# PLANNING TOOLS FOR TRANSIT MANAGERS TO IMPROVE EFFICIENCIES AND PREPARE FOR THE POST-COVID ENVIRONMENT

## FINAL PROJECT REPORT

by

Brandon Bullard, Jake Wagner
Timur Dincer, Danna Moore
Washington State University

Pacific Northwest Transportation Consortium (PacTrans)
Washington State University Transportation Services

for

Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology (OST-R)

## DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government, and matching sponsor assume no liability for the contents or use thereof.

# TECHNICAL REPORT DOCUMENTATION PAGE

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| | 01784878 | |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| PLANNING TOOLS FOR TRANSIT MANAGERS TO IMPROVE EFFICIENCIES AND PREPARE FOR THE POST-COVID ENVIRONMENT | 2023-02-01 |
| | 6. Performing Organization Code |

| 7. Author(s) and Affiliations | 8. Performing Organization Report No. |
|---|---|
| Brandon Bullard, Jake Wagner, 0000-0002-1614-4282; Timur Dincer, and Danna Moore; 0000-0001-5171-4546<br>Washington State University | 2021-S-WSU-3 |

| 9. Performing Organization Name and Address | 10. Work Unit No. (TRAIS) |
|---|---|
| PacTrans<br>Pacific Northwest Transportation Consortium<br>University Transportation Center for Federal Region 10<br>University of Washington More Hall 112 Seattle, WA 98195-2700 | 11. Contract or Grant No.<br>69A355174110 |

| 12. Sponsoring Organization Name and Address | 13. Type of Report and Period Covered |
|---|---|
| United States Department of Transportation<br>Research and Innovative Technology Administration<br>1200 New Jersey Avenue, SE<br>Washington, DC 20590 | Final Project Report |
| | 14. Sponsoring Agency Code |

**15. Supplementary Notes**

Report uploaded to: www.pactrans.org

**16. Abstract**

Washington State University Transportation Services is a self-sustaining unit responsible for managing the parking and transportation facilities and operations at WSU. They manage over 8,300 parking spaces, covered garages, paved lots and unpaved gravel lots.

Project Goals:

-Analyze parking revenues and costs at WSU to identify high margin opportunities for revenue growth and catalogue existing shortfalls in cost recovery

-Identify parking lot management strategies (lot locations, payment types, prices, level of service, etc.) to improve operational efficiencies and provide a path towards financial sustainability.

-Develop a transferable parking demand model that can be used to conduct scenario analyses and evaluate the effects of proposed parking policies.

| 17. Key Words | 18. Distribution Statement |
|---|---|
| Transit | |

| 19. Security Classification (of this report) | 20. Security Classification (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassified. | Unclassified. | 22 | N/A |

**Form DOT F 1700.7 (8-72)** **Reproduction of completed page authorized.**

# SI* (MODERN METRIC) CONVERSION FACTORS

| APPROXIMATE CONVERSIONS TO SI UNITS | | | | |
|---|---|---|---|---|
| **Symbol** | **When You Know** | **Multiply By** | **To Find** | **Symbol** |
| | | **LENGTH** | | |
| in | inches | 25.4 | millimeters | mm |
| ft | feet | 0.305 | meters | m |
| yd | yards | 0.914 | meters | m |
| mi | miles | 1.61 | kilometers | km |
| | | **AREA** | | |
| $in^2$ | square inches | 645.2 | square millimeters | $mm^2$ |
| $ft^2$ | square feet | 0.093 | square meters | $m^2$ |
| $yd^2$ | square yard | 0.836 | square meters | $m^2$ |
| ac | acres | 0.405 | hectares | ha |
| $mi^2$ | square miles | 2.59 | square kilometers | $km^2$ |
| | | **VOLUME** | | |
| fl oz | fluid ounces | 29.57 | milliliters | mL |
| gal | gallons | 3.785 | liters | L |
| $ft^3$ | cubic feet | 0.028 | cubic meters | $m^3$ |
| $yd^3$ | cubic yards | 0.765 | cubic meters | $m^3$ |
| | NOTE: volumes greater than 1000 L shall be shown in $m^3$ | | | |
| | | **MASS** | | |
| oz | ounces | 28.35 | grams | g |
| lb | pounds | 0.454 | kilograms | kg |
| T | short tons (2000 lb) | 0.907 | megagrams (or "metric ton") | Mg (or "t") |
| | | **TEMPERATURE (exact degrees)** | | |
| °F | Fahrenheit | 5 (F-32)/9<br>or (F-32)/1.8 | Celsius | °C |
| | | **ILLUMINATION** | | |
| fc | foot-candles | 10.76 | lux | lx |
| fl | foot-Lamberts | 3.426 | $candela/m^2$ | $cd/m^2$ |
| | | **FORCE and PRESSURE or STRESS** | | |
| lbf | poundforce | 4.45 | newtons | N |
| $lbf/in^2$ | poundforce per square inch | 6.89 | kilopascals | kPa |

