: Merging the goals of traffic monitoring and direct-demand modeling. *Journal of Transport and Health*, *7*, 149–159. https://doi.org/10.1016/j.jth.2017.08.009

Hanson, S., & Hanson, P.O. (1977). Evaluating the Impact of Weather on Bicycle Use. *Transportation Research Record, 629*, 43-48.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography*, *75*, 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013

Huang, R., Moudon, A. v., Zhou, C., & Saelens, B. E. (2019). Higher residential and employment densities are associated with more objectively measured walking in the home neighborhood. *Journal of Transport and Health*, *12*, 142–151. https://doi.org/10.1016/j.jth.2018.12.002

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Kothuri, S., Broach, J., McNeil, N., Kate Hyun, M., Mattingly, S., Mintu Miah Krista Nordback, M., & Proulx, F. (2022). *Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes Photo by Lacey Friedly*. www.nitc-utc.net

Kuzmyak JR, Walters J, Bradley M, Kockelman KM. (2014). *"Estimating Bicycling and Walking for Planning and Project Development: A Guidebook." NCHRP Rep. No. 770. Washington, DC:Transportation Research Board.*

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing Journal*, *86*. https://doi.org/10.1016/j.asoc.2019.105836

Laib, M., & Kanevski, M. (2019). A new algorithm for redundancy minimisation in geo-environmental data. *Computers & Geosciences*, *133*, 104328. https://doi.org/https://doi.org/10.1016/j.cageo.2019.104328

Le, H. T. K., Buehler, R., & Hankey, S. (2018). Correlates of the built environment and active travel: Evidence from 20 US metropolitan areas. *Environmental Health Perspectives*, *126*(7). https://doi.org/10.1289/EHP3389

Lee, K., & Sener, I. N. (2021). Strava Metro data for bicycle monitoring: a literature review. *Transport Reviews*, *41*(1), 27–47. https://doi.org/10.1080/01441647.2020.1798558

Lin, Z., & Fan, W. (David). (2020). Modeling bicycle volume using crowdsourced data from Strava smartphone application. *International Journal of Transportation Science and Technology*, *9*(4), 334–343. https://doi.org/10.1016/j.ijtst.2020.03.003

Lindsay, G., Macmillan, A., & Woodward, A. (2011). Moving urban trips from cars to bicycles: Impact on health and emissions. *Australian and New Zealand Journal of Public Health*, *35*(1), 54–60. https://doi.org/10.1111/j.1753-6405.2010.00621.x

Martin, A., Goryakin, Y., & Suhrcke, M. (2014). Does active commuting improve psychological wellbeing? Longitudinal evidence from eighteen waves of the British Household Panel Survey. *Preventive Medicine*, *69*, 296–303. https://doi.org/10.1016/j.ypmed.2014.08.023

Miah, Md. M., Hyun, K. K., Mattingly, S. P., Broach, J., McNeil, N., & Kothuri, S. (2022). Challenges and Opportunities of Emerging Data Sources to Estimate Network-Wide Bike Counts. *Journal of Transportation Engineering, Part A: Systems*, *148*(3). https://doi.org/10.1061/jtepbs.0000634

Miranda-Moreno, L. F., & Fernandes, D. (2011). Modeling of pedestrian activity at signalized intersections: Land use, urban form, weather, and spatiotemporal patterns. *Transportation Research Record*, *2264*, 74–82. https://doi.org/10.3141/2264-09

Miranda-Moreno, L., & Nosal, T. (2011). Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. In *Transportation Research Record* (Issue 2247, pp. 42–52). https://doi.org/10.3141/2247-06

Miranda-Moreno, L., Nosal, T., Schneider, R., & Proulx, F. (2013). Classification of bicycle traffic patterns in five North American cities. *Transportation Research Record*, *2339*, 68–79. https://doi.org/10.3141/2339-08

Moudon, A. V., Lee, C., Cheadle, A. D., Collier, C. W., Johnson, D., Schmid, T. L., & Weather, R. D. (2005). Cycling and the built environment, a US perspective. *Transportation Research Part D: Transport and Environment*, *10*(3), 245–261. https://doi.org/10.1016/j.trd.2005.04.001

Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., de Nazelle, A., Dons, E., Gerike, R., Götschi, T., Int Panis, L., Kahlmeier, S., & Nieuwenhuijsen, M. (2015). Health impact assessment of active transportation: A systematic review. In *Preventive Medicine* (Vol. 76, pp. 103–114). Academic Press Inc. https://doi.org/10.1016/j.ypmed.2015.04.010

Munira, S. (2017). *Use of Direct-Demand Modeling in Estimating Nonmotorized Activity: A Meta-analysis Project Number: UTC Safe-D 01-003 Project Title: Data Mining to Improve Planning for Pedestrian and Bicyclist Safety*.

