NONMOTORIZED TRAFFIC MONITORING AND CRASH ANALYSIS

Final Report

PROJECT SPR 813

Oregon Department of Transportation

NONMOTORIZED TRAFFIC MONITORING AND CRASH ANALYSIS

Final Report

PROJECT SPR 813

by
Josh Roll,
ODOT Research Coordinator and Data Scientist

for

Oregon Department of Transportation
Research Section
555 13th Street NE, Suite 1
Salem OR 97301

and

Federal Highway Administration
1200 New Jersey Avenue SE
Washington, DC 20590

January 2021

Technical Report Documentation Page

1. Report No. FHWA-OR-RD-21-08	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Nonmotorized Traffic Monitoring and Crash Analysis		5. Report Date January 2021 6. Performing Organization Code
7. Author(s) Josh Roll , https://orcid.org/0000-0002-9617-5045		8. Performing Organization Report No.
9. Performing Organization Name and Address Oregon Department of Transportation Research Section 555 13th Street NE, Suite 1 Salem, OR 97301		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Oregon Dept. of Transportation Research Section 555 13th Street NE, Suite 1 Salem, OR 97301 Federal Highway Admin. 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract: This report documents the collaboration between ODOT and Bend MPO to develop and execute a nonmotorized traffic data collection program. Using automated technology and data processing techniques, this research demonstrates a least cost path to developing a nonmotorized traffic count program for a small urban community. Data processing techniques are developed to ease the burden of labor on staff through automatic data retrieval, processing and quality control using open source tools. The research also demonstrates how to utilize collected data for system wide monitoring using various data fusion techniques. Finally, these estimates of activity are used to demonstrate the disparate risk to nonmotorized users by performing crash analyses. The program and related elements in this report should highlight a nonmotorized data collection model for other urban areas to follow.

17. Key Words	18. Distribution Statement
bicycle, pedestrian, traffic counts, data imputation, machine learning, data fusion, crash rates, traffic count program, disparity, inequity, data processing, traffic count program costs	Copies available from NTIS, and online at www.oregon.gov/ODOT/TD/TP_RES/

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	183	

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	millimeters squared	mm^2
ft^2	square feet	0.093	meters squared	m^2
yd^2	square yards	0.836	meters squared	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	kilometers squared	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	ml
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	meters cubed	m^3
yd^3	cubic yards	0.765	meters cubed	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	Mg
		TEMPERATURE (exact)		
°F	Fahrenheit	(F-32)/1.8	Celsius	°C

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	millimeters squared	0.0016	square inches	in^2
m^2	meters squared	10.764	square feet	ft^2
m^2	meters squared	1.196	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	kilometers squared	0.386	square miles	mi^2
		VOLUME		
ml	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	meters cubed	35.315	cubic feet	ft^3
m^3	meters cubed	1.308	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.205	pounds	lb
Mg	megagrams	1.102	short tons (2000 lb)	T
		TEMPERATURE (exact)		
°C	Celsius	1.8C+32	Fahrenheit	°F

*SI is the symbol for the International System of Measurement

ACKNOWLEDGEMENTS

Oregon Department of Transportation would like to acknowledge the Bend Metropolitan Planning Organization staff, especially Jovi Anderson, do their contribution to this research. This research effort would not have been possible without close collaboration with these partners.

DISCLAIMER

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

This research report documents the implementation of a nonmotorized traffic counts program in Bend, Oregon as a part of a partnership between Oregon Department of Transportation and Bend Metropolitan Planning Organization. ODOT's research staff partnered with Bend MPO staff to devise data collection protocols for nonmotorized traffic data collection followed by three years of data collection. In addition to collected traffic counts data, this research developed a high quality fully attributed bicycle network dataset used for determine count locations but more importantly used in data fusion modeling and crash analysis.

The relevant content of the report is broken into 11 chapters covering every element of the work performed in this research. Chapters 2 through 4 cover core components of the data collection and processing phase of the research. Chapters through 5 through 8 document daily traffic count imputation methods and data fusion techniques developed for this research. Finally, chapters 9 through 13 report on the crash analysis using the counts based travel activity estimation.

This work will ideally paint a complete picture for why nonmotorized traffic count and network data are important data elements for public agencies to collect as they can be used to highlight the disparate risk faced by nonmotorized users of the transportation system. Raw frequencies of nonmotorized crash injuries tell only a small part of the story and only when these injuries are normalized using estimated exposure measures is the true state of the nonmotorized system revealed. When the crash injury risk for nonmotorized users are orders of magnitude higher than motorized traffic public agencies will continue to convince more than the most dedicated or vulnerable people to use the transportation system. The reality of the disparate risk makes meeting safety, air quality, livability, and climate goals unlikely and should be a key component of communication strategies for why projects to improve safety for nonmotorists are so important.

1.1 DATA COLLECTION AND PROCESSING INTRODCUTION

Chapters report 2 through 7 document the following tasks of the SPR 813 adopted work plan including:

- Task 1: Data Collection Strategy
- Task 2: Data Collection and Compilation
- Task 3: Data Processing

These tasks have been completed though data collection is ongoing and will continue after the completion of this research project by local agency staff. These tasks are fundamental building blocks for the latter analysis tasks but should be helpful on their own for practitioners looking for guidance on these elements of nonmotorized data collection programs.

This research project is being executed in conjunction with agency staff from the Bend Metropolitan Planning Organization (MPO) and has relied on those staff and their contractor for much of the data development and collection. This arrangement represents a novel approach to conducting a project for the ODOT Research Program. Benefits from this arrangement include additional funding and staff resource from the Bend MPO for elements of the project which allow the ODOT research funded effort to do more work with less direct funding. A second major benefit of this arrangement is having a clear line to implementation since the Bend MPO is currently working on major planning efforts including a Transportation System Plan and a Transportation Safety Action Plan. This uncommon arrangement is not without its limitations however since Bend MPO began some of the associated work before the beginning of this research project some efforts are underway and so the research project has less ability to make changes. These limitations are not deal breakers however and the current state of the project is yielding significant benefits for advancing nonmotorized count programs and related analyses in the state of Oregon

The Oregon Department of Transportation's recently released Bicycle and Pedestrian Mode Plan recognized the lack of data in non-motorized transportation planning. Recent bicycle and pedestrian safety research completed by ODOT's Research Section found it difficult to interpret final results for many elements of their efforts due to a lack of traffic counts for these modes. The recently published report SPR 778 Safety Effectiveness of Pedestrian Crossing Enhancements found that "the estimation of the safety effectiveness of pedestrian treatments was challenging due to… the general lack of reliable pedestrian counts". Another recently published ODOT Research Report SPR 779 Risk Factors for Pedestrian and Bicycle Crashes concluded that, "The identification of risk factors and the magnitude of their influence on the likelihood of future crashes were significantly constrained by limited roadway information.

1.2 STUDY AREA DETAILS

The Bend MPO is located near the center of the state and comprises a population of roughly 94,500 people as of 2017 according to U.S. Census figures. The MPO area is bisected by two main highways, highway 97 and highway 20 which are access controlled in many places. The region has 50.5 miles of off-street bicycle and pedestrian path as well as over 120 miles of on-street bike lanes. Figure 1.1 below shows the Federal Urbanized Boundary (FAUB) which coincides with the MPO boundary and constitutes this research projects study area. The map also shows the location of bicycle facilities across the transportation network. Table 1.1 below summarizes the number of miles of bicycle facilities by federal functional classification.

Table 1.1: Summary of Study Area Travel Network

Functional Classification	No Bicycle Facility	Bike Lane	Off-street path	Total
Local	463.1	3.4	50.5	517
Minor Collector	2.8	1.1	-	3.9
Major Collector	17.8	31.2	-	49
Minor Arterial	5.8	54.9	-	60.7
Other Principle Arterial	4.2	38.3	-	42.5
Total	493.7	128.9	50.5	673.1

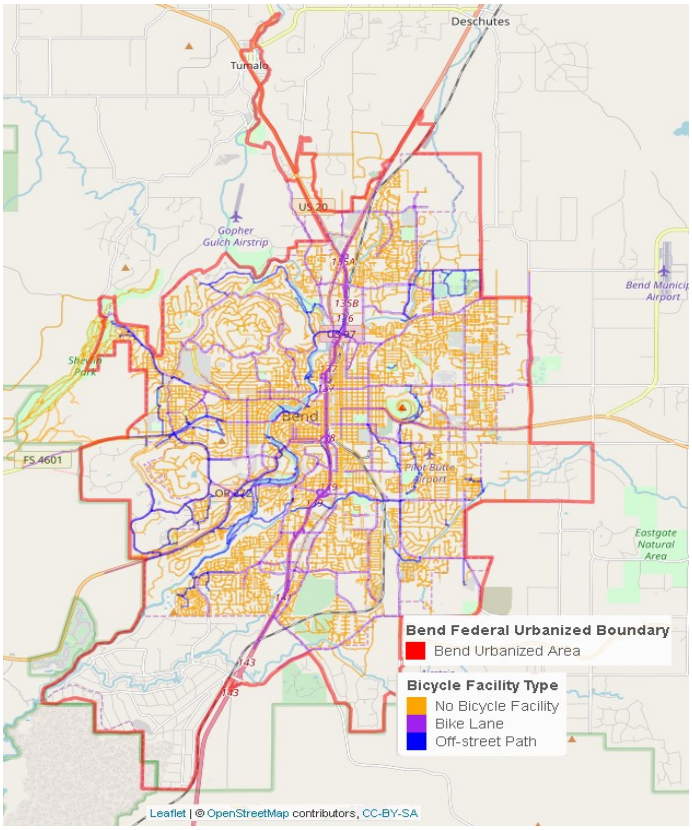

Figure 1.1: Study area boundary with bicycle facilities

The city and MPO are currently going through a Transportation System Plan (TSP) update with some focus developing a project list of bicycle and pedestrian improvements. The aim of these planning efforts is in part to increase the number of people who walk and bicycle for travel purposes. In order to achieve those goals additional information on nonmotorized travel behavior is needed for the area.

1.3 RESEARCH OBJECTIVES

This research seeks to fill gaps in key measures of performance for walking and bicycling by furthering methods for estimating total activity for these modes using traffic counts. These measures of activity can fill basic metrics of performance across the system and help monitor changes over time including those occurring in response to system upgrades.

In addition to fundamental measures of travel activity for people who walk and bicycle, this research would seek to analyze safety outcomes for these modes by utilizing the activity measures in crash rate development. Crash rates allow engineers, planners and other practioners better metrics for understanding facilities and street configurations with higher risk, and help deliver a key performance measure of safety outcomes.

2.0 DATA COLLECTION STRATEGY

This section describes the data collection strategy employed by this research effort and executed in conjunction with Bend MPO staff and the ongoing data collection program housed with that agency. Since the Bend MPO staff had begun data collection before the start of this research project it was necessary to adapt the data collection strategy to some of the work already underway. Some of the initial data collection elements including equipment purchases were funded in part by Oregon Department of Transportation grants through the Transportation Records Coordinating Committee (TRCC). Overall, this collection of efforts represents a novel partnership between the state DOT and MPO in Oregon and can serve as a model for initiating future non-motorized traffic count programs.

2.1 EXISTING AND FUTURE BEND MPO COUNT PROGRAM GOALS AND PRIORITIES

2.1.1 Existing Count Program Priorities

A plan was developed and adopted before the start of this research effort and guided the formation of the traffic count program including where data would be collected (KA 2016). Before the start of this count program no systemic nonmotorized traffic data collection was being performed though vehicle counts are collected for Highway Performance Monitoring System (HPMS) purposes by ODOT. This Bend MPO traffic count plan lays out five key needs the data collection plan would satisfy including:

- Monitor use and trends
- Measure project success
- Plan for the future
- Prioritize maintenance activities and operations
- Improve safety analysis.

With these uses in mind the locations selected for data collection included streets that existed at traffic bottlenecks like bridges and underpasses as well as locations with planned infrastructure improvements.

2.1.2 Count Program Priorities Going Forward

In addition to current priorities count data collection should help inform higher level analyses including modeling of total nonmotorized traffic activity. A primary feature of this research will be to offer additional information on data collection strategies for that purpose. Future steps in this research project will utilize nonmotorized counts in a direct demand model which seeks to

estimate total bicycle and pedestrian travel activity across the transportation network. The direct demand approach uses a statistical model where traffic counts are a function of a number of inputs including the type of roadway, the presence of a bicycle facility, and the accessibility to jobs and employment as well as some connectivity measures.

The use of counts data for this purpose was not originally envisioned for the Bend MPO count program but can support it nonetheless. Below is a map of the current distribution of traffic count locations and one measure of access to jobs. The direct demand model will use the observed relationships at these sites and the underlying access measure to forecast travel activity to the rest of the street network where no counts have been collected.

Future priorities for traffic count data collection would aim to collect enough count observations I where other attributes, like accessibility, vary to a degree that supports strong statistical strength. In estimated models. For instance in the below figure, the counts seem reasonably distributed across the region where access to jobs varies, as opposed to the counts being concentrated in the high access core of the downtown. This should help make the direct demand models more reliable. In future interim reports the direct demand model method will be discussed at length with a lengthy examination of past approaches, including one analysis completed by the ODOT Research Program where the method was applied in the Central Lane MPO (ODOT 2018).

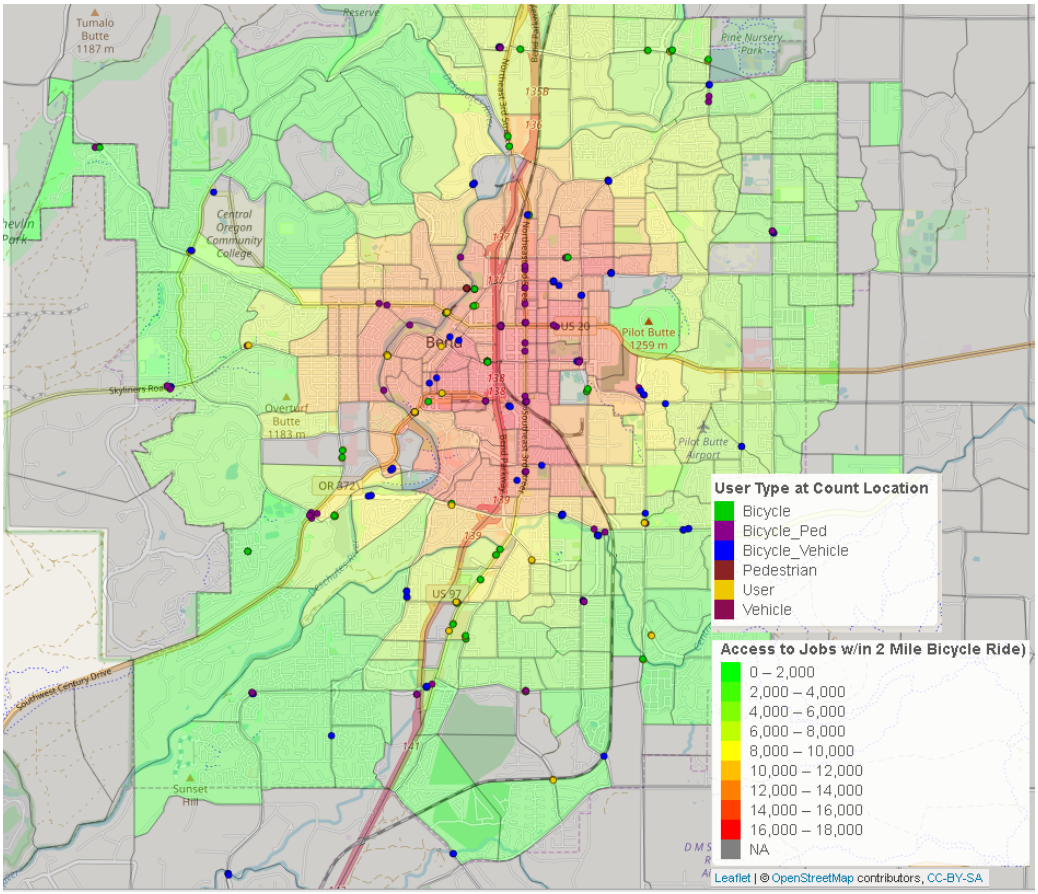

Figure 1.1: Access to jobs and nonmotorized count locations

3.0 DATA COLLECTION AND COMPILATION

This section describes the equipment, procedures, and data processing protocols established in the Bend MPO for collecting and storing traffic counts. The count program uses a combination of permanent and mobile counts that utilize various technologies including inductive loops, infrared, tubes, and pressure sensors. Paid contract staff are utilized to collect data using mobile counting equipment and a procedure was constructed that uses a cloud based spreadsheet to record the deployment and pick up dates and times for each device deployed. Traffic data is sent via cell phone connection from the counting device to a cloud based data repository and later combined with deployment information to process the data and combine with spatial attributes for a final usable format. This process was designed to reduce the potential for human error and also reduces data handling which should save staff time and associated costs.

3.1 NONMOTORIZED TRAFFIC COUNTING EQUIPMENT

This research and the developing data collection program utilize multiple types of data collection devices and are summarized below. These devices collect vehicle, bicycle or pedestrian traffic though some collect both without distinguishing between the user types. These counts are termed 'users' and denote the aggregation of both bicycle and pedestrian users for the purposes of this report. The following describes the equipment used in the study area's count program with deployment examples. Additionally, discussion of the likely inaccuracy of the devices is presented along with the results of validation tests performed on some of the devices in the study area.

3.1.1 Inductive Loop

Inductive loops detectors use induced current detection system that detects when metal objects cross over the in-ground loop or wire permanently installed in the ground. Bicycle traffic count data are recorded with these devices but does not have the capability of recording pedestrian traffic. Inductive loop hardware made by Eco-Counter© have been found to be accurate with as little as 0.4% error when counting in off-street conditions (Munro 2015) and up to 5.0% (Norback 2011) in on-street conditions. The inductive loop hardware for this research has all been manufactured by Eco Counter and is installed in both an off-street and on-street setting. When the loops are collecting in an on-street setting they have been placed in the bike lane to minimize issues where the loops count vehicles as bikes. Figure 3.1 below shows an installation of an inductive loop detector in an off-street setting at Franklin Avenue where no vehicle traffic can access the loops while Figure 3.2 shows the installation of the inductive loops in a bike lane on Galveston Avenue.

Figure 3.1: Separated inductive loop detector at Franklin Avenue in Bend MPO

Figure 3.2: On-street inductive loop detector at Galveston Ave. in Bend MPO

In addition to counting bicycles, inductive loops that count vehicles are also present in the study region.

3.1.2 SLABs

The SLAB detector uses pressure to detect the presence of both pedestrians and people riding bicycles and does not distinguish between users. The picture below shows the installation of the SLAB system on Colorado Avenue in the study area. No published validation studies could be found for this device type but a short validation evaluation was performed and is described in Table 3.1 and were shown to be relatively accurate.

Figure 3.3: SLAB detector at Colorado Avenue in Bend MPO

3.1.3 Passive Infrared

Passive infrared detector devices detect uses of a facility by measuring changes in ambient temperature of users compared to background radiation (heat) as the user moves through the detection zone. This study includes data from two vendors of passive IR devices including TRAFX and Eco Counter. The figure below shows a permanently installed IR Eco Counter device installed in the study area.

Figure 3.4: IR detector at Galveston Avenue in Bend MPO

The TRAFX trail counter is used by the Bend Parks and Recreation Department to collect users on trails and off-street paths in parks within Bend. The devices collect both pedestrians and bicycle users but does not distinguish between the two user types. Similarly, the Eco Counter PYRO infrared device does not distinguish between bicycle and pedestrian users but does offer an integrated product that pairs with either an inductive loop for permanent count sites or pneumatic tubes for mobile counting devices. Minnesota DOT (2015) performed validation evaluations of Eco Counters integrated bicycle and pedestrian counting devices and found the IR component was within 10% of the observed counts. The difference in this evaluation was an under count most likely due to the inability of the device to detect all pedestrians in a group, referred to as occlusion. The MnDOT report recommends developing correction factors. NCHRP Project 797 (Ryus, et al, 2014) tested two brands of passive IR sensors finding that the accuracy was ranges -3.1% and -16.7% for each product though did not describe which products were tested. Validation tests for the study area equipment are presented below and compare well with previous results.

3.1.4 Pneumatic tubes

Pneumatic tube counters detect bicycles using sensors that measure the pressure change in the tubes by an instrument in the recording device. These device types have been in practice for many decades for vehicle counting and are now being deployed for bicycle counting. This research relies on Eco Counter's pneumatic tube counters for all of the mobile site data collection. The Eco Counter pneumatic tube was tested by MNDOT and found to have error of 1.6% in an off-street setting (MNDOT 2015). Oregon Department of Transportation (ODOT) tested the Eco Counter tube counter using both bicycle only tubes and standard roadway tubes

finding the devise counted bicycle traffic with a reported 1.7% mean absolute percent error (MAPE).

Eco Counter also produces a product that combines the pneumatic tube device with an IR device in order to count both bicycle and pedestrian traffic. The hardware does the subtraction of bicycle users from the total user counts in order to calculate the pedestrian traffic at the location. These Eco Counter Multi units are also deployed in the study area and shown in the figure below.

Figure 3.5: IR and pneumatic detector combo device at Greenwood undercrossing in Bend MPO

3.1.5 Summary of Device Types and Deployment Locations

The map below shows the spatial distribution of the count locations and also details the device types and collection method with details views of the downtown and south east sections of town.

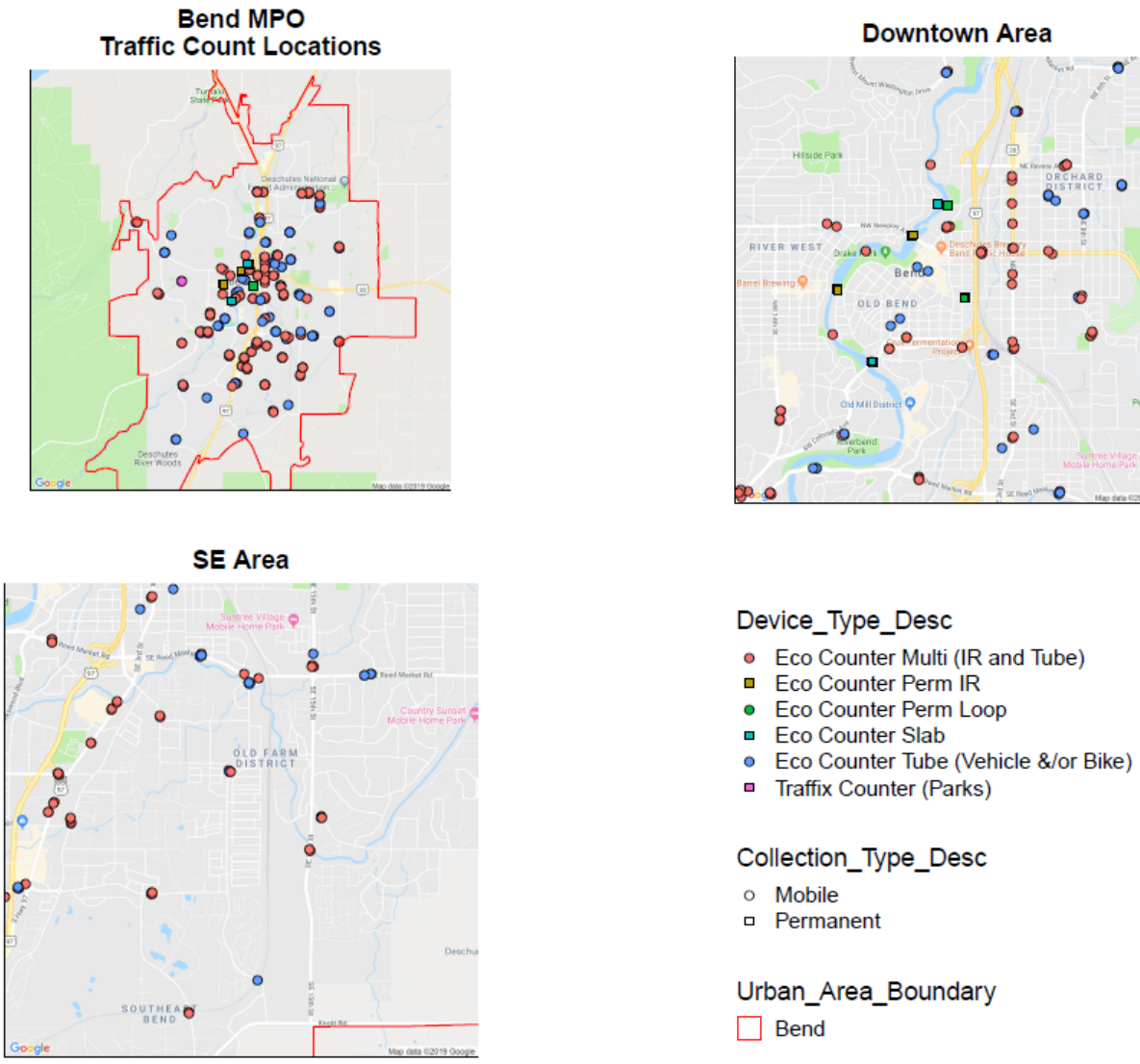

Figure 3.6: Count locations by device type and collection type

Table 3.1 below summarizes the number of locations by device type, collection type, user type and whether the traffic data is collected from a permanent count location or from a mobile device that can be moved to alternative locations. This table shows there are five types of devices being used including the Eco Counter Multi which collects bicycle, pedestrian, and vehicle counts data using tubes and IR in a combination system. There are also locations that use just an Eco Counter IR or Traffix device to collect traffic data for multiple user types. Where bicycle and pedestrian counts are collected with these devices a tube device is deployed as well in order to separate out the bicycle users form the pedestrians. The table also shows that loop devices as well as slabs are used in permanent locations. Total there are 195 device deployments in either a permanent or mobile basis. This does not mean that there are 195 locations being counted however, since multiple devices are needed at certain locations to collect all the traffic moving along the roadway. This is explained in more detail in section 3.3.1.

Table 3.1: Summary of Traffic Count Device, Collection, User Type

Device Type	Collection Type	User Type	# Locations
Eco Counter Multi (IR and Tube)	Mobile	Bicycle	12
Eco Counter Multi (IR and Tube)	Mobile	Bicycle/Ped	56
Eco Counter Multi (IR and Tube)	Mobile	Bicycle/Vehicle	9
Eco Counter Multi (IR and Tube)	Mobile	User	33
Eco Counter IR	Permanent	Bicycle	2
Eco Counter IR	Permanent	Pedestrian	2
Eco Counter IR	Permanent	User	8
Eco Counter Loop	Permanent	Bicycle	12
Eco Counter Loop	Permanent	Pedestrian	2
Eco Counter Loop	Permanent	Vehicle	10
Eco Counter Slab	Permanent	Pedestrian	2
Eco Counter Slab	Permanent	User	2
Eco Counter Tube (Vehicle &/or Bike)	Mobile	Bicycle/Vehicle	43
Traffix Counter (Parks)	Mobile	User	1
Traffix Counter (Parks)	Permanent	User	1

3.2 EQUIPMENT VALIDATION AND ACCURACY

In the sections above published validation evaluations were summarized for each device type along with their description. It would be expected that the equipment used in this research and supporting traffic counts program would function similarly. To be sure and to certify that permanent sites were constructed properly, limited validation tests were also performed for select locations. The table below summarizes the accuracy for four sites where validation evaluations were performed. The results show that accuracy for of the devices work well enough for Newport Bridge, Colorado Bridge and Galveston Bridge locations with minimum error of 0.0% up to 21.4% error but that error for the Franklin Underpass site were considerable with 533% for pedestrian traffic. These locations also collect vehicle counts and were shown to be relatively accurate with the Colorado Bridge location being nearly perfectly accurate while the Galveston location revealing 7.6% error. The Franklin Underpass location did not have directly comparable data since the counter was not online at the time of the observed data collection. However, a comparison of a similar time period reveals reasonable similarity in traffic counts. For this location and the bicycle counts and vehicle counts will be evaluated again to better understand how well the devices are performing.

Table 3.2: Validation Evaluation of Bicycle and Pedestrian Traffic Counters in Study Area

Location	Equipment Description	Data Collection Date/Time	North Sidewalk				South Sidewalk				Bikes in Road			Vehicles in Road		
			Mode	Observed	Eco	% Diff.	Mode	Observed	Eco	% Diff.	Observed	Eco	% Diff.	Observed	Eco	% Diff.
Newport Bridge	Loops in roadways, loops in bike lane, Eco Multi on sidewalk	5.16.2017 2-6 PM	Ped/bikes	35	31	-11.4%	Ped/bikes	56	42	-25.0%	27	26	-3.7%			
Colorado Bridge	Loops in roadways, loops in bike lane, slabs on sidewalk	5.16.2017 12-4 PM	Peds	7	6	-14.3%					11	10	-9.1%	5268	5266	0.0%
Franklin Underpass	Loops in roadways, loops on edge of roadway (bikes), Eco Multi and loops on sidewalk	5.16.2017 12-4 PM	Peds	37	39	5.4%	Peds	39	38	-2.6%	6	38	533.3%	3785	4570	20.74%
			Bikes	17	23	35.3%	Bikes	23	34	47.8%						
Galveston	Loops in roadways, loops in bike lane, Eco Multi on sidewalk*	5.16.2017 12-4 PM	Ped/bikes	32	32	0.0%	Ped/bikes	5	5	0.0%	28	22	-21.4%	4170	4486	7.6%

3.3 RAFFIC DATA COLLECTION

Traffic data is collected in two methods depending on whether the device is permanently installed equipment or mobile and able to be moved around the study area to collect at multiple locations. All of the Eco Counter devices transmit their data via wireless cell phone connection where the data is stored in a cloud based data repository available through a web portal or application programming interface (API). The mobile traffic counting devices are deployed using paid contract staff with details on the deployment maintained in a Google spreadsheet. The traffic data and the deployment information are combined in a custom R based software program that processes, cleans and stores the data for use by agency and research staff. The below section describes these processed in more detail.

3.3.1 Location and Site Setup

Counting bicycle and pedestrian traffic is more complicated than collecting vehicle counts. The expected zone of travel for vehicles is more certain, with vehicles traversing the roadway in a predicable fashion on a limited area of the right-of-way. For that reason bicycle and pedestrian traffic cannot be collected in locations without specific conditions.

Devices that use IR cannot face the device towards vehicle traffic where it may erroneously count a moving vehicle as a user. The same is true for pointing the IR devices towards parking lots where the warmth from the engine of a parked vehicle may also register as a pedestrian. The pneumatic tube counters that collect bicycle traffic data require are only able to accurately collect data on roads with widths of 30 feet or less which makes the deployment of multiple devices at a single location necessary. To count the total bicycle throughput at a given location, it's often necessary to collect traffic counts on the sidewalk and in the roadway which includes a bicycle facility like a bike lane. In order to do both sides where the roadway is greater than 30 feet in width, it's common to deploy up to four devices at a single location.

In order to fully account for all traffic on a travel network link the seemingly more complicated approach described above was necessary. There was not any published approach to managing bicycle and pedestrian traffic counts in the Traffic Monitoring Guide (2013) and so a system was devised and is presented below. It balances the inherent complications of collecting these data with an eye on simplicity for users that will be required to operate the data collection program in the future.

Traffic count locations are composed of sub locations where various modes of travel are collected for available directions of movement. The figure below attempts to describe a potential setup where bicycle, pedestrian, and vehicle traffic are all being collected simultaneously. In the example below the location (Location Id 001) has six sub locations collecting data using two loops, two IR, and two pneumatic tube devices. The two IR devices collect users (both people on bike and walking) that use the sidewalk while the two embedded loops collect bicycle traffic. These data are combined and the pedestrian traffic is calculated by removing the bicycle users from the total users collected by the IR device. Roadway bicycle and vehicle traffic is collected using the two tube counters.

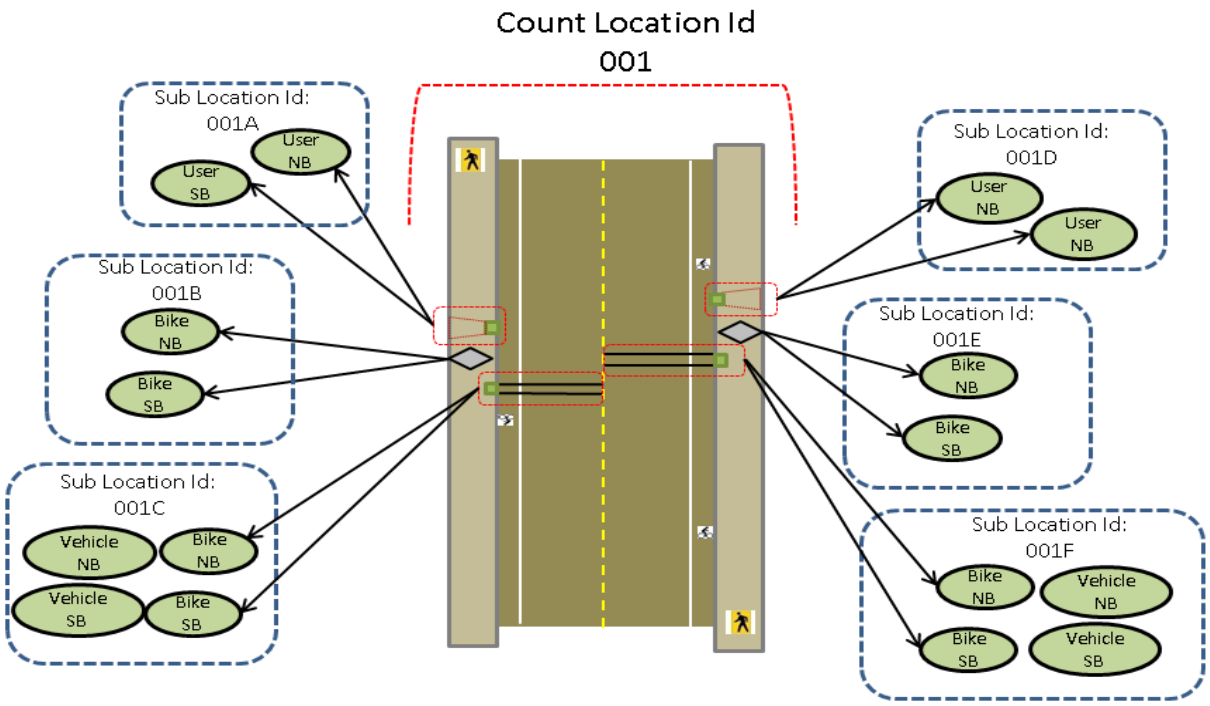

Figure 3.7: Example count location setup

Data from all of the counting devices is streamed on a daily basis to an online data repository (named Eco Visio) managed by the hardware vendor. These counts data can be retrieved through a web based platform with some reporting functions including the ability to make charts and compare locations. The counts data are also available through an application programming interface (API) which can be accessed through various computer programing tools. This research project has constructed a custom data processor built in the R open source statistical and programming language which has API functionality. For permanent count locations the processing is straight forward and only involves aggregating data by direction (for total link flows) and mode. For locations where mobile equipment is used exclusively the processing is more complex.

This process starts by first retrieving data retrieved from the online data repository through R using the API call and an API key purchased from the vendor. The counts data are then combined with information about a given device deployment so that the appropriate counts data are retrieved for the right time for the respective location since the online data base is agnostic about the location of the device. To clarify, a given mobile device might collect data in 10 locations throughout the year and those counts data are all stored in the online database without the location information or anything related to the deployment. The deployment information has details on when the location was at a given locations and for what time period so data can be extracted and assigned to the appropriate location. The R software also employ information stored in a geodatabase to help sort and process the locations properly in addition to adding attributes like the facility type and link level attributes such as the functional classification of the adjacent roadway. This process all happens automatically using the custom R software written for this effort.

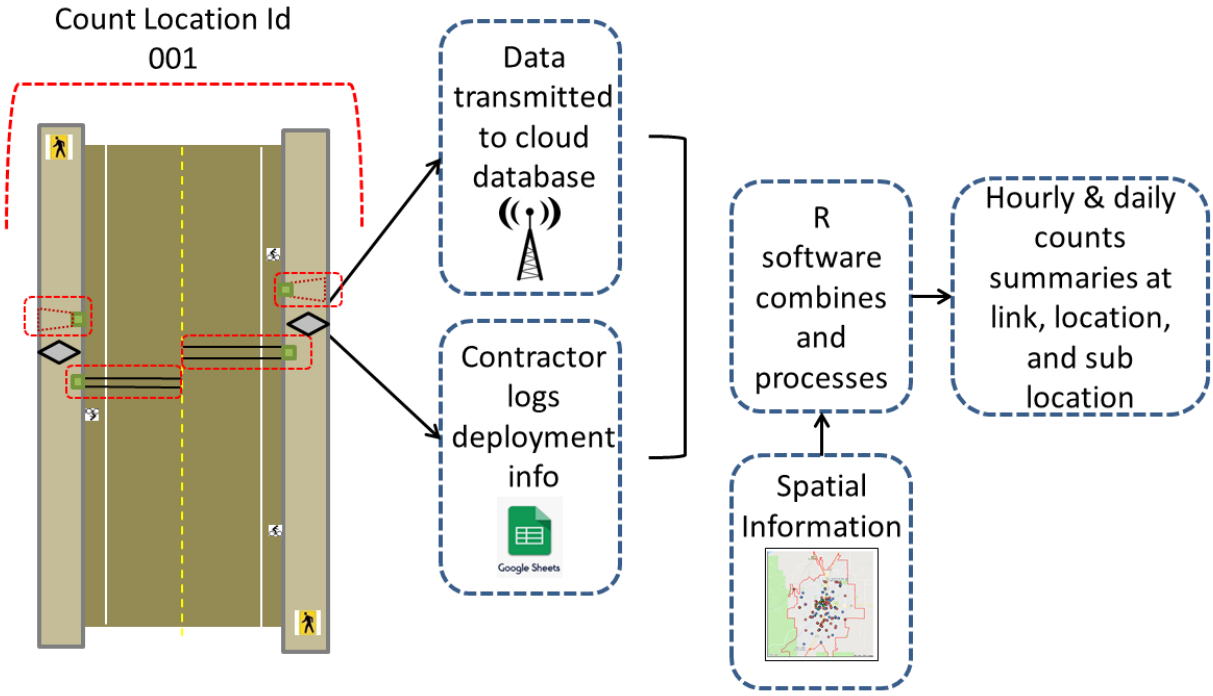

Figure 3.8: Traffic data transmission and processing schematic

Figure 3.9 below shows an excerpt of the deployment information entered into Google Sheets by the deployment contractors. The information includes the timestamp of the latest information entered for that row, the Location Id and description, device name (vital for linking to the counts data), the deployment date and time, the collection site type, either a sidewlak, roadway, or bike lane, and the pickup date and time as well as any notes of interest to the deployment. Other elements collected in the Google Sheets include the email address of the user submitting the information and a picture of the deployment for verifying its setup.

Timestamp	Location ID	1 Location Description (Street or area)	2 Deployment Date	3 Deployment Time	4 Type of Collection	6 Equipment ID	7 Pick up date	8 Pick up Time	9 Comments at Pickup (issues with equipment, collection or site)
8/20/2018 9:37:25	160A	Brosterhaus wb bike and sidewalk at Brent	8/12/2018	17:42	Combo (Sidewalk/Bikelane)	MULTI543	8/20/2018	15:00	post
8/25/2017 14:05:26	171A	Butler Mkt RAB Sidewalk east side of road	8/9/2017	10:30	Sidewalk	MULTI546	8/16/2017	10:30	None
9/12/2018 17:02:05	171A	Butler market eb at wb approach	9/4/2018	17:45	Combo (Sidewalk/Bikelane)	MULTI550	9/17/2018	14:50	
8/25/2017 14:06:14	171B	Butler Mkt. Roadway leaving RAB NB	8/9/2017	10:30	Roadway	TUBE542	8/16/2017	10:30	verify data for 8/14 - 8/15
9/12/2018 17:05:48	171B	Butler market eb at wb approach	9/4/2018	17:45	Combination (street and bike lar	TUBE539	9/17/2018	14:50	
8/25/2017 14:03:49	171C	Butler Mkt RAB Sidewalk west side of roac	8/9/2017	10:00	Sidewalk	MULTI545	8/16/2017	10:00	None

Figure 3.93: Example of Google Sheets deployment information

To expand the information available on the count locations a spatial database of characteristics has been constructed that carries the location and sub locations attributes so that a linkage can be made. Attributes such as the facility type, e.g. presence of a bike lane or off-street path, can then be appended to the traffic counts data in order to perform later analyses. These data are all stored in a geodatabase titles Bend_Spatial_Data.gdb.