| APPROXIMATE CONVERSIONS FROM SI UNITS | | | | |
|---|---|---|---|---|
| **Symbol** | **When You Know** | **Multiply By** | **To Find** | **Symbol** |
| | | **LENGTH** | | |
| mm | millimeters | 0.039 | inches | in |
| m | meters | 3.28 | feet | ft |
| m | meters | 1.09 | yards | yd |
| km | kilometers | 0.621 | miles | mi |
| | | **AREA** | | |
| $mm^2$ | square millimeters | 0.0016 | square inches | $in^2$ |
| $m^2$ | square meters | 10.764 | square feet | $ft^2$ |
| $m^2$ | square meters | 1.195 | square yards | $yd^2$ |
| ha | hectares | 2.47 | acres | ac |
| $km^2$ | square kilometers | 0.386 | square miles | $mi^2$ |
| | | **VOLUME** | | |
| mL | milliliters | 0.034 | fluid ounces | fl oz |
| L | liters | 0.264 | gallons | gal |
| $m^3$ | cubic meters | 35.314 | cubic feet | $ft^3$ |
| $m^3$ | cubic meters | 1.307 | cubic yards | $yd^3$ |
| | | **MASS** | | |
| g | grams | 0.035 | ounces | oz |
| kg | kilograms | 2.202 | pounds | lb |
| Mg (or "t") | megagrams (or "metric ton") | 1.103 | short tons (2000 lb) | T |
| | | **TEMPERATURE (exact degrees)** | | |
| °C | Celsius | 1.8C+32 | Fahrenheit | °F |
| | | **ILLUMINATION** | | |
| lx | lux | 0.0929 | foot-candles | fc |
| $cd/m^2$ | $candela/m^2$ | 0.2919 | foot-Lamberts | fl |
| | | **FORCE and PRESSURE or STRESS** | | |
| N | newtons | 0.225 | poundforce | lbf |
| kPa | kilopascals | 0.145 | poundforce per square inch | $lbf/in^2$ |

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

## TABLE OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

## EXECUTIVE SUMMARY

Public transit plays a crucial role in reducing the externalities associated with automobile use, such as pollution, congestion, and traffic accidents. Encouraging bus ridership is especially important because other modes of public transport have large startup costs and require greater population density, making small towns infeasible locations for them.

In this project, predictive models were developed to serve as the foundation of a decision support tool to help local transit authorities make better transit service decisions.

# CHAPTER 1. INTRODUCTION

Pullman Transit is the leading rural transit system throughout Washington state and within the region, and it provides over 1.4 million rides annually. In addition to serving Pullman residents, Pullman Transit also provides contracted service to Washington State University and Pullman Public Schools. As the sector leader, Pullman Transit serves as a model transit system and provides leadership and information to other rural transit systems throughout the state and the region.

Like most transit agencies, Pullman Transit is faced with challenging questions in its effort to meet service demands and community needs in a financially sustainable way. Where should bus stops be located? How frequently should each stop be serviced? Which routes should be driven to ensure all stops are serviced while minimizing rider travel time? What should the rider fare be? The answers to these questions are vital to the smooth and efficient operation of any transit network, but transit planners are typically left without the tools they need to make these important decisions.

The objective of this project was to develop a spatial transit demand model that will serve as the foundation of a transit planning decision support tool and will empower local transit planners with the information they need to make informed transit planning decisions. This model and decision support tool will help planners optimize daily operation decisions, identify transit service gaps, and efficiently respond to both demand shocks (such as the Covid-19 pandemic) and supply shocks (such as a temporary reduction in fleet size).

## CHAPTER 2. LITERATURE REVIEW

Predicting transit demand has proved to be challenging, as ridership and service levels work in concert (Taylor et al. 2008, Dill et al. 2013, Beberri et al. 2021, Boisjoly 2018, Chen 2015). The supply of transit influences demand, just as peak commuting times influence greater transit availability. Beberri et al (2021) used a Poisson fixed-effects model to estimate the elasticity of ridership demand with respect to frequency. Frequency was measured as the number of stops on a route, and ridership demand was given by the sum of boardings and alightings (at each stop/on each route). Using local stop-level data, they found that increased service frequency resulted in increased ridership, but that there were diminishing returns where a route was already popular.

Several studies have examined which additional variables are most important in predicting ridership. Taylor et al. (2008) used two-stage simultaneous equation regression models with data from hundreds of urbanized areas throughout the U.S. The researchers investigated the effects of transit supply on demand as well as which variables had the most influence. They examined geographic, economic, population, and auto system characteristics and found that population, household income, percentage of college students, recent immigration status, and lack of access to a car were important in explaining levels of ridership. Chakrabarti (2015) focused on how transit reliability affects ridership and found that routes with greater adherence to an established schedule were associated with greater ridership. This effect was more pronounced on routes with larger headways, presumably because of the greater consequence to riders of missing a route.

An advantage of using disaggregated stop-level data is the ability to explore how the built environment around a stop influences ridership. Chakour and Eluru (2016) examined the city of Montreal to determine how both stop-level infrastructure and the built environment influenced bus ridership. They found that transit service characteristics such as frequency and accessibility had the greatest impact, while enhancements to the land such as parks had a small but positive impact, and inhibitors such as major roads had a negative impact. With respect to spatial measurement of built environment variables, Pulugurtha and Agurla (2012) utilized spatial modeling methods to capture several attributes surrounding bus stops. They found that a quarter-mile buffer distance yielded the most meaningful estimates of ridership, and many subsequent

studies have used the same heuristic when gathering spatial data (Dill et al. 2013, Chakrabarti 2015, Li 2020).