National Bicycling And Walking Study. (1992). https://ntlrepository.blob.core.windows.net/lib/6000/6300/6341/CASE1.pdf

Nelson, T., Roy, A., Ferster, C., Fischer, J., Brum-Bastos, V., Laberee, K., Yu, H., & Winters, M. (2021). Generalized model for mapping bicycle ridership with crowdsourced data. *Transportation Research Part C: Emerging Technologies*, *125*. https://doi.org/10.1016/j.trc.2021.102981

Nordback, K., Kothuri, S., Johnstone, D., Lindsey, G., Ryan, S., & Raw, J. (2019). Minimizing Annual Average Daily Nonmotorized Traffic Estimation Errors: How Many Counters Are Needed per Factor Group? *Transportation Research Record*. https://doi.org/10.1177/0361198119848699

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*. https://doi.org/10.1016/j.petrol.2021.109244

PeopleforBikes Bicycle Network Analysis. (n.d.). Retrieved May 19, 2021, from https://bna.peopleforbikes.org/#/methodology

Peter G. Furth, Maaza C. Mekuria, & Hilary Nixon. (2012). *LOW-STRESS BICYCLING AND NETWORK CONNECTIVITY*. http://transweb.sjsu.edu

Plaut, P. O. (2005). Non-motorized commuting in the US. *Transportation Research Part D: Transport and Environment*, *10*(5), 347–356. https://doi.org/10.1016/j.trd.2005.04.002

Porter, C., Suhrbier, J., & Schwartz, W. L. (1999). Forecasting Bicycle and Pedestrian Travel State of the Practice and Research Needs. *TRANSPORTATION RESEARCH RECORD , No. 1674: 94-101*.

Prelog. (2015). *Bicycle Equity: The Equity of Access to Bicycle Infrastructure*. https://bikeleague.org/sites/default/files/bike_equity_index_final_web.pdf

Rastogi, R. (2011). Promotion of non-motorized modes as a sustainable transportation option: policy and planning issues. In *Current Science* (Vol. 100, Issue 9). https://about.jstor.org/terms

Roll, J. (2018). *BICYCLE COUNT DATA: WHAT IS IT GOOD FOR? A STUDY OF BICYCLE TRAVEL ACTIVITY IN CENTRAL LANE METROPOLITAN PLANNING ORGANIZATION Final Report PROJECT 304-761*.

Roll, J. F., & Proulx, F. R. (2018). Estimating Annual Average Daily Bicycle Traffic without Permanent Counter Stations. *Transportation Research Record*, *2672*(43), 145–153. https://doi.org/10.1177/0361198118798243

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science*, *3*(2), 62. https://doi.org/10.3390/urbansci3020062

Runa, F. (2020). *The Effect of Weather on Pedestrian Activity at Signalized The Effect of Weather on Pedestrian Activity at Signalized Intersections in Utah Intersections in Utah*. https://digitalcommons.usu.edu/etd/7973

Salon, D., & Handy, S. (2014). *Estimating Total Miles Walked and Biked by Census Tract in California*. https://dot.ca.gov/-/media/dot-media/programs/research-innovation-system-information/documents/f0016815-final-report-task-2200.pdf

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*, *116*, 10–17. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2018.05.006

Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. *Transportation Research Record*, *2605*(1), 32–44. https://doi.org/10.3141/2605-03

Schneider, R., Henry, T., Mitman, M., Stonehill, L., & Koehler, J. (2012). Development and application of volume model for pedestrian intersections in San Francisco, California. *Transportation Research Record*, *2299*, 65–78. https://doi.org/10.3141/2299-08

Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2009). Pilot model for estimating pedestrian intersection crossing volumes. *Transportation Research Record*, *2140*, 13–26. https://doi.org/10.3141/2140-02

Schoner, J. E., Levinson, D. M., Schoner, J., & Levinson, D. (2014). The Missing Link Bicycle Infrastructure Networks and Ridership in 74 US Cities. *Transportation*, *41*, 1187–1204. https://doi.org/10.1007/s11116-014-9538-1

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Smith, A. (2015). Crowdsourcing for Active Transportation. In *Journal* (Vol. 85).

St-Louis, E., Manaugh, K., van Lierop, D., & El-Geneidy, A. (2014). The happy commuter: A comparison of commuter satisfaction across modes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *26*(PART A), 160–170. https://doi.org/10.1016/j.trf.2014.07.004

Strava. (2020). *Strava Metro Frequently Asked Questions*. https://metro.strava.com/faq

Sunde, E. (2019). *Tracking the rise of bike commuting around the world*.

Tabeshian, M., & Kattan, L. (2014). Modeling Nonmotorized Travel Demand at Intersections in Calgary, Canada. *Transportation Research Record: Journal of the Transportation Research Board*, *2430*(1), 38–46. https://doi.org/10.3141/2430-05

Thomas, T., Jaarsma, C. F., Tutert, S. I. A., Thomas, T., Jaarsma, R., & Tutert, B. (2008). *Temporal variations of bicycle demand in the Netherlands: The influence of weather on cycling*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. In *J. R. Statist. Soc. B* (Vol. 58, Issue 1).

Turner, S., Ipek Sener, Michael Martin, Subasish Das, Eva Shipp, Robert Hampshire, Kay Fitzpatrick, Lisa Molnar, Ravi Wijesundera, Mike Colety, & Stewart Robinson. (2017). *Synthesis Of Methods for Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels And On Specific Transportation Facilities*.

Wang, X., Lindsey, G., Hankey, S., & Hoff, K. (2013). *Estimating Mixed-Mode Urban Trail Traffic Using Negative Binomial Regression Models*. https://doi.org/10.1061/(ASCE)UP

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2010). Built environment influences on healthy transportation choices: Bicycling versus driving. *Journal of Urban Health*, *87*(6), 969–993. https://doi.org/10.1007/s11524-010-9509-6

Xu, Z., Huang, G., Weinberger, K. Q., & Zheng, A. X. (2014). Gradient boosted feature selection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 522–531. https://doi.org/10.1145/2623330.2623635