3.3.2 Data Schema

The data schema is that shows the various data elements from the three data sources are presented below. In the appendix, a data dictionary has been provided for each element in the data schema.

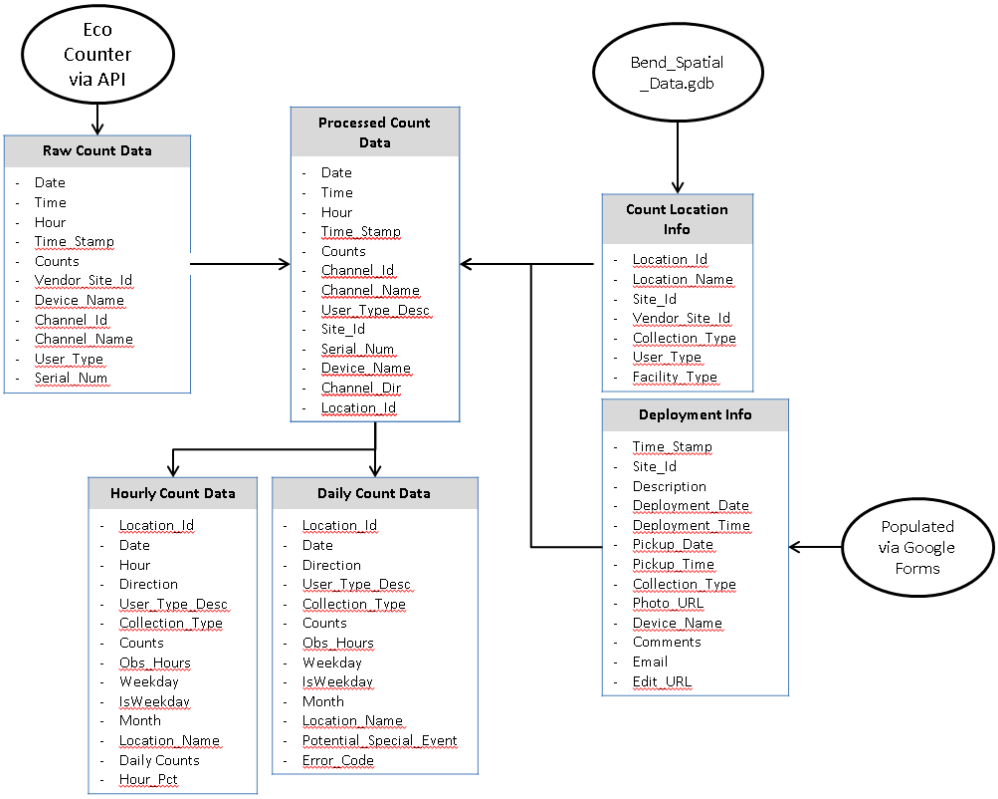

Figure 3.40: Data schema for traffic counts processing

4.0 TRAFFIC COUNTS DATA PROCESSING

This section describes the data processing that occurs to clean and annualize the traffic counts data. As of the publication of this report data collection devices have gathered nearly 29,000 daily records from nearly 200 sub locations in the study region. In order to ensure high quality data for further analyses, these data need to be checked and any data not fully representative of traffic conditions should be removed to avoid entering bias into any results that employ these data. However, doing manual review of this data by staff would be costly and take too much staff time so what follows includes a description of a multi-stage process that looks for certain data problems through an automated method. Data problems include consecutive zeros and excessively high values and well as other outliers. All data anomalies are flagged and retained so that any analyses using these data may still have access to suspect data if needed and to ensure transparency for other data users.

Annualization of data is necessary when a full year of observed counts is not available. This research uses two techniques including the traditional factors as well as a method proposed by Roll and Proulx (2017) called the Seasonal Adjustment using Regression Method (SARM). The traditional traffic factor method creates extrapolation factors where a full year of counts is available and applies them to short term counts. The SARM approach utilizes the established relationship between daily conditions such as weather and the daily traffic counts to estimate traffic during days when no data was collected. Both of these methods are utilized and compared.

4.1 FLAGGING SUSPECT DATA

The data cleaning algorithm looks for and created flags for the following error types:

- At least 3 days of consecutive zeros
- Rolling Mean Error (Outside specified error boundary)
- Excessively large value over 2,500 (For nonmotorized only)
- Manual Error Check Required

This process is explained using the flow chart below. The process is applied to each sub location of data, as opposed to the parent location, so that errors can be found at the most disaggregate level. This will allow utilization of other data from the related sub locations, provided counts can be ‘filled in’ at the suspect sub location.

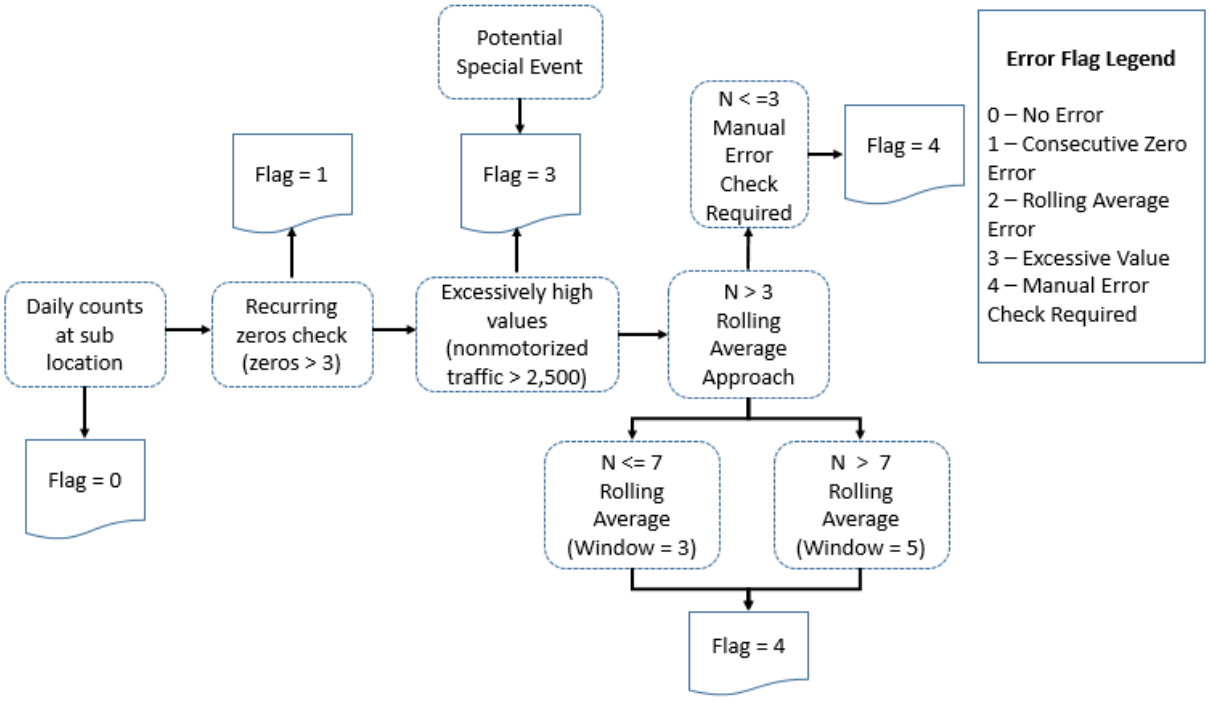

Figure 4.1: Data error flagging process

4.1.1 Consecutive Zeros

As shown in Figure 4.1 above, the first check performed on the sub location data is to look for daily records where the traffic counts were zero for three or more consecutive days. Based on knowledge of some of the locations where these errors have occurred, these records are presumed to be the result of equipment failure. These data are flagged with a tag indicating this type of error.

4.1.2 Rolling Mean

The primary element of the data flagging process uses a simple approach of calculating the rolling mean of the daily observations and calculating a confidence boundary where observed values are compared with and if the observed daily values fall within the boundary the record is not flagged with an error tag. If the observation falls outside the confidence boundary then it is given an error flag. This process separates weekdays and weekends since those conditions alone relate to significant variation at many locations.

4.1.3 Excessively High Values

For nonmotorized traffic counts, records are flagged when they exceed 2,500 counts per day. This value was determined by manual inspection of these kinds of events and expert judgment regarding the reasonableness of extreme high values for the study region. Some high values are expected on holidays and special events that would induce nonmotorized travel such as a marathon or organized bicycle ride. In order to avoid incorrectly flagging data collected on days with an error flag, an additional process was created that looks across locations to determine days

where high values are detected and creates a lookup table of dates. If a given day is flagged as either excessively high, or outside the upper bounds of the confidence boundary but is on a day where a special event may have occurred, the error flag is moved.

4.1.4 Manual Error Check

Lastly, if there is not sufficient data to calculate a rolling average, such as when there are three or fewer days of data, a manual inspection of the data is carried out.

The results of these error checks and the application of flags are shown for a sub location below in Figure 4.2**.** In this example only four daily counts fall outside the rolling mean confidence boundary and two of those are potential special events. One of those days, the December 8th of 2018 date, was checked and in fact an event called the *Holiday Lights Ride* took place on that date and likely led to the higher than expected value.

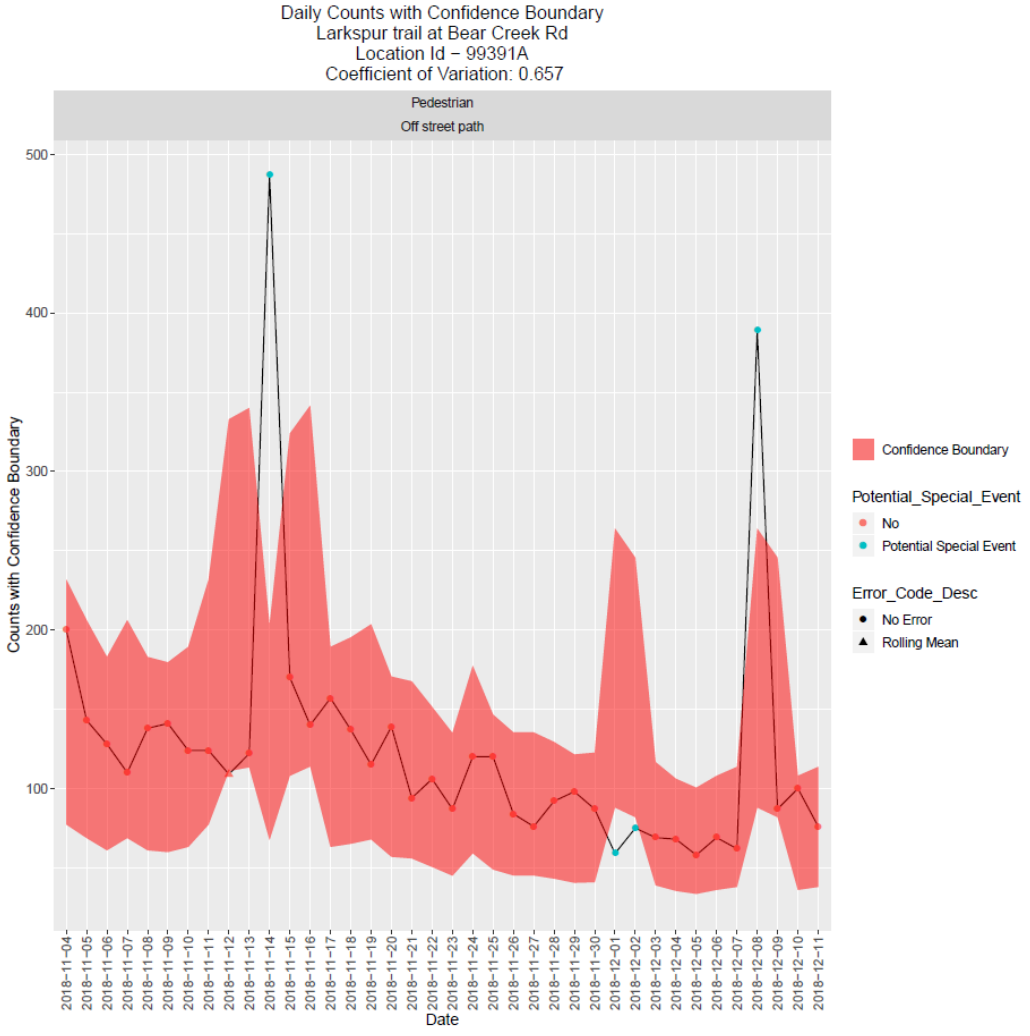

Figure 4.2: Example of rolling mean and potential special event flag

4.2 RESULTS OF FLAGGING ALGORITHM

The results of applying these flags are shown below in table. For the bicycle user daily traffic counts, the most common errors are consecutive zeros followed by observations falling outside the rolling mean confidence boundary and then observations needing manual review and only four observations with an excessive value flag (greater than 2,500 daily bicycles). A similar outcomes is shown with for the pedestrian, user, and vehicle traffic with the greatest number of error flags being assigned to observations with excessive zeros following by the rolling mean and excessive values. For all of the nonmotorized traffic counts, about 75% had no errors detected and are considered usable. Of the 24% with a detected error, 14% were due to the detection of consecutive zeros which is associated with known equipment failures. The annualization process will be able to interpolate these missing data when, as is the case for most of the permanent count sites, sufficient data exists to estimate seasonal adjustment models using the SARM approach.

Table 4.1: Summary of Study Area Travel Network

User Type	Error Code Type	Observations with Error	Total Observations	% of Daily Observations with Error Flag
Bicycle	Manual Error Check	130	10,605	1.0%
	Excessive Value	4	10,605	0.0%
	Rolling Mean	1,069	10,605	10.0%
	Consecutive Zeros	1,996	10,605	19.0%
	No Error	7,406	10,605	70.0%
Pedestrian	Manual Error Check	60	5,513	1.0%
	Excessive Value	201	5,513	4.0%
	Rolling Mean	395	5,513	7.0%
	Consecutive Zeros	646	5,513	12.0%
	No Error	4,211	5,513	76.0%
User	Manual Error Check	34	6,587	1.0%
	Excessive Value	261	6,587	4.0%
	Rolling Mean	420	6,587	6.0%
	Consecutive Zeros	556	6,587	8.0%
	No Error	5,316	6,587	81.0%
Vehicle	Manual Error Check	53	6,727	1.0%
	Rolling Mean	319	6,727	5.0%
	Consecutive Zeros	1,745	6,727	26.0%
	No Error	4,610	6,727	69.0%
Total Nonmotorized	Manual Error Check	224	22,705	1.0%
	Excessive Value	466	22,705	2.1%
	Rolling Mean	1,884	22,705	8.3%
	Consecutive Zeros	3,198	22,705	14.1%
	No Error	16,933	22,705	74.6%

4.3 SPLITTING USER DATA INTO BICYCLE AND PEDESTRIAN COUNTS

Traffic count hardware used in Bend collect pedestrian and bicycle separately in most instances, however numerous sites have data where user counts are collected as well as bicycle and pedestrian counts. In these instances, the ratio of bicycle and pedestrian counts from the bicycle and pedestrian specific sensors are used and applied to the user counts in order to estimate bicycle and pedestrian counts separately. This process is applied to counts data by weekend and weekday separately, since the ratios appear to fluctuate depending on the day of the week, as well as by month.

4.4 ESTIMATED ANNUAL TRAFFIC VOLUMES

Once data is retrieved, processed, and cleaned, it is most useful as an annual and average daily value since most analyses including in crash and health, seek these comprehensive values. For sub locations where data was collected intermittently, annual and annual average daily values take into account seasonal differences and ensure the reported values are not too high if the counts were collected in part of the year more favorable to nonmotorized traffic, or two low if the data was collected during cold and rainy parts of the year. Data collected at permanent sites that experienced equipment issues, the annualization process will fill in the gaps since the facility was still operating as normal.

Methods for matching short term sites with permanent sites using land use characteristics were tried but ultimately it was decided that a single factor would be used. Because the literature indicates. Esawey (2014) demonstrated that using a single daily factor can minimize error compared to the tradition day of week by month factors so this approach was applied for this research.

5.0 DATA IMPUTATION AND MODELING INTRODUCTION

Chapters 5 through 8 document work done for ODOT Research Project SPR 813 – Methods for Estimating Nonmotorized Travel Activity. Specifically this report covers work required in Task 4 – Data Analysis and Modeling. This project seeks to develop nonmotorized traffic data collection practices applicable across the state of Oregon by demonstrating data collection protocols and processes in partnership with the Bend MPO. Following these sets of tasks, the data is employed in estimating traffic activity across the network. This report will focus on the implementation of machine learning tools for data imputation of daily traffic counts as well as the use of machine learning in data fusion modeling.

Machine learning has quickly become a commonly used tool in a number of domains including image and speech recognition, medical diagnosis, genetics, finance, and marketing. This form of artificial intelligence allows data scientists to harvest more information from data and take full advantage of larger datasets with sizable number of features and interaction effects among features. The transportation domain has also been utilizing machine learning techniques but most examples remain in the research side of the field with fewer examples found in practice. Two applications of machine learning are explored in this report including its deployment in data imputation of traffic counts data and data fusion modeling or direct demand modeling. The report will first summarize literature related to imputation of traffic counts followed by a literature review of data fusion or direct demand modeling. Both motorized and nonmotorized research will be included in these reviews. Following the literature review, this report will document the traffic count data imputation process developed, tested, and implemented using a variety of analytic techniques. Lastly, this report will describe the data developed and deployed in a data fusion models for vehicle, bicycle, and pedestrian traffic in the study region.

6.0 TRAFFIC COUNT IMPUTATION AND DATA FUSION MODELING LITERATURE REVIEW

The literature review will cover two topics including traffic counts data imputation as well as a review of direct demand modeling and related data fusion modeling work.

6.1 MOTORIZED TRAFFIC COUNT IMPUTATION

In the past traffic count data imputation was relatively widespread with Albright (1991) documenting at least 23 states using some procedure for imputation on their permanent counter devices. Imputation is often necessary because of the common occurrence of missing data in automatic traffic recorders or ITS data collection devices (Zhong et al 2005). Though widespread, data imputation became a questionable practice as public agencies neglected to flag data that was imputed from those that were actually observed leading to a small crisis of confidence in the traffic counts data. In the early 1990s, the American Society for Testing and Materials (ASTM) and the Association of State Highway Transportation Official (AASHTO) adopted a Base Data Integrity principle that highlighted the significance of raw traffic measurements being retained without modification or adjustment. Further, the principle of Truth in Data directs highway agencies to clearly document any procedures used in any imputation process. (ASTM International 2018) As ITS systems that collect volume data have expanded, imputation methods are needed both to fill in missing data but also to predict traffic conditions on a short term basis for operational needs. Missing data for these systems have been reported to be as much as 15% (Chandra and Al-Deek 2004) and 14% (Ni et al 2005). Most of the recent literature documents more statistically principled techniques for data imputation and seems to shed the simplistic methods of the past except for base method comparisons. Some studies are hard to compare with others because they report estimation results for hourly count estimation while others look at monthly or annual estimation quantities.

Traffic count imputation uses three broad categories of methods including historical and factor based, time series analysis, and machine learning. Historical or factor based methods use historic observations of traffic at a given location to fill in missing data or develop factors using traditional factoring approaches to estimate missing data. Moving average techniques use varying levels of sophistication to employ larger sets of observations to inform imputed values for missing data. Machine learning approaches may utilize a variety of algorithmic techniques and will be the approach reviewed in most details below, followed by moving average approaches with only cursory review of historic and factor based approaches.

In a survey of state DOT monitoring programs from 1990, it was found that at least seven states used simplistic procedures of imputing missing traffic count records. For instance, it was found that South Dakota DOT would use the previous three years of counts for the same period needed for imputation to inform their missing values while Delaware DOT would look at the same period during the previous and following month to inform their missing data estimate. (Albright 1990). Montana DOT would use historical approach and apply a change factor based on reduction or growth observed in nearby sites. Some of these approaches were reviewed by

Zhong et al. 2005 and found that for estimating hourly traffic counts, these simplistic historical approached resulted in more error compared to a relatively more sophisticated moving average approach. Turner and Park tested historical factor approaches on a number of scenarios where data were missing at random and also not at random finding that results, even when missing up to eight months of data, error was low at less than 5% (Turner and Park 2008).

A time series is a chronological series of data on a given variable, in this case traffic counts data collected on a five minute, hourly, or daily basis. Time series data are analyzed hoping to find a historical pattern for use in forecasting unknown, typically future, values. Time series modeling is based on the assumption that previous trends offer information to predict future values (Box and Jenkins, 1970). Numerous techniques exist for modeling univariate time series analysis such as Holt-Winters, exponential smoothing, and Box-Jenkins. Exponential smoothing should not be applied to data with seasonal variation and instead the Holt-Winters procedure should be applied. Box-Jenkins procedure is a common tool for time series analysis and is more commonly referred to as autoregressive integrated moving average (ARIMA) model using Box-Jenkins methodology. Autoregressive and moving average components are considered in these models, thus the name *integrated* model since the stationary model that is estimated to the differenced data has to be summed or integrated to provide a model for the non-stationary data (Chatfield 1989). It has been noted that ARIMA (0, 2, 2) and Holt-Winters approaches are equivalent (Castro-Neto et al. 2009). Sharma el al. (2004) used ARIMA and found it worked better for predicting hourly volumes compared with time delay neural network, and factor approach. (Sharma et al. 2004).

Some kind of moving average procedure has been used in traffic imputation since at least the 1990s where it was employed by London's Department of Transportation (Redfern et al. 1993). Zhong et al (2005) found the moving average approach employed by London DOT performed better than the historic and factor based approaches of some state DOTs.

6.2 NONMOTORIZED TRAFFIC COUNT IMPUTATION

Esawey (2018a) tested a Monte Carlo Markov Chain (MCMC) multiple imputation model to impute missing data including data missing completely at random (MCR) and data not missing at random (NMR). The idea behind this approach is to take advantage of information from historical information from the count station, patterns in data from neighboring stations, and weather to develop an estimate of missing data. The tests found that in the MCR tests results of were better than NMR but only tested missing data scenarios of up to four months. The work also found the MCMC was significantly better than the baseline method of using monthly factors. . Beitel et al (14) experimented with a process to automatically flag anomalous bicycle traffic counts, remove them, and impute replacement observations using a DOY of year factor from sites that exhibited similar day of factor year patterns. This research illustrated the effectiveness of the day of year factoring approach for data imputation when traffic count sites can be matched with other permanent sites. This approach however, requires enough data and counters to match the traffic count site to a site with similar day-of-year factors which is not always possible. The author's use a correlation coefficient threshold of 0.75 to determine sites to match and average the DOY for situations where multiple sites are matched. Additionally, Beitel et al. (14) did not examine the ability of the method to impute pedestrian counts and how often the ability to match any site with a set of sites to use for factors.

6.3 DATA FUSION AND DIRECT DEMAND MODEL LITERATURE REVIEW

There are numerous methods for estimating traffic volumes with the most common methods being a travel demand model, statistical model, geospatial analysis, machine learning, or image processing. The focus of this review will center on statistical modeling and machine learning techniques. This section of the literature review will focus first on motorized traffic and then be followed by the literature found on nonmotorized traffic.

6.3.1 Motorized Traffic Volume Estimation

Numerous attempts to use statistical models to estimate traffic counts are present in the literature. These models use the general functional form, aiming to find relationships between roadway characteristics like number of lanes, functional classification, and access to jobs and people and vehicle counts. Mohamad et al. (1998) used data in 40 of Indiana's 92 counties to estimate a multivariate regression model to predict AADT for vehicles. Validation of the models was done using additionally collected data in 8 randomly selected counties. Results showed that prediction error in the model ranged from 2% to 34% with a 17% mean percent difference.

Xiu et al. (1999) used data from the Florida DOT's traffic count database including 89 count stations across 40 counties to estimate a model relating roadway features, surrounding land use and socio economic factors to the traffic counts. The final selected model produced estimates that ranged from 1% absolute present different to 57% difference with an average error of 22.7 percent. Zhao and Chung (2001) used over 800 counts from Broward County, Florida to estimate a multiple regression model employing roadway features like number of lanes and accessibility measures. The authors also tried using spatially weighted regression techniques in their analysis procedures. The range of error for the best model was between 0.3% and 155.6 percent with no mean error reported though the authors state that 73% of the comparisons possessed 30% error or less. Tang et al. (2014) used a number techniques including neural network machine learning Gaussian maximum likelihood (GML), and non-parametric regression. The results of estimating near-future volumes on roadways showed that the GML approach worked best though all techniques had mean error of 2% or less. Sekula et al. (2018) tests multiple machine learning algorithms to estimate hourly traffic volumes on the Maryland highway network. Machine learning techniques include a fully connected feed forward multi-layer artificial neural network (ANN), linear regression, k-nearest neighbor, support vector machines with linear kernel, and random forests. The ensemble ANN works best with 22% median absolute percent error.

6.3.2 Nonmotorized Traffic Volume Estimation

Significant parts of the following literature review are derived from the previous report on nonmotorized traffic modeling Bicycle Count Data: What is it good for? A Study of Bicycle Travel in Central Lane Metropolitan Planning Organization (Roll, 2018) though has been updated to reflect recent research. Facility demand models are an increasingly common method for analyzing non-motorized travel but were tried as early as 1977 with Benham and Patel (1977). These models use counts of people walking or people riding bicycles as dependent variables and employ weather, built environment, sociodemographic and network characteristics

as independent variables to estimate statistical models. These models are simpler than travel demand models and do not include a behavior components or data from travel surveys. Some of the research below, especially the more recent research, attempts to estimate network wide demand while other research only estimates models to determine how each dependent variable relates to the counts without ever applying the model to the rest of the study area.

Lindsey et al. (2007) uses mixed-mode (bicycle and pedestrian) counts collected by infrared devices in Indianapolis, Indiana to correlate weather, temporal, sociodemographic and urban form variables with non-motorized travel activity. The authors use a log-linear model specification to determine the effect that these variables have on observed daily traffic volumes. Findings suggest reasonable relationships between dependent and independent variables across four model specifications with high explanatory power, with adjusted R^2 of 0.7966. This research uses gross measures of socio-demographics and urban form, assigning Census tract information where counts are collected to the count location. Counts for this research were collected on off-street paths and were not applicable to on-street locations.

Hankey et al. (2012) use two-hour evening peak period (4:00 - 6:00 pm) counts of bicyclists and pedestrians from 259 locations in Minneapolis, Minnesota to estimate models relating counts to weather, built environment, socio-demographics, and infrastructure variables. Measures of socio-demographics and some of the built environment variables' areal unit is at the Census block group level. The authors tried two model specifications, ordinary least squares (OLS) and negative binomial regression to understand the relationship between the dependent and independent variables concluding that due to the over dispersion of the count data the negative binomial distribution is best. For the bicycle count models, Hankey et al. produce results using the negative binomial regression technique with pseudo R^2 value of 0.476 with eight of the independent variables not significant at the 0.05 level. The authors attempt some validation, comparing estimated counts with observed counts though with no hold out and no discussion of absolute error just a visual inspection. Additionally, the authors expand the two-hour counts up to 12-hour counts using some locally derived factors which however substantiated, would likely introduce some error into any application of these models to the entire network. This application of the model to the entire network results in citywide estimates of 12-hour non-motorized traffic for each link of the network.

Wang et al. (2014) estimate models relating weather and sociodemographic variables to mixed-mode counts from six off-street counters. The authors compare the use of OLS and negative binomial regression techniques, concluding that the latter is a better specification based on the distribution of the counts data and resulting error from validation tests which was as low as 16.6% for the general model (pooled data from all six locations). The authors suggest that the models could be used to estimate non-motorized volumes at locations where trails construction is proposed. Hankey and Lindsey (2016) build on past research using additional mixed-mode count data from the Minneapolis, Minnesota which include afternoon peak period (4:00 pm – 6:00 pm) counts from 954 locations for years 2007 through 2014. The authors use linear regression models to relate weather, sociodemographic, and infrastructure to collected counts data experimenting with models using varying numbers of independent variables hoping to find a reduced form specification usable in areas with less available data. This research is the first to try network density variables where the total length of certain network characteristics (e.g. on-street bicycle facilities) are employed as independent variables with results yielding intuitive

results in some but not all cases. For example in the statistically optimal model off-street trail network meters within the vicinity of the bicycle count location are positively correlated with more bicycle volume but local roads are negatively correlated. The authors perform robust validation steps for their core and time-averaged models where they hold out (10%) a random sample of their data, estimate and apply their model, and compare with the held out data and do this process 100 times to assess predictive capability. Using R^2 as a performance metric the authors report measures no higher than 0.55 suggesting the models work moderately well as predictive models.

Fagant and Kockelman (2015) used 340 three-hour peak period counts collected in the Puget Sound area of Washington to test how the features of the Highway Capacity Manual's bicycle level of service (BLOS) have an impact on bicycle traffic. The authors tested the impact of the BLOS features by using a negative binomial and Poisson regression model finding that vehicle volume is negatively correlated with bike volumes, as is the number of lanes, and higher speed limits. Bike lane width was associated with an increase in bicycle volume as were many of the control variables such as mean daily temperature and if the count was taken on an off-street path.

Wang et al. (2016) use mixed-mode counts from multiple places in the U.S. including Minneapolis, Columbus, and the Central Ohio to test the transferability of the facility demand approach across these areas. The authors estimate separate models for each city using AADT as the dependent variable which was possible because counts data were collected from 17 (from all areas) permanent counters collecting year round. Independent variables included sociodemographic and built environment variables from U.S. Environmental Protection Agency (EPA) 2010 Smart Location Database (SLD) in addition to accessibility measures from the National Accessibility Observatory based at the University of Minnesota. The models used a negative binomial specification but did not include any infrastructure variables. The resulting models for each city had pseudo R^2 values of 0.64, 0.576, and 0.318 for Minneapolis, Columbus, and Central Ohio region respectively. Validation tests were performed similar to Hankey and Lindsey (2016) where some data is held out and later compared to estimated counts. Different tests applying the models within each of the cities and also across cities were performed with error of 27% 22% for Minneapolis and Columbus respectively. The cross city validation resulted in considerably higher reported error suggesting transferability of models across cities results in less much less reliable estimates. Since most studies are done using slightly different methods and data it's hard to directly compare the outcomes.

Proulx and Pozdnukhov (2017) used geographic weighted regression to fuse crowd sourced bicycle data from Strava Metro and the local bicycle share system, as well as outputs from the regional travel models to train a model on 536 directional bicycle counts at intersections. The models were rated based on root mean squared deviation (RMSD with no measure of error reported, finding that the models that used the travel model outputs which employed a more sophisticated route choice bicycle that better accounted for actual bicyclist's behavior worked best. Additionally the authors found the use of bike share data decreased RMSD and that using a Gaussian based weighting matrix for the geographically weighted regression outperformed the ordinary least squares regression approach.

Hankey et al. (2017) developed a nonmotorized count program specifically to feed data into a direct demand model. They performed a stratified random sample using functional

classification and network centrality as features to guide the stratification. Collecting one week of counts at 101 bicycle locations and 71 pedestrian locations the authors use a Monte Carlo hold out validation technique (20% holdout) to determine the best model based on r-squared. Covariate data included 199 different features including functional classification, bike facility type, distance to population and commercial area as well as measures of centrality. The final selected model included five variables for the bike model and 6 variables for the pedestrian traffic model with r-squared values of 0.52 and 0.71 respectively. These final models are applied to the entire network but no total bike or pedestrian miles traveled are reported.

Ermagun et al. (2018) used permanent count data from 32 sites from 14 urban areas in the U.S. with an aim to further develop an off-street trail forecast tool. The authors developed econometric models using generalized linear form with land use data from the Environmental Protection Agency's Smart Location Database (SLD) to understand how land use impacts nonmotorized traffic activity. This research uses McFadden's pseudo r-squared and mean absolute percent error as performance metrics for assessing model quality. No discussion is offered about the various models tried before concluding on model specifications that include network density, a measure of higher education, accessibility, water density, lower education, and worker age with different modes using different combinations of variables. The pseudo r-squared measures for the bike, pedestrian and mixed-mode models were 0.63, 0.61, and 0.71 respectively. Mean absolute percent error was 65%, 85% and 46% for bike, pedestrian and mixed mode models. The authors tested a correction factor by regressing error against select SLD variables and were able to improve model result to 48% 58% and 39% mean absolute percent error for the bike, pedestrian, and mixed mode models respectively. Griswold et al (2019) estimated a direct demand model using pedestrian volumes collected at 1,270 intersections across the state. These pedestrian volume counts were collected on a short term basis of 12 hours or less and we factored using permanent count stations matched to short term sites based on surrounding land use. The authors developed a feature set of 75 variables and concluded the use of just eight variables based on step-wise selection process. The results of the model were rated based on r-squared and residual sum of squares with the final model r-squared of 0.714 and no reported value of RSS in the paper. The results of this research have been applied to all 12,414 intersections on the CalTrans state system in order to be used in crash analysis.

6.3.3 Machine Learning Literature Review and Overview

Machine learning has become a more common analytic approach when analyzing data sets containing complex interactions among covariates or features and has been shown to compare well with traditional methods (Diaz-Uriarte & Alvarez de Andres 2006; Heidema et al. 2006). Many kinds of machine learning algorithms exist and include supervised learning methods where a response variable is defined by the user as well as unsupervised where the algorithm determines patterns of importance. Machine learning tasks are typical categorized as either classification, where the model is learning to predict a binary or categorical variable, or a regression problem where a continuous variable is being predicted. This review will focus supervised learning algorithms for a regression problem (traffic counts) using tree based ensemble methods including random forest and extreme gradient boost (XgBoost).

Random forest algorithms work by drawing random samples of data from the input data set and fit a single classification tree to each sample. Classification trees are constructed recursively by selecting the next splitting variable by locally optimizing a criterion such as GINI gain (Strobl et al 2008). Because of the random nature of the samples single classification trees can be unstable but when multiple trees are combined into a forest or ensemble, prediction accuracy increases as those predictions are averaged. Multiple studies have been done to demonstrate the accuracy of these predictions across various fields (Bauer and Kohavi al; 1999; Breiman 1996; Dietterich 2000). Ensembles or forests help to smooth hard edges of decision trees because of random selection, some features to enter the set of predictor variables that may otherwise be outperformed other features. This characteristics of random forests may reveal important interaction effects with other variables that would have otherwise been missed (Strobl et al. 2008). Gradient boosted decision trees have become increasingly popular machine learning algorithms because of their speed and performance. Extreme gradient boosting (XgBoost) uses a more regularized model formalization to control over-fitting but is still built on the gradient boosting framework proposed by Freidman (2000). Boosting constructs the model in a stage-wise process and then generalizes them by allowing optimization of a determined differentiable loss function. Tree based and machine learning techniques with gradient boosting frameworks will be the methods used in this research.

6.3.4 Feature/Variable Importance Overview

A diagnostic measure for machine learning algorithms used in this research include a measure of variable importance. Because many audiences are not familiar with machine learning generally variable importance deserves its own explanation and review below to better acquaint readers with what information this measure can provide. The below section will discuss how variable importance is calculated for random forest and XgBoost machine learning algorithms.

For this research the measure used for describing variable importance is limited to Gini impurity and is essentially a measure of the number of times a feature is used to make node split for a given tree in a given forest. In most calculations of feature importance using Gini impurity the sum of the GINI decrease for every tree in the forest is aggregated each time that feature is chosen as a splitting variable. This aggregate value is then divided by the number of trees in the forest for an average. The scale of the final measure is not important but its comparison to other features gives model users the relative importance of that variable compared to the others. This research will document feature importance as a way to diagnose how the model is utilizing input features.

6.3.5 Cross Validation Overview

This research utilizes statistical models and machine learning to solve analytic problems important to meeting the research objectives. A key element of testing the predictive performance of these algorithms is the use of cross-validation. Cross validation is the process of dividing data into training and testing sets where the training set is used to develop a model which is then applied to the training set. Since the observed values being predicted by the model are available in the training set, a performance metric can be computed by comparing the observed and estimated values. Common measures are percent error and root mean squared error (RMSE), both of which are used in this research.

Cross validation can be used to gauge the performance of any kind of model, whether a traditional statistical model or a machine learning model. In machine learning, cross validation is used to gauge model performance, as previously described, but also to train a model by iteratively specifying a model, testing its performance, and then adjusting elements of the model such as which variables are used to make splits in trees in the random forest algorithm. There are many kinds of cross-validation but for this research we will use 10-fold, leave-one-out, and exhaustive. Below are detailed descriptions of these techniques:

- **10-fold –** In 10-fold cross-validation the data is partitioned into 10 equally sized subsamples at random or using a stratified random selection. Models are then estimated on nine of the 10 partitions holding one partition aside to test the model application. The model is applied and then compared to the one partition that was held out and the estimated values are compared with the observed and a measure of error or model performance is computed. This is done until each of the 10 partitions are held out of the model estimation.

- **Leave-one-out –** In leave-one-out cross validation each data point is used as a test set and the model estimated on the remaining data then the model is applied to see how well can estimate the data that was not used in the model training. This is done until all the data points have been left out. A summary of the error, either mean or median, is then computed and used to gauge the model performance.

- **Exhaustive –** Exhaustive cross validation tests all possible combinations of data being divided into training and sample sets. A deeper explanation is provided in the section below on traffic counts imputation and is deployed to tell which months the imputation modeling process works best for by testing all possible combinations of months being in the training and test set.

7.0 TRAFFIC COUNTS IMPUTATION

It is not uncommon for a traffic count sensor to stop collecting data due to a variety of reasons related to counting hardware or data transmission issues. These outages are usually for continuous blocks of time but as noted in the ITS literature, can often times be intermittent as well, with just a few days or hours missing. The existing data are still valuable and the missing data can be imputed with confidence, though the uncertainty should be characterized. This section describes the tests performed to understand the best imputation procedures to deploy for bicycle, and pedestrian counts in the Bend, MPO study region. Using nonmotorized traffic counts data from across Oregon where a full year of data are available, various machine learning techniques are tested to see how well daily, monthly, and yearly volumes can be imputed. As a baseline to compare the machine learning algorithms a negative binomial regression statistical model will be estimated and applied as well.

7.1 IMPUTATION EXPERIMENTAL DESIGN

To test the efficacy of data imputation using machine learning the experimental approach will reflect practical imputation needs using a not missing at random hold out of counts data. What seems to be most common in the nonmotorized counts data for Oregon, are extended periods of time when the traffic sensor is either not working or not transmitting data (and forever being lost). Figure 7.1 below shows an example for a count location in Bend MPO study area where 119 days of day are missing from the traffic counts for 2018 for two separate periods.