With the advent of automated passenger count systems and the Global Positioning System (GPS) there is much greater data availability at the stop and route levels. These systems are primarily used by transit authorities to evaluate changes in performance, but researchers can also use these technologies to estimate demand at a much lower level of aggregation than previously available. Some of the earlier literature aggregated data over the course of a day or entire route. At lower levels of aggregation, researchers have found smaller elasticities with respect to transit service characteristics such as frequency and headway. Frei and Mahmassani (2013), for example, estimated ridership by using stop-level transit data from the Chicago, Illinois, transit system. They found much lower transit service elasticities with respect to ridership when their results were compared to those from similar studies with larger levels of aggregation.

## CHAPTER 3. DATA AND METHODS

Pullman Transit (PT) operates in the city of Pullman, Washington, and includes 223 stops along 41 routes. These stops and routes have mixed purposes, transporting elementary school students, college students, and the broader community. Pullman Transit provided the latitude, longitude, names, and service details (boardings and alightings) for each bus, stop, and route within the transit network for 2019 through 2021. Boardings and alightings were summed at each stop by the hour. Figure 3-1 shows the spatial distribution of locations where riders typically boarded the bus.

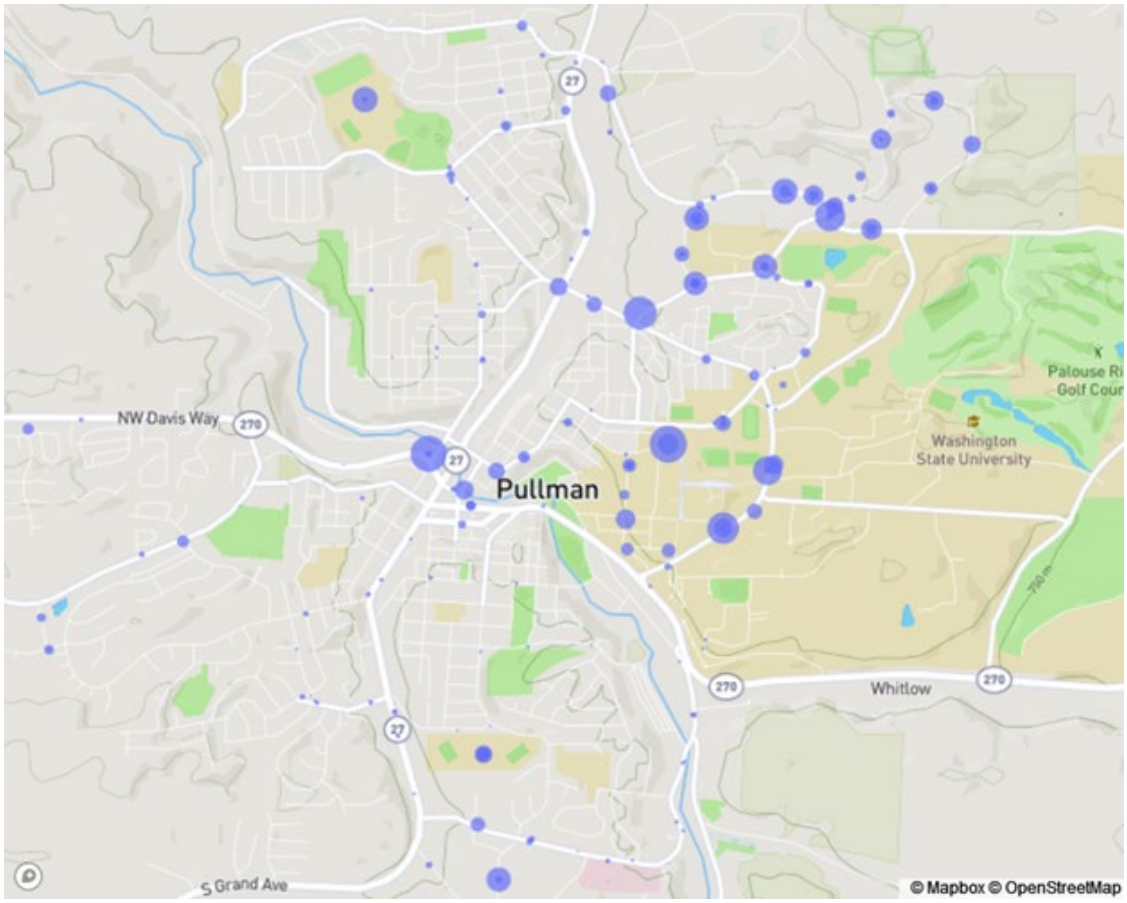


**Figure 3-1** Total Boardings, 2019-2021

Figure 3-2 illustrates how ridership varied by the hour and confirms that ridership had two distinct peaks, in the morning and afternoon. Both morning and afternoon peaks were a function of riders traveling to and from campus or a place of work.

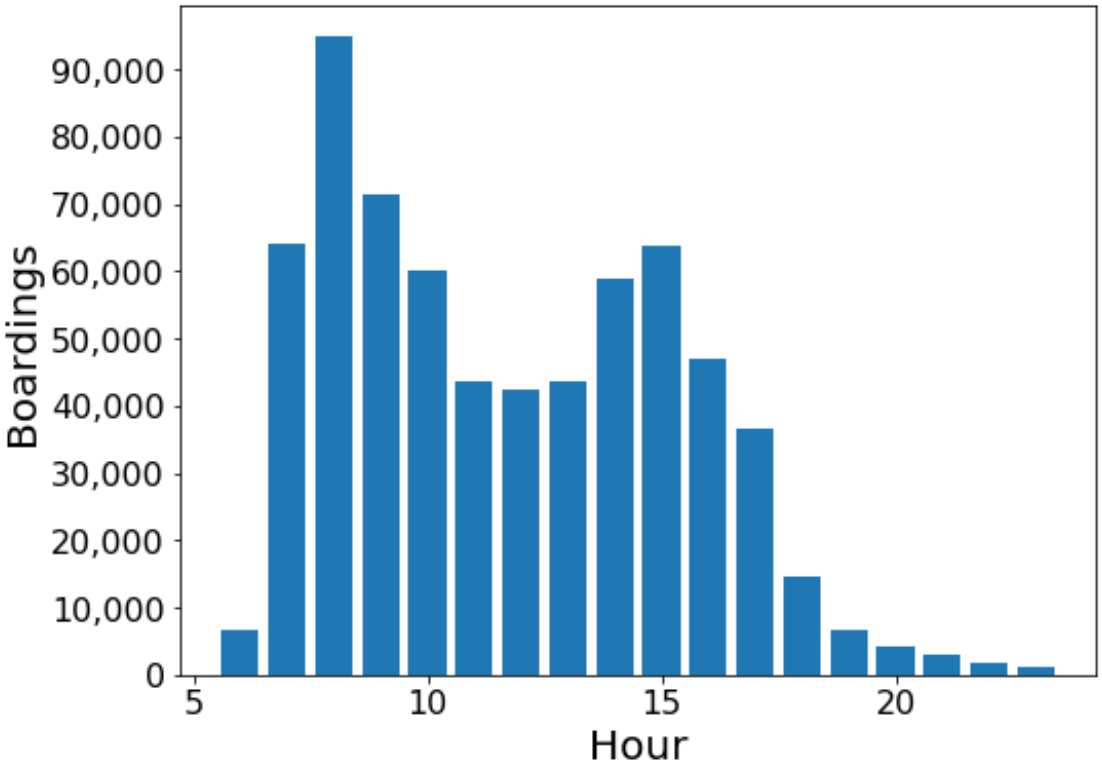


**Figure 3-2** Total Hourly Boardings, 2019-2021

Figure 3-3 shows how ridership changed throughout the entire sample. First, the Covid-19 pandemic had an immense, negative effect on ridership; the dashed line illustrates when WSU transitioned to online classes in response to pandemic shut-downs. Therefore, this analysis focuses solely on 2019 ridership because it more closely resembled normal ridership levels. Before this negative shock, ridership appeared to follow the flow of students throughout WSU's 15-week semester system. Semesters began in January, June, and August, with most students enrolling in the August semester. Ridership decayed after its peak in August from student attrition, and large negative troughs coincided with school holidays. Peaks throughout the semesters coincided with examinations and the beginning of new semesters in the fall, spring, and summer.