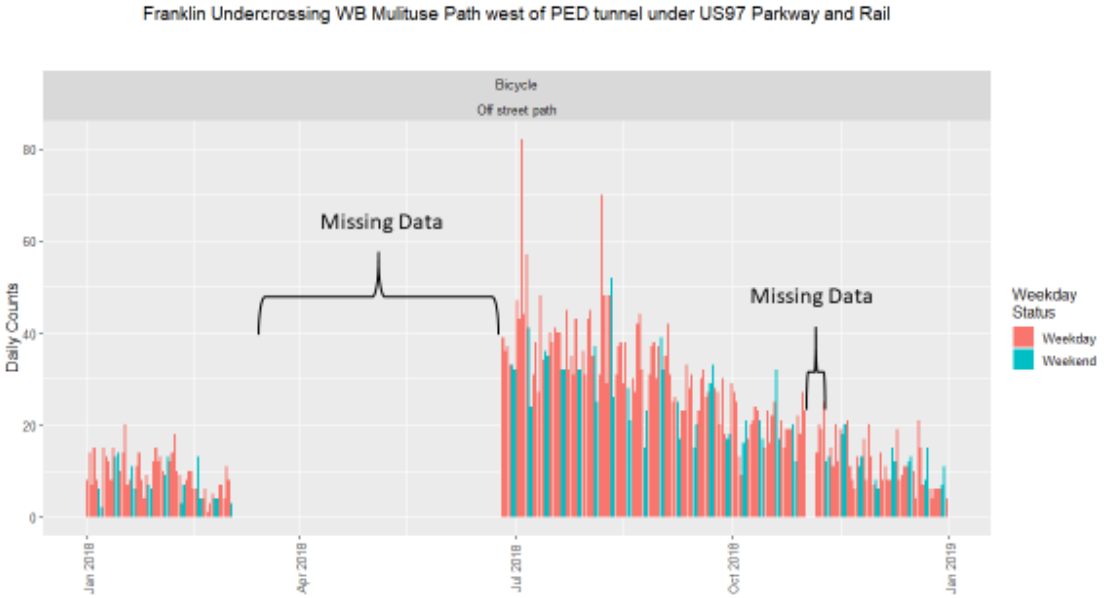

Figure 7.1: Period of missing data example

These continuous periods of missing data will inform the experimental design of our imputation tests but in order to simplify reporting on results whole months will be removed. In the imputation tests, various months of counts will be imputed so that we can document the likely error under different scenarios of missing data. Using full year of counts from traffic sensors across Oregon, we can simulate these outages and understand well, the likely error under different data outage circumstances. The work flow for the experiment is described in the Figure 7.2 below showing three examples of how the imputation procedure will be tested.

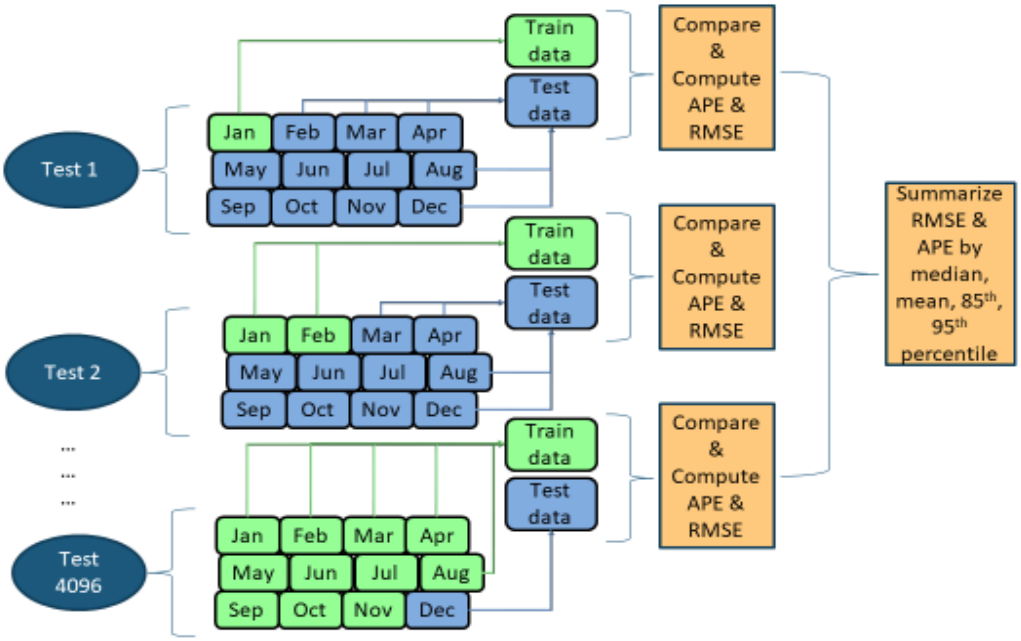

Figure 7.2: Missing data experimental design

For instance in Test 1, we simulate a scenario when traffic counts were available for January and those data are then used in the model training process and then applied to estimate February through December (11 months) traffic counts. We then compare those estimated counts to the actual counts and compute the absolute percent error (APE). APE is calculated using the following equation:

$$APE = \left| \frac{AADT_{obs} - AADT_{est}}{AADT_{obs}} \right| \quad \textbf{(7-1)}$$

In Test 2 both January and February to train the model and then estimate the remaining 10 months, compare and compute error. This is done for all possible combinations representing an exhaustive cross-validation design. A summary of absolute percent error by median and 95th percentiles will be computed as performance measures. All possible monthly combinations will be tested so that during the application of the final imputation procedure, confidence interval for the likely error can be assigned. There are 4,096 possible combinations of months to use in the test, all of which will be tested in this experiment.

Multiple machine learning algorithms were tested including conditional inference, random forest, and recursive partition. These algorithms were implemented in the R statistical computing environment using the caret package (Kuhn, et al, 2020). For more discussion of these methods please consult the literature review above. For baseline comparisons a negative binomial statistical regression model is estimated and applied. Based on the success of using a negative binomial regression model documented in Roll and Proulx (2017), where reliable annual estimates were achieved with as little as six weeks of daily counts, the statistical model is likely simpler to understand for some practioners as it uses more common statistical methods.

7.2 IMPUTATION EXPERIMENT DATA DESCRIPTION

Table 7.1 below summarizes the daily traffic counts data used in this imputation experiment. Two years of data are used utilizing nearly 9,100 daily traffic count records from 23 unique count locations throughout Oregon. Because complete annual datasets are hard to achieve, only 18 of 25 locations (combination of location and year) have a full year of data while the other eight locations have at least 351 days, or 98 percent. All of the count locations in this research are featured on multi-use paths. The mean values below show that the nonmotorized traffic volumes are generally on the lower end with Bend exhibiting the lowest counts and Eugene with the highest of the data used in this research.

Table 7.1: Imputation Experiment Data Summary

City	User Type	Summary Statistics				
		Mean	Median	Standard Deviation	Number of Sites	Daily Records
Bend	**Bicycle**	56.3	43	54.9	5	2167
Bend	**Pedestrian**	148	99	150	7	2907
Eugene	**Bicycle**	340	275	240	3	1095
Eugene	**Pedestrian**	594	281	576	2	729
Springfield	**Bicycle**	185	125	182	4	1460
Springfield	**Pedestrian**	103	97	41.8	1	365
Total	**Bicycle**	153	81	190	13	5087
Total	**Pedestrian**	225	105	328	10	4001

This research is utilizing supervised machine learning algorithms and regression models, utilizing the documented relationships between weather, day of week, and lighting conditions, to predict the traffic counts sunlight (Miranda-Moreno & Nosal 2011; Tin et al. 2012; Thomas et al 2012; Rose et al. 2011; Lewin 2011; Nosal and Miranda-Moreno 2012). Historical climate data used as features in the machine learning and negative binomial regression approaches come from the National Oceanic and Atmospheric Administration (NOAA) and are accessed using the rnoaa library (Chamberlain 2019). Climate data stations for each city, typically the nearest airport, are queried and assigned to the traffic count locations nearest the station. It was considered to use PRISM data that interpolates weather conditions between stations using a gridded system, potentially giving better localized weather conditions but this approach is not currently being applied.

7.3 IMPUTATION EXPERIMENT RESULTS

Using the experimental design described above results from the experiments will be shown followed by a number of diagnostic summaries of the machine learning algorithms and negative binomial regression model. Figure 7.3 below shows the 95th percentile APE for daily, monthly, and annual aggregations, meaning that for 95 percent of the tests the APE was at or below the indicated value. For example, the first column within the bicycle panel, indicates that for estimating daily counts using the conditional inference machine learning algorithm, 95 percent of the daily count estimates were 24.6 percent or less. The daily aggregation level shows the 95th percentile of the median error of the daily comparisons, meaning that for each hold out test, the *median* APE of all the days in that test were used to calculate 95th percentile as opposed to just using the APE in the monthly and annual summaries. Monthly aggregation directly compares entire months or groups of months estimates with observed, while the annual aggregation compares the entire year of estimates plus observed months, not in estimate, to an observed annual count. For instance, in Test 1 from Figure 7.2, the experiment estimates counts from February through December and the annual error measures the difference between the observed annual amount and the estimated counts from February to December plus the observed counts from January. This way we can show the overall annual error when we add imputed data for missing data plus remaining observed data.

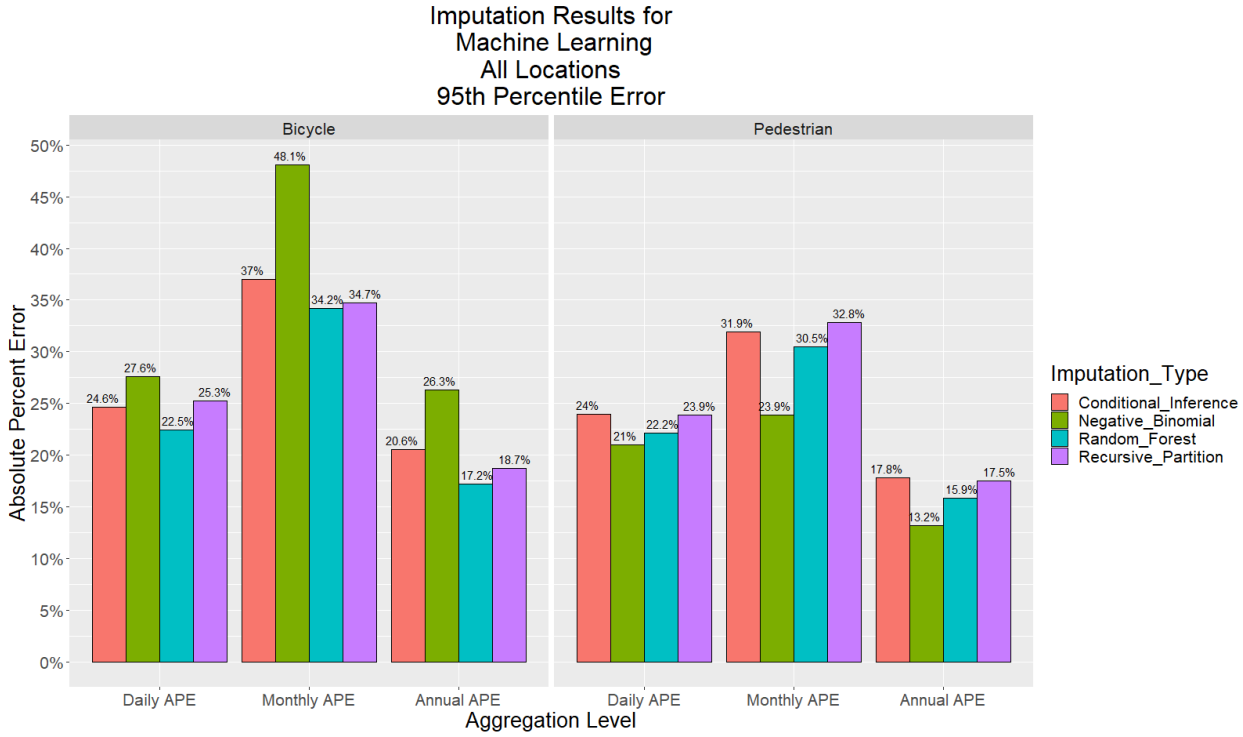

Figure 7.3: Imputation results for all machine learning aAlgorithms – 95th percentile error summary

Nevertheless, the Figure 7.3 shows the upper end of the possible error using each of the machine learning approaches across all tests using all years of data. For the bicycle traffic counts imputation, the random forest appear to work best with random forest demonstrating the lowest

error for all of the bicycle tests with 22.5 %, 34.2% and 17.2% for daily, monthly, and annual aggregations respectively. For the pedestrian traffic counts, the negative binomial regression model works best, with 21%, 23.9%, and 13.2% for daily, monthly, and annual aggregations respectively. For the bicycle counts, the random forest is significantly better than the negative binomial model for each aggregation, and the other machine learning algorithms are next best, most often the recursive partitioning algorithm. For pedestrian counts, the random forest and negative binomial approaches are similar in error for daily and annual error while the negative binomial outperforms the random forest significantly in the monthly aggregation.

The error shown in the figure above shows the worse outcomes, or at least the imputation scenario with the most error 95% of the time, but imputation results can differ depending on amount of data used in the model training and the particular days or months used in the training. Generally, the more data used in the training, the better the imputation estimation and it highlighted in Figure 7.4. This figure shows the 95th percentile error and the median error by annual estimation for each of the scenarios of months used in training data. In the left panel the bicycle counts tests are summarized and related error summaries are shown while the pedestrian count tests are in the right panel.

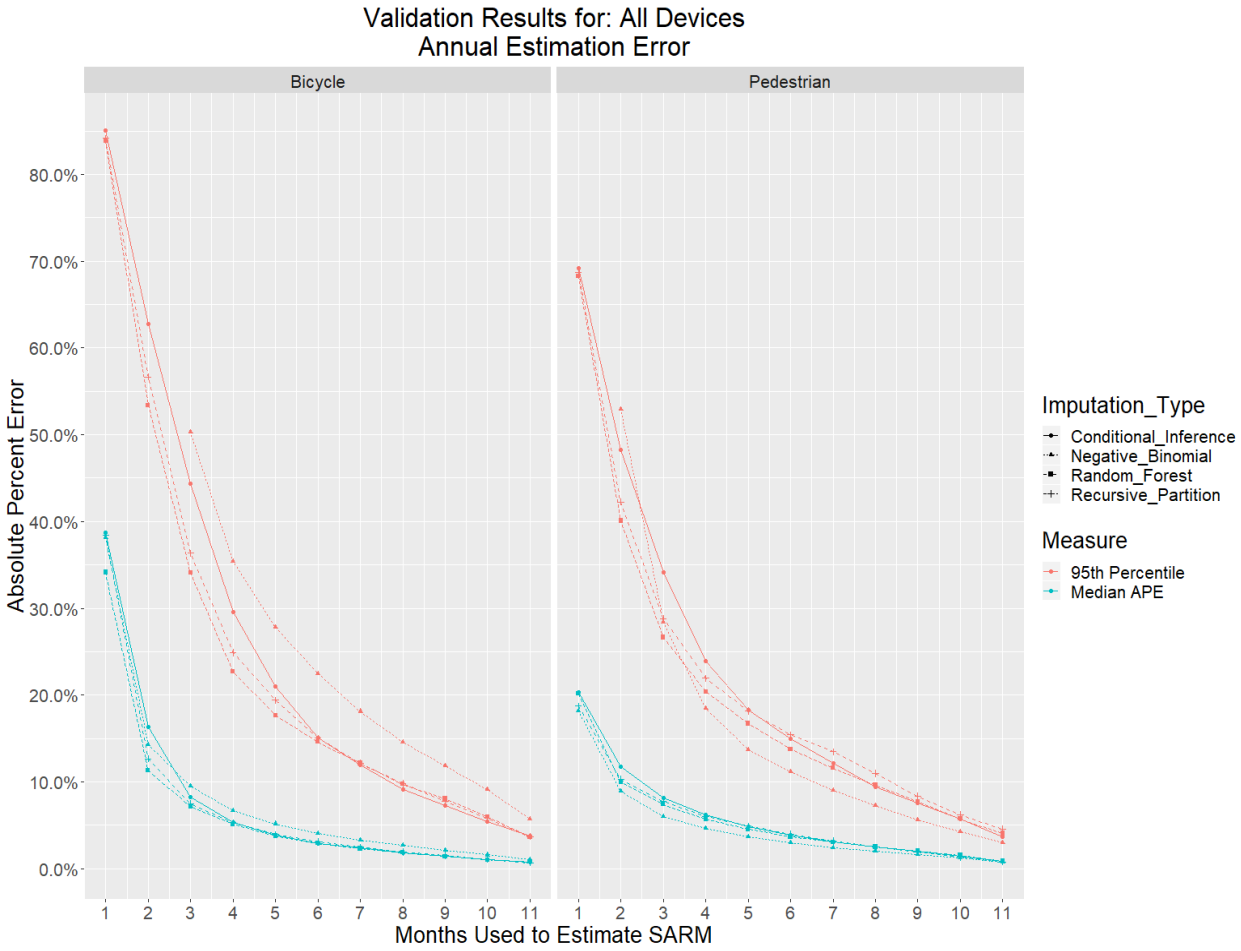

Figure 7.4: Imputation results for all machine learning algorithms – 95th percentile and median error summary by months used to train model

Table 7.2 below shows the results from the chart for the negative binomial and random forest models (since they were consistently the best performing overall) in table format so readers can examine the information in more detail. It should be noted that some results are not shown in the chart since a small number of negative binomial model tests had such high error that the chart was unreadable. For instance the 95th percentile error results for one month of data for the negative binomial model was over 26,000 percent! Of the 442,152 tests done for the 21 count locations, over five imputation techniques, across the 4,096 monthly holdout scenarios, 82 estimates for the annual aggregation APE was over 500% so this seems to be a rare outcome overall. Also, all of these incredible outliers were from the negative binomial regression model and usually resulted from training the model on a single month, usually December, but in some cases, using two months, usually, some two month combination of January, November, and December. It appears that even though the negative binomial regression model does pretty well overall (Figure 7.3) it struggles when data little data informs the training data (i.e. one or two months of winter data feeding the model).

Table 7.2: Imputation Experiment Results by Number of Months Used to Train Model

Number of Months Used in Training	Bicycle				Pedestrian			
	Negative Binomial		Random Forest		Negative Binomial		Random Forest	
	95th Pct.	Median	95th Pct.	Median	95th Pct.	Median	95th Pct.	Median
1	26,288%	38%	84%	34%	244%	18%	68%	20%
2	131%	14%	53%	11%	53%	9%	40%	10%
3	50%	10%	34%	7%	28%	6%	27%	7%
4	35%	7%	23%	5%	19%	5%	20%	6%
5	28%	5%	18%	4%	14%	4%	17%	4%
6	22%	4%	15%	3%	11%	3%	14%	4%
7	18%	3%	12%	2%	9%	2%	12%	3%
8	15%	3%	10%	2%	7%	2%	10%	3%
9	12%	2%	8%	1%	6%	2%	8%	2%
10	9%	2%	6%	1%	4%	1%	6%	2%
11	6%	1%	4%	1%	3%	1%	4%	1%

Overall, in either of the models shown in Table 7.2, given at least nine months of training data or more, the annual counts can be estimated within 8 percent in almost any combination of training months. In 50% of training tests (median error) as little as three months of data can be used to impute missing data and arrive within 7% or better of the actual annual total for the random forest and 10% or better with the negative binomial regression model. These results should lend significant confidence in either of these approaches for estimating annual total counts for bicycle or pedestrian traffic volume.

7.4 IMPUTATION EXPERIMENT DIAGNOSTICS

The above results are presented to demonstrate the prediction power of each of the imputation procedures. However, it's necessary to unpack some of the underlying modeling procedures to understand more about each imputation procedure works. Information about the estimation results from the negative binomial regression models will be shared and then a few diagnostic elements from the machine learning algorithms will follow.

7.4.1 Negative Binomial Regression Model Diagnostics

As noted above, covariates for each of the regression model are selected based on their predictive power using a 10-fold cross validation procedure. This process selects daily covariates most useful in predicting daily nonmotorized traffic counts. This results in different covariates being used in different models. Figure 7.5 below summarizes the variable used in each location for both the bicycle and pedestrian models. Max daily temperature and the weekend variable are the most common covariates used, followed by snow fall, daily precipitation and minutes of daylight for both user types.

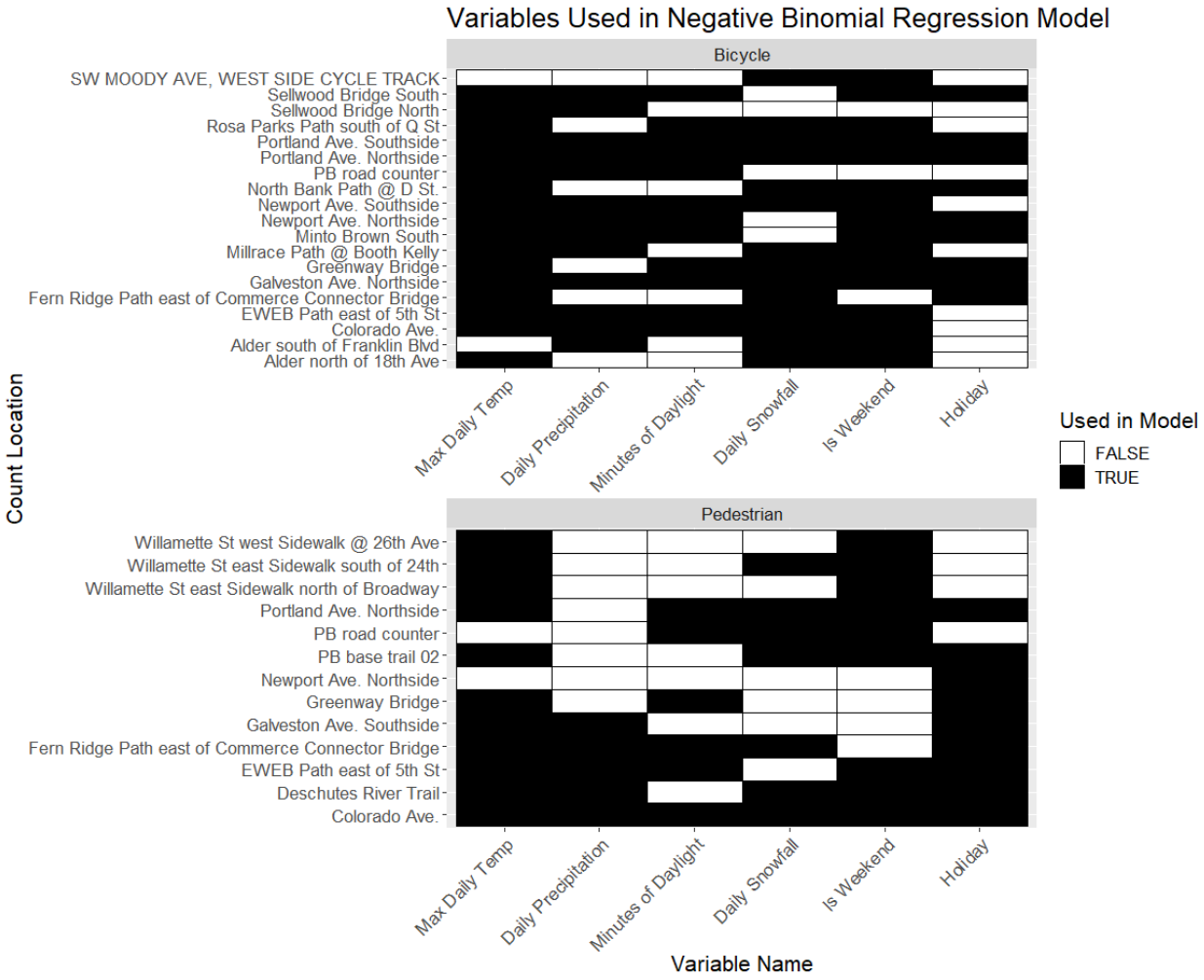

Figure 7.5: Variables used in negative binomial regression imputation procedures

The model coefficients from the holdout experiment offer useful information as to why in scenarios with fewer months used in the model training, predictions are poorer than in months with more data. Figure 7.6 shows the standardized beta-coefficients of all 4,096 tests charted for each of the variables used in the specification for a single location (Millrace Path) in order to highlight how the coefficients converge as more data is used in the training of the model. Assuming the coefficients in the tests with more data (Months Used equals 9-11 months) are closer to the real values, it's easy to see why the tests only one to three months perform so poorly since their model coefficients are so much different.

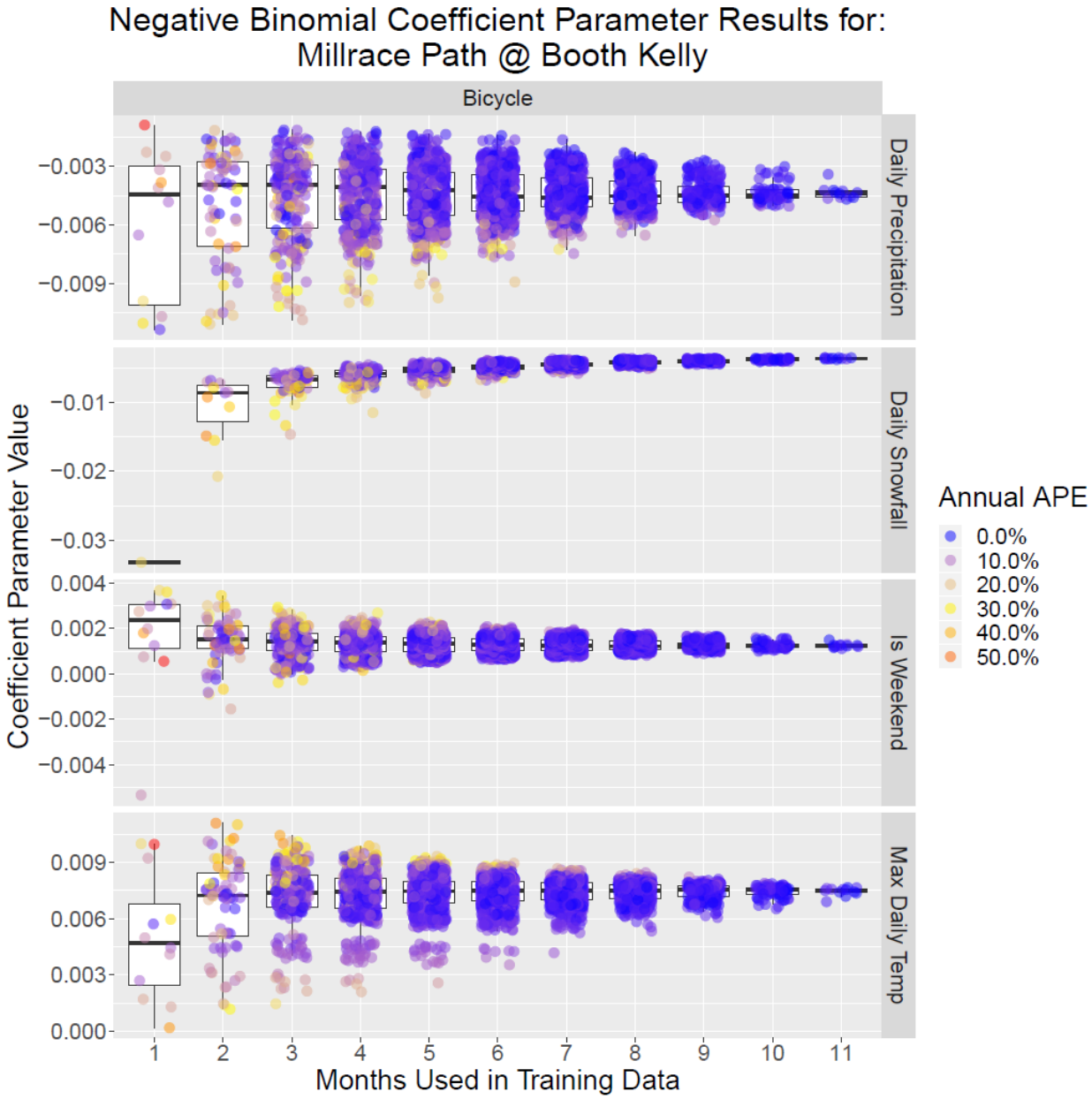

Figure 7.6: Example of negative binomial model coefficients perturbation

7.4.2 Machine Learning Algorithm Diagnostics

Machine learning algorithms have fewer measures to help users understand how the model is working but information about variable importance can be a helpful guide to understand how the model is working. The variable importance measure for tree based learning algorithms essentially measures the usefulness of each feature in the construction of the trees or more simply the number of times a features is used to make split in a tree at a node. For illustrative purposes, an example decision tree is presented below for a single location in the study area in Figure 7.7. The tree shows how the recursive partitioning tree determines splitting criteria including features and feature values. For example, starting at node one (denoted by the value at the top of the node) where all 357 observations in this dataset are present (n = 357, mean value = 37), the data is split by the TMAX (max daily temperature) variable based on whether the counts were taken on a day with less than 63 degrees (F) or more than 63 degrees. If the max daily temperature is less than 63 degrees the decision tree moves left to node two (n = 176, mean value = 20) where temperature is used to split the data further, this time at 53 degrees (node 5, n = 74, mean value = 28). If the temperature is more than 53 degrees the decision tree uses another feature, a dummy variable for weekday or weekend, to make a splitting decision. If the counts were taken on a weekend, the branch moves left (to node 10, n = 26, mean = 18) or if the count was on a weekday branch moves right (to node 11, n = 48, mean = 33).

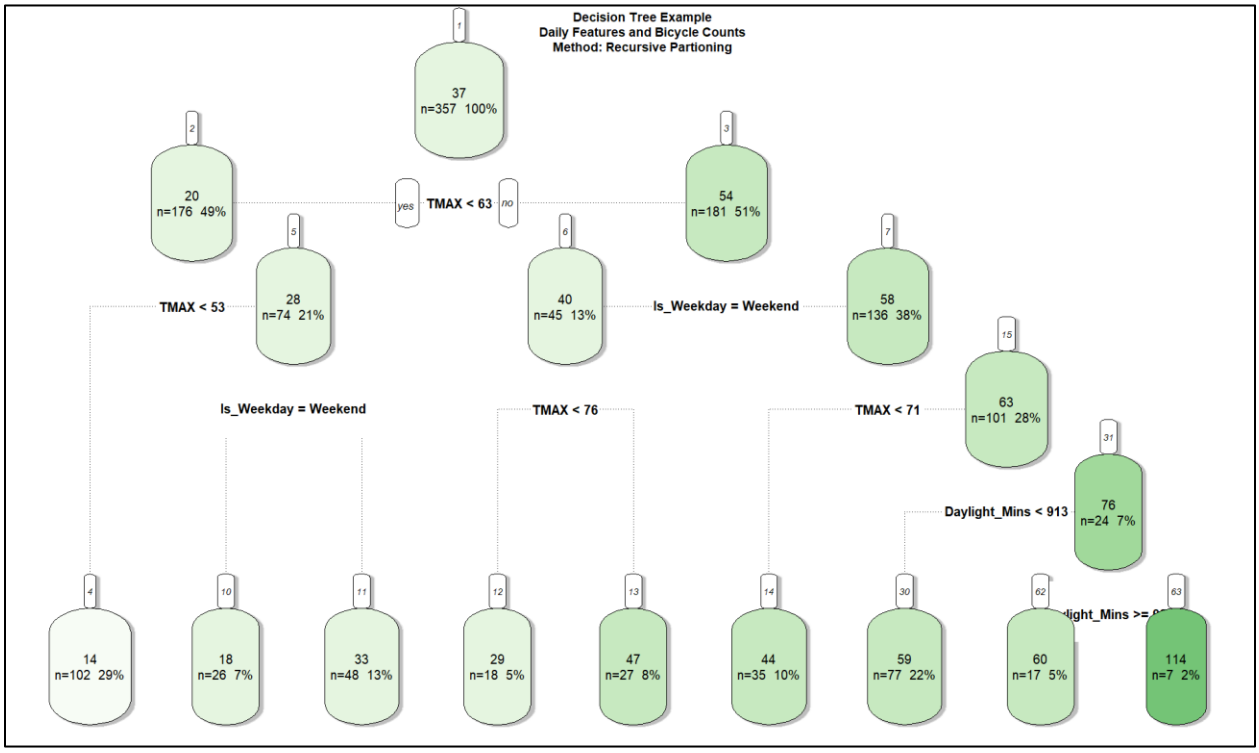

Figure 7.7: Example decision tree

From this description it's clear that TMAX (max daily temperature) and *Is_Weekday* (dummy for weekend or weekday) are important variables used for splitting data at nodes. These outcomes can now be quantified in the chart below in Figure 7.8 showing the relative importance of each

variable highlighting that max daily temperature is the most important variable in determining daily traffic counts for bikes, followed by minutes of daylight, and weekday dummy variable, precipitation (PRCP), and a dummy for if the count is on a federal holiday. Inches of snow on the day of the count was not important, likely due to low number of days with snow and also the impact of snow might be dealt with using the temperature and precipitation variables.

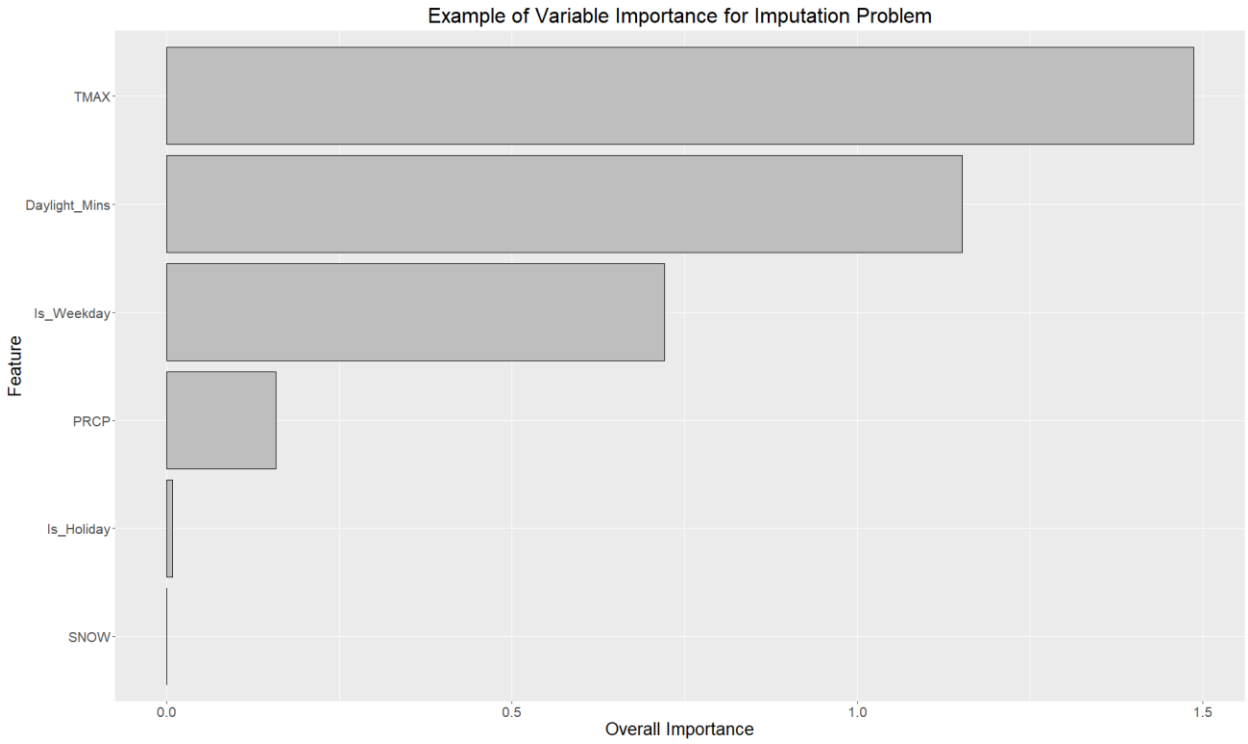

Figure 7.8: Example of feature/variable importance for single recursive partition tree

Now that an explanation of the variable importance measures has been described, these measures will be presented for the working models from the data imputation experiment. Because there are models estimated for multiple locations, the Figure 7.9 summarizes these using relative representations based on color. The variable importance summary is broken out by bicycle (top panel) and pedestrian (bottom panel) traffic counts. This figure shows how max daily temperature is important for both bicycle and pedestrian traffic for all sites, albeit less so for the pedestrian Newport Avenue location. Daily precipitation and minutes of daylight are also important variables for most location specific models.

Using variable importance we can check whether our models are working well by assessing whether the decision tree splitting variables align with the documented research and theoretical foundation. Based on the results below and what has previously been documented as daily conditions affecting daily nonmotorized traffic counts, the models seem to be working as expected.

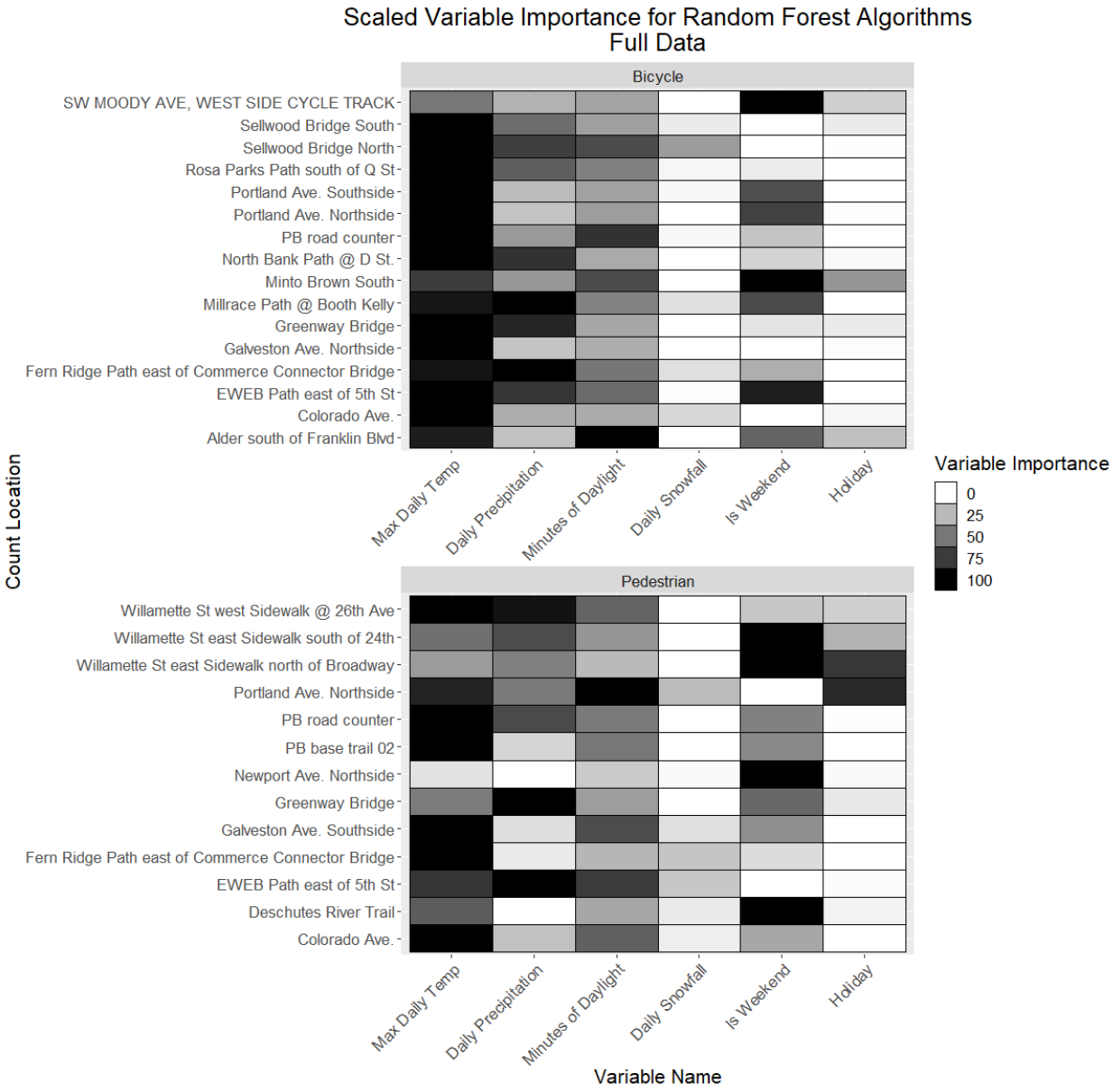

Figure 7.9: Variable importance for random forest models by count location

7.5 IMPUTATION EXPERIMENT DISCUSSION

Accuracy of the prediction is an important element when deciding on an imputation procedure but it is not the only thing to consider. Ease of implementation and acceptance by practitioners are other important considerations. The difference in the negative binomial statistical model and random forest machine learning algorithm along these two elements will now be discussed along with an examination of the accuracy results for daily, monthly and annual estimates.