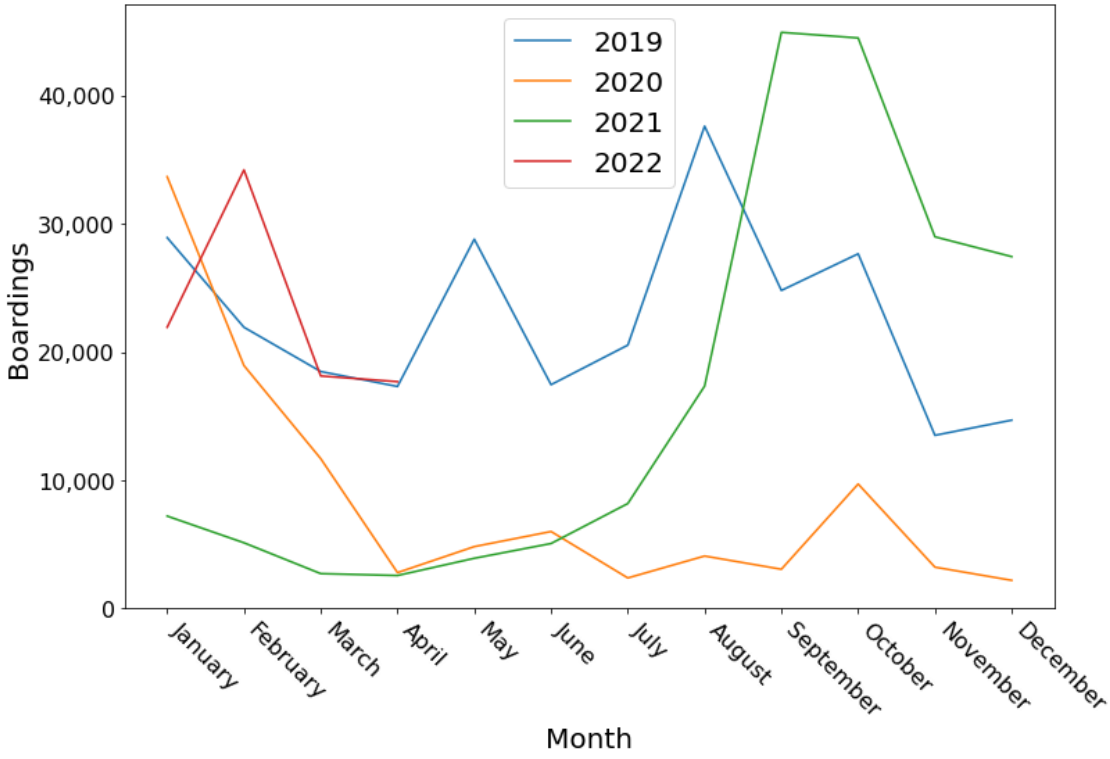


**Figure 3-3** Total Monthly Boardings, 2019-2021

Characteristics of the bus network are important because the transit authority can change them, and they also have the greatest direct influence on ridership (Berrebi et al. 2021). Frequency was a variable that captured the number of times a bus serviced a stop, aggregated by the hour. Each stop also had a travel time and distance for a bus or car to reach locations of interest. To identify locations of interest, counts of alightings were aggregated by stop. After the most popular stops for alightings had been identified, the Google Maps API was used to calculate the time and distance necessary to reach each location of interest. Distance and time are two of the most important factors when travelers choose a mode of transport, and these controls were designed to capture this effect. To better understand how they worked, Table 3-1 provides an example.

**Table 3-1** Travel Time Variables

| Prefixes | Description |
|---|---|
| busTime/driveTime | Travel time in seconds for bus or car |
| busDist/driveDist | Distance in meters for bus or car to travel |
| Morn/Aft | Morning or afternoon |
| Origin/Dest | Whether time/distance is to or from a location |
| busTimeAftOriginChinook | Afternoon bus travel time from Chinook to given bus stop |
| Locations | Description |
| Beasly | Beasly Coliseum |
| Sloan | University building |
| GrandMain | Intersection of streets Grand and Main |
| Spark | University building |
| Dissmores | Grocery store |
| TerreViewFairway | Intersection of streets TerreView and Fairway |
| Vogel | University building |
| Walmart | Grocery store |
| SRC | University building |
| ValleyStadium | Intersection of streets Valley and Stadium |
| CUB | University building |
| Safeway | Grocery store |
| Merman Valley | Intersection of streets Merman and Valley |
| SEL | Place of work, engineering firm |
| Highschool | Pullman High School |

Socio-demographic data are often used when bus ridership trends are analyzed. For instance, population density near a stop is believed to have a direct relationship with ridership. Most socio-demographic variables used in this analysis came from the U.S. Census Bureau American Community Survey. A total of 12 types of Census data were used in this analysis to describe different characteristics of each respective block group. The rationale for the inclusion of these variables is in the literature review, but they were broadly thought to have a relationship with ridership. From this selection other characteristics were created to explore each block group further. For example, the count of unemployed people divided by the labor force yielded the unemployment rate for each block group. Employment characteristics are thought to be especially important for predicting ridership, given that public transport serves as a means of commuting. Additional data that described the number of jobs at the block group level came

from the U.S. Census Bureau at Longitudinal Employer Household Dynamics (LEHD) web page. These data were enumerated by the 2010 census block.

To control for the effects of the natural and built environments on ridership at each bus stop, three different data sources were utilized. Walkability and bike-ability are important factors when mode of transport is considered. Those factors include the presence of sidewalks, bike lanes, and the distance and density of nearby amenities or locations of interest. The WalkScore index, developed by a private company of the same name, provides a number from 0 to 100 for any address summarizing these factors. For each bus stop, WalkScore and BikeScore numbers were obtained through the company's API. Another feature at each stop is seating and shelter. These variables were provided by PT, and they are thought to be positively associated with ridership. Another type of environmental variable considered in this analysis was weather. Weather data were gathered from the National Oceanic and Atmospheric Administration and included hourly precipitation, wind speed, daily snow, and daily snow depth. The last set of environmental data considered included counts of types of places near the bus stops. The Google Maps API was utilized to count the number of cafes, grocery stores, etc. within 1/8-, 1/4-, and 1/2-mile radiuses around each stop. Table 3-2 details which places were used.

**Table 3-2** Ridership and Socio-Economic Variables

| Variable | Measurement | Description |
|---|---|---|
| rid | Stop | Sum of boardings and alightings |
| income | Census block | Average income |
| population | Census block | Total population |
| incomePovertyRatio | Census block | Income to poverty ratio |
| degreePopOver25 | Census block | Number of people with a degree above age 25 |
| enrolledOver3 | Census block | Number of people enrolled in school above age 3 |
| ownerNoVehicle | Census block | Number of homeowners without a vehicle |
| renterNoVehicle | Census block | Number of renters without a vehicle |
| owner | Census block | Number of homeowners |
| renter | Census block | Number of renters |
| employed | Census block | Number of employed (place of residence) |
| unemployed | Census block | Number of unemployed |
| labor_frc | Census block | Employed + Unemployed/Population |
| med_age | Census block | Median age |