The regression model uses weather, day of the week, and minutes of daylight as covariates to predict the traffic counts for a missing day of data. It's important to determine the best variables, for instance, if snow fall should be used in a city with little snow fall, may not be obvious and so

some testing needs to be done to determine which covariates to use in the regression specification. Prior to the execution of the tests conducted in the experiments above, a k-fold (k=10) cross validation procedure was performed to determine which of the possible daily covariates were best at predicting daily counts in order to specify the final model used in the experiments. This required some programming in R to properly set up and execute but was generally straight forward. The algorithms for the machine learning algorithm on the other hand, make the implementation of these imputation procedures a little simpler, doing the cross validation internally (within the package functions) and the nature of machine learning means model specification is not necessary in the traditional way. Computation time for the regression model, even considering the cross validation for covariate selection, took just a few hours while the machine learning test training took about four days to do all of the tests for each of the different algorithms, even when parallel processing was utilized which cut each test's run time by about 60% when utilizing 7 processors at once on a four core system with eight 3.4 Ghz setup. For testing accuracy like what was done in the experiment above, this is not much of an issue though since the data is not needed in real time for operations or some similar function. Both of these two approaches require some technical skills that might not be in high supply at transportation agencies and so may be difficult to implement. This is likely why simplistic historic and factoring methodologies persist. Factoring can work if the data exists but as will be shown in the next section, oftentimes sensors fails and there is not a full year of data to use in the development of daily factors so these statistical and machine learning methods offer a more flexible, albeit more complicated, approach to traffic data imputation.

It should be noted that the imputation experiment results might overstate potential error when actually deploying an algorithm since in practice. Since the experiment only looked at a single year in isolation, if any given month was missing the pattern in that month was also gone but in practice if a given month is missing from one year the likelihood is high that that month is present for the preceding or following year. This would likely improve the machine learning performance in terms of error.

7.6 IMPUTATION APPLICATION

With a tested approach documented above, this section will now summarize the application of the imputation process for count sites in the Bend MPO study area. A description of the sites and the missing data will be described, followed by the results produced by the imputation procedure and a short discussion of the potential error in these annual estimates. The random forest machine learning algorithm was selected due to its low error and ease of implementation.

7.6.1 Missing Data Description

Figure 7.10 below shows the number of days missing for the permanent count stations that had data lost due to hardware issues or data filtered due to the error check process discussed in an earlier section. Nearly 80% of the daily counts imputed in the imputation process were consecutive zeroes likely due to equipment problems, with another 16% due to the values falling outside the rolling mean error boundaries. The maximum number of days missing is 179 at location 267B with the average number of days missing at 54 days and a median of 50 days.

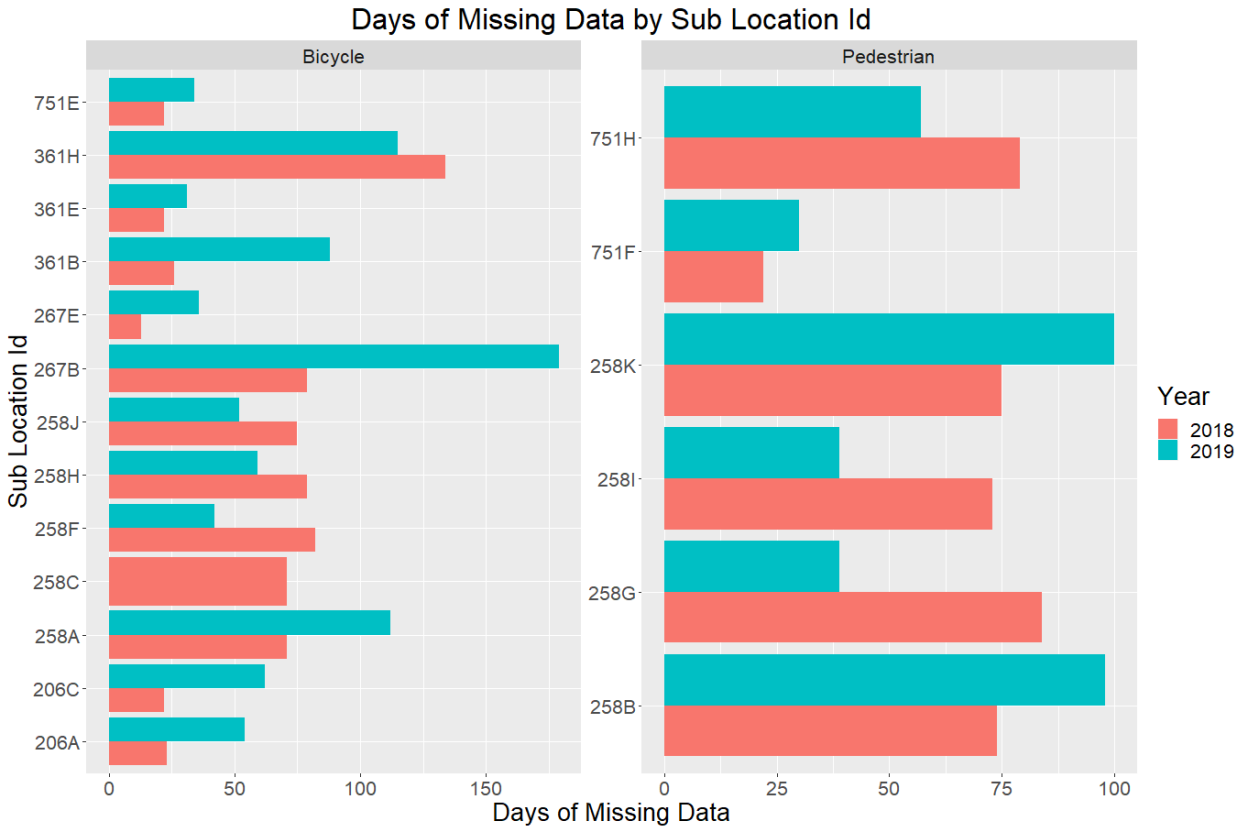

Figure 7.10: Days of missing data by sub location Id

7.6.2 Imputation Application Results

The results of filling in missing data are shown below in Figure 7.11 and show the annual average daily traffic from the combined observed and imputed daily traffic counts for the years where data was collected.

Figure 7.11: AADT estimates from imputation

The figure shows how the imputed AADT are mostly stable across years and also shows error bars for imputed AADTs where more than at least 15% of the of the daily records for any given month were missing. The information for the error bars comes from the imputation experiments described above. For instance, application site 258F (for bicycle) site used data from months January, February, July, August, and September, to impute the missing data so information from the imputation experiment above can be used to assess the potential error by using the 95th percentile error APE. This is appended as a lower and upper bound error to give some confidence intervals. For sites without error bars, the missing data was sparse enough (less than 15% for any one month) that it did not align with results from the imputation experiment so were not appended. In these cases the point estimates should be pretty close to actual considering the missing data is pretty small.

7.7 IMPUTATION DISCUSSION

The above section summarizes the relevant literature on traffic counts data imputation, performs experiments to test a number of imputation procedures, and then applies the selected random forest machine learning imputation procedure. Results for bicycle counts from the imputation experiment show that results for annual estimates of traffic counts can be quite good to the

actual, with 95th percentile error of just 18% when missing up to seven months of data and as little as 4% if using the proper combination of months in the training set. Results are similar for pedestrian imputation experiment results. The machine learning algorithms tested in the experiment and deployed in the application are simply implemented if users have working knowledge of the R statistical software environment that is free for any agencies to use. Results of the random forest algorithm on data in the study region appear internally consistent (from year to year) with assigned confidence intervals from the experiment results showing these results fluctuate within a reasonable amount each year. Another useful application of the algorithm might be to estimate counts using shorter term equipment deployments. As can be seen in the 2017 results above, where the sensors were actually not installed until mid-year, a full year of data can be estimated in the absence of hardware. Agencies should feel comfortable with this approach to traffic data imputation for bicycle and pedestrian traffic.

8.0 DATA FUSION MODELING

The above sections documented the processes involved in collecting, cleaning, and preparing annual estimates of nonmotorized traffic volumes. There are many ultimate uses of these counts data on their own but a focus this research effort is to employ the annual counts in statistical and machine learning modeling in order to mine the relationships that the traffic counts have with other features in the study area to then estimate vehicle, bicycle and pedestrian volume across the entire network, even in places where no counts have been taken. This section will document the procedures developed to estimate vehicle, bicycle and pedestrian volumes in the Bend MPO study area. Models will be estimated using statistical models including negative binomial and Poisson regression specifications in addition to the model estimation using two machine learning techniques including random forest and extreme gradient boosting (XgBoost).

For each of the models, the estimated annual average daily traffic will be predicted based on features or inputs available across the network allowing for application of the model estimation for the whole system. These features include network characteristics such as functional classification, speed limit, and network centrality as well as access to jobs and population. For the bicycle and pedestrian models, additional features will be tested that aim to improve model performance and include samples of 'probe' data from a smart phone app that tracks bicycle rider trips and transit data. The data sources and processing procedure for these features will be described in the sections below.

8.1 VEHICLE TRAFFIC DATA FUSION MODELS

This section will document the development and application of a data fusion model for vehicle traffic in the bend MPO study area. There are two primary objectives in developing vehicle traffic models even though this research is directed at nonmotorized travel estimation techniques.

- **Objective 1 –** Demonstrate accuracy of data fusion models compared to establish reporting protocols

  The first objective is to demonstrate the use of parametric and machine learning methods for the purposes of network wide volume estimation. Because vehicle counts data and the associated collection methods are more established and more data exists to test models, some confidence can be established as to how these methods work with varying amounts of data. Since vehicle counts and the vehicle miles traveled metrics they inform are standard elements of ODOT's annual reporting to the FHWA, a comparison dataset for vehicle miles traveled estimates is available for validation of the final data fusion models. The results will give a sense as to how accurate the data fusion technique is compared to traditional approaches using the Highway Performance and Monitoring (HPMS) procedures with varying amounts of data. Additional test results will be presented for select model specifications where only a subset of vehicle counts data are used to train the model with the object of showing aggregate VMT stability even when less data is available.

- **Objective 2 –** Employ estimates of network wide vehicle traffic counts in bicycle and pedestrian models

  The second objective of this section is to develop vehicle AADT measures for the entire study area network to use in the bicycle and pedestrian models. Nonmotorized traffic volumes are sensitive to the presence of vehicle volumes since they make users feel less safe and in fact do lead to higher risk for nonmotorized users **(CITATION)**. Therefore having motorized volumes for the entire network will be important information in the nonmotorized models developed in the other sections of this report. The objectives of the two sections on nonmotorized traffic data fusion modeling is to develop a working prototype that Bend MPO could use for travel monitoring and planning purposes including in safety analyses featured as a later chapter in this report.

8.2 DATA DESCRIPTION FOR VEHICLE TRAFFIC FUSION MODELS

The data fusion models utilize a number of data sets to train and apply models including annual average daily traffic (AADT) estimates of traffic counts, network attributes, access to jobs and population and a measure of network centrality. Figure 8.1 below depicts the different network features that come together in the data fusion models to estimate the network wide AADT for vehicles. Each of these data are explained in more detail below

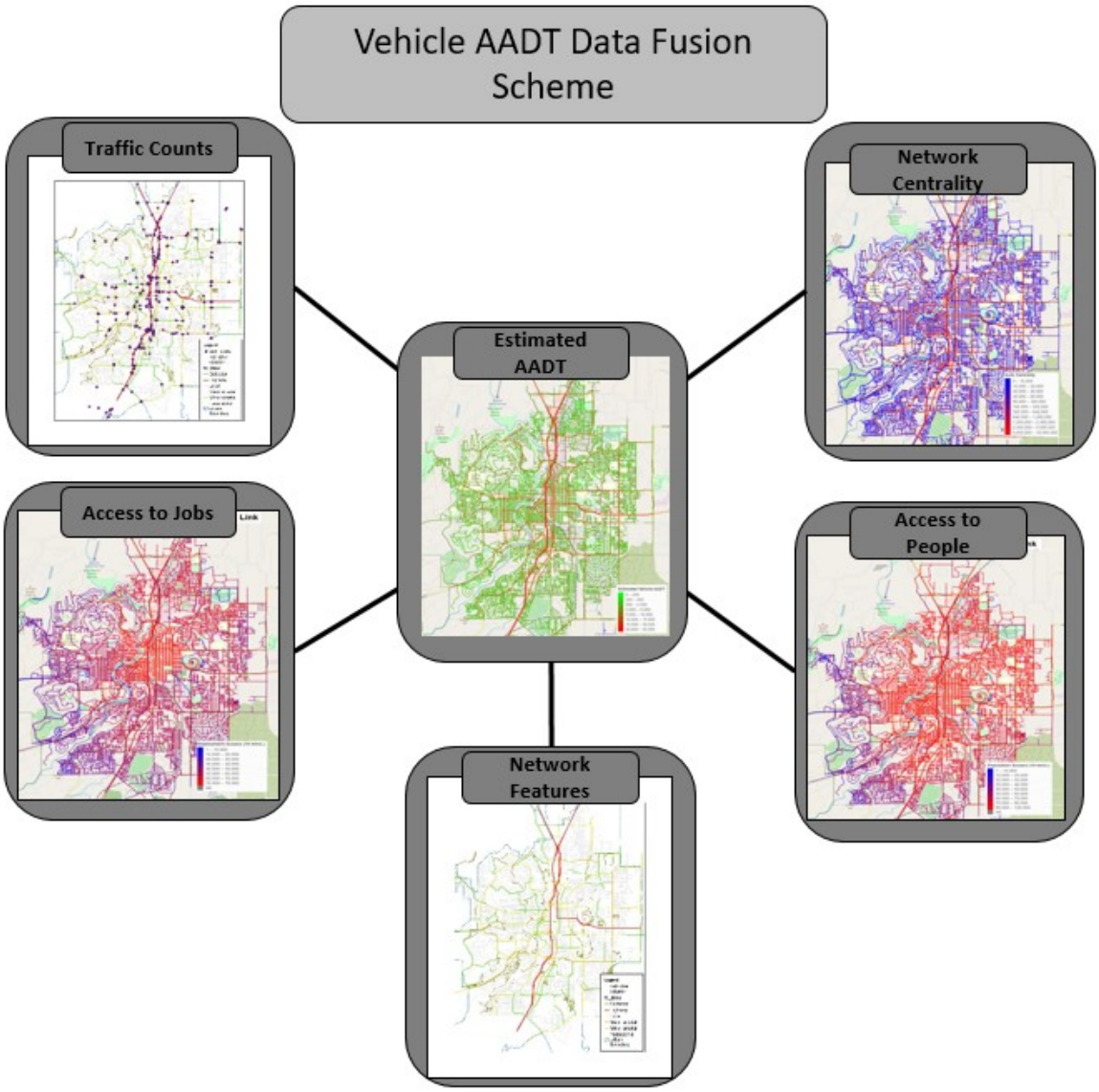

Figure 8.1: Vehicle AADT data fusion model schema

Table 8.1 below summarizes the AADT data used as the response variable in the data fusion models. The data represents two years of data, with a number of summary statistics available in the table by functional classification. As can be observed from the table, vehicle traffic volumes are generally stable with minor increases in most functional classifications. Also, most of the counts are on higher functional classification roads such as principle arterials with some counts taken on local roads. The counts taken on local roads are usually done on network links with higher volumes than most local facilities. These sites are selected because they are importance connector roads to intermodal freight facilities or some other important regional destination. That being said, these volumes are still lower than most of the other functional classifications.

Table 8.1: Vehicle Data AADT Summary

Functional Classification	Year	Vehicle AADT Summary Data					
		Minimum	Mean	Median	Std. Dev.	Max	Observations
Local	**2017**	133	719	519	893	3822	15
	2018	123	773	559	897	3957	16
Minor Collector	**2017**	250	670	560	484	1200	3
	2018	240	673	580	487	1200	3
Major Collector	**2017**	380	4173	3500	30147	9600	50
	2018	360	4077	3500	3048	10000	49
Minor Arterial	**2017**	430	9262	8900	4184	19700	101
	2018	540	9256	9050	4230	21800	100
Principal Arterial - Other	**2017**	40	16639	15800	14107	54000	87
	2018	40	16856	16300	14262	55100	88

Network data attributes used to both train and apply the model are derived from a data set created for this research project. The network data set is a fully routable graph and has a number of attributes including functional classification and posted speed that are useful as prediction features but also useful to help generate the accessibility to jobs and population data described below. Table 8.2 below summarizes the miles of network by functional classification and posted speed limit. A significant portion of the street network is represented by local streets even though fewer traffic counts are collected on those types of facilities since counts are typically very low. Standard practice for agencies is to assume a static value for local streets and apply that value to all streets of local functional classification, typically a value between 500 to 1,000 AADT per day.

Table 8.2: Network Miles by Functional Classification and Posted Speed

Functional Classification	Posted Speed							
	20	**25**	**30**	**35**	**40**	**45**	**50**	**55**
Local	0.0	423.0	0.0	0.0	0.0	0.0	0.0	0.0
Collector	1.6	16.6	12.2	10.2	6.2	6.2	0.0	0.0
Minor Arterial	1.5	10.0	3.9	26.5	7.6	10.7	0.6	0.0
Principal Arterial - Other	0.0	5.7	0.0	6.6	1.0	26.0	0.0	3.2

Figure 8.2 shows the count site locations and layout of the functional classification system across the study are in Bend MPO. As described in the table above, many of the count sites are on higher functional classification roads with many concentrated along Highway 97 corridor and supporting arterials. For the purposes of reporting, the highways in the study region, though controlled access in many parts, is classified as a principal arterial - other.

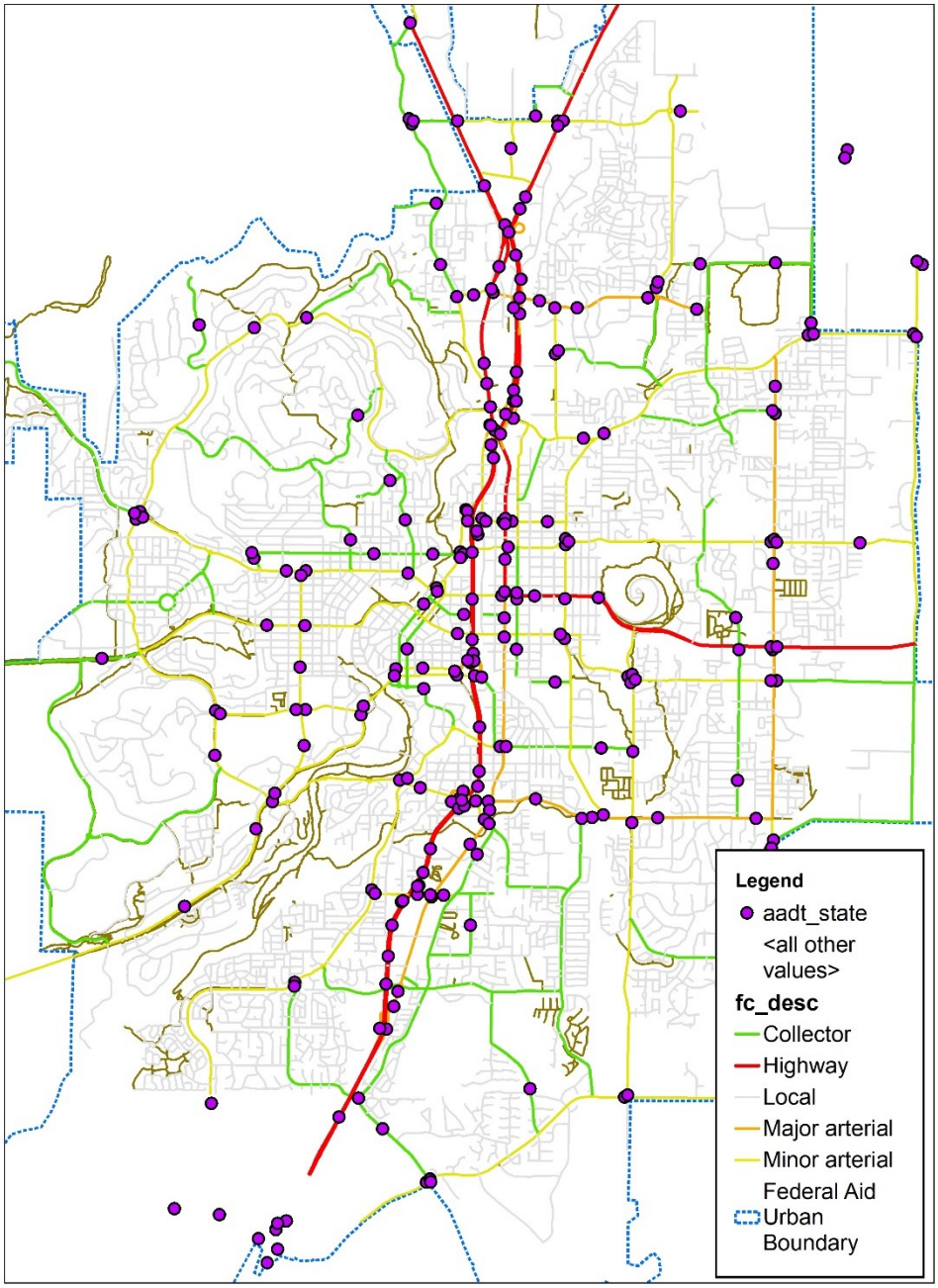

Figure 8.2: Count site location and functional classification for Bend MPO study area

Accessibility data are created by using analytic methods that combine the functionality of a routable network with data on population and employment location at a zonal level, either transportation analysis zone or Census block. Each link in the network is assigned accessibility to population and employment measures based on the number of each opportunities that can be reached by either travel time or shortest path distance. Accessibility measures are created by first calculating the drive time from each network node to the network node nearest the centroid of the Census block or transportation analysis zone (TAZ) using the igraph library (Csardi et al. 2006) within the R statistical computing environment. Link cost is either the travel time to

traverse the link based on the link length (for shortest distance) or the length and posted speed limit for driving travel time. The number of opportunities (either jobs or people) is calculated for various shortest path and drive time thresholds and then summing the number of people and jobs within these different thresholds. Different thresholds are used because different trip purposes have different trip lengths and these thresholds aim to simulate that heterogeneity in trip making decisions. However, because the travel network is relatively small most jobs and population are reachable within a low travel time and distance threshold. Figure 8.3 below shows the results for total jobs accessible within a 10 minute drive time from all given network links. The core area of the region, downtown Bend, has significantly higher access to jobs due to its proximity to jobs concentrated in this area. Figure 8.4 below shows the results of link level accessibility to population for the study region. Since population density is higher near the downtown of the study region the accessibility to these people from the links near these inner areas is also higher than the outlying areas of the region with lower population density.

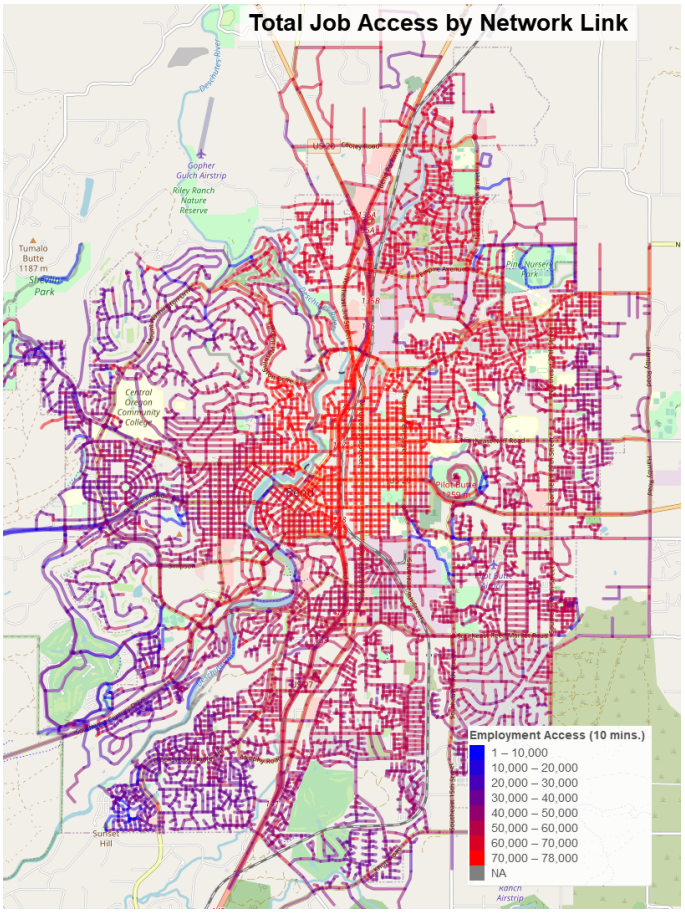

Figure 8.3: Total jobs accessible within 10 minute drive

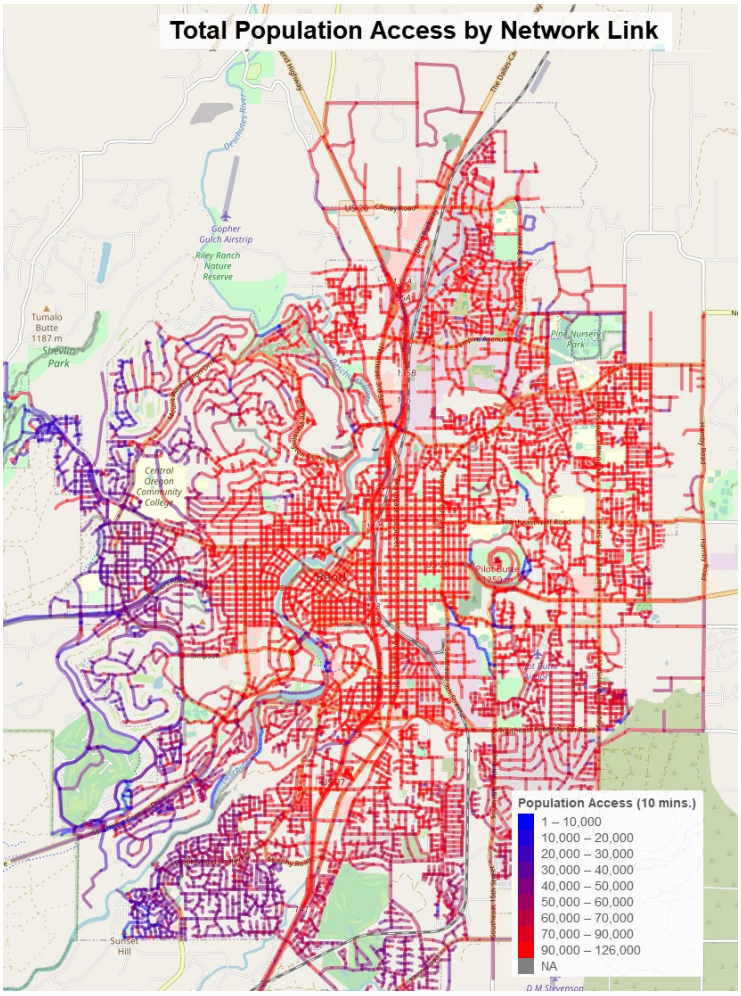

Figure 8.4: Total population accessible within 10 minute drive

The above figures show just two of the accessibility measures used in the model training but many more were created and utilized. Since the employment accessibility data is based on LEHD, there are nearly 40 types of employment [1] with many job types (manufacturing, healthcare, retail) as well as worker types (worker sex, race, and educational attainment) represented. For the population accessibility measures, total population, households, and park acres, are also included. These opportunity measures are also computed at multiple thresholds (as mentioned earlier) resulting in hundreds of features usable in the machine learning algorithms.

The next measure used for the vehicle data fusion model are measures of network centrality. In graph or network analysis, centrality is a measure of importance of nodes and their respective edges (links) to one another. This type of analysis is very common in understanding social media data, commerce and logistics but is also key to understanding traffic flow. High measures of centrality in transportation networks are nodes and links that are commonly used to traverse the network, such as one of just a few bridges over a river or high speed facilities like highways

[1] LEHD information - https://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.4.pdf

and arterials that provide quick traversal across a network. This research employed edge betweenness as is define as the number of the shortest paths, or in this case the lowest cost paths that go through a link in a network (Zhang 2013). This measure of centrality was calculated using the R statistical computing software package igraph's *edge_betweenness* function which calculates the shortest path from each node to all other nodes in the network and returns the count of trips on the traversed links. For this calculation, weights were assigned to the links to represent travel time by multiplying the link distance by the posted travel speed. Figure 8.5 below shows the results of the least cost path network centrality measure used in the vehicle data fusion model. As can be seen in the figure, higher measures of centrality are assigned to the Highway 97 and Highway 20 corridors. The next most important links are the principal arterials which also have relatively high measures of centrality. Local streets have very low measures of centrality because the frequency of their use is low when traversing the network from a given origin node.

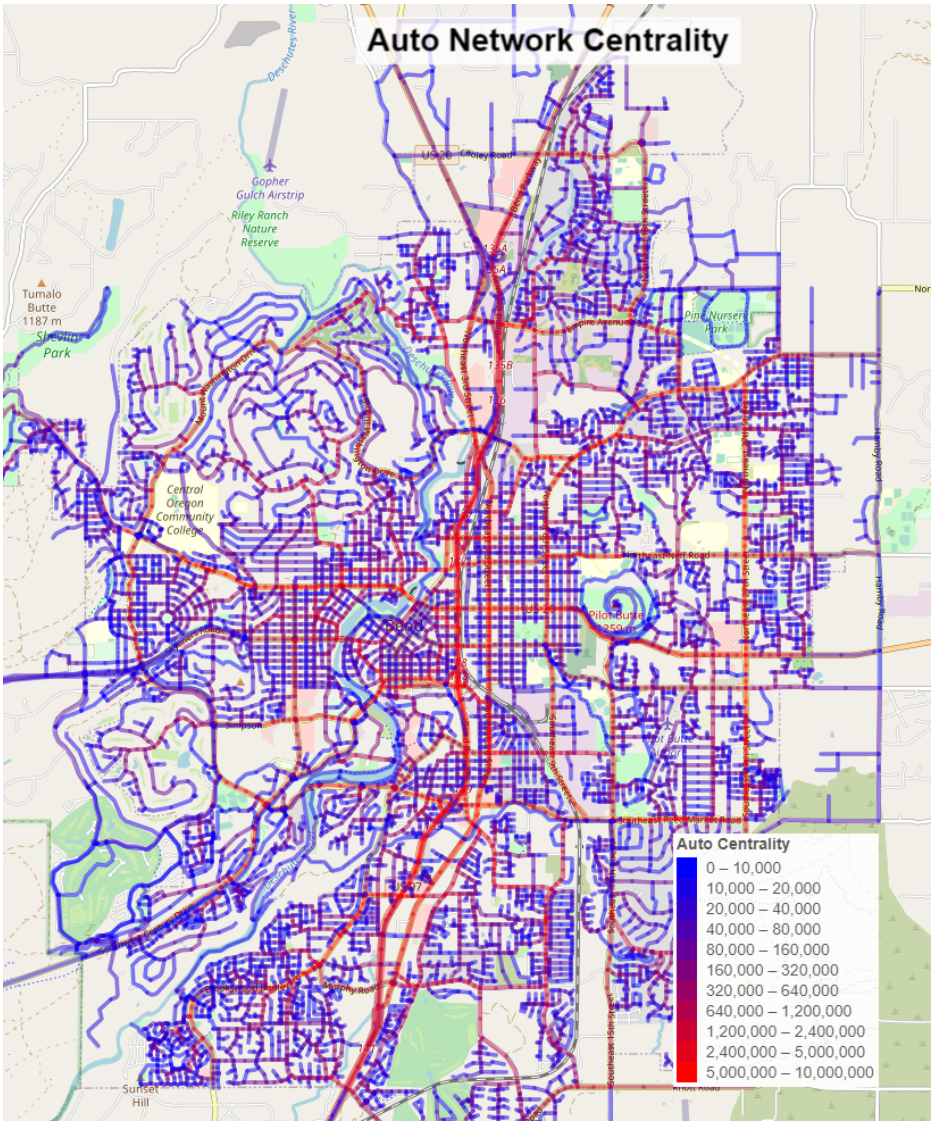

Figure 8.5: Network centrality using least cost path

8.3 VEHICLE TRAFFIC DATA FUSION MODEL RESULTS

The results of the vehicle data fusion models will be presented in four sections below. The first section will describe and summarize the machine learning based data fusion models including the features used and the cross-validation results. The second section will describe and summarize the parametric based data fusion models including the final model covariates and results of the cross validation results. For the machine learning and regression approaches root mean squared error (RMSE) and r-squared values are used to measure model performance. The third section will compare applied machine learning and regression models to known estimates of vehicle miles traveled for the study area from the Highway Performance Monitoring System (HPMS). The third section will also show model results from subset models, models in which only a subset of the AADT data are used to train the model with an objective of showing aggregate VMT estimate stability even when less data is available. The last section will offer a discussion of the model approaches and discuss the tradeoffs and opportunities for each approach.

8.3.1 Machine Learning Based Vehicle Traffic Data Fusion Model Cross-Validation Results

This section summarizes model features and cross-validation results of machine learning based data fusion models. Cross validation was done through both an internal and external cross validation process. The results presented below are based on two machine learning algorithms including extreme gradient boosting (XgBoost) and random forest. Two sets of cross validation are performed, one that is characterized as internal that uses random partitions in a 10-fold cross validation and is done as a part of the model training process within the caret package. The second cross validation process, characterized as external, is performed on a select set of model specifications with high accuracy from the first validation and uses a stratified partition to do another 10-fold cross-validation. The internal cross validation uses 10 folds and was performed twice. The internal cross validation executes rather quickly for each specification taking about 12 minutes to run using seven cores running in parallel on a four core system with eight total processors each with 3.4 Ghz processor speed. Multiple model specifications are tested in the internal validation step using two type of algorithms (XgBoost and Random Forest) with a set of selected model specification being put forward to the external cross validation process.

A key feature of machine learning algorithms are the ability to change input parameters specific to the machine learning algorithm. The purpose of tuning parameters is to find the optimal trade-off between model complexity and the training set size. For this research parameters are held constant for all the different cross validation tests with ranges of inputs described below Table 8.3. These parameters are summarized in the Appendix for select models.

Table 8.3: Hyper Parameter Description and Input Range

Parameter	Package Parameter Name	Values Used	Algorithm	Description
Boosting Rounds	nrounds	50, 75, 100	XgBoost	Corresponds to the number of boosting rounds or trees to build. Its optimal value highly depends on the other parameters, and thus it should be re-tuned each time you update a parameter. You could do this by tuning it together with all parameters in a grid-search, but it requires a lot of computational effort.
Learning Rate	eta	0.05, 0.075, 0.1	XgBoost	Step size shrinkage used in update to prevent overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.
Maximum Depth	max_depth	6 through 8	XgBoost	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
Minimum Child Weight	min_child weights	2.0, 2.25, 2.5	XgBoost	Defines the minimum sum of weights of all observations required in a child. Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting hence, it should be turned using CV.
Subsample Ratio of Columns	Colsample_ bytree	0.36, 0.4, 0.5	XgBoost	Subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.
Gamma	gamma	0	XgBoost	A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split. Higher values make the algorithm more conservative. The values can vary depending on the loss function and should be tuned.
Subsample ratio	subsample	1	XgBoost	Subsample ratio of the training stances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees, and this will prevent overfitting. Subsampling will occur once in every boosting iteration.
Split Variable Count	mtry	2 through 6	Random Forest	Number of drawn candidate variables in each split
Trees to Grow	ntree	2000	Random Forest	Number of branches will grow after each time split.

Many different kinds of training features were tested but selected scenarios are described in Table 8.4 below. The primary difference in the feature scenarios is that the features used to describe the functional classification differ. In the *Base Features + Is Ramp* scenario the functional classifications include both local and federal classifications which differ slightly with federal classifications having more classes, including a split for collector classes into a minor and major classification. In both scenarios there is a dummy variable called *Is Ramp* included to distinguish highway on and off ramps separately from the principle arterials they are classified as in the classification schemes (both local and federal).

Table 8.4: Vehicle Model Feature Scenario Description

Feature Specification	Description
Federal Fc	Accessibility calculated using least cost paths based on travel time; Auto Centrality, Local & State Functional Classification
Local Fc	Accessibility calculated using least cost paths based on travel time; Auto Centrality, Local Functional Classification only with minor/major collector

Model performance is based on RMSE and r-squared values while the number of features used in the model is also presented. The internal validation results are a product of the initial model training using the caret package in R and uses a random partitioning process, using 10 folds and performed two times. The results from the internal cross validation tests show that the XgBoost algorithm significantly out performs the random forest algorithm with a minimum r-squared value of 54% versus a 30% in the random forest. The maximum r-squared value for XgBoost is 73% while the maximum for random forest was only 43 percent. The number of features used in the XgBoost is generally fewer than the random forest.

Table 8.5: Internal Cross Validation Results for Vehicle Model

Algorithm Specification	RMSE	R-squared	Algorithm	Feature Count	Year
Federal Fc	7637	43%	Random Forest	352	2017
Federal Fc	8160	42%	Random Forest	352	2018
Local Fc	8303	30%	Random Forest	338	2017
Local Fc	8752	33%	Random Forest	338	2018
Federal Fc	5025	73%	XgBoost	163	2017
Federal Fc	5529	70%	XgBoost	143	2018
Local Fc	6631	54%	XgBoost	156	2017
Local Fc	6806	58%	XgBoost	137	2018

One way to diagnose how the machine learning algorithms are using the input features is to use a measure of variable importance. In Table 8.5 the number of features that were ultimately found to be useful in predicting AADT were summarized for each specification and algorithm. Of all of the features used in each algorithm, the top 10 most important are displayed in Figure 8.6.

This chart summarizes the relative number of times a feature is used in the splitting of trees. The top panel shows both model specifications (Federal Fc and Local Fc) for the random forest algorithm and the bottom panel shows variable importance summary for the XgBoost algorithm. The random forest results show that speed, shortest path and auto centrality, street classes with principle arterial are some of the more important features in the decision tree splitting. Access to employment features that were relatively importance include access to retail trade, and native American and Hispanic workers as well as workers ages 30 to 54 and 55 and older all within 5 minute drive time from the network link in which the count location resides.

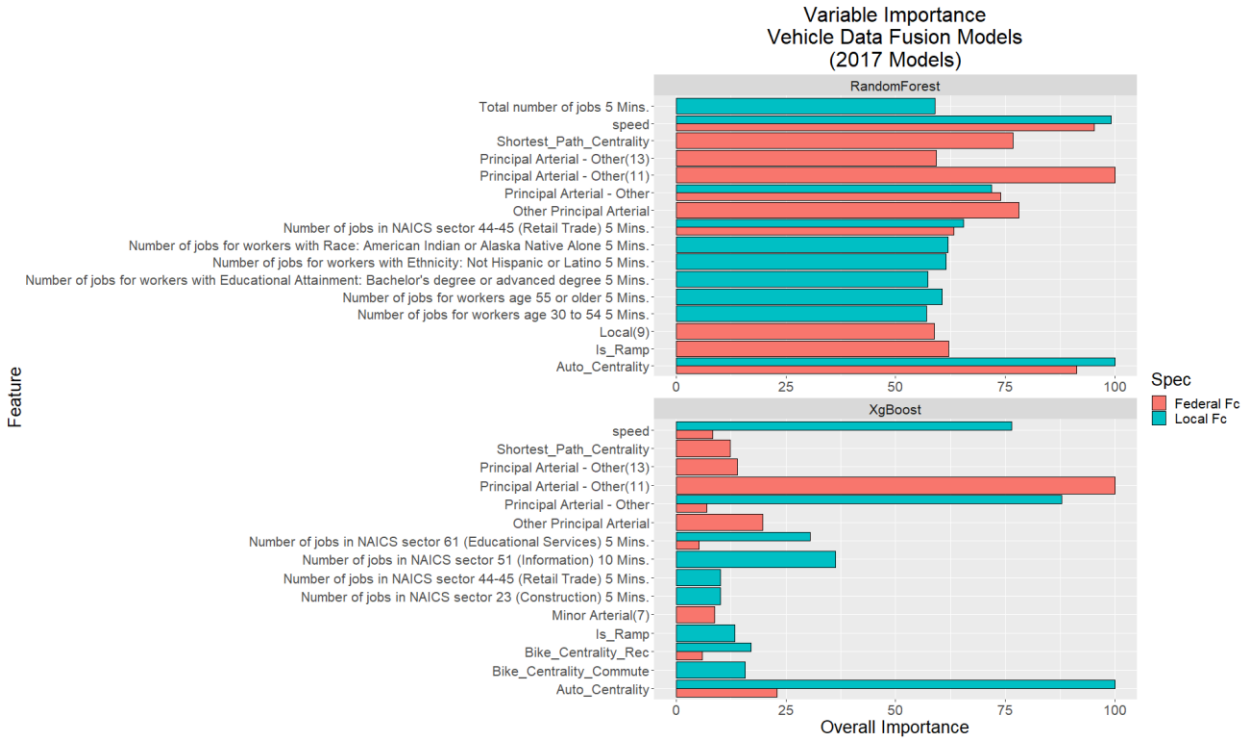

Figure 8.6: Variable importance for select vehicle data fusion models

For the XgBoost algorithm some of the same features were of most relative importance such as speed and auto centrality but also includes measures of bike centrality. Streets classified as principal arterials were also shown to be important. Access to workers variables differed compared to the random forest algorithm and include access to educational, information, retail trade, and construction jobs.

The variable used in the machine learning algorithms were selected based on a theoretical relationship to vehicle traffic counts and the variables highlighted in the variable importance charts make intuitive sense for being important in the prediction of vehicle counts. Centrality measures would be expected to be important since high centrality are places on the network with many important connections to other parts of the network. Any measure of network classification, like principal arterial, would also be expected considering those designations are in fact based on the expected volume at that location. And worker access being important is not surprising considering vehicle traffic is a proxy for economic activity, which requires workers. It should be noted that the variables displayed and discussed in the above section only include the

top 10 variables for each algorithm and specification but many others are used in these machine learning approaches. In the random forest models the number of features is up to 352 and 163 for the XgBoost algorithm.

External validation tests are performed using both a 10-fold and a leave-one-out (LOO) process. The purpose of the external validation tests are twofold with the first motivation looking to understand in more detail the prediction error by volume bin and functional classification which is not possible to extract from the internal cross validation results. The second motivation is to try and determine how much the model results might be biased by spatial autocorrelation making earlier test results somewhat biased because sites used in training may be near tests where the model is applied. To control for this, the LOO cross validation only uses sites in the training that are at least 1,000 feet from the test site.

Results from the external 10-fold cross validation analysis are presented below in Figure 8.7 and shows the mean absolute percent error by volume bin for the two model specifications and algorithm types. These results demonstrate that XgBoost model works better than the random forest for most volume bin predictions. Additionally, the Federal Fc specification seem to perform better than the Local Fc model specification, likely due to the additional categories available in the federal functional classification scheme. Generally, for all model specifications and algorithm types, the error diminishes as the volume increases. Estimating volumes at low volume of less than 1,000 AADT results in a lot of error in percentage terms, likely due to a low number of observations for roads with low volume in the training data.

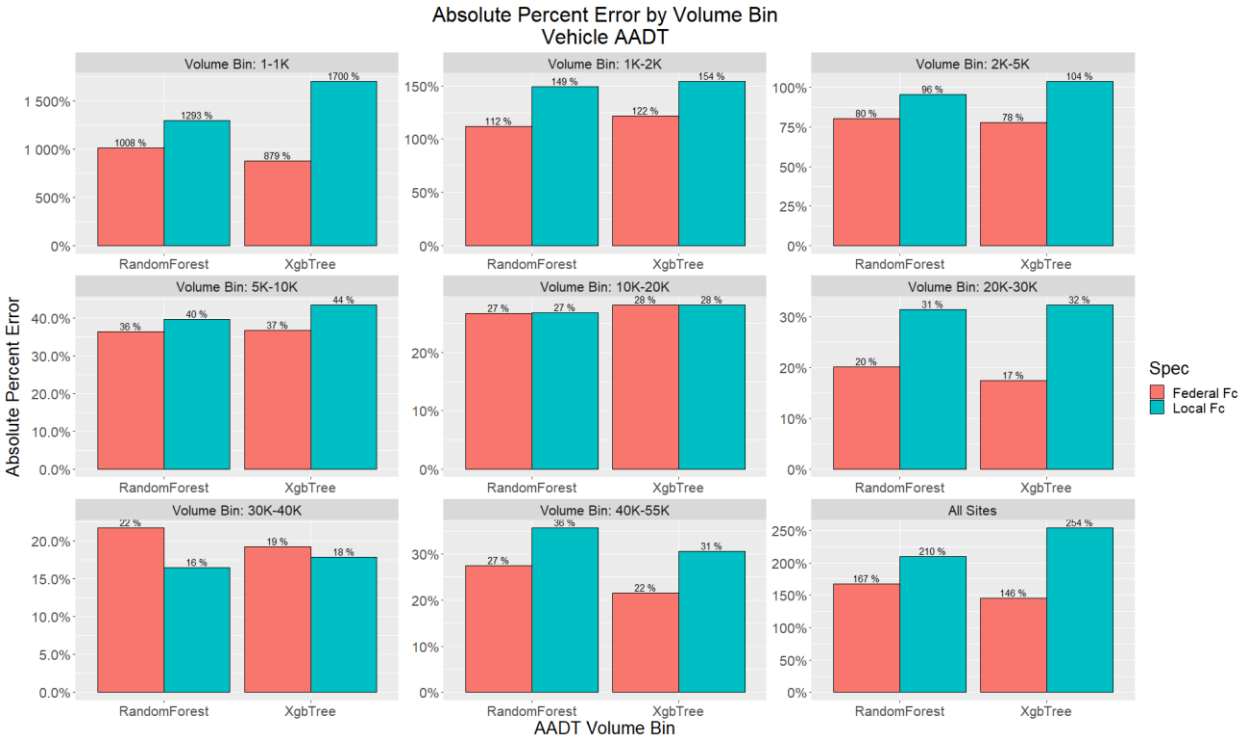

Figure 8.7: External 10-fold cross validation for vehicle models

Because the XgBoost algorithm worked best based on the internal validation and the 10-fold external validation the LOO cross validation process only tested this approach. The Local Fc specification was selected for the LOO process due to its better performance in matching HPMS data total shown in the next section. Table 8.10 summarizes the results of the LOO cross validation. Generally the error decreases with increasing model volume with the model struggling to predict well the lowest volume bin categories.

Table 8.6: External Leave-One-Out Cross Validation Results for Vehicle Model

Algorithm	Volume Bin	Absolute Percent Error		Number of Sites
		Mean	Median	
XgBoost	1-1K	711%	547%	29
	1K-2K	245%	170%	13
	2K-5K	136%	109%	47
	5K-10K	46%	30%	78
	10K-20K	30%	28%	51
	20K-30K	28%	23%	15
	30K-40K	24%	26%	5
	40K-55K	30%	29%	11
	All Sites	145%	39%	249

Figure 8.8 shows the correlation between the observed AADT and the estimated AADT from the LOO tests showing the general relationship and the trend toward over and under predicting. Higher volume roads look to be under estimated while the mid volume sites look to be about split. It would not be surprising to have the higher volume roads under predict considering this model does not account for external traffic other than in the response feature (AADT). For instance there are not training features that account for access to population and jobs outside the study area.

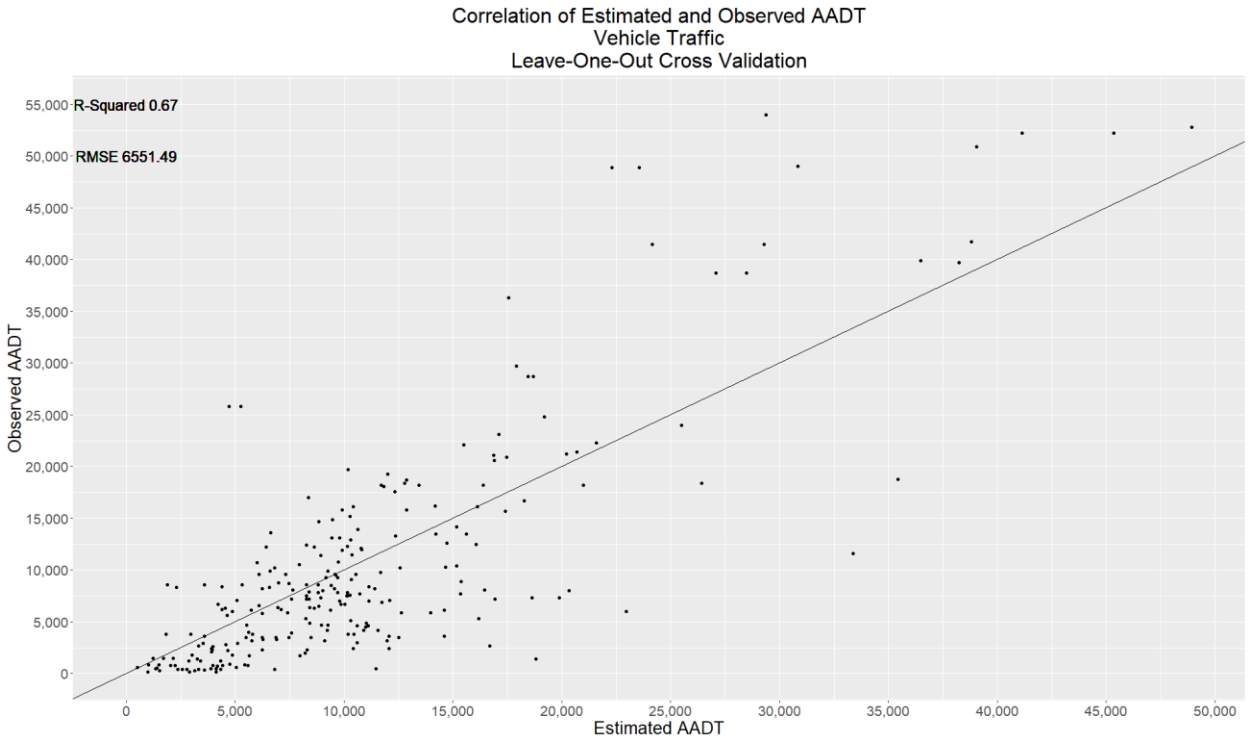

Figure 8.8: Observed and estimated AADT from LOO tests

These results reveal that error generally decreases as the volume increases. These results should be considered more rigorous compared to the 10-fold cross validation because only sites at least 1,000 feet away from the test site are used to train the model used in prediction with an aim to alleviate issues of spatial autocorrelation. Surprising when compared to the 10-fold cross validation results, the LOO results are slightly better in some cases. Table 8.7 compares the results from these two validation approaches showing that in some cases the LOO results are slightly better but based on overall (All Sites) median error the two approaches are generally telling the same story that low volume roads remain difficult to predict and error diminishes as volume increases.

Table 8.7: Comparison of 10-Fold and LOO Cross Validation Results

Volume Bin	Mean Absolute Percent Error		Median Absolute Percent Error	
	10-Fold	LOO	10-Fold	LOO
1-1K	1700%	711%	619%	547%
1K-2K	154%	245%	128%	170%
2K-5K	104%	136%	70%	109%
5K-10K	44%	46%	31%	30%
10K-20K	28%	30%	22%	28%
20K-30K	32%	28%	29%	23%
30K-40K	18%	24%	18%	26%
40K-55K	31%	30%	27%	29%
All Sites	254%	145%	40%	39%

8.3.2 Statistical Vehicle Traffic Data Fusion Model Cross-Validation Results

This section will describe the development of statistical models to estimate vehicle AADT including an exploration of the individual effects of the covariates used in the final model. Since the number of available covariates for estimating a statistical model for vehicle traffic are numerous it was necessary to use a testing procedure to determine the variables with the best model prediction accuracy. This process uses 10-fold cross-validation to test the prediction accuracy of thousands of possible model specifications. For the vehicle model 9,408 specifications are tried based on a grid of all possible combinations of select variables including population access, total employment access, retail, health, and warehouse workers, intersection density, auto centrality and shortest path centrality. All the accessibility measures use drive time network distance thresholds of 5-30 minutes with 5 minute increments. All models are estimated using a negative binomial regression specification due the counts data featuring over dispersion where the dependent variable (vehicle AADT) variance is greater than the mean of the counts which is generally the case for traffic counts data. The model is specified as linear-in-parameters with a log-link function:

$$\boldsymbol{Y_{id} \sim NegBinom(\mu_{id})} \tag{8-1}$$

$$\mathbf{log(\mu_{id}) = \beta_i X_{id}} \tag{8-2}$$

Where:

Y_{id} = Average annual daily traffic (AADT) volume at site *i*

$\boldsymbol{\beta_i}$ = Vector of parameters for count site *i*

$\boldsymbol{X_{id}}$ = Vector of observed covariates for count site *i*

A custom process was developed in R where for each year of vehicle counts available the data is partitioned into 10 folds using a stratified random sample ensuring functional classification designations are equally distributed among the folds. A negative binomial regression model is estimated on each of the k-1 groups (training data) and then estimated on the k-9 (test data) and then compared to the observed data. For each selection of variables three performance metrics are computed include RMSE, mean absolute percent error (MAPE) and adjusted r-squared. Based on these metrics models top performing models are selected for further examination. For the vehicle models the final estimated parameters are presented in Table 8.10 for three select models. Model results below present the estimated coefficient and the associated standard error and p-value for selected models with the highest r-squared, the lowest RMSE, and lowest MAPE for the two periods available including 2017 and 2018 data.

Table 8.8: Regression Results for Vehicle Model

Coefficient	Std. Error	z value	P-value	Feature_Update	Year	Metric
8.741E-04	0.0003	2.6276351	0.0086	Population_30	2017	Highest R Squared
9.700E-04	0.0004	2.603193	0.0092	Num_Intersections_15		
1.8315	0.2107	8.6922012	0.0000	Major Collector		
2.5670	0.1954	13.139571	0.0000	Minor Arterial		
-0.0130	0.5332	-0.02435	0.9806	Minor Collector		
3.1780	0.2028	15.672757	0.0000	Principal Arterial – Other		
-1.30279	0.1538	-6.6984048	0.0000	Is_RampTRUE		
7.501E-04	0.0003	2.2773467	0.0228	Population_30	2018	
1.247E-03	0.0004	3.3523657	0.0008	Num_Intersections_15		
1.7426	0.2038	8.5509856	0.0000	Major Collector		
2.5017	0.1877	13.330093	0.0000	Minor Arterial		
-0.0624	0.5243	-0.119005	0.9053	Minor Collector		
3.1341	0.1945	16.11221	0.0000	Principal Arterial – Other		
-1.0519	0.1522	-6.909997	0.0000	Is_RampTRUE		
-3.613E-05	2.40E-05	-1.507485	0.1317	Total number of jobs 15 Min.	2017	Lowest RMSE
9.332E-04	3.75E-04	2.4875365	0.0129	Population_30		
9.513E-04	0.0004	2.4672678	0.0136	Num_Intersections_15		
1.245E-07	2.39E-08	5.2207689	0.0000	Auto_Centrality		
1.6891	0.2027	8.3345421	0.0000	Major Collector		
2.3093	0.1942	11.892813	0.0000	Minor Arterial		
0.0734	0.5095	0.1440024	0.8855	Minor Collector		
2.8938	0.2040	14.185107	0.0000	Principal Arterial – Other		
-0.9837	0.1526	-6.447746	0.0000	Is_RampTRUE		
-3.522E-05	2.38E-05	-1.450019	0.1389	Total number of jobs 15 Min.	2018	
8.474E-04	0.0004	2.2815215	0.0225	Population_30		
1.087E-03	0.0004	2.8286788	0.0047	Num_Intersections_15		
1.250E-07	2.37E-08	5.2638814	0.0000	Auto_Centrality		
1.5959	0.1964	8.1277663	0.0000	Major Collector		
2.2314	0.1870	11.931655	0.0000	Minor Arterial		
8.747E-03	0.5009	0.0174619	0.9861	Minor Collector		

2.8423	0.1963	14.481998	0.0000	Principal Arterial – Other		
-0.9677	0.1515	-6.387225	0.0000	Is_RampTRUE		
-5.552E-05	2.56E-05	-2.169612	0.0300	Total number of jobs 15 Min.	2017	Lowest MAPE
1.902E-03	0.0010	1.9532807	0.0508	Population_30		
-5.684E-03	0.0085	-0.670942	0.5023	Num_Intersections_30		
2.704E-04	0.0001	2.0563793	0.0397	Number of jobs in Retail Trade 10 Min.		
8.253E-05	4.69E-05	1.7608205	0.0783	Jobs in Health Care and Social Assistance 10 Min.		
-1.232E-03	0.0005	-2.408753	0.0160	Jobs in Transportation and Warehousing 10 Min.		
1.259E-07	2.41E-08	5.2277651	0.0000	Auto_Centrality		
1.6464	0.2018	8.1600414	0.0000	Major Collector		
2.2584	0.1943	11.623813	0.0000	Minor Arterial		
-0.0548	0.5039	-0.108825	0.9133	Minor Collector		
2.8102	0.2039	13.783979	0.0000	Principal Arterial – Other		
-0.9897	0.1515	-6.534239	0.0000	Is_RampTRUE		
-5.154E-05	2.55E-05	-2.022899	0.0431	Total number of jobs 15 Min.	2018	
1.358E-03	0.0010	1.4277074	0.1534	Population_30		
-2.080E-03	0.0082	-0.252644	0.8005	Num_Intersections_30		
3.294E-04	0.0001	2.4860195	0.0129	Jobs in Retail Trade 10 Min.		
5.305E-05	4.60E-05	1.1529826	0.2489	Jobs in Health Care and Social Assistance 10 min.		
-1.226E-03	0.0005	-2.40941	0.0160	Jobs in Transportation and Warehousing 10 Min.		
1.263E-07	2.40E-08	5.2576776	0.0000	Auto_Centrality		
1.5637	0.1964	7.9613189	0.0000	Major Collector		
2.1749	0.1878	11.579602	0.0000	Minor Arterial		
-0.1139	0.4965	-0.229366	0.8186	Minor Collector		
2.7408	0.1967	13.937237	0.0000	Principal Arterial – Other		
-0.9881	0.1510	-6.543958	0.0000	Is_RampTRUE		

Most of the selected variables are significant within the 0.05 level of significance though some variables not commonly found to be significant at this level include categorical variable for minor collector and a few of the access to jobs variables. The minor collectors are probably not significant because there are so few observations in the counts data on this functional classification. Table 8.9 below summarizes the three select models error measures. These error measures omitted the lowest volume bin (1-1K) since the error for sites with this range of volume were very high.

Table 8.9: Model Diagnostic Information for Vehicle Regression Models

Specification	Performance Metric	MAPE	RMSE	Adjusted R-Squared
Population_30 + Num_Intersections_15 + AADT + Fc_Desc + Is_Ramp	Highest R-Squared	63%	7821.1	0.422
C000_15 + Population_30 + Num_Intersections_15 + Auto_Centrality + AADT + Fc_Desc + Is_Ramp	Lowest RMSE	56%	7746.3	0.419
C000_15 + Population_30 + Num_Intersections_30 + Retail_10 + Warehouse_10 + Healthcare_10+	Lowest MAPE	55%	7896.9	0.397

The 10-fold holdout analysis results are further summarized by volume bin (this time including AADT within the 1-1K range) detailing the median APE for each of the models. The model with the lowest median APE for all sites is the same model with the highest r-squared while the model with the lowest mean APE has the highest median APE of the three models compared.

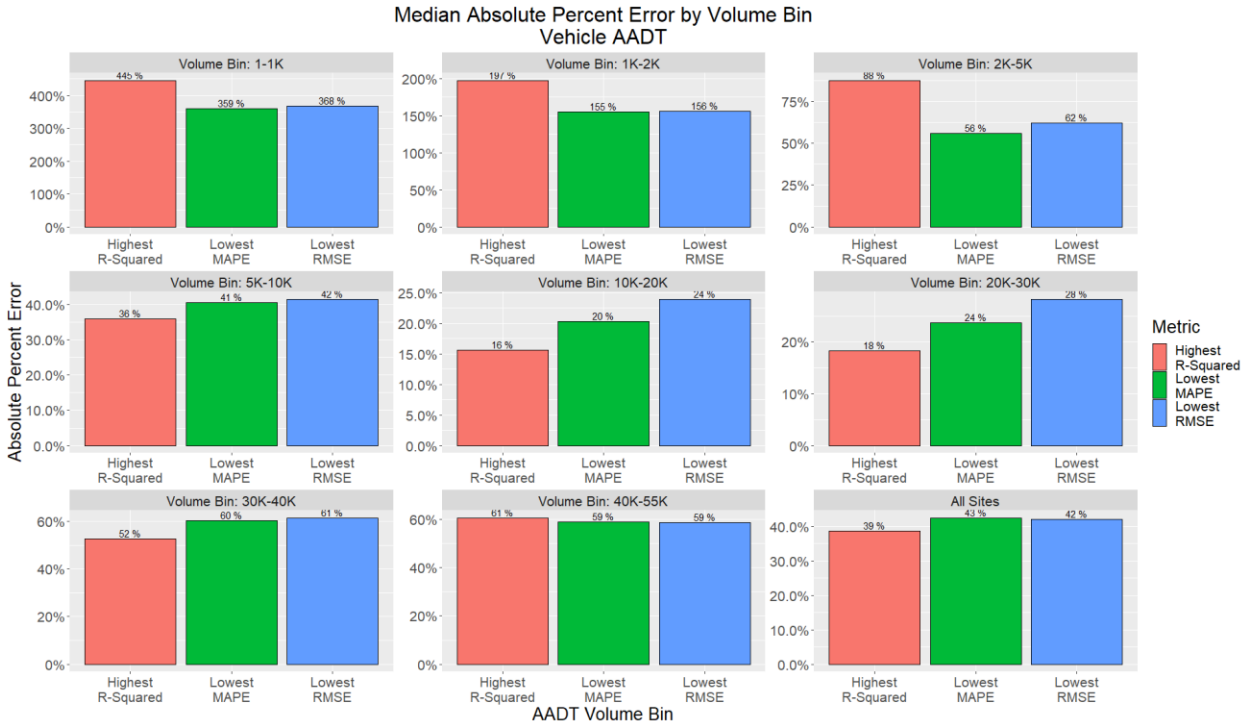

Figure 8.9: Top vehicle regression model median absolute percent error by volume bin

8.3.3 Vehicle Model Comparison with Federal Reporting Data (HPMS)

This section will compare the results from a network wide application of the various data fusion models against aggregate VMT estimates from the Highway Performance Monitoring System (HPMS) to help gauge total vehicle activity estimation value. The HPMS VMT data is submitted by state DOTs on an annual basis for each urban area within the state. VMT estimates are submitted for each federal functional classification. For the purposes of comparing data fusion model system wide VMT estimates, functional classifications will be reduced to just four categories, including collector (combine minor and major) minor arterial, and principal arterial (classified as principal arterial – other) and total VMT. Local streets are used in the data fusion model training but because HPMS reporting for local streets assumes a blanket average for all streets the two outcomes are not comparable. Additionally, even though models were trained for 2017 and 2018, since the employment data used in the training was for 2017 only the estimate from that year will be compared.

Figure 8.10 below shows the results from this comparison for two machine learning model specifications, *Federal Fc* and *Local Fc* demonstrating that in both models specifications, VMT can be estimated within a relatively low margin of error compared to the HPSM estimate. The XgBoost algorithm appears to outperform the random forest in this comparison, with percent differences of -6%, 10%, and 2% and 2% for total VMT, principal arterial, minor arterials, and collectors respectively. Even though the random forest model produces a total VMT estimate near 0% in both model specifications shown in the figure, it appears that the over- estimate of the collector and minor arterials helps to offset the under-estimate of the principle arterial, making the total VMT look pretty close to the HPMS estimate.

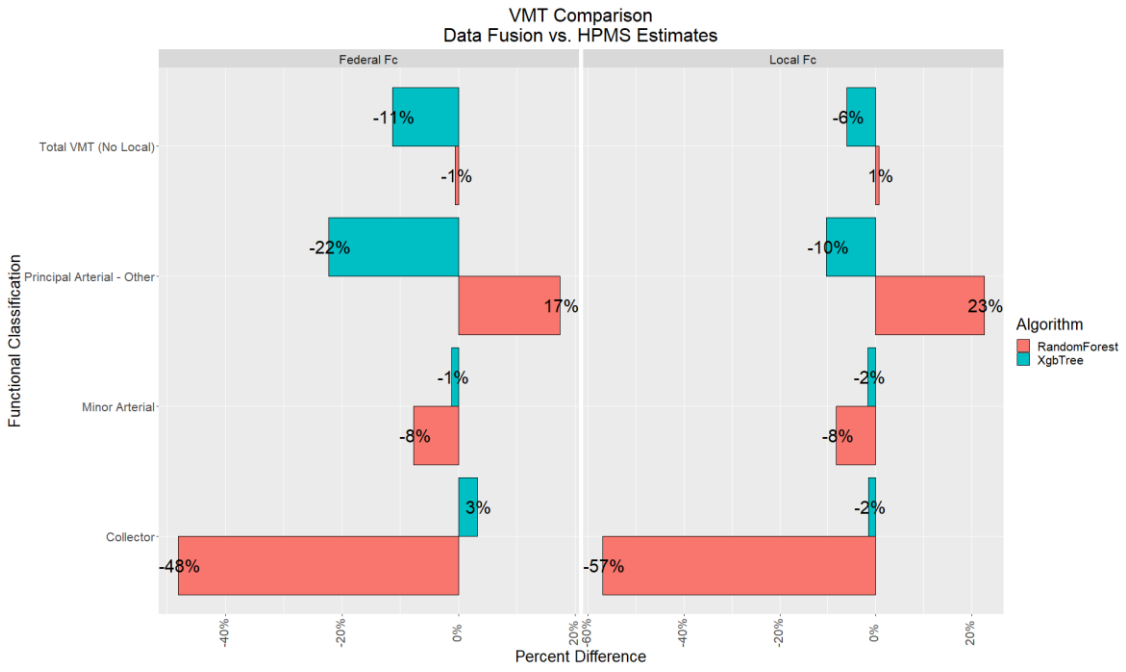

Figure 8.10: Comparison of data fusion and HPMS VMT estimates

The next figure shows the top selected regression model results when compared to the HPMS VMT estimates and shows the three models perform similarly when compared to the HPMS figures. Overall error is lowest for the model where the r-squared was highest but that is partially because the model over estimates in the principal and minor arterials and then under estimates in the collectors. However the highest r-squared model does the best for the collector and principal arterial.

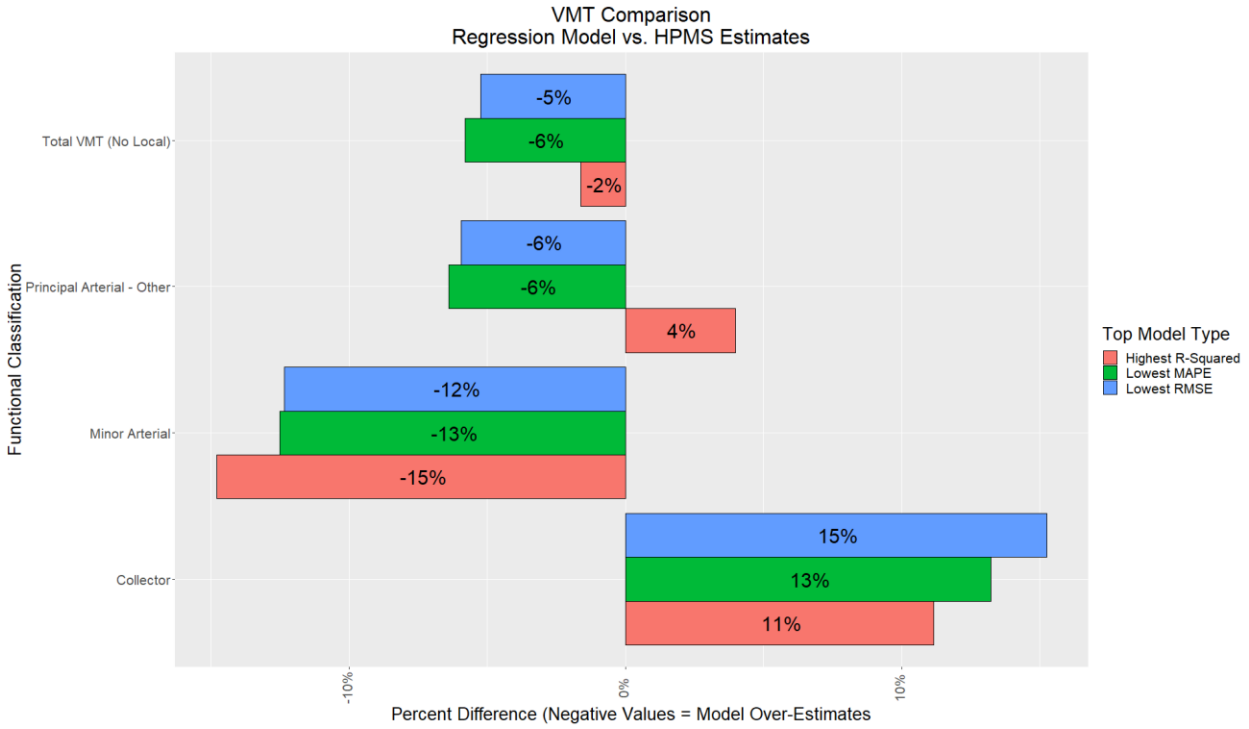

Figure 8.11: Comparison of VMT estimates by regression specification

The figure below compares the top model from the machine learning tests and the regression models, (*XgBoost; Local Fc & Lowest MAPE* respectively) to demonstrate how each performed when estimating network wide VMT. Both models perform well and though the negative binomial model looks best when comparing the total VMT, as mentioned above this results looks like this partially because of over and under estimation within the other functional classifications. The machine learning model consistently over estimates within each functional classification performing better in the collector and minor arterial category but then does worse than the regression model for the total VMT on principal arterials.

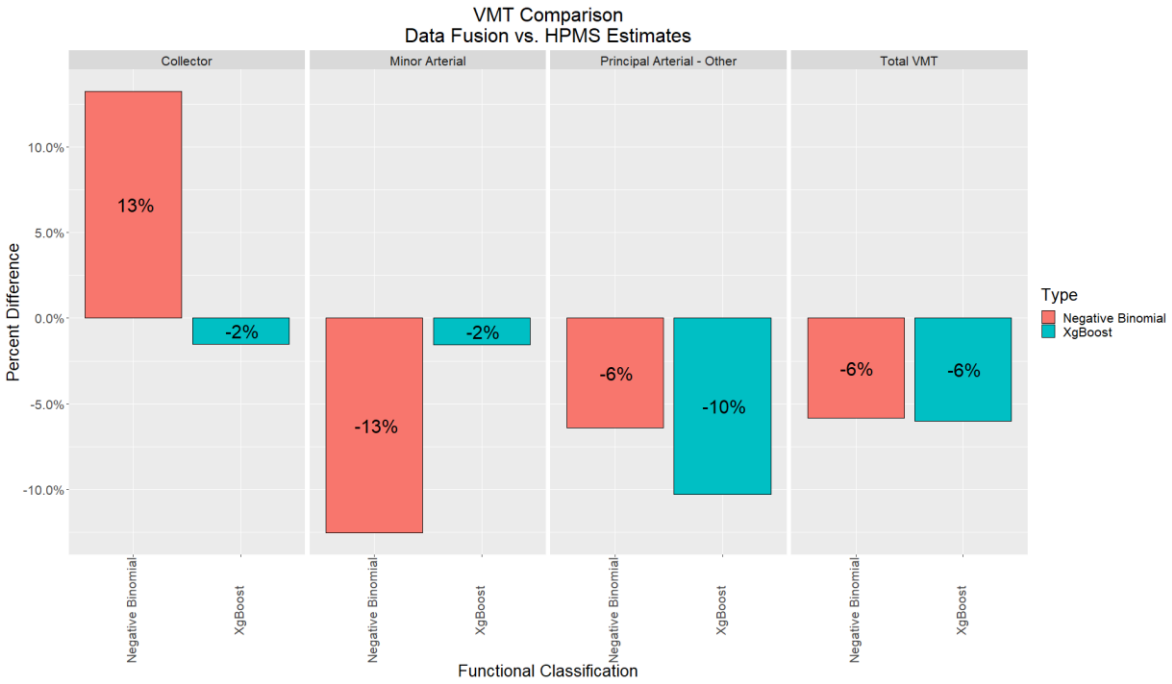

Figure 8.12: Comparison of VMT estimates by estimation method

The comparison above shows that the data fusion model technique works well to estimate the reported HPMS VMT when data from all 250 count stations are used in the model training. In one sense the two estimates are not all that independent from one another. In the HPMS estimate the vehicle count on a given roadway segment is multiplied by the distance to get the VMT for that segment. In the data fusion approach however, network links, or edges, are not summarized to segments the way they are in the HPMS reporting. In HPMS reporting roadway links are aggregated together to form reporting segments when the segment is believed to have the same volume along all of its component links. In the data fusion model all links are assigned an AADT value in a more disaggregate fashion with links ending at each intersection in the network. Figure 8.13 shows an example on the study area network where a set of links (right) is aggregate to represent a segment (left) when doing HPMS reporting. The point of describing these differences is to point out that the VMT comparisons are comparing an aggregate HPMS network represented by 312 segments with a much more disaggregate network of 13,458 links or edges in the data fusion model.

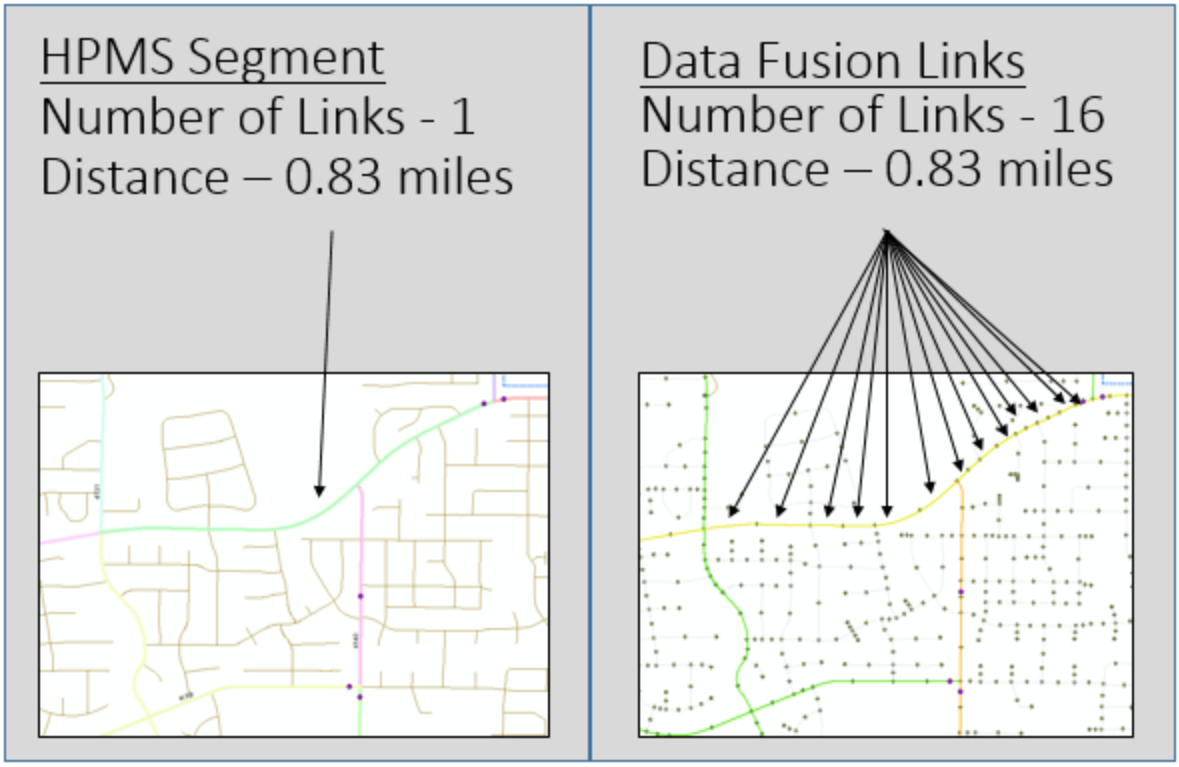

Figure 8.13: Comparison of HPMS segments and data fusion model links

In addition to showing how the machine learning approach works for estimating network wide VMT when using all count sites available, roughly 250 sites with an AADT estimate. However, for the bicycle and pedestrian models below the AADT data are more sparse and so it's of interest to test the VMT estimate stability when only a subset of the 250 sites (roughly 83 sites) are used to train the model. This scenario is more analogous to the count data circumstances the nonmotorized models will be working within.

Figure 8.14 below shows the VMT estimates from the selected machine learning model, aggregated to functional classification, compared to the VMT estimates from HPMS. The mean difference of all the subset models is also shown in the top panel. The total VMT estimate error ranges from 3% in subset model 2 to -14% in subset model 3 though model 2 performance is impacted by an underestimate in principal arterials which then offsets the over estimates in minor arterial and collectors. Similar to the comparison results using full data presented in Figure 8.10 above, the collector facilities have the most error with error ranging from -20% to -59 percent. Minor arterials error is lowest with -8%, 0% and 18% for subset models 1 through 3 respectively. These results indicate that for high level reporting, VMT estimates using subsets of the full data available are relatively stable. These results should lend additional confidence to the results presented in the nonmotorized models below.