| Variable | Measurement | Description |
|---|---|---|
| med_house_val | Census block | Median house value |
| n_jobs | Census block | Number of employed people (workplace) |
| n_jobs\textless{}29 | Census block | Number of employed people under age 29 |
| n_jobs30_54 | Census block | Number of employed people between 30 and 54 |
| n_jobs\textgreater{}55 | Census block | Number of employed people over age 55 |
| walkscore | Stop | Walkability |
| bikescore | Stop | Bike-ability |
| stopRouteVehicleFreq | Stop | Number of times bus services bus stop by day |
| stopRouteFreq | Stop | Number of time route services bus stop by day |
| stopFreq | Stop | Number of busses servicing bus stop by day |
| Shelter | Stop | 1 if shelter, else 0 |
| Simme Seat | Stop | 1 if seat, else 0 |
| gas_cpi | U.S. City avg | Gas Consumer Price Index |
| gas_pct_diff | U.S. City avg | Month over month percent change in gas_cpi |
| daily_snowfall | City Weather station | Daily snowfall (Inches) |
| daily_snowdepth | City Weather station | Daily snow depth (Inches) |
| DailyDryBulbTemperature | City Weather station | Daily temperature (Fahrenheit) |
| DailyPrecipitation | City Weather station | Daily precipitation (Inches) |
| DailyWindSpeed | City Weather station | Daily Windspeed (MPH) |
| Month | Time | Month of observation |
| Week | Time | Week of observation |
| DOY | Time | Day of year |
| DOW_num | Time | Day of week number (0 – M, 4 – F) |
| DOM | Time | Day of month |

The objective of this analysis was to estimate a predictive transit demand model to 1) better understand network characteristics that affect transit demand (wait time, headway time, stop proximity, etc.) and 2) provide an endogenous demand model for use in transit network optimization.

Before the machine learning models were developed, it was necessary to understand the models used and their benefits. There are three major types of algorithms: supervised learning, reinforcement learning, and unsupervised learning. Supervised learning models are used where the variables are labeled and can be predicted in regression or classification problems, given another set of variables. Unsupervised learning models are more useful when data are unlabeled and the model can self-discover any naturally occurring patterns. Reinforcement learning

methods assign positive values to the desired attributes to encourage the model and assign negative values to undesired attributes (Ray 2017). Before a model is chosen, one must consider the objective of the study, the nature of the data being used, and the desired accuracy of the model. The objective of this study was to predict ridership, so the appropriate analysis for this paper was a supervised regression algorithm because the target variable, ridership, was known.

To test accuracy and avoid overfitting, data were split into testing and training sets. A problem can occur when a model is trained heavily on the data of a training set, causing it to achieve excellent in-sample prediction but poor out-of-sample prediction. It is imperative that predictive models perform well for both in-sample and out-of-sample predictions.

To help improve prediction accuracy, many feature transformations were applied to the census data. These transformations resulted in a more complex and likely more predictive model, but they also increased the potential for overfitting. One method used to combat overfitting is called regularization. The regularization model used in this study was called the "least absolute shrinkage and selection operator" or LASSO (Tibshirani 1996). LASSO, or L1, regularization works by applying a penalty function to a model's loss term. This has the effect of reducing the coefficients of non-important variables to zero, leading to a simpler model.

To tune hyperparameters, a Bayes search with cross-validation was used. In cross-validation, several splits or "folds" are made on the data; the model is run on each fold, and then an average of the folds is taken to obtain an overall error estimate. Briefly, a Bayes search finds the minimum to an objective function in a large problem-space. In this case, the objective was to arrive at the best model output given the variables included, so it randomly tried different combinations and returned the combination with the greatest validation score. The validation score used was mean absolute error, which was obtained by comparing predicted and actual estimates within the training set. Grouping was used to prevent the same set of stops from being used in each of the folds, which might bias the estimates toward a particular set of stops. These steps make up the foundation for machine learning models to be fitted to the data.

Decision-tree (DT) and random forest (RF) algorithms were used in this analysis. Tree-based methods involve segmenting the predictor space into a number of simple regions (Venables and Ripley 1999). The motivation for using regression trees is that they are easily interpretable while they also vastly improve prediction accuracy. DTs work by taking each observation and partitioning an explanatory variable into different subsets. In the example in

Figure 3-4, the decision tree first splits on income and predicts that ridership will equal 11 for observations where income is less than $50,000. Then, another split is made for when the day of the week is Monday. The prediction here finds observations on Monday in block groups that have more than $50,000 income, and a prediction is generated given these two conditions. This process goes on until a stopping criterion, such as minimum number of observations per leaf or maximum depth of the tree, is met.