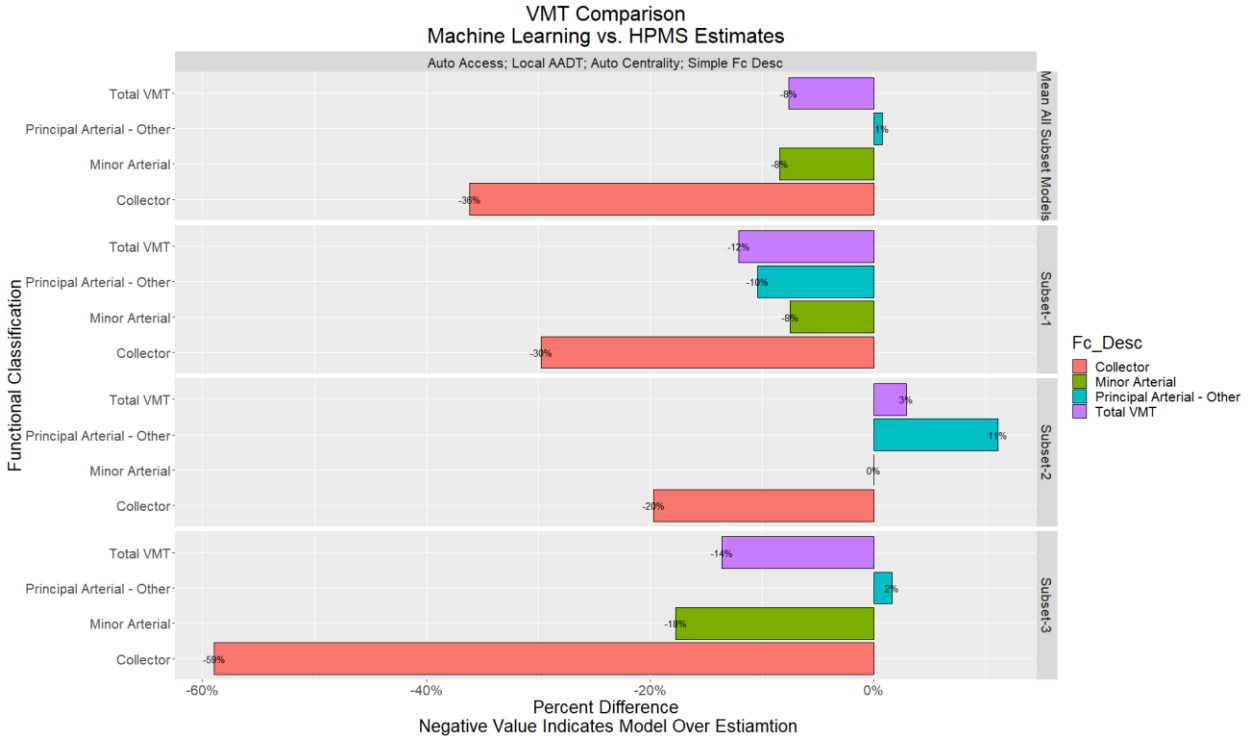

Figure 8.14: Subset model comparisons with HPMS

8.3.4 Vehicle Traffic Data Fusion Model Discussion

The above sections describe the data and modeling results from the estimation and application of the data fusion approach using machine learning and regression. Results from three cross validation are presented for the machine learning model testing in order to present the performance of different model specifications and machine learning algorithms. Generally, the results show that predicting accurately on lower volume roads is a challenge with error diminishing as volume increases. Results from the 10-fold and LOO cross validation are comparable but the LOO is more rigorous because it removes near sites from the training data to ensure that cross validation results are not biased by having near neighbors in the estimation process. Cross-validation using the regression approach are comparable with the machine learning for the overall median error. But as demonstrated in Figure 8.15 below the machine learning model performs worse on roads with volumes of between 1K and 10K and then 40K+ while the regression model performs between in the other volume categories.

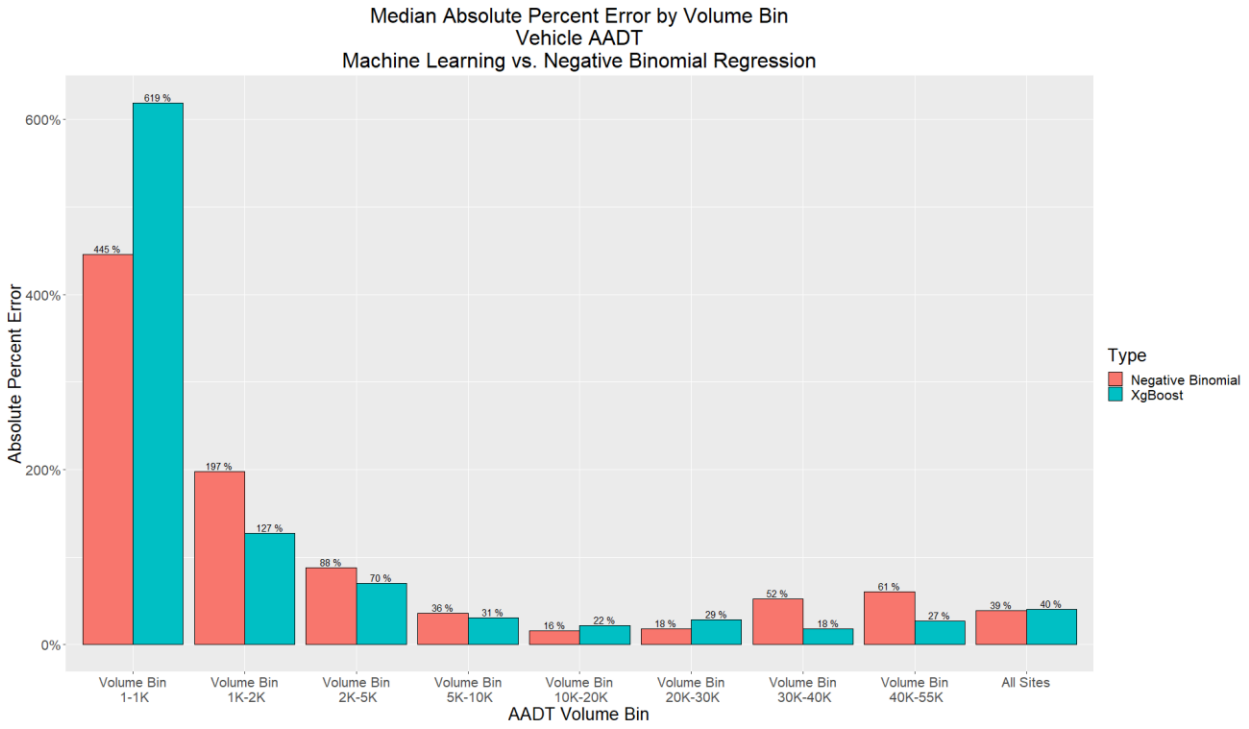

Figure 8.15: Median error by volume bin by estimation type for vehicle data models

These results would likely be better is some kind of probe data were used to help train the models but no probe data were available for this effort. Future research should explore the use of probe data in improving these modeling approaches.

Even though median error in the cross validation tests range from 16% to 619%, models applied for total VMT estimation reveal similar results with the selected model matching all functional classification summaries by 13% or less. The high accuracy when compared to HPMS is the best evidence that these data fusion modeling approaches may work well for estimating bicycle miles and pedestrian miles traveled when deployed using nonmotorized specific data. Evidence of the stability of these approaches is provided in Figure 8.14 where the results of three subset models, with roughly 80 count locations per model, are presented. The results from each model compare relatively well with the VMT estimates from HPMS though collector streets continue to perform worse than desired.

8.4 BICYCLE TRAFFIC DATA FUSION MODEL

This section on bicycle data fusion modeling will be divided into two parts with the first part describing the data used in the machine learning and regression based data fusion modeling while the second part details the cross-validation tests and final application results of the two models. The second part will feature a discussion of the trade-offs between the two data fusion modeling approaches.

8.5 DATA DESCRIPTION FOR BICYCLE TRAFFIC FUSION MODELS

A number of features used in the bicycle data fusion model are described in the section below. Figure 8.16 shows the overall data fusion schema and presents key network features used to train the bicycle data fusion model. As noted above in the vehicle model data description, this schema representation does not show all features used, for instance the access to jobs feature shown in the figure below actually has over 600 different versions when all worker industry, demographic, and access threshold combinations are computed.

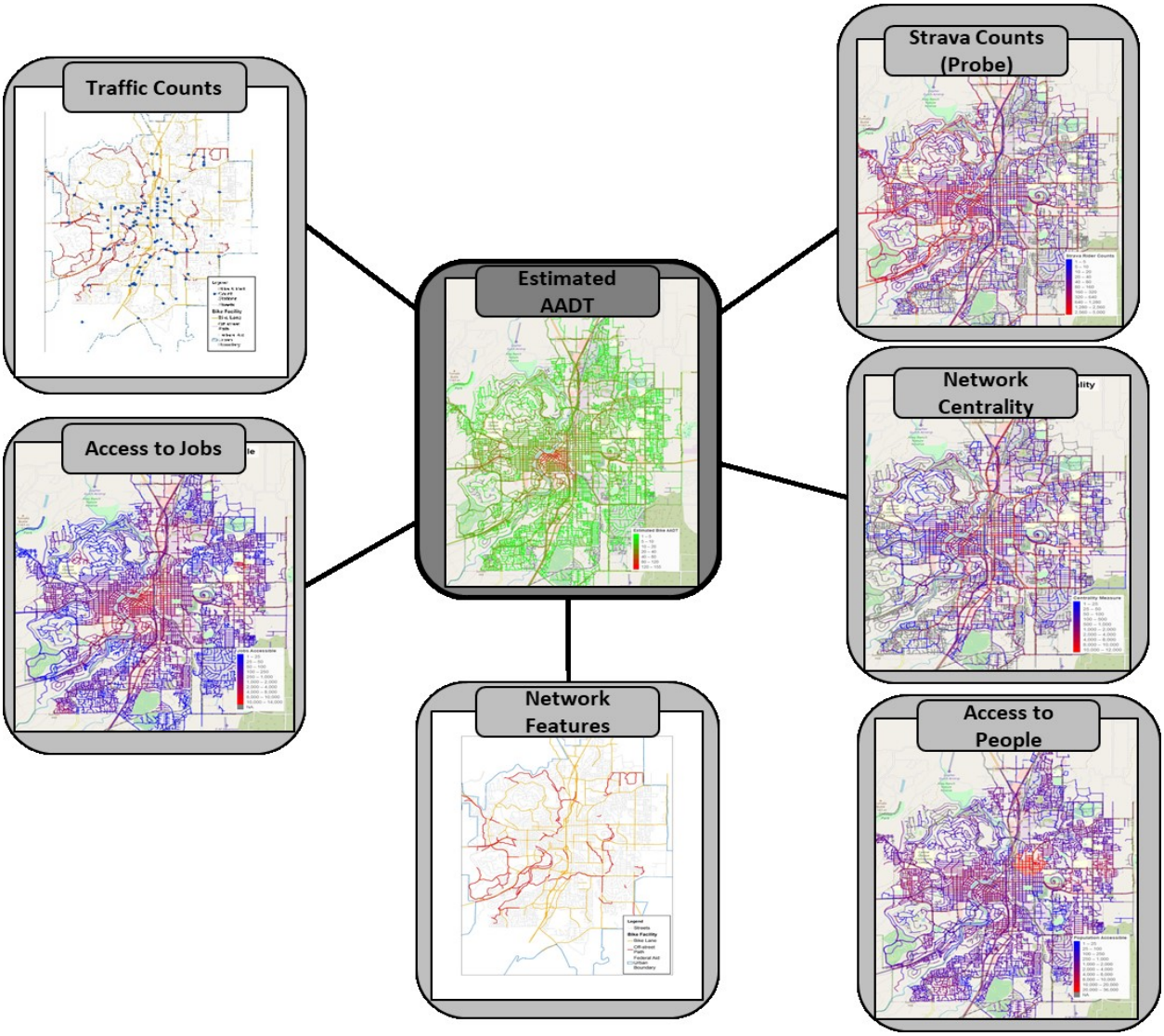

Figure 8.16: Bicycle data fusion model schema

Table 8.10 below summarizes the AADT estimates for bicycle traffic by bicycle facility type for three years of data. In addition to the summaries by year, summaries for rolling averages are also presented where *2017/2018* denoted the average of those years of data. These summaries are constructed and presented below to take advantage of as much of the counts data as possible for data fusion modeling and are used in place of applying a growth factor. Observations from

the table below show that bicycle traffic volumes on off-street paths are higher (on average) than streets with bike lanes which are generally higher than places with bike lanes. The counts appear to be trending downwards for each facility type though the averaging of years helps to smooth those declines out reducing the influence of any single year.

Table 8.10: Bicycle Traffic Count Summary

Bicycle Facility	Year	Bicycle AADT Summary Data					
		Minimum	Mean	Median	Std. Dev.	Max	Observations
No Facility	2017	10	42.4	21	55.1	140	5
	2018	15	53	25	51.9	170	11
	2019	2	24.2	23.5	13.5	57	20
	2017/2018	12	37.8	23.5	43.9	170	12
	2018/2019	2	24.5	19.5	31	170	26
Bike lane	2017	9	64.3	55	45.6	151	13
	2018	3	43.4	30.5	42.8	187	38
	2019	2	36.8	27	39	183	33
	2017/2018	3	22.6	20	15.6	84	38
	2018/2019	3	23.4	20.5	17	82	48
Off-street path	2017	39	89	101	34.2	115	4
	2018	4	63.3	45.5	56.4	205	18
	2019	5	56.5	45	42.3	159	13
	2017/2018	4	57.5	43.6	54.3	205	18
	2018/2019	4	47.8	34.6	44.1	182	20

Bicycle traffic have been collected at nearly 100 locations over the three years where data was actively collected. Those locations are displayed below in Figure 8.17. Many of the locations are on facilities where bicycle users would be expected to use and thus inserts a certain amount of bias where the model would likely be biased upward, especially at sites with very low or zero bicycle activity. Later in the report an approach is proposed to handle the issue of having no zero counts in the observed bicycle traffic counts data. Another feature shown in Figure 8.17 is the bicycle specific network elements including the location of bicycle lanes and off-street paths. These will also be used in the model training process.

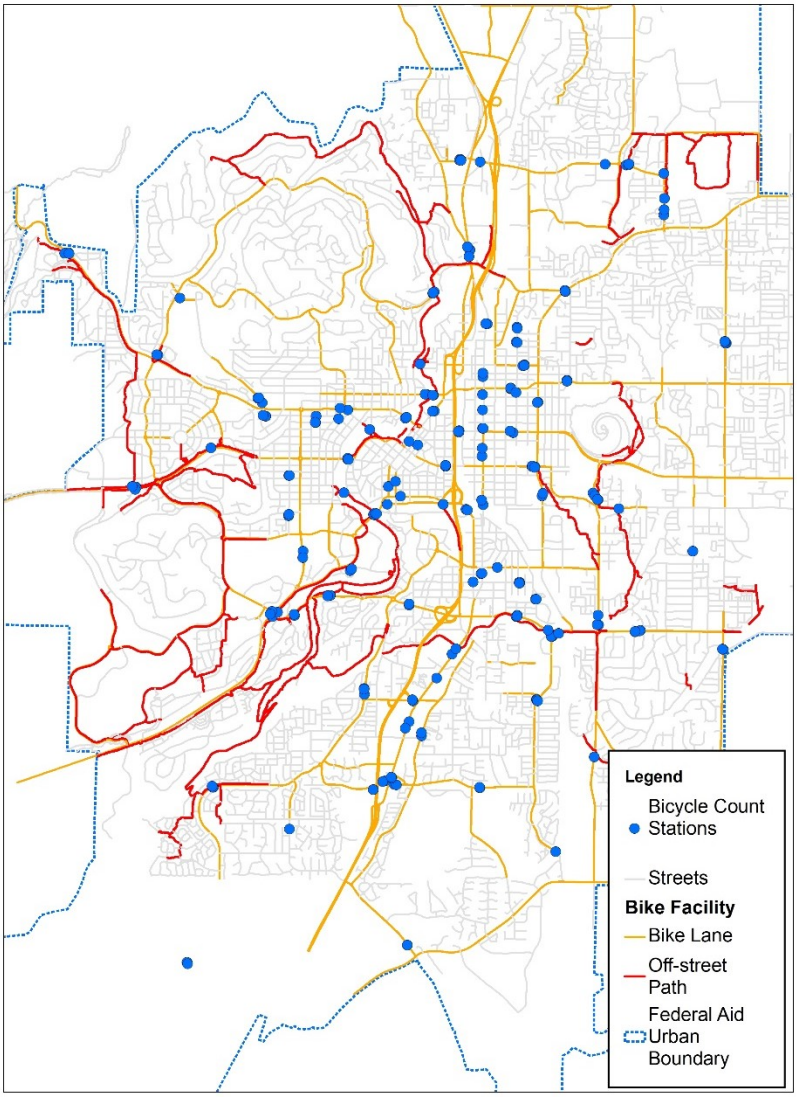

Figure 8.17: Bicycle Count Locations

Table 8.11 below summarizes the number of miles of network in a cross classification table of bicycle facility and functional classification. The region is currently implementing bicycle boulevards, or neighborhood greenways, but currently this designation will not be used in this research. The two bicycle facilities are bicycle lanes either on one or both sides of the street and an off-street multi use path. It should be noted that the two highways that intersect the study region, Highway 97 and Highway 20, are technically classified as principal arterial – other but are summarized below as highway to emphasize that these facilities have bike lanes, including on the on and off ramps. No direct bicycle traffic count measurement of these facilities has been taken but activity would be expected to be limited due to the high speed, high vehicle volume conditions.

Table 8.11: Bicycle Network Summary

Functional Classification	Bicycle Facility Type			
	No Bike Facility	Bicycle Lanes	Off-street path	Total
Highway*	0.0	23.5	0.0	23.5
Principal Arterial - Other	1.6	14.8	0.0	16.5
Minor Arterial	5.8	54.9	0.0	60.7
Collector	20.6	32.3	0.0	52.9
Local	419.7	3.4	0.0	423.0
Off-street path	29.1	0.0	50.8	79.8
Total	476.8	128.9	50.8	656.4

*Officially a Principal Arterial - Other but functions very much like a controlled access freeway/highway

As mentioned above the network wide training features differ from the vehicle models in many cases to better account for how people on bicycles use the system. For instance, when calculating the network centrality and accessibility measures we do not assume the link costs are based on posted speed and so different links appear more important in the 'bicycle' centrality measure. Similarly, access to jobs and population are more limited because people on bikes are not willing to travel as far as someone in a vehicle to access amenities. Additionally, instead of using a drive time metric for measuring access a distance threshold is used. Figure 8.18below displays one of the nonmotorized specific measures and shows the total jobs accessible within a 0.5 mile bicycle ride. Similar to some of the observations found above in the vehicle model inputs for total jobs access, employment centers can be seen in the figure with concentrations of jobs in the downtown core and north east where a large hospital resides. It should be noted similar to the accessibility measures created for the vehicle data fusion model above, multiple measures of accessibility have been created for the bicycle models using all available job types in the LEHD data. Additionally, accessibility was calculated using multiple distance thresholds of half-mile increments from 0.5 to 6.0 miles. All of these features are tried in the machine learning training though not all end up being important.

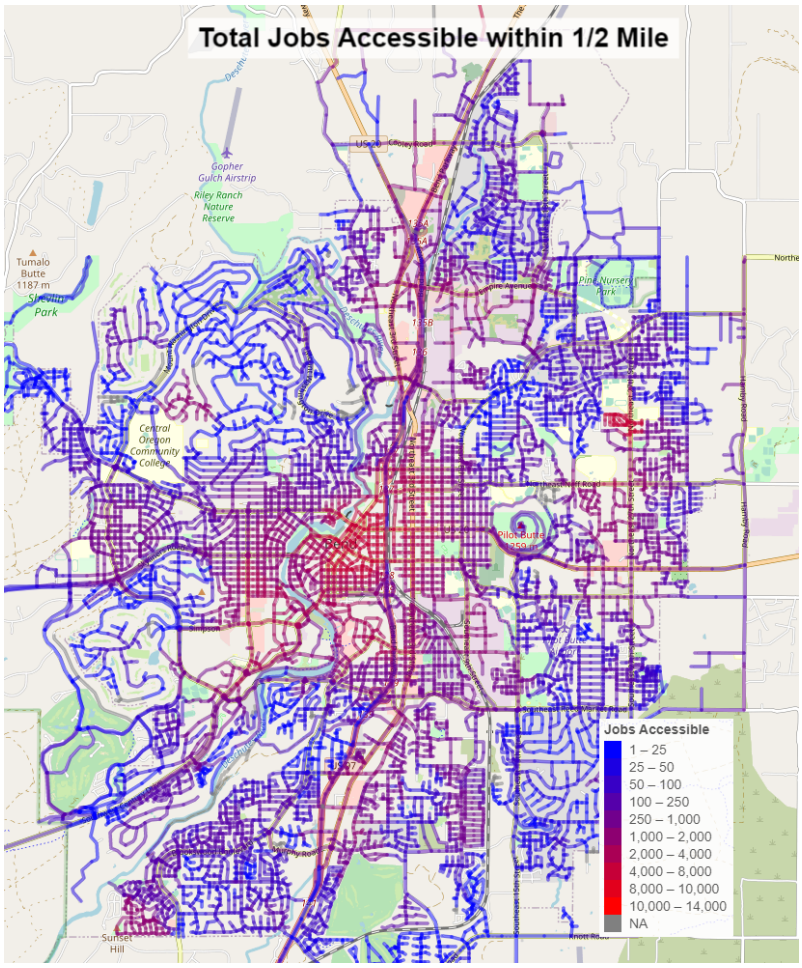

Figure 8.18: Jobs accessible within a ½ mile bicycle ride

Another unique feature in the bicycle data fusion models include counts of users of a bicycle specific smart phone app called Strava which allows people who ride bikes to download the app and use the GPS functionality of their phones to record their trip. Data for this study is available for each year in which there are counts. The data for 2018 are shown in Figure 8.19 below. These data are likely just a subset of total bicycle users and some research has shown that they do not reflect the general bicycle rider population. These data may be thought of as probe data similar to data from vendors such as INRIX that many DOTs use to monitor traffic speeds. Based on a review of concentrations of Strava user trips are in the study area, there appears to be high level bicyclists that are less sensitive to streets with higher speed limits with lots of vehicles. This observation is based on the relatively high number of Strava users on minor and major arterials. Many of the local streets have low to zero counts of Strava rider counts, and a lot of the activity is concentrated in the western portion of the study region, perhaps due to that part of the regions access to mountain biking trails west of the urban area.

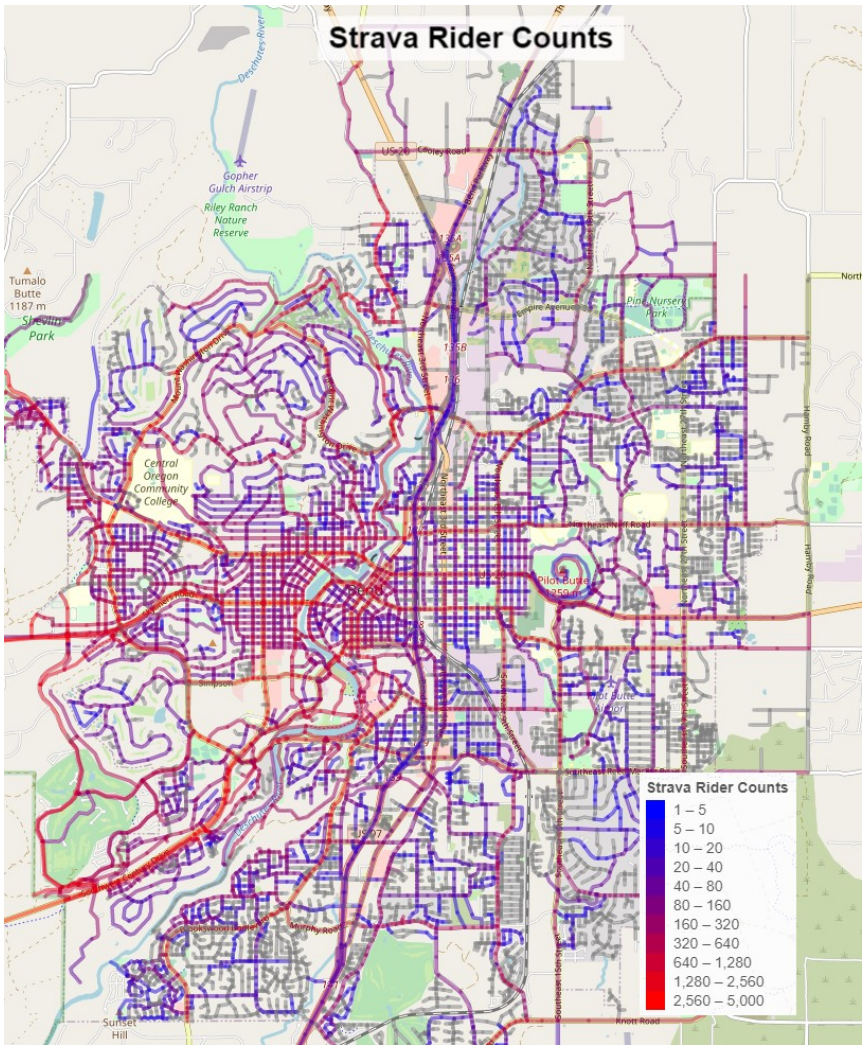

Figure 8.19: Strava rider counts 2018

A bicycle specific measure of centrality has also been created for the bicycle data fusion model using the approach developed by Broach et al. (2009) that accounts for out of direction travel required to remain on streets with characteristics that make the average bicycle rider feel safe. A summary of the bicycle centrality measure is shown in Figure 8.20 below. This measure of centrality emphasizes trips on bike lanes and off-street paths as opposed to just the shortest distance. Vehicle volume is also taken into account which is why the bicycle centrality measure does not have Highway 97 as an important link like it is in the vehicle model data.

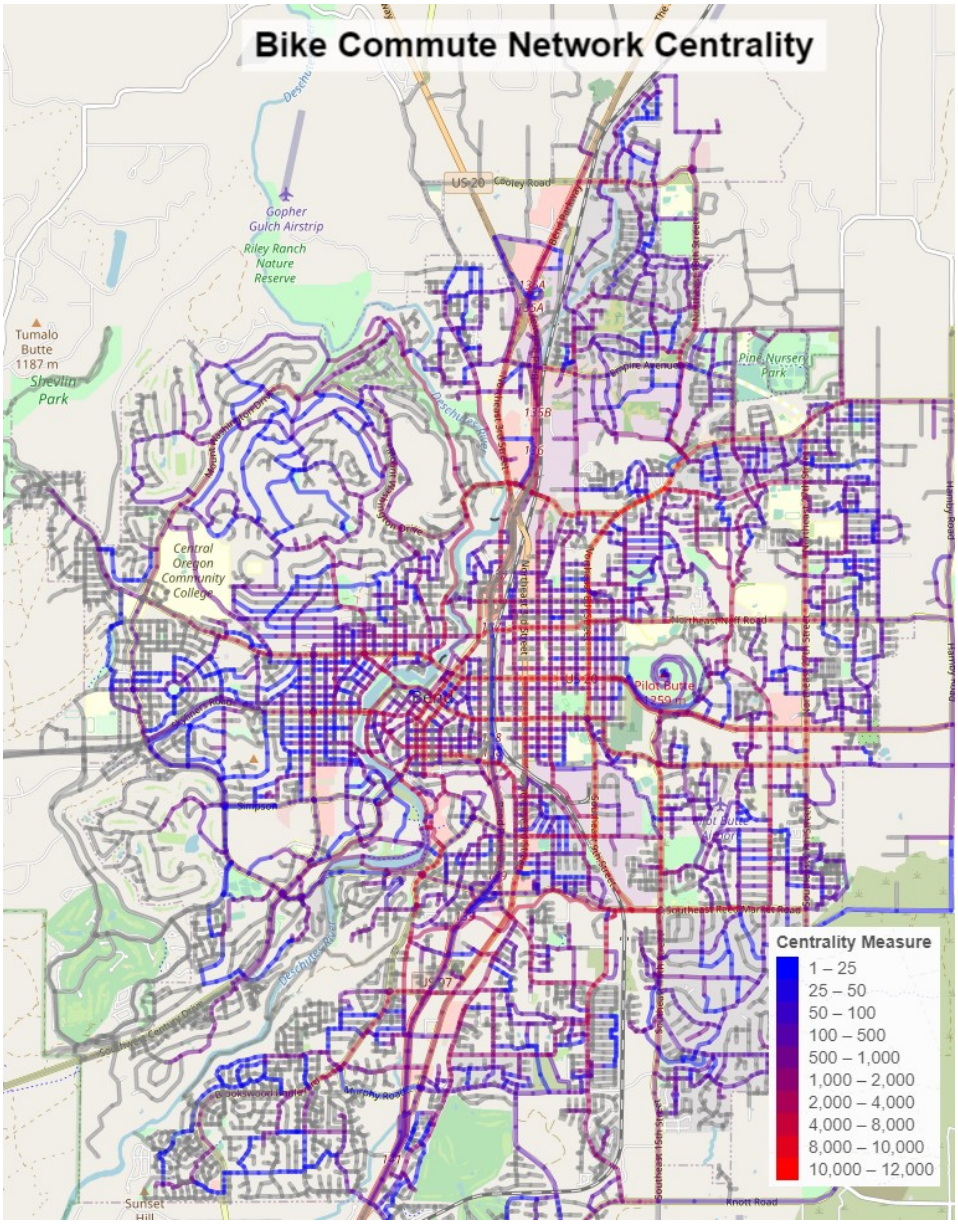

Figure 8.20: Bicycle specific network centrality

A number of features used in the bicycle data fusion model are described above. Below in Figure 8.16 the overall data fusion conceptual model is presented to summarize the network features used to train the bicycle data fusion model. As noted above, this conceptual representation does not show all features used, for instance the access to jobs feature shown in the figure below actually has over 600 different versions when all worker industry, demographic, and access threshold combinations are computed.

8.6 BICYCLE TRAFFIC DATA FUSION MODEL RESULTS

The results of the bicycle data fusion models will be presented in four sections below. The first section will describe and summarize the machine learning based data fusion models including

the features used and the cross-validation results. The second section will describe and summarize the regression based data fusion models including the final model covariates and results of the cross validation results. For the machine learning and regression approaches root mean squared error (RMSE), absolute percent error, and r-squared values are used to measure model performance. The third section will then compare total bike miles traveled estimates when applying select models to the entire network. In addition the third section will discuss an approach to handling upwardly biased estimates of bicycle traffic on low density local streets. The fourth section summarizes these results and offers a discussion about the two methods.

8.6.1 Machine Learning Based Bicycle Traffic Data Fusion Model Cross-Validation Results

This section summarizes the cross validation procedures applied in the bicycle machine learning model development element of this research as well as describes the features used in each of the machine learning algorithms. Similar to the vehicle model training, cross-validation was done through both an internal and external cross validation. The results presented below are based on two machine learning algorithms including extreme gradient boosting (XgBoost) and random forest. Two sets of cross validation are performed, one that is characterized as internal that uses random partitions in a 10-fold cross validation and is done as a part of the model training process within the caret package. The second cross validation process, characterized as external, is performed on a select set of model specifications with high accuracy from the first validation and uses a stratified partition to do another 10-fold cross-validation. The internal cross validation uses 10 folds and was performed twice. The internal cross validation executes rather quickly for each specification taking about 10 minutes to run using seven cores running in parallel on a four core system with eight total processors each with 3.4 Ghz processor speed. Multiple model specifications are tested in the internal validation step using two type of algorithms (XgBoost and Random Forest) with a set of selected model specification being put forward to the external cross validation process.

Many different kinds of training features were tested but selected scenarios are described in Table 8.12 below. The primary difference in the feature scenarios is that the *All + Strava* specification includes Strava data rider counts. Models were tested separately to determine how the use of Strava impacts the model performance. Otherwise, both models use a number of network features described in more detail in the data description section above.

Table 8.12: Bicycle Model Feature Specification

Feature Specification	Description
All	Uses all network features including multiple measure of centrality, accessibility, and network characteristics
All + Strava	Uses all the network features described in "All" specification plus the Strava rider counts

Diagnostic information includes RMSE and r-squared values while the number of features used in the model is also presented. The internal validation results are a product of the initial model training using the caret package in R and uses a random partitioning process, using 10 folds and

performed two times. The results from the internal cross validation tests show that the XgBoost algorithm and random forest algorithm are similar in performance with a minimum r-squared value of 28% for XgBoost versus a 27% in the random forest. The maximum r-squared value for XgBoost is 32% while the maximum for random forest was 47 percent. The number of features used in the XgBoost is generally fewer than the random forest with at most 277 features while the random forest used nearly double with as many as 511 features being used.

Table 8.13: Internal Cross Validation Results for Vehicle Model

Algorithm Specification	RMSE	R-squared	Algorithm	Feature Count	Year
All + Strava	33.1	47%	Random Forest	514	2017+2018
All + Strava	26.8	33%	Random Forest	514	2018+2019
All	32.3	39%	Random Forest	511	2017+2018
All	24.6	27%	Random Forest	511	2018+2019
All + Strava	34.7	29%	XgBoost	226	2017+2018
All + Strava	25.7	32%	XgBoost	274	2018+2019
All	34.6	32%	XgBoost	238	2017+2018
All	25.8	28%	XgBoost	277	2018+2019

One way to diagnose how the machine learning algorithms are using the input features is to use a measure of variable importance. In Table 8.5 the number of features that were ultimately found to be useful in predicting bicycle AADT were summarized for each specification and algorithm. Of all of the features used in each algorithm, the top 20 most important are displayed in Figure 8.6. This chart summarizes the relative number of times a feature is used in the splitting of trees. The top panel shows both model specifications All and All + Strava) for the XgBoost algorithm.

The XgBoost model using the Strava feature (*All + Strava*) used the Total Strava commute riders. Access to total jobs and jobs within specific job categories (accommodation and food services, professional, scientific, and technical, as well as manufacturing) male workers all at various thresholds with common thresholds being at half-mile (2640), one mile (5280) and one and a half mile (7920). Bike centrality was also in the top 20 most important features in the XgBoost model for both model specifications.

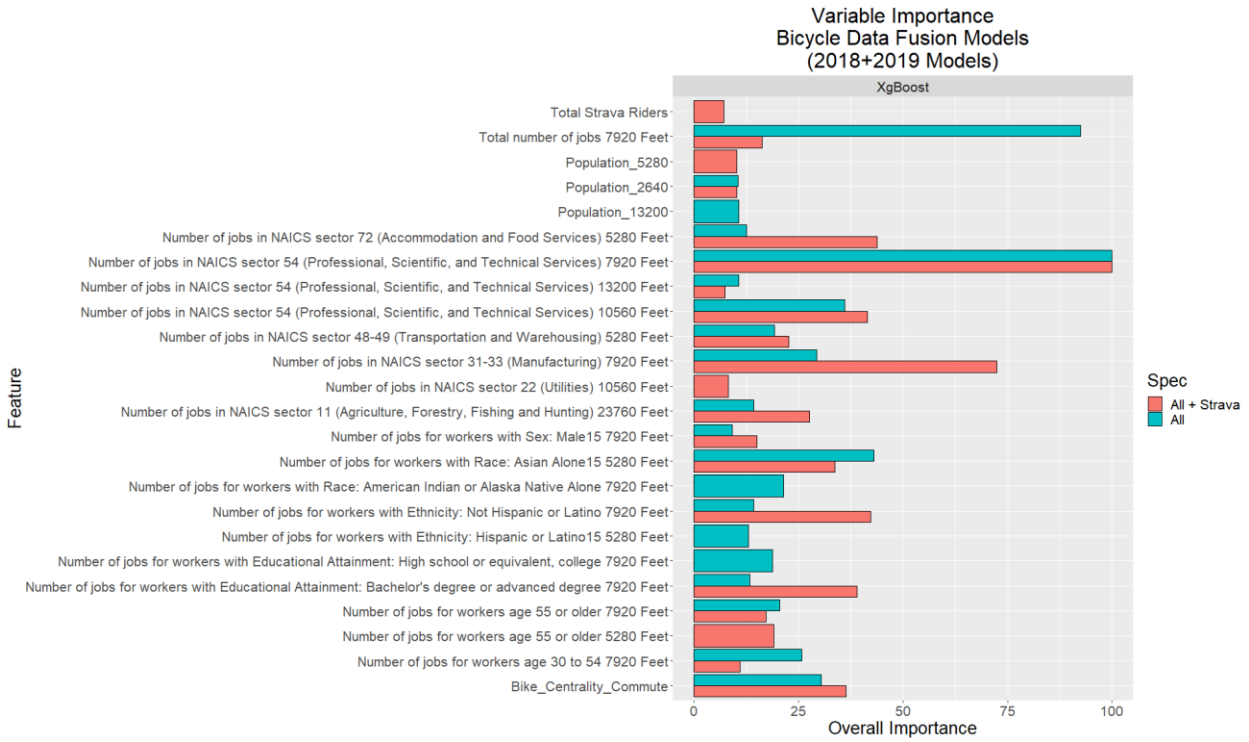

Figure 8.21: Variable importance for select bicycle data fusion models

External validation tests are performed using both a 10-fold and a leave-one-out (LOO) process. The purpose of the external validation tests are twofold with the first motivation looking to understand in more detail the prediction error by volume bin and functional classification which is not possible to extract from the internal cross validation results. The second motivation is to try and determine how much the model results might be biased by spatial autocorrelation making earlier test results somewhat biased because sites used in training may be near tests where the model is applied. To control for this, the LOO cross validation only uses sites in the training that are at least 1,000 feet from the test site.

Results from the external 10-fold cross validation analysis are presented below in Figure 8.22 and shows the median absolute percent error by volume bin for the two model specifications (*All and All + Strava)* and both algorithm types. These results demonstrate that XgBoost model works better than the random forest with in both specifications with 84% 87% error for the *All + Strava and All* models respectively and 115% and 113% using the random forest algorithm. The addition of Strava data to the training features seems to make modest improvement in the median APE for all models and in all volume bins. The best model is the XgBoost using the *All + Strava* specification with 84% error. In this model the error varies depending on volume bin with the lowest volume bin exhibiting the highest error of 240% for the XgBoost and the lowest error in the 11-20 bin with 34% error.

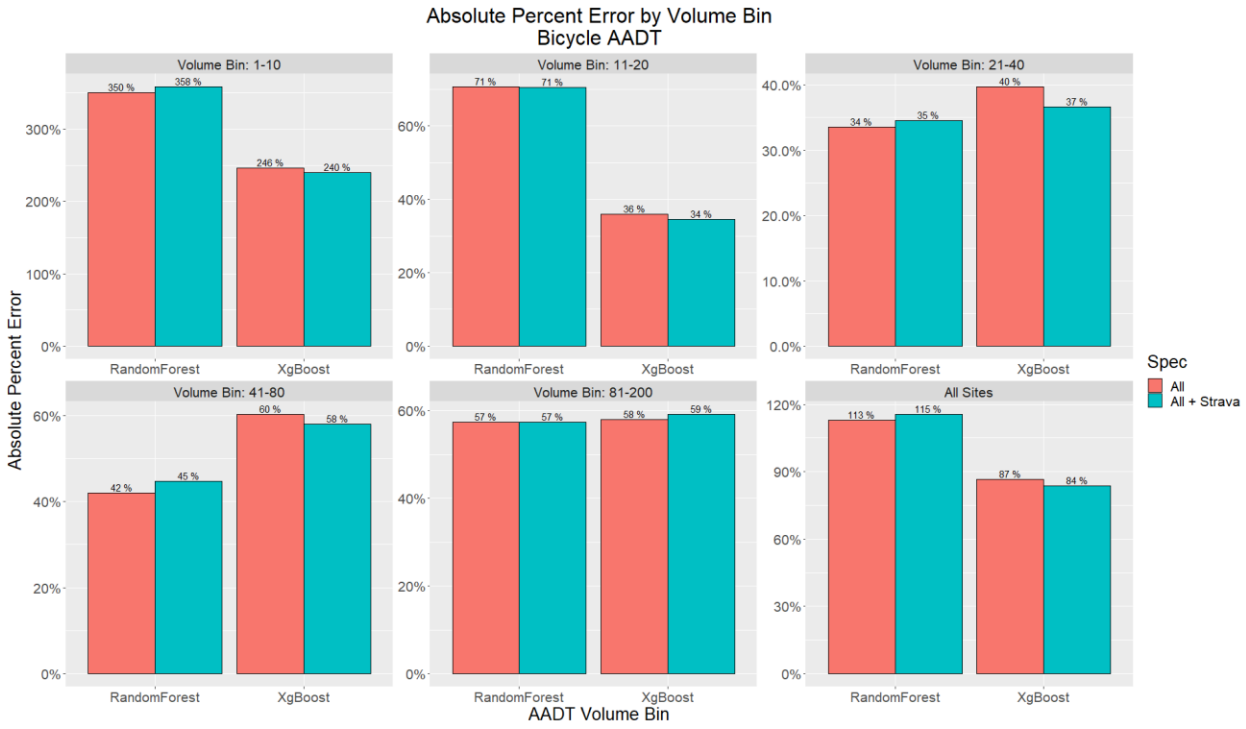

Figure 8.22: External 10-fold cross validation for bicycle models

Because the XgBoost algorithm worked best based on the internal validation and the 10-fold external validation the LOO cross validation process only tested this approach. The *All + Strava* specification was selected for the LOO process due to its better performance the earlier validation tests. Table 8.14 summarizes the results of the LOO cross validation. These validation tests ensure that sites near a validation site are not included in the estimation by only using sites outside a 1,000 buffer. Performing tests this way helps to reduce bias in the cross validation results with mean error of the LOO external validation rising to 80% from 19% mean APE in the 10-fold process summarized above. Error was lowest in the 21-40 volume bin with just 34% and highest in the lowest volume bin with 228% mean APE.

Table 8.14: External Leave-One-Out Cross Validation Results for Vehicle Model

Algorithm Type	Volume Bin	Absolute Percent Error		Number of Sites
		Mean	Median	
XgbBoost	1-10	228%	169%	19
XgbBoost	11-20	40%	25%	23
XgbBoost	21-40	34%	28%	32
XgbBoost	41-80	56%	59%	13
XgbBoost	81-200	61%	68%	5
XgbBoost	All Sites	80%	44%	92

8.6.2 Statistical Bicycle Traffic Data Fusion Model Cross-Validation Results

This section will describe the development of statistical models to estimate bicycle AADT including an exploration of the individual effects of the covariates used in the final model. Since the number of available covariates for estimating a statistical model for bicycle traffic are numerous it was necessary to use a testing procedure to determine the variables with the best model prediction accuracy. This process uses 10-fold cross-validation to test the prediction accuracy of thousands of possible model specifications. Identical to the process used in the vehicle model development above, a large number of specifications are tried though in the bicycle model the total was much greater and included 497,664 possible specifications based on a grid of all possible combinations of select variables including population access, total employment access, retail, health, and warehouse workers, intersection density, auto centrality, shortest path centrality a two measures of the Strava data including the total rider counts and the proportion of the Strava rider counts that were tagged as commute. All the accessibility measures use shortest network distance thresholds of either one-quarter mile, half-mile, or one and a half miles. All models are estimated using a negative binomial regression specification due the counts data featuring over dispersion where the dependent variable (bicycle AADT) variance is greater than the mean of the counts which is generally the case for traffic counts data.

A custom process was developed in R where for the 2018/2019 counts period data is partitioned into 10 folds using a stratified random sample ensuring functional classification and bike facility designations are equally distributed among the folds. A negative binomial regression model is estimated on each of the k-1 groups (training data) and then estimated on the k-9 (test data) and then compared to the observed data. To do this for all 497,664 models the total runtime is 6.9 hours even using parallel processing. For each selection of variables three performance metrics are computed include RMSE, mean absolute percent error (MAPE) and adjusted r-squared. Based on these metrics models top performing models are selected for further examination. For the bicycle models the final estimated parameters are presented in **Error! Reference source not found.** for three select models using these model performance measures. Model results below present the estimated coefficient and the associated standard error and p-value for selected models with the highest r-squared, the lowest RMSE, and lowest MAPE for 2018+2019 data.

These results show that many of the covariates are correlated with an increase in bicycle traffic including the presence of off-street path, total job and retail job access, bike centrality (commute), and Strava riders and the proportion of Strava riders flagging their trip as commute. Features associated with a decreased traffic volume include population access and access to jobs with less than a high school education access and functional classification. Functional classification was selected in the in the Lowest RMSE and Highest R-squared models and is operationalized as a factor variable with the reference set as off-street path. The coefficients for this variable reveal that compared to off-street path facilities, highways and minor arterials have the biggest effect on reducing bicycle volume followed by local streets and minor arterials. The effect of the local streets is surprising and might be capturing some of the lack of connectivity of the local streets network but that effect would ideally be captured with the centrality measures.

Table 8.15: Regression Results for Bike Model

Coefficient	Std. Error	z value	P-value	Variable	Year	Metric
0.0001594	1.10E-04	1.4463	0.1481	Total number of jobs 7920 Mi.	2018+2019	Highest R-Squared
-0.0017	1.53E-03	-1.1406	0.2540	Number of jobs for workers with Educational Attainment: Less than high school 7920 Mi.		
2.29E-04	5.97E-05	3.8308	0.0001	Bike_Centrality_Commute		
7.13E-04	2.24E-04	3.1795	0.0015	Strava Commute Riders		
-1.41E-04	4.65E-05	-3.0343	0.0024	Bike_Centrality_Rec		
-0.6525	0.2647	-2.4645	0.0137	Local (Reference - Off-street path)		
-0.3264	0.1960	-1.6654	0.0958	Collector		
-0.8497	0.1766	-4.8106	0.0000	Minor arterial		
-0.6418	0.2484	-2.5836	0.0098	Major arterial		
-1.2156	0.3888	-3.1264	0.0018	Highway		
-0.0462	0.1812	-0.2548	0.7989	No Facility (Reference - Bike Lane)		
1.62E-04	1.16E-04	1.4023	0.1608	Total number of jobs 7920 Mi.	2018+2019	Lowest RMSE
6.82E-07	9.01E-06	0.0757	0.9397	Population_2640		
-0.0018	0.0016	-1.1111	0.2665	Number of jobs for workers with Educational Attainment: Less than high school 7920 Mi.		
2.28E-04	6.04E-05	3.7806	0.0002	Bike_Centrality_Commute		
7.12E-04	2.25E-04	3.1585	0.0016	Strava Commute Riders		
-1.41E-04	4.71E-05	-2.9905	0.0028	Bike_Centrality_Rec		
-0.6553	0.2674	-2.4507	0.0143	Local (Reference - Off-street path)		
-0.3258	0.1960	-1.6622	0.0965	Collector		
-0.8511	0.1793	-4.7462	0.0000	Minor arterial		
-0.6408	0.2486	-2.5779	0.0099	Major arterial		
-1.2132	0.3911	-3.1020	0.0019	Highway		
-0.0455	0.1817	-0.2502	0.8024	No Facility (Reference - Bike Lane)		
2.47E-04	1.18E-04	2.0849	0.0371	Total number of jobs 7920 Mi.	2018+2019	Lowest MAPE
9.19E-06	5.95E-06	1.5452	0.1223	Population_5280		
-3.19E-03	1.66E-03	-1.9210	0.0547	Number of jobs for workers with Educational Attainment: Less than high school 7920 Mi.		
6.33E-04	4.21E-04	1.5045	0.1325	Number of jobs in NAICS sector 44-45 (Retail Trade) 2640 Mi.		
2.58E-04	6.01E-05	4.3007	0.0000	Bike_Centrality_Commute		
7.13E-04	2.24E-04	3.1846	0.0014	Strava Commute Riders		
-1.91E-04	4.70E-05	-4.0603	0.0000	Bike_Centrality_Rec		
-0.0476	0.1578	-0.3016	0.7629	No Facility (Reference - Bike Lane)		
0.6914	0.1623	4.2610	0.0000	Off-street path		

The bicycle facility variable is also operationalized as a factor variable with the bike lane set as the reference and is used in all models. In the models with functional classification the off-street paths are included and so do not show up in the coefficient table whereas in the Lowest MAPE model functional classification is not used so the result for off-street path is shown. In either specification the lack of bike facility is correlated with lower bicycle volumes.

Not all variables are significant within the 0.05 level of significance but about three-quarters of the variables in each of the statistical model scenarios are significant at the 0.10 level. Table 8.16 below summarizes the three select models error measures.

Table 8.16: Summary Information for Bicycle Regression Model

Specification	Performance Metric	MAPE	RMSE	Adjusted R-Squared
C000_7920 + CD01_7920 + Bike_Centrality_Commute + Commute_Counts + Bike_Centrality_Rec + Fc_Desc + Bike_Facility	Highest R-Squared	93.7%	21.42	0.498
C000_7920 + Population_2640 + CD01_7920 + Bike_Centrality_Commute + Commute_Counts + Bike_Centrality_Rec + Fc_Desc + Bike_Facility	Lowest RMSE	94.7%	21.39	0.496
C000_7920 + Population_5280 + CD01_7920 + CNS07_2640 + Bike_Centrality_Commute + Commute_Counts + Bike_Centrality_Rec + Bike_Facility	Lowest MAPE	86.2%	24.4	0.348

The 10-fold holdout analysis results are further summarized by volume detailing the median APE for each of the models. The model with the lowest median APE for all sites is the same model with the lowest mean APE, as would be expected, and is better by about 4 percent overall median APE. The Lowest MAPE model has lower error in all the volume bins except for the 1-10 and 81- 200 volume bins.

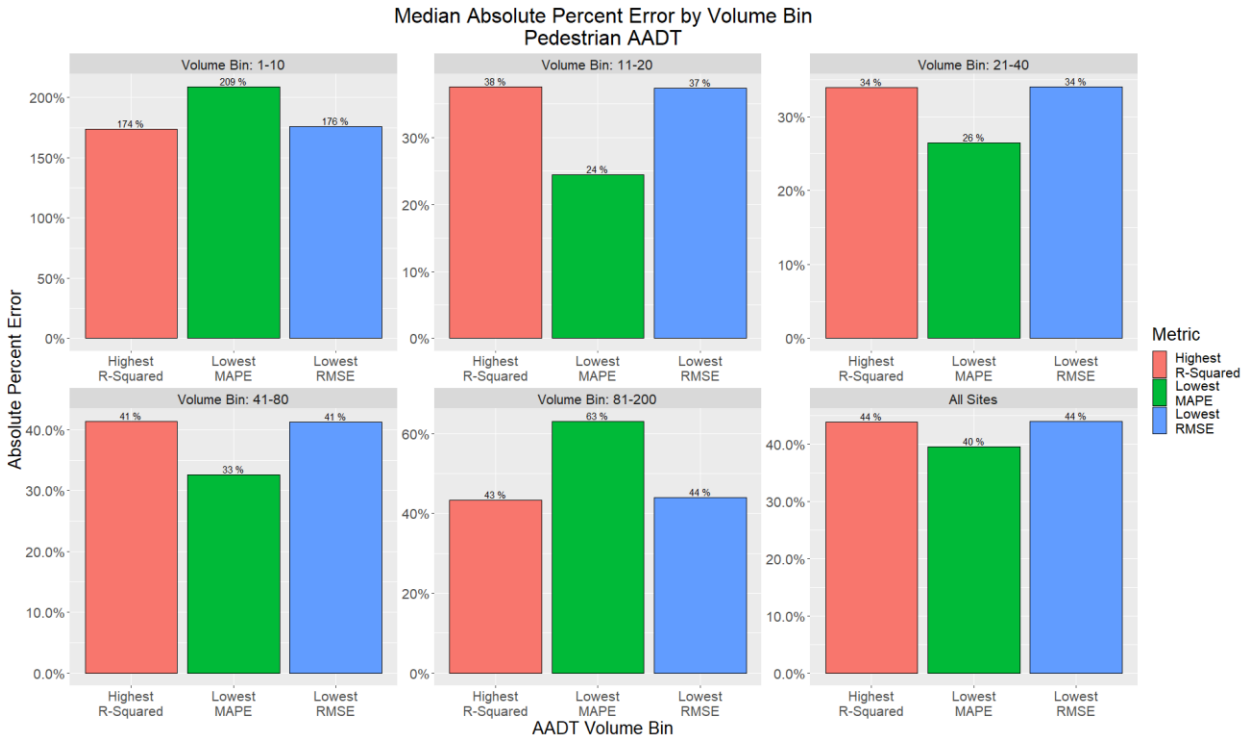

Figure 8.23: Top bicycle regression model median absolute percent error by volume bin

8.6.3 Select Bicycle Data Fusion Model Application

A primary objective of this research is to develop an estimation framework to apply network wide that will provide information about nonmotorized travel activity for the entire study area. This section will summarize the application results of select bicycle data fusion models by applying the models to the entire network in order to generate system wide bicycle activity estimates. Additionally, an approach is suggested to handle over inflated counts on low volume, low density residential streets that make up significant lane miles of most urban networks. The issues and a proposed solution will be discussed below.

As summarized in Table 8.11 above, there are over 650 miles of network in the study region transportation system, including nearly 129 miles of bicycle lanes and over 50 miles of off-street paths. A prime objective of this research is deploying the models estimated and validated above on the entire system in order to estimate a system wide measure of bicycle activity. The results below in Table 8.17 show the total annual bicycle miles estimated using the XgBoost algorithm and the selected regression models. These results show that in the first estimate period using counts from 2017 and 2018 (2017+2018), the estimated total bicycle miles traveled in the study region was 5.22 and 5.54 million miles for the *All + Strava* and *All* machine learning models respectively. The regression model estimates are 5.20, 5.14, and 5.18 million miles for Highest R-Squared, *Lowest MAPE* and *Lowest* RMSE models respectively in the 2017+2018 estimation period. For the second estimate period, from 2018 and 2019 (2018+2019) the total BMT estimate is 4.44 and 4.84 million miles for the *All + Strava* and *All* machine learning models respectively.

Table 8.17: Total Bicycle Miles Traveled for Select Models

Model Specification	Algorithm Type	Total Annual Bicycle Miles Traveled	Bend Population	Per Capita BMT	Year
All + Strava	XgbBoost	5,225,730	96,058	0.15	2017+2018
		4,444,592	99,171	0.12	2018+2019
All		5,544,053	96,058	0.16	2017+2018
		4,847,462	99,171	0.13	2018+2019
Highest R-Squared	Negative Binomial	5,203,217	96,058	0.15	2017+2018
		3,828,137	99,171	0.11	2018+2019
Lowest MAPE		5,141,664	96,058	0.15	2017+2018
		3,861,726	99,171	0.11	2018+2019
Lowest RMSE		5,187,065	96,058	0.15	2017+2018
		3,825,118	99,171	0.11	2018+2019

The estimates from the regression models are within 1% of one another in both estimation periods. The machine learning model that uses Strava as a training feature appears to moderate the total estimate for the 2017/2018 period by about 6% and 2018/2019 period by 9 % compared to the *All* model that does not use this training feature. This might be expected considering the Strava feature is not present on most local roads and so moderates estimated volume on those facilities. Considering local roads make up over 60% of the network this moderation can have a significant impact on total BMT.

Figure 8.24 below displays the total annual BMT estimates by selected model scenario and shows that the BMT summary aggregated by functional classification and bicycle facility for a *Strava + All* machine learning model and the *Lowest MAPE* regression model. *Lowest MAPE* is selected because MAPE was the performance measure used to select which of the machine learning model specifications to focus on. The figure below shows that many BMT estimates are similar though some significant differences exist including the local streets where no bike facility exists. The *All + Strava* machine learning model estimates 3.29 and 2.72 million BMT for the two estimation periods while the regression model only estimates 3.03 and 2 million BMT in each estimation period. The 2018+2019 estimation period is different by just over one million BMT which seems significant.

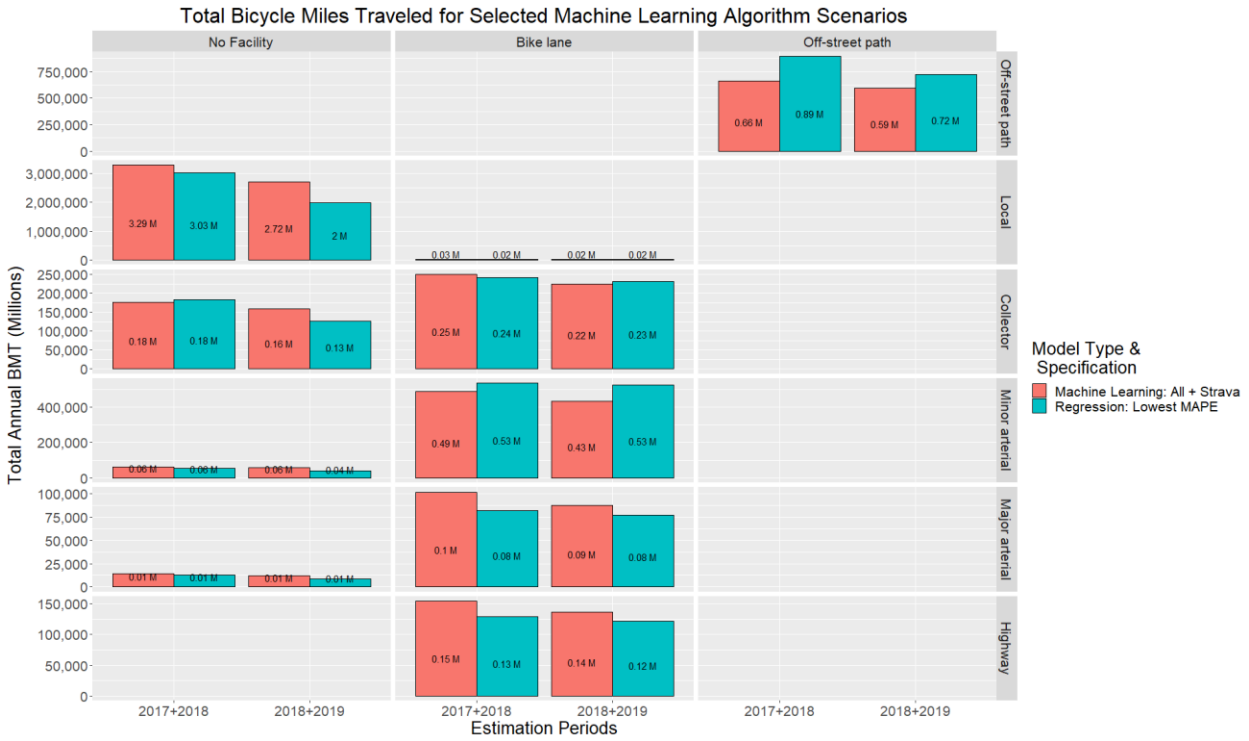

Figure 8.24: Bicycle miles traveled estimates for selecteds by bicycle facility type and functional classification

Of note is the significant number of BMT that are being estimated on the local road system. The local road system, even without a bicycle facility may be an attractive facility for people to bicycle due to its low vehicle and speed and volume and relative proximity to residential areas (population access) and parks. However, many of these streets are likely to have zero counts given their low accessibility to key destinations and because of the nature of the traffic count programs where streets with likely bicycle users were counted, the available counts are likely biased upwards and using them in a network wide application is likely biasing the total BMT results upward. In order to handle this issue, a proposed solution is offered where zero counts locations are introduced into the counts data at locations where zero bicycle traffic is likely. The criteria for the random selection of these zero count locations are described below:

- Local street functional classification with no bicycle lane
- Population access within 0.5 miles must be 400 people or less
- Bicycle centrality must be zero
- No Strava rider counts

Using this criteria about 41 miles or 10% of the local street network, become eligible for having a zero count assigned to it. Of these local streets, 30 links are randomly selected and those 30 locations are added to the counts data and the machine learning algorithms are retrained with the

inclusion of the simulated zero counts data. The remainder of this section will detail the BMT results of the modeling with the inclusion of these randomly selected zero count locations.

With the introduction of the zero counts the distribution of the data is altered and the negative binomial model is no longer appropriate and instead a Poisson model is used to estimate the model using the simulated zero counts. Future research should explore the use of zero inflated hurdle models to see if that specification changes the final BMT results. With about 25% of the counts now being zeros it's likely this would be a more proper specification. Table 8.18 below details the results for the new BMT estimate scenario where 30 zero count locations were inserted into the model training data. On an aggregate basis, the total BMT decreases to 65% of initial estimate for the 2017/2018 estimation period, and 67% for the 2018/2019 estimation period when estimated using the XgBoost machine learning algorithm with the *All + Strava* specification. Using the Poisson regression approach but including the simulated zeros the estimated BMT drops 55% of initial estimate for the 2017+2018 estimation period and 71% for the 2018+2019 estimation period.

Table 8.18: Total Bicycle Miles Traveled Comparison with Simulated Zero Counts Scenario

Model Type and Specification	Estimation Periods	Total Annual Bicycle Miles Traveled		Percent Difference
		No Zero Counts	Simulated Zero Counts	
Machine Learning: All + Strava	2017+2018	5,225,730	3,385,390	65%
	2018+2019	4,444,592	2,985,239	67%
Regression: Lowest MAPE	2017+2018	5,141,664	2,803,758	55%
	2018+2019	3,861,726	2,727,744	71%

Figure 8.25 below details the aggregate BMT by functional classification and bicycle facility for both modeling approaches (machine learning vs. regression) and without simulated zero counts and with those simulated zero counts. The insertion of zero counts into the machine learning training data depress the estimated BMT for the local streets with no bike facility, as designed, reducing the estimated BMT on those facilities from 2.72 million BMT to 1.21 million BMT for the 2018/2019 estimation period, a reduction of roughly 55 percent. When the zero counts are included in the regression model approach the BMT on local streets with no bike facility goes from 2 million BMT to 1.21 million for the 2018/2019, a change of about change is about 40% percent. Most facility types have a diminished BMT estimate in both periods but highway facilities with bike lanes see a marginal increase in the machine learning model.

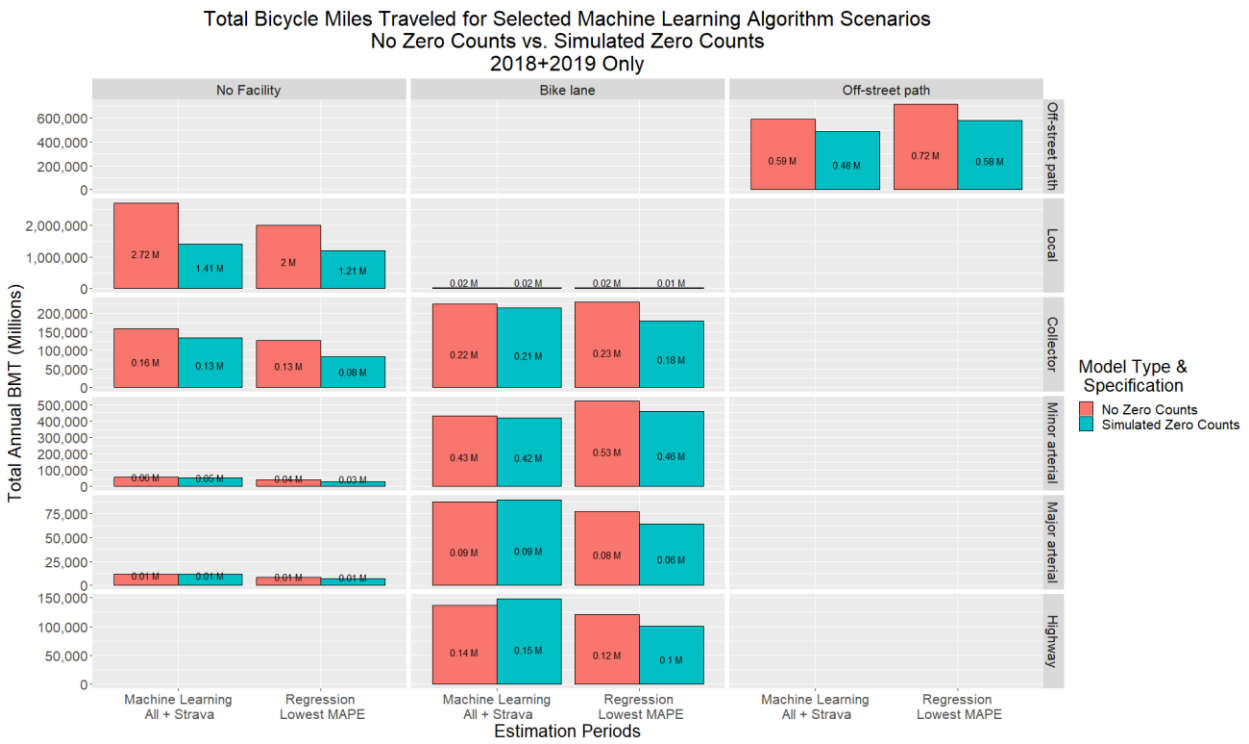

Figure 8.25: Bicycle miles traveled estimates comparison of zero counts scenario by bicycle facility type and functional classification

The insertion of zero counts at locations with low density and low network connectivity appear to have the desired effect of moderating the overall BMT estimates. Figure 8.26 and Figure 4.25 below shows the results of the network wide application of both model approaches and the scenarios using counts data and counts data with simulated zeros. The left panel shows the results of the model applied to the network with all observed data while the right panel shows the model with simulated zero counts at low density locations. Where as in the left panel there are no locations where zero counts are estimated (denoted by grey) while the right hand panel shows a small number of links in far flung parts of the network with no estimated bicycle activity. Additionally, the simulated zero counts scenario moderates bicycle volumes throughout the low density areas surrounding the core of the study region, with many more links in the 1-5 AADT volume bin. In fact there only 11 links in the No Zero Counts scenario with 1-5 bicycle AADT while in the Simulated Zero Counts scenario there are 4,074 links with volume in this range for the XgBoost based model.

Even though the aggregate measure of BMT similar, different by only 15% between the two model approaches in the 2018+2019 period, the network level estimates reveal a number of differences. The XgBoost model appears to spread the activity out in the downtown area while the regression model targets the activity to a discrete corridors. Those corridors are more pronounced in the scenarios where the zero counts were injected into the training data. The XgBoost results do create about 30 links where the estimate is a negative value which are then converted to a zero for the purposes of aggregation and network visualization.

The table below presents some summary statistics of the estimated bicycle volumes on the 14,000 links that make up the study region network and which were presented below in Figure 8.26 and Figure 8.27 via map. As expected the mean estimated summary statistics all decrease with the injection of simulated zero counts with the XgBoost model estimating a negative values on about 30 links, which are converted to zero.

Table 8.19: Summary Statistics of Estimated Counts for Total Network Application of Bicycle Fusion Models

Model Specification	Scenario	Estimated AADT Summary Statistics				
		Minimum	**Maximum**	**Mean**	**Median**	**Std. Dev.**
All + Strava	No Zero Counts	2.2	142	19	16	10
Lowest MAPE		6.3	165	15	13	9
All + Strava	Simulated Zero Counts	0*	141	12	9	13
Lowest MAPE		4.4	150	11	7	10

*30 links had an estimated AADT of between -0.6 & 0.0005 and were assigned a zero value

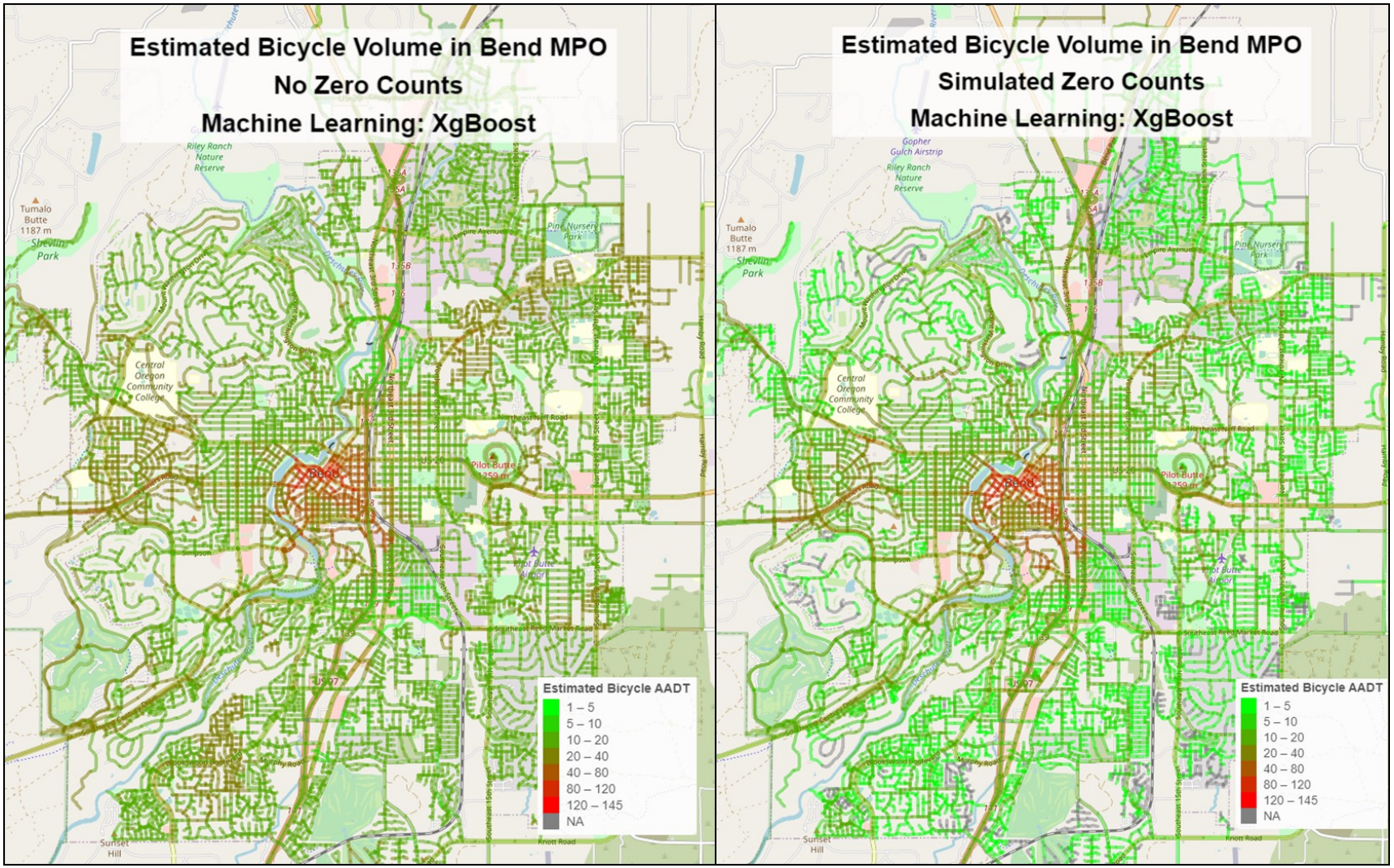

Figure 8.26: XgBoost - comparison of bicycle miles traveled scenarios – network level estimates

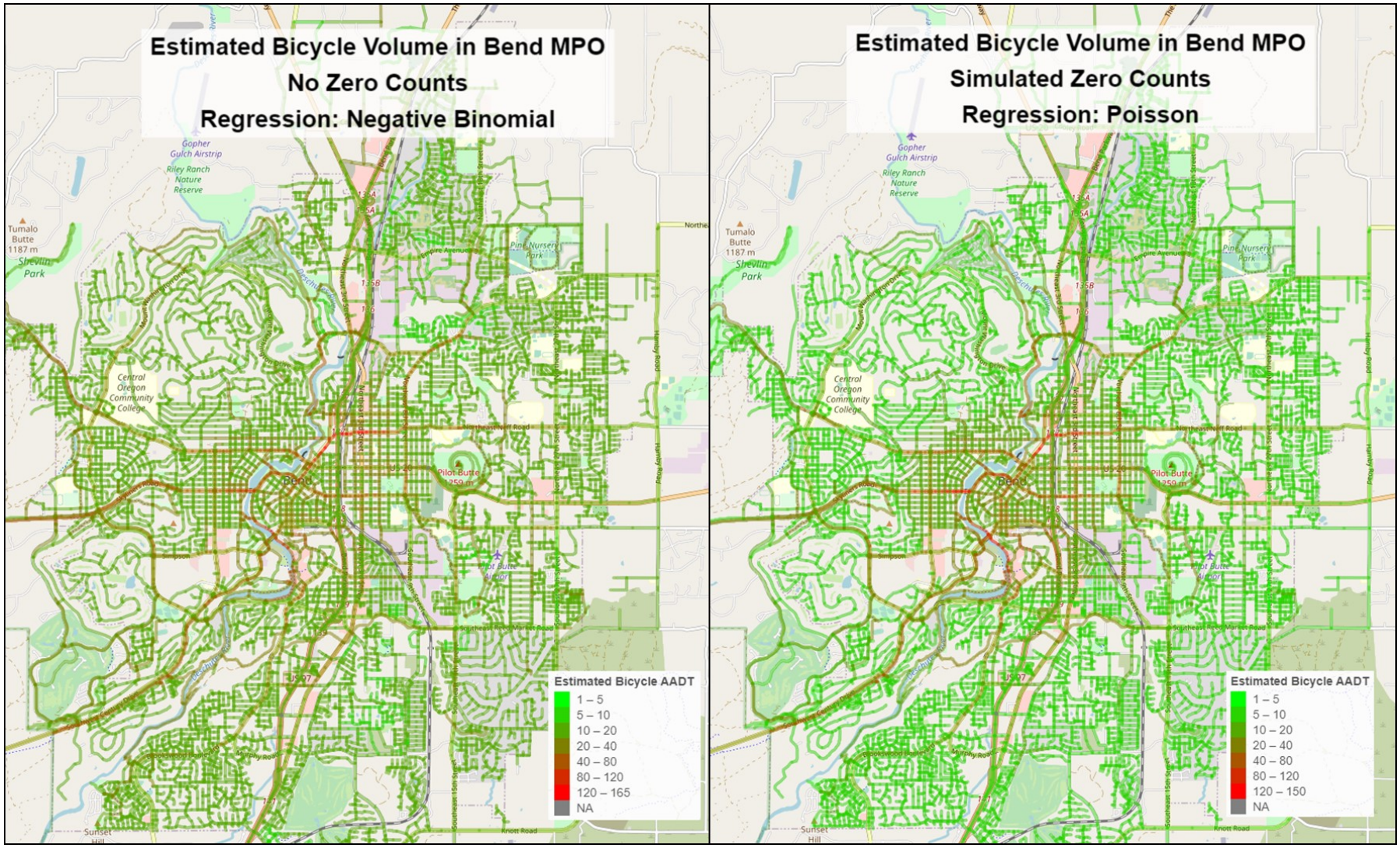

Figure 8.27: Regression - comparison of bicycle miles traveled scenarios – network level estimates

8.6.4 Bicycle Data Fusion Discussion

The above section detailed the data, estimation procedures, validation, and results of data fusion models for bicycle traffic volumes in the study region. The validation results showed that the XgBoost machine learning algorithm worked better than random forest across three separate cross-validation procedures that tested the different machine learning algorithms. In order to specify a regression model, nearly 500,000 models are estimated and tested using 10-fold cross validation. Of these models top performing models based on MAPE, r-squared, and RMSE are selected for further examination. These validation tests also showed that including the Strava data helped to improve model accuracy, albeit only marginally in the machine learning tests though the Strava variable was found to be important in all of the regression models.

In the case of the machine learning model application on the entire network in order to produce a BMT estimate, the specification with Strava data feature appeared to be useful, helping to moderate overall activity estimates. However, using just the observed data in the data fusion models is likely biasing the BMT estimate upward, due to the selection of count locations where bicyclists are expected. To handle this bias, an approach is suggested whereby zero counts are injected into the training data at locations where zero bicycle riders would be expected. The results of this approach present the expected outcomes, further moderating estimated bicycle activity across the network, especially at locations a high likelihood of low bicycle ridership. Continued discussions are necessary with potential model users about an application ready bicycle data fusion model so model users completely understand the advantages and limitations of using either of the models examined in this research as tradeoffs exists.

The use of machine learning in estimating network wide bicycle activity is novel, based on the current status of the literature. Machine learning offers significant advantages for predicting important quantities such as bicycle volumes where inferential data is less important for model users. Additionally, the selected machine learning algorithms offer powerful mechanisms for accounting for the interaction of many complicated relationships between network variables and are likely important tools for monitoring the system and understanding network wide activity.

However, the results from the application of the machine learning model seem less reasonable that the regression model, spreading demand across the downtown area instead of focusing the activity to certain corridors. It's likely that with many fewer features in the training data for the regression models, the centrality and strava features have more impact than the machine learning approach where the effect might be getting washed out some by the large number of employment features.

Either of these model approaches will only improve as more data is collected and the data collected and fed into the model estimation process. Additionally, model results would be improved with updated data for certain data elements. For instance, the decrease in bicycle miles traveled from the first estimation period to the second could be because the employment data used in training and application was a single year, representing 2017 since 2018 data has yet to be released by Census Bureau. Other data from LEHD could be harnessed, including origin-destination information that connects worker residential locations and their place of work. A major issue in the training feature data is the use of population data from 2011. These data were used because of their ease of availability but more updated data from American Community

Survey could be used to better reflect the conditions when traffic counts were collected. Other model estimation and application improvements could be to evaluate the Strava data in more detail and correct places where potential issues are present. Strava also offers an origin-destination product that could be useful however in order to completely take advantage of these data a larger travel shed would likely be needed that expands beyond the boundaries of the urban area. It's generally accepted that a lot of the bicycle activity in the Bend study area is related to recreational travel and Bend's proximity to path and trail networks outside the urban area. One reason the cross validation results are lower than the vehicle models is because this out of area travel is not accounted for in any of the access measures.

8.7 PEDESTRIAN TRAFFIC DATA FUSION MODEL

The results of the pedestrian data fusion models will be presented in four sections below. The first section will describe and summarize the machine learning based data fusion models including the features used and the cross-validation results. The second section will describe and summarize the regression based data fusion models including the final model covariates and results of the cross-validation results. For the machine learning and regression approaches root mean squared error (RMSE), absolute percent error, and r-squared values are used to measure model performance. The third section will then compare total pedestrian miles traveled estimates when applying select models to the entire network. In addition the third section will discuss an approach to handling upwardly biased estimates of bicycle traffic on low density local streets. The fourth section summarizes these results and offers a discussion about the two methods.

8.7.1 Data Description for Pedestrian Traffic Fusion Models

A number of features used in the pedestrian data fusion model are described in the section below. Figure 8.28 shows the overall data fusion schema and presents key network features used to train the pedestrian data fusion model. As noted above in the vehicle and bicycle model data description, this schema representation does not show all features used, for instance the access to jobs feature shown in the figure below actually has over 600 different versions when all worker industry, demographic, and access threshold combinations are computed.

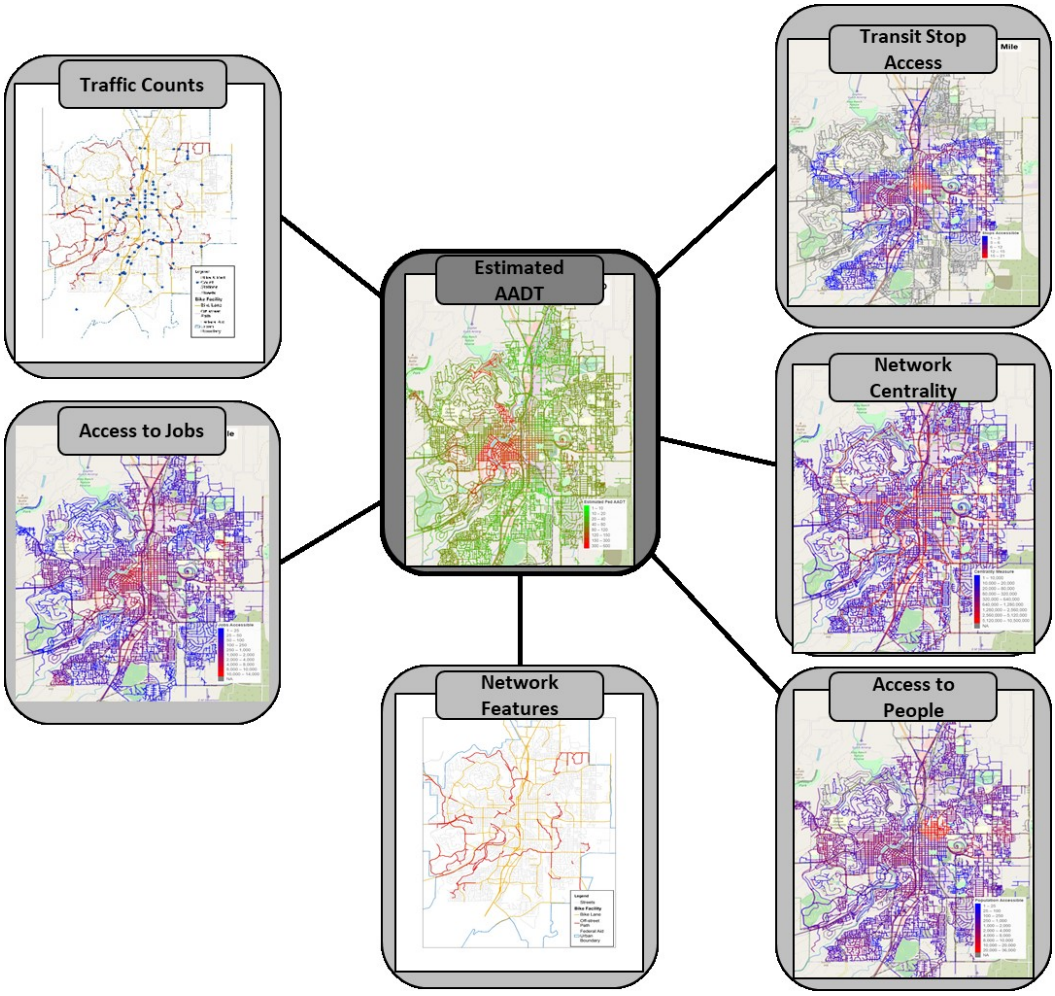

Figure 8.28: Pedestrian data fusion model schema

Table 8.20 below summarizes the AADT estimates for pedestrian traffic by functional classification for two time periods denoted below as *2017+2018* and *2018+2019*. These period represent average volumes for each year denoted in count locations where counts are available for both years. These averages are constructed and presented below to take advantage of as much of the counts data as possible for data fusion modeling and are used in place of applying a growth factor. Observations from the table below show that median pedestrian traffic volumes on off-street paths are higher than other streets followed by local streets, minor arterials, and principal arterials with collectors demonstrating the least pedestrian volume. The counts appear to be trending slightly upward for each facility type across aggregation periods.