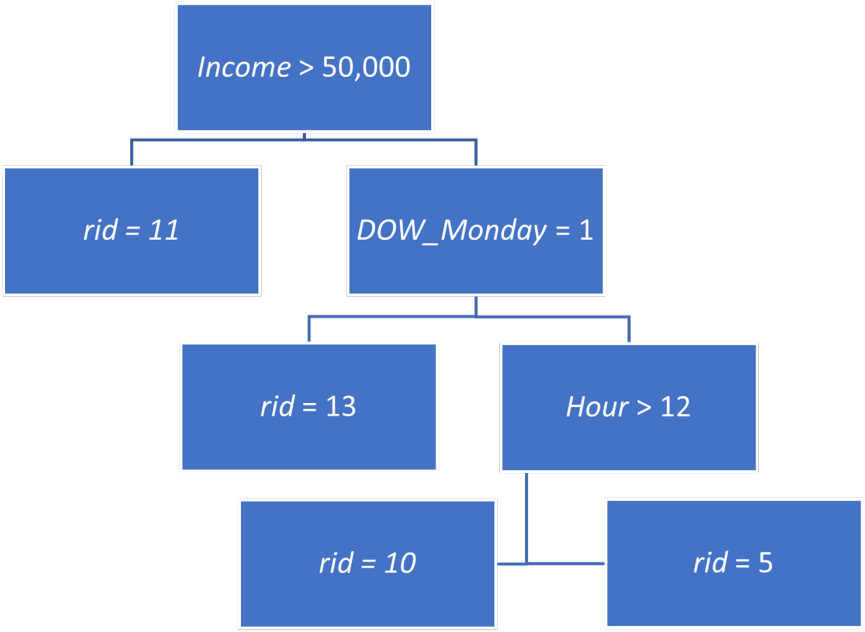


**Figure 3-4** Decision Tree

Finally, the DT stops growing when a leaf or branch node has less than a minimum number of observations, or a maximum depth has been reached. Setting the minimum number of observations required at a leaf node or setting the maximum depth of the tree are necessary to avoid overfitting the model. This process is called hyperparameter tuning, and in this model, it was performed by defining a set of values for the Bayes search algorithm to search over. After running hundreds of times, the model will have tried many different combinations for parameters such as maximum depth, and it outputs the best parameters to use when testing for out-of-sample predictions. A final model with exact hyperparameters is selected when training and testing accuracy are roughly equivalent.

The RF model is a bagging method that utilizes the aggregation of several decision trees to make a final prediction (Breiman 2001). Bagging is short for “bootstrap aggregation,” and it

works by randomly sampling from the training data with replacement, which further prevents overfitting. RF is a meta-estimator, meaning it simply uses the process of creating DTs but aggregates the predictions of each one. However, an additional feature of the RF model that makes it distinct is that it limits the number of features that can be split at each node to some percentage of the total. This hyperparameter ensures that no one feature is relied on too heavily.

## CHAPTER 4. FINDINGS

To assess each model, it was necessary to evaluate its in-sample and out-of-sample predictive accuracy. Accuracy was measured in two ways: pseudo-R2 and root mean squared error (RMSE). A pseudo R-squared is useful only when it is compared to another pseudo R-squared predicting the same outcome with the same data. A higher value for pseudo R-squared indicated better prediction of ridership. RMSE is another useful tool for examining predictive power. It is defined as the square root of the squared difference between observed and predicted values.

After the decision tree had been tuned by allowing a Bayes search cross-validation to search over hundreds of possible hyperparameters, its predictive performance was found to be better than that of the Poisson model. However, to quote from *Elements of Statistical Learning*, "Trees have one aspect that prevents them from being the ideal tool for predictive learning, namely inaccuracy" (Hastie et al. 2009). In other words, the in-sample accuracy was much better, but the DT proved to not be as flexible out-of-sample, with a pseudo R-squared of .31 and an RMSE of 42.87. This coefficient for RMSE means that where ridership was predicted, it was off on average by 42.87 riders at a stop.

**Table 4-1** In-Sample Prediction Performance

| Accuracy | Decision Tree | Random Forest |
|---|---|---|
| RMSE | 23.12 | 22.39 |
| Pseudo R2 | 0.72 | 0.73 |

**Table 4-2** Out-of-Sample Prediction Performance

| Accuracy | Decision Tree | Random Forest |
|---|---|---|
| RMSE | 42.87 | 37.63 |
| Pseudo R2 | .31 | .47 |

SHAP values, an acronym for Shapley Additive Explanations, help break down a prediction to show the impact of each feature (Figure 4-1). For machine learning models like DTs and RF, this is useful because the depth of a tree can make it hard to interpret which features have the greatest impact on prediction. SHAP values interpret the impact of having a chosen value for a given feature in comparison to the prediction made if that feature was some baseline

value. The top features used for prediction in the RF model were *busTimeAftDestSafeway* and *renterNoVehicle*. The first variable captured the time for a bus to reach the local Safeway supermarket, while *renterNoVehicle* described the number of renters in a block group that did not own a vehicle.


**Figure 4-1** SHAP (Shapley Additive Explanations) Values

# CHAPTER 5. CONCLUSIONS

Public transit plays a crucial role in reducing the externalities associated with automobile use such as pollution, congestion, and traffic accidents. Encouraging bus ridership is especially important because other modes of public transport have large startup costs and require greater population density, making small towns infeasible locations for them.

In this project predictive models were developed to serve as the foundation of a decision support tool to help local transit authorities make better transit service decisions. From this analysis it was clear that machine learning is a viable approach for ridership prediction when complex datasets make estimation difficult. The random forest algorithm was demonstrated to be most effective at out-of-sample predictive accuracy in comparison to the alternatives. However, the algorithms did not produce high enough predictive accuracy to be useful in the context of a decision support tool. Alternative levels of aggregation and other methods for prediction have proved to be more effective in terms of predictive accuracy (Li 2020; Dill et al. 2013; Frei and Mahmassani 2013; Taylor et al. 2009).

## REFERENCES


Berrebi, Simon J, Sanskruti Joshi, and Kari E Watkins. “On Bus Ridership and Frequency.” Transportation Research. Part A, Policy and Practice 148 (2021): 140–54. https://doi.org/10.1016/j.tra.2021.03.005.