Table 8.20: Bicycle Traffic Count Summary

Functional Classification	Year	Pedestrian AADT Summary Data					
		Minimum	Mean	Median	Std. Dev.	Max	Observations
Off-street	2017+2018	15	267	144	286	900	14
	2018+2019	19.5	263	170	256	808	14
Local	2017+2018	62	83	79	23.3	108	3
	2018+2019	9	59.8	62	31.6	103	9
Collector	2017+2018	3	13.7	13.4	10.8	25	4
	2018+2019	3	15.7	15.3	8.82	26.5	5
Minor Arterial	2017+2018	11.5	57.2	40.5	49.1	163	16
	2018+2019	11.5	72.5	47.5	61.3	234	21
Principal Arterial - Other	2017+2018	7	75.5	50	68.2	206	7
	2018+2019	12.8	76.1	58.8	64.4	196	7
All Sites	2017+2018	3	99.2	50	113.1	900	44
	2018+2019	3	97.4	58.8	98.6	808	56

Pedestrian traffic have been collected at nearly 60 locations over the three years where data was actively collected. Those locations are displayed below in Figure 8.29. Many of the locations are on facilities where pedestrian users would be expected to use and thus inserts a certain amount of bias where the model would likely be biased upward, especially at sites with very low or zero pedestrian activity. Later in the report an approach is proposed to handle the issue of having no zero counts in the observed pedestrian traffic counts data. Another feature shown in Figure 8.17 is the bicycle specific network elements including the location of bicycle lanes and off-street paths. These will also be used in the model training process.

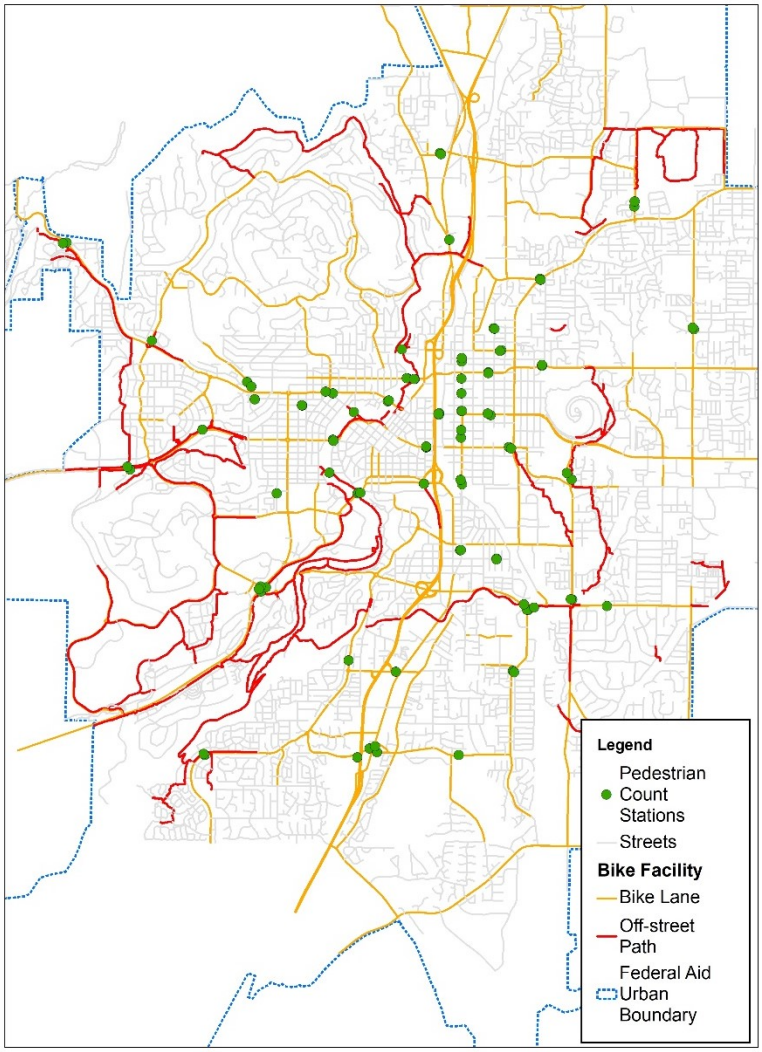

Figure 8.29: Pedestrian count locations

Unlike with the vehicle and bicycle network system data, this research does not have access to high quality network data for the pedestrian system, other than the off-street path network data. The presence of sidewalk information is available in a regional data set but is not used in this research. Pedestrian network data such as sidewalk quality and width as well as crossing treatments like mid-block crossing and cross walk type would be ideal features to use in the pedestrian models.

Most of the training features in the pedestrian model are also used in the other user type models but one training feature novel to the pedestrian models is the access to transit stops. A preferable transit related metric would be to use transit ridership but those data were not available at the time of the publication of this report. However, Figure 8.28 below shows the transit stops accessible within one-half mile walk and represents an access to transit measure. Transit access was developed using multiple thresholds from one-half mile to six miles in half-mile increments. These distance measures are network based using shortest path assumptions, not buffer or Euclidean based. Figure 8.30 shows the areas of the study region where transit access is

available highlighting the density of access in the core of the region and revealing that much of the network has no access to transit within one-half mile walk trip. These transit access measures do not account for frequency of service or other service quality measure.

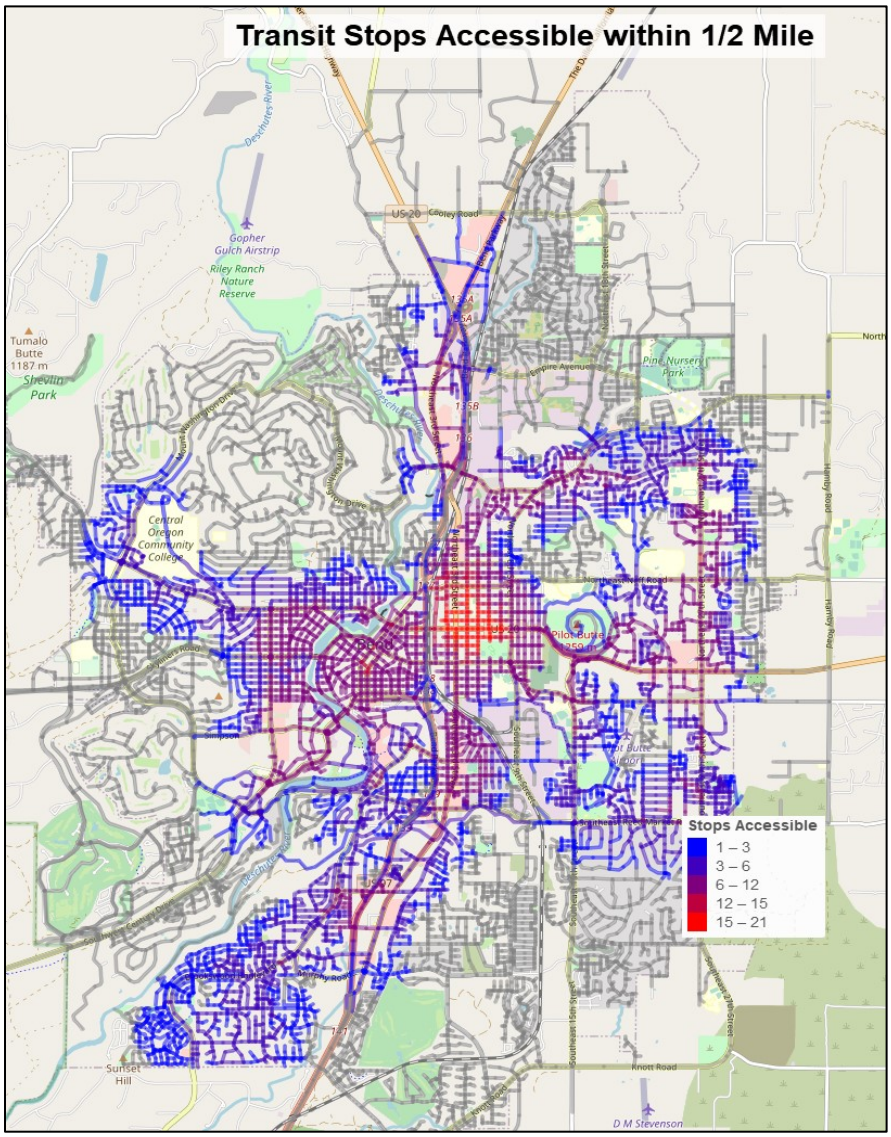

Figure 8.30: Transit stops accessible within ½ mile walk

Other data used in the pedestrian data fusion model mirror those used in the bicycle models and can be reviewed in the section above describing those features.

8.8 PEDESTRIAN TRAFFIC DATA FUSION MODEL RESULTS

The results of the pedestrian data fusion models will be presented in three sections below. The first section will describe cross-validation procedures and the model specifications for each model presented in this section. The second section will summarize the internal and external cross-validation processes which use information including root mean squared error (RMSE) and

r-squared values to measure model performance. Additionally in this section, the top ten most important features will be presented from the full model estimation. The third section will then apply selected models across the network in order to examine model performance. This section will also detail a proposed method to handle the bias in pedestrian traffic counts data due to the lack of zero counts being collected but surely exist at some locations on the network. A comparison of aggregate network wide estimates of pedestrian activity will be performed to assess the performance of the modeling approaches.

8.8.1 Machine Learning Based Pedestrian Traffic Data Fusion Model Cross-Validation Results

This section summarizes the cross validation procedures applied in the pedestrian data fusion model development element of this research as well as describes the features used in each of the machine learning algorithms. Similar to the vehicle and bicycle model training, cross-validation was done through both an internal and external cross validation. The results presented below are based on two machine learning algorithms including extreme gradient boosting (XgBoost) and random forest. Two sets of cross validation are performed, one that is characterized as internal that uses random partitions in a 10-fold cross validation and is done as a part of the model training process within the caret package. The second cross validation process, characterized as external, is performed on a select set of model specifications worthy of further investigation from the first validation and uses a stratified partition to do another 8-fold cross-validation. Eight folds are used because the data set is too small when trying to partition based on a specified stratification using functional classification and with 10 folds some partitions do not have all of the functional classifications making application the training data impossible. The internal cross validation uses 10 folds and was performed twice. Multiple model specifications are tested in the internal validation step using two type of algorithms (XgBoost and Random Forest) with a set of selected model specification being put forward to the external cross validation process.

Diagnostic information includes RMSE and r-squared values while the number of features used in the model is also presented. The internal validation results are a product of the initial model training using the caret package in R and uses a random partitioning process, using 10 folds and performed two times. The results displayed below in Table 8.21 summarize the internal cross validation tests and show that the XgBoost algorithm and random forest algorithm are similar in performance with a minimum r-squared value of 36% for XgBoost versus a 39% in the random forest. The maximum r-squared value for XgBoost is 53% while the maximum for random forest was 53 percent. The number of features used in the XgBoost is generally fewer than the random forest with at most 208 features while the random forest used nearly double with as many as 527 features being used. The impact of using Strava data denoted by *All + Strava* is not consistent across model estimations. Using the random forest algorithm, the Strava data does not improve the model as measured by r-squared but does reduce RMSE in the 2017+2018 estimation period. For the XgBoost algorithm, r-squared and RMSE is improved for the 2017+2018 estimation period.

Table 8.21: Internal Cross Validation Results for Vehicle Model

Algorithm Specification	RMSE	R-squared	Algorithm	Year	Feature Count
All + Strava	122.9	49%	Random Forest	527	2017+2018
All + Strava	129.6	39%	Random Forest	527	2018+2019
All	138.1	51%	Random Forest	524	2017+2018
All	128.4	53%	Random Forest	524	2018+2019
All + Strava	139.8	53%	XgBoost	192	2017+2018
All + Strava	129.3	36%	XgBoost	208	2018+2019
All	148.6	51%	XgBoost	165	2017+2018
All	119.4	44%	XgBoost	213	2018+2019

The top 20 most importance features are displayed below in Figure 8.31 for the *All + Strava* model for a select year and shows that employment features were commonly important variables in both model specifications, along with population measures and shortest path centrality measures. Network features include in the top 20 most important variables are limited to designation as an off street path and the estimated vehicle volumes.

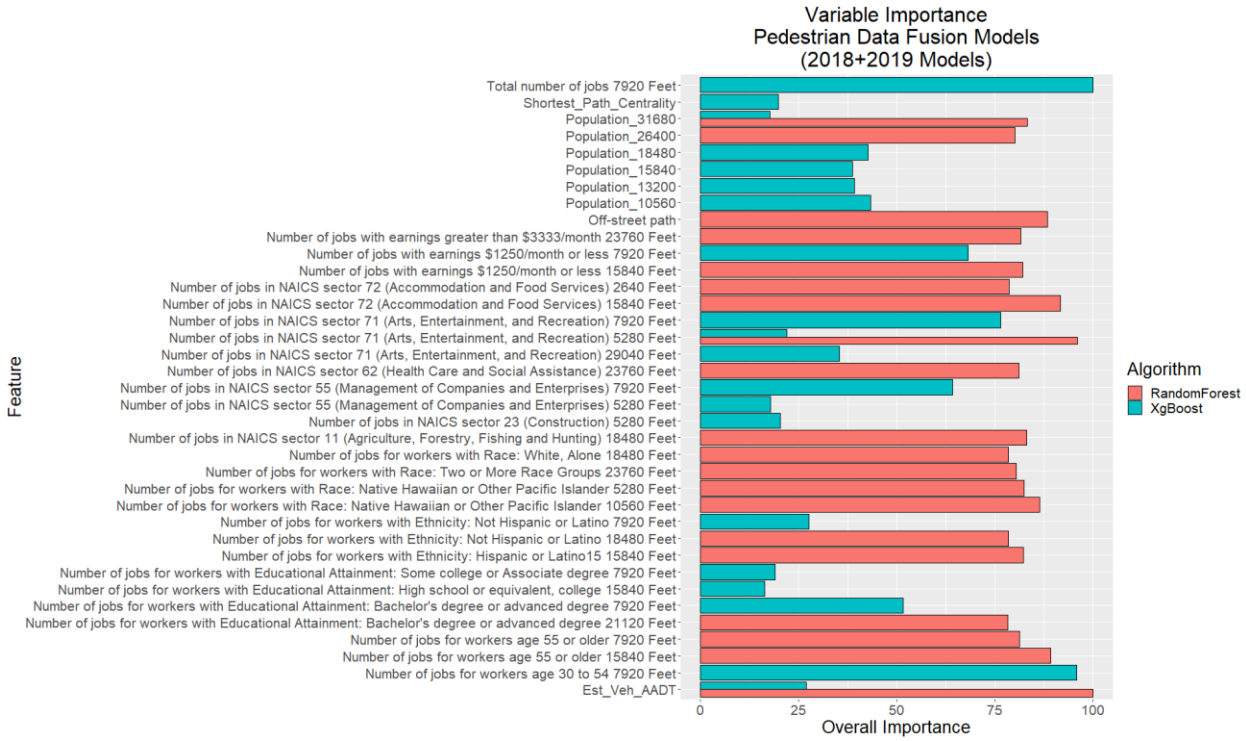

Figure 8.31: Variable importance for select pedestrian data fusion models

External validation tests are performed using both an 8-fold and a leave-one-out (LOO) process as was done in the vehicle and bicycle models validation above. An 8-fold test was done because of the lower amount of data available for the pedestrian model prevented partitioning into 10 folds. Results from the external 8-fold cross validation analysis are presented below in

Figure 8.32 and shows the median absolute percent error by volume bin for the two model specifications (*All and All + Strava)* and both algorithm types. These results demonstrate that XgBoost model works better than the random forest with in both specifications with having just 57% and 60% error for the *All + Strava and All* models respectively and 81% and 78% using the random forest algorithm. The addition of Strava data to the training features seems to make modest improvement in the median APE for all models and in all volume bins. The best model is the XgBoost using the *All + Strava* specification. In this model the error varies depending on volume bin with the lowest volume bin exhibiting the highest error of 251% for the XgBoost and the lowest error in the 81-160 bin with 34% error.

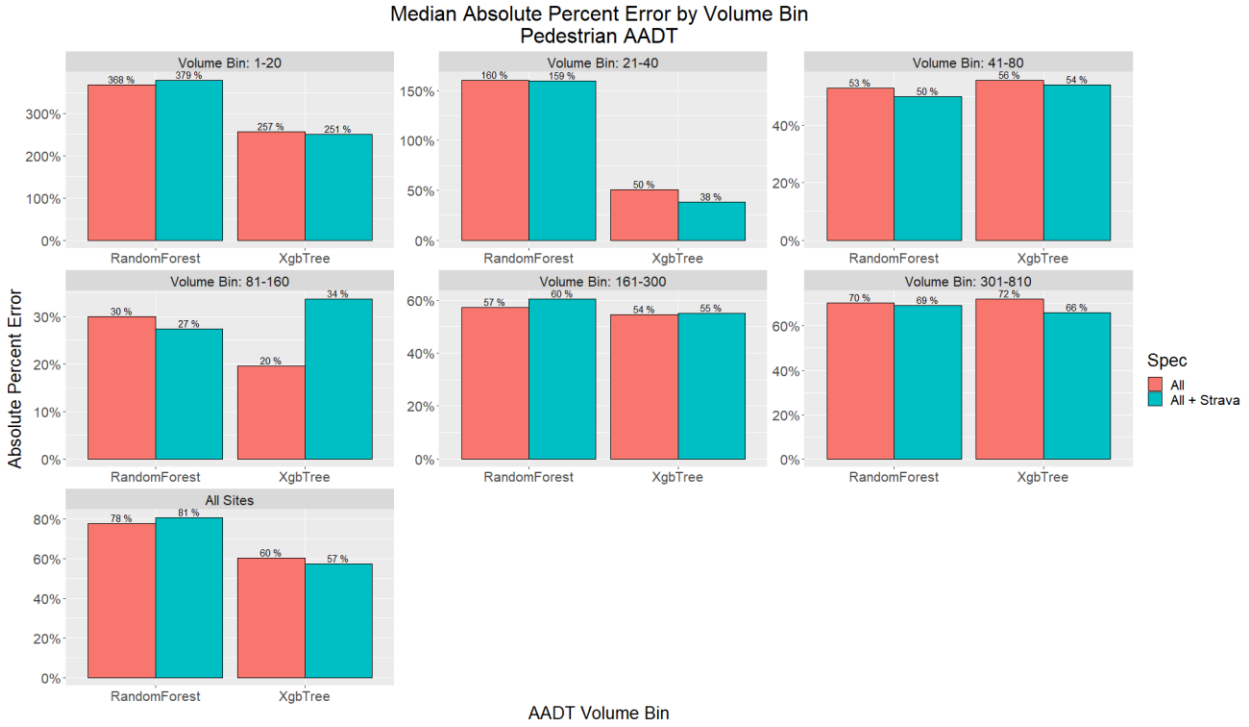

Figure 8.32: External 8-fold cross validation for bicycle models

Because the XgBoost algorithm worked best based on the internal validation and the 10-fold external validation the LOO cross validation process only tested this modeling approach. Because both model specifications (*All; All* + Strava) performed about the same both specifications are tested in the LOO cross-validation. Figure 8.33 summarizes the results of the LOO cross validation. These validation tests ensure that sites near a validation site are not included in the estimation by only using sites outside a 1,000 buffer, though due to the geographic sparseness of the pedestrian data this condition is not hard to attain. Performing tests this way helps to reduce bias in the cross validation results with median error of the LOO external validation rising to 66% from 67% mean APE in the 8-fold process summarized above. Error was lowest in the 81-160 volume bin with just 43% (random forest) and highest in the lowest volume bin with 297% median APE.

Figure 8.33: LOO Cross validation for pedestrian models

The last summary below shows estimated AADT compared to observed AADT for both estimation periods using the full model (without any data withheld) showing the correlation between the two values. It would be expected that the performance in this summary is high considering the estimation data is not separate from the application data. In fact, showing model performance in this way demonstrates the machine learning model does very well in predicting the observations in its estimation data with a high r-squared and relatively low RMSE.

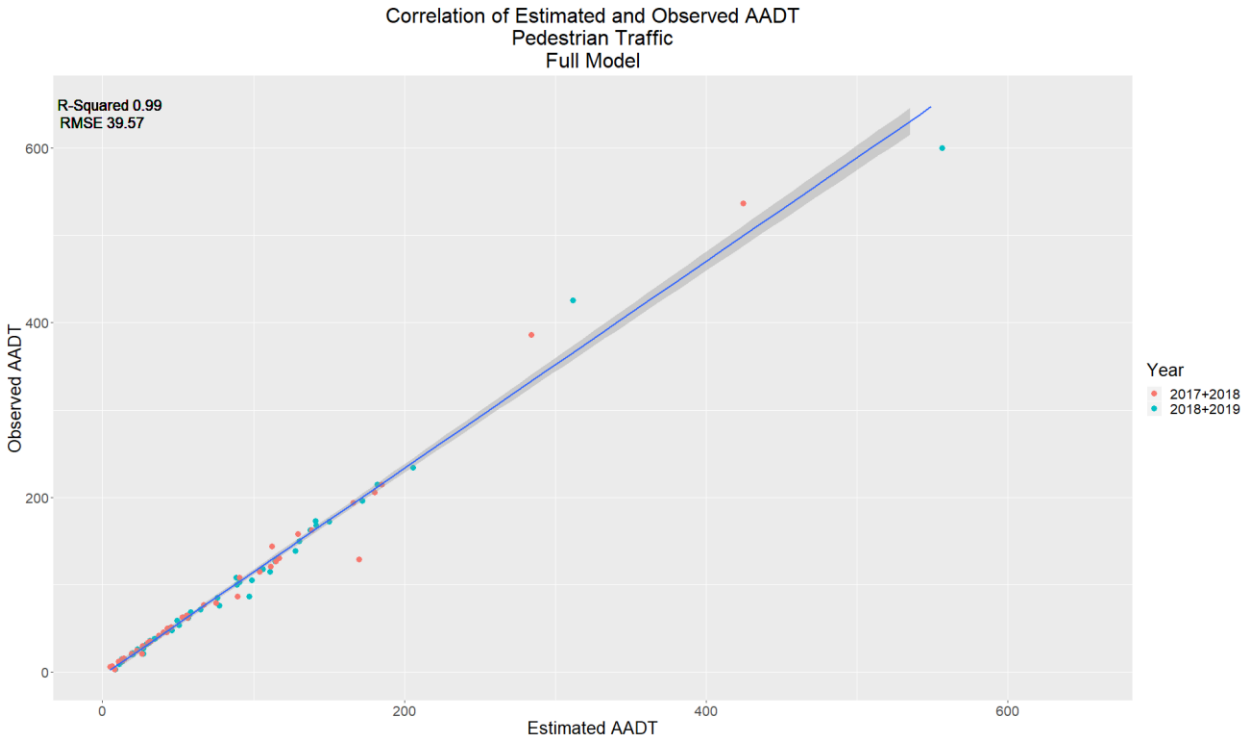

Figure 8.34: Correlation of estimated and observed AADT pedestrian traffic

This chart does not show any out of sample predictive capability since all of the data in the comparison data set are included in the estimation data. However, this comparison shows that with a full model estimated values are very close the observed data used to train the model. The full model does appear to do less well with higher AADT values, likely due to the lower number of observations in those volume ranges

8.8.2 Statistical Pedestrian Traffic Data Fusion Model Cross-Validation Results

This section will describe the development of statistical models to estimate bicycle AADT including an exploration of the individual effects of the covariates used in the final model. Since the number of available covariates for estimating a statistical model for bicycle traffic are numerous it was necessary to use a testing procedure to determine the variables with the best model prediction accuracy. This process uses 8-fold cross-validation to test the prediction accuracy of thousands of possible model specifications. Identical to the process used in the vehicle and bicycle model development above, a large number of specifications are tried and included 31,104 possible specifications based on a grid of all possible combinations of select variables including population access, total employment access, retail, health, and warehouse workers, intersection density, auto centrality, shortest path centrality a two measures of the Strava data including the total rider counts and the proportion of the Strava rider counts that were tagged as commute. The pedestrian model also included transit access measure at various distance thresholds. All the accessibility measures use shortest network distance thresholds of either one-quarter mile, half-mile, or one and a half miles. All models are estimated using a negative binomial regression specification due the counts data featuring over dispersion where

the dependent variable (pedestrian AADT) variance is greater than the mean of the counts which is generally the case for traffic counts data.

A custom process was developed in R where for the 2018/2019 counts period data is partitioned into 8 folds using a stratified random sample ensuring bike facility designations are equally distributed among the folds, especially the off-street paths. Eight folds are used because the amount of data is limited and the stratification process limits the number of folds. A negative binomial regression model is estimated on each of the k-1 groups (training data) and then estimated on the k-9 (test data) and then compared to the observed data. To do this for all 82,944 models the total runtime is about 12 hours even using parallel processing. For each selection of variables three performance metrics are computed include RMSE, mean absolute percent error (MAPE) and adjusted r-squared. Based on these metrics models top performing models are selected for further examination. For the pedestrian models the final estimated parameters are presented in Table 8.22 or three select models using these model performance measures. Model results below present the estimated coefficient and the associated standard error and p-value for selected models using full data with the highest r-squared, the lowest RMSE, and lowest MAPE for the 2018+2019 data.

These results shown in Table 8.22 below reveal that many of the covariates are correlated with an increase in pedestrian traffic including the presence of off-street path facility, shortest path centrality, total jobs, retail jobs, streets without a bike lane and Strava riders on a commute trip. Features associated with a decreased traffic volume include warehouse jobs, vehicle volumes and higher functional classification roads except that the highest classification, principal arterial, has a positive sign.

Table 8.22: Regression Results for Pedestrian Model

Coefficient	Std. Error	z value	P-value	Variable	Year	Metric
1.58E-04	0.0000	5.2343849	0.0000	Total number of jobs 5280 Mi.	2018+ 2019	Lowest MAPE
1.03E-05	0.0000	1.7906993	0.0733	Population_7920		
-0.0019	0.0007	-2.5847896	0.0097	Number of jobs in NAICS sector 48-49 (Transportation and Warehousing) 7920 Mi.		
6.00E-08	0.0000	1.3289739	0.1839	Shortest_Path_Centrality		
-1.03E-04	0.0000	-2.2675452	0.0234	Est_Veh_AADT		
-1.8150	0.3374	-5.3786179	0.0000	Local (Reference off-street path)		
-2.0746	0.3475	-5.9702033	0.0000	Collector		
-0.6852	0.4064	-1.6860395	0.0918	Minor Arterial		
0.3606	0.8560	0.4212569	0.6736	Principal Arterial - Other		
0.6965	0.2517	2.7678192	0.0056	No Bike Lane		
7.57E-05	0.0000	3.9013077	0.0001	Total number of jobs 7920 Mi.	2018+ 2019	Highest R-Squared
4.04E-04	0.0006	0.7102569	0.4775	Number of jobs in NAICS sector 44-45 (Retail Trade) 2640 Mi.		
-0.0057	0.0018	-3.1146703	0.0018	Number of jobs in NAICS sector 48-49 (Transportation and Warehousing) 2640 Mi.		
4.87E-08	0.0000	0.9902439	0.3221	Shortest_Path_Centrality		
-1.04E-04	0.0000	-2.4472025	0.0144	Est_Veh_AADT		
2.44E-04	0.0003	0.7908876	0.4290	Strava Commute Riders		
-2.3454	0.3665	-6.3989161	0.0000	Local (Reference off-street path)		
-1.9675	0.3476	-5.6605423	0.0000	Collector		
-0.9260	0.4068	-2.2766232	0.0228	Minor Arterial		
0.5487	0.8030	0.6833119	0.4944	Principal Arterial - Other		
1.0425	0.2648	3.9362692	0.0001	No Bike Lane		
7.55E-05	0.0000	3.8993362	0.0001	Total number of jobs 7920 Mi.	2018+ 2019	Lowest RMSE
-6.25E-06	0.0000	-0.543102	0.5871	Population_2640		
4.41E-04	0.0006	0.7687996	0.4420	Number of jobs in NAICS sector 44-45 (Retail Trade) 2640 Mi.		
-0.0058	0.0018	-3.1534368	0.0016	Number of jobs in NAICS sector 48-49 (Transportation and Warehousing) 2640 Mi.		
5.09E-08	0.0000	1.0225663	0.3065	Shortest_Path_Centrality		
-9.74E-05	0.0000	-2.2029157	0.0276	Est_Veh_AADT		
2.41E-04	0.0003	0.7811447	0.4347	Strava Commute Riders		
-2.3094	0.3744	-6.1680651	0.0000	Local (Reference off-street path)		
-1.9950	0.3507	-5.6883413	0.0000	Collector		
-0.9484	0.4074	-2.3278852	0.0199	Minor Arterial		
0.4291	0.8275	0.5185853	0.6040	Principal Arterial - Other		
1.0161	0.2696	3.7695985	0.0002	No Bike Lane		

The functional classification variable is operationalized in the models as a factor variable with the reference set to an off-street path so all coefficient estimates are in reference to this classification. For instance compare to an off-street path, all else being equal, local streets have less pedestrian volume but collector streets have even less. The bicycle facility variable is also operationalized as a factor variable with the bike lane set as the reference and is used in all models. And since most arterials have a bike lane this variable might be picking up on some of the vehicle traffic conditions since the results for most models show that compared to streets with a bike lane, pedestrian traffic is greater on streets without a bike lane (No facility) and off-street paths.

Not all variables are significant within the 0.05 level of significance but the proportion of variables that are significant at the 0.05 level ranges from just over half to two-thirds while the proportion of variables significant at the 0.10 level ranges from about 60% to just over three-quarters of the variables Table 8.23 below summarizes the three select models error measures.

Table 8.23: Model Diagnostic Information for Bicycle Regression Models

Specification	Performance Metric	MAPE	RMSE	Adjusted R-Squared
C000_5280 + Population_7920 + CNS08_7920 + Shortest_Path_Centrality + Est_Veh_AADT + Fc_Desc + Est_AADT + Bike_Facility	Lowest MAPE	88.9%	88.6	0.708
C000_7920 + CNS07_2640 + CNS08_2640 + Shortest_Path_Centrality + Est_Veh_AADT + Commute_Counts + Fc_Desc + Est_AADT + Bike_Facility	Highest R-Squared	111.2%	71.93	0.804
C000_7920 + Population_2640 + CNS07_2640 + CNS08_2640 + Shortest_Path_Centrality + Est_Veh_AADT + Commute_Counts + Fc_Desc,Est_AADT + Bike_Facility	Lowest RMSE	113.0%	71.89	0.799

The 10-fold holdout analysis results are further summarized by volume detailing the median APE for each of the models. The model with the lowest median APE for all sites is the same model with the lowest mean APE (*Lowest MAPE)*, as would be expected, and has lower median APE than the next model by about 4 percent. The Lowest MAPE model has lower error in all the volume bins except for the 11-20 and 601- 810 volume bins.

Figure 8.34: Top pedestrian regression model median absolute percent error by volume bin

8.8.3 Select Pedestrian Data Fusion Model Application

A primary objective of this research is to develop an estimation framework to apply network wide that will provide information about nonmotorized travel activity for the entire study area. This section will summarize the application results of select pedestrian data fusion models by applying the models to the entire network in order to generate system wide bicycle activity estimates. Additionally, an approach is suggested to handle over inflated counts on low volume, low density residential streets that make up significant lane miles of most urban networks. The issues and a proposed solution will be discussed below.

A prime objective of this research is deploying the models estimated and validated above on the entire system in order to estimate a system wide measure of pedestrian activity. The results below in Table 8.24 show the total annual pedestrian miles estimated using the XgBoost algorithm and the selected regression models. These results show that in the first estimate period using counts from 2017 and 2018 (2017+2018), the estimated total pedestrian miles traveled in the study region was 18.3 and 17.3 million miles for the *All + Strava* and *All* machine learning models respectively. The regression model estimates are 16.96, 11.4, 16.90 million miles for *Highest R-Squared*, *Lowest MAPE* and *Lowest RMSE* models respectively. For the second estimate period, from 2018 and 2019 (2018+2019) the total PMT estimate is 15.9 and 15.2 million miles for the *All + Strava* and *All* machine learning models respectively. The regression model estimates are 14.1, 13.8, and 13.9 million miles for the *Highest R*-Squared, *Lowest MAPE* and *Lowest RMSE* models respectively.

Table 8.24: Total Pedestrian Miles Traveled for Select Models

Model Specification	Algorithm Type	Total Annual Pedestrian Miles Traveled	Bend Population	Per Capita BMT	Year
All + Strava	XgbBoost	18,338,312	96,058	0.52	2017+2018
		15,876,191	99,171	0.44	2018+2019
All		17,300,492	96,058	0.49	2017+2018
		15,220,786	99,171	0.42	2018+2019
Highest R-Squared	Negative Binomial	17,007,324	96,058	0.49	2017+2018
		14,009,875	99,171	0.39	2018+2019
Lowest MAPE		11,393,385	96,058	0.32	2017+2018
		13,758,836	99,171	0.38	2018+2019
Lowest RMSE		16,958,069	96,058	0.48	2017+2018
		13,921,306	99,171	0.38	2018+2019

The estimates from the regression models for the 2017+2018 study period deviate somewhat substantially, especially in the case of the *Lowest MAPE* model that is about one-third less the other regression model. There is more consistency in the latter study period models with all the result being within 3% of one another. The machine learning model that uses Strava as a training feature appears to increase the total estimate for the 2017/2018 period by about 6% and 2018/2019 period by 4 % compared to the *All* model that does not use this training feature. It's not clear how why the machine learning algorithm is making use of the Strava data but has demonstrated in the cross-validation model accuracy improve when using Strava data and the *Lowest RMSE* regression model also uses a measure derived from Strava data.

Figure 8.35 below displays the total annual PMT estimates by selected model scenario including the *All + Strava* machine learning model and the *Lowest* MAPE regression models. These specifications were chosen for their performance in low APE. These results show that the BMT summary aggregated by functional classification for a *Strava + All* machine learning model and the *Lowest MAPE* regression model. *Lowest MAPE* is selected because MAPE was the performance measure used to select which of the machine learning model specifications to focus on and so was followed for the selection of regression models. The figure below shows that PMT estimates are higher in the application of the machine learning models on half the functional classifications including the local, collector, and minor arterial streets whereas the regression model estimates higher PMT on off-street paths, major arterials, and highways. The collector classification has the largest percentage difference followed by off-street paths.

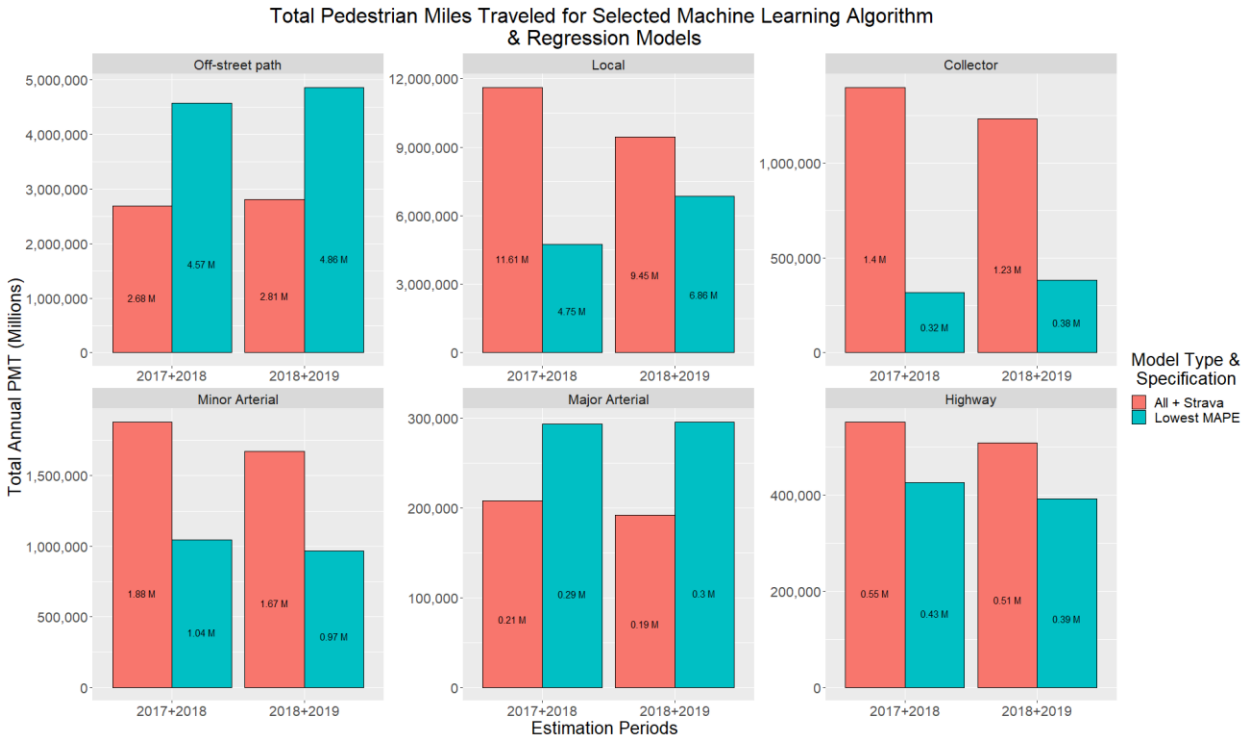

Figure 8.35: Pedestrian miles traveled estimates for selected scenarios by functional classification

Of note is the significant number of PMT that are being estimated on the local road system. The local road system may be an attractive facility for people to walk due to its low vehicle and speed and volume and relative proximity to residential areas (population access) and parks. However, many of these streets are likely to have zero counts given their low accessibility to key destinations and because of the nature of the traffic count programs where streets with likely pedestrian users were counted, the available counts are likely biased upwards and using them in a network wide application is likely biasing the total PMT results upward. In order to handle this issue, a proposed solution is offered where zero counts locations are introduced into the counts data at locations where zero pedestrian traffic is likely. The criteria for the random selection of these zero count locations are described below:

- Local street functional classification with no bicycle lane
- Population access within 0.5 miles must be 400 people or less
- Shortest