Boisjoly, Geneviève & Grisé, Emily & Maguire, Meadhbh & Veillette, Marie-Pier & Deboosere, Robbin & Berrebi, Emma & El-Geneidy, Ahmed, 2018."Invest in the ride: A 14 year longitudinal analysis of the determinants of public transport ridership in 25 North American cities," Transportation Research Part A: Policy and Practice, Elsevier, vol. 116(C), pages 434-445. https://ideas.repec.org/a/eee/transa/v116y2018icp434-445.html

Breiman, Leo. “Random Forests.” Machine Learning 45, no. 1 (2001): 5–32. https://doi.org/10.1023/A:1010933404324.

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

Chakrabarti, Sandip. “The Demand for Reliable Transit Service: New Evidence Using Stop Level Data from the Los Angeles Metro Bus System.” Journal of Transport Geography 48 (2015): 154–64. https://doi.org/10.1016/j.jtrangeo.2015.09.006.

Chen, C., Varley, D., & Chen, J. 2011. “What Affects Transit Ridership? A Dynamic Analysis involving Multiple Factors, Lags and Asymmetric Behaviour,” Urban Studies, 48(9), 1893–1908. https://doi.org/10.1177/0042098010379280

Dill, Jennifer, Marc Schlossberg, Liang Ma and Cody Meyer. “Predicting Transit Ridership at Stop Level: Role of Service and Urban Form.” (2013).

Frei, Charlotte, and Hani S Mahmassani. “Riding More Frequently: Estimating Disaggregate Ridership Elasticity for a Large Urban Bus Transit Network.” Transportation Research Record 2350, no. 2350 (2013): 65–71. https://doi.org/10.3141/2350-08.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. Elements of Statistical Learning. New York: Springer, 2009.

Li, L. Bai, W. Liu, L. Yao and S. T. Waller, 2022. "Graph Neural Network for Robust Public Transit Demand Prediction," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4086-4098. doi: 10.1109/TITS.2020.3041234.

Schwab, D., Ray, S. 2017. “Offline reinforcement learning with task hierarchies,” Mach Learn 106, 1569–1598. https://doi.org/10.1007/s10994-017-5650-8

Taylor, Brian D, Douglas Miller, Hiroyuki Iseki, and Camille Fink. “Nature And/or Nurture? Analyzing the Determinants of Transit Ridership Across US Urbanized Areas.” Transportation Research. Part A, Policy and Practice 43, no. 1 (2009): 60–77. https://doi.org/10.1016/j.tra.2008.06.007

Tibshirani, Robert. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society. Series B (Methodological) 58, no. 1 (1996): 267–88. http://www.jstor.org/stable/2346178.

Venables, W. N., and Ripley, Brian D. Modern Applied Statistics with S-PLUS. 3rd ed. New York: Springer, 1999.

## BIBLIOGRAPHY


Desaulniers, Guy, and Mark D Hickman. “Chapter 2 Public Transit.” Handbooks in Operations Research and Management Science 14 (2007): 69–127. https://doi.org/10.1016/S0927-0507(06)14002-5.

Ding, Chuan, Donggen Wang, Xiaolei Ma, and Haiying Li. “Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees.” Sustainability (Basel, Switzerland) 8, no. 11 (2016): 1100. https://doi.org/10.3390/su8111100.

Fabrikant, “Predicting bus delays with machine learning”. Google AI Blog. (2019) https://ai.googleblog.com/2019/06/predicting-bus-delays-withmachine.htm

Fontes, T^ania, Ricardo Correia, Joel Ribeiro, and Jos´e Lu´ıs Borges. “A Deep Learning Approach for Predicting Bus Passenger Demand Based on Weather Conditions.” Transport and Telecommunication 21, no. 4 (2020): 255–64. https://doi.org/10.2478/ttj-2020-0020.

Kawatani, Takuya, Tsubasa Yamaguchi, Yuta Sato, Ryotaro Maita, and Tsunenori Mine. “Prediction of Bus Travel Time over Intervals Between Pairs of Adjacent Bus Stops Using City Bus Probe Data.” International Journal of ITS Research 19, no. 2 (2021): 456–67. https://doi.org/10.1007/s13177-021-00251-8.

National Academies of Sciences, Engineering, and Medicine. 2020. Analysis of Recent Public Transit Ridership Trends. Washington, DC: The National Academies Press. https://doi.org/10.17226/25.

Pedregosa, Fabian, Gaeel Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. “Scikit-Learn: Machine Learning in Python.” Journal of Machine Learning Research 12 (2011): 2825–30. https://doi.org/10.5555/1953048.2078195.

“Rural Transit Fact Book.” SURCOM - Rural Transit Fact Book, North Dakota State University, 2021, https://www.ugpti.org/surcom/resources/transitfactbook/.

Seabold, Skipper, and Josef Perktold. “statsmodels: Econometric and statistical modeling with python.” Proceedings of the 9th Python in Science Conference. 2010.

Stover, Victor W, and Edward D McCormack. “The Impact of Weather on Bus Ridership in Pierce County, Washington.” Journal of Public Transportation 15, no. 1 (2012): 95–110. https://doi.org/10.5038/2375-0901.15.1.6.

Wooldridge, Jeffrey M., 1960-. Introductory Econometrics: a Modern Approach. Mason, Ohio :South-Western Cengage Learning, 2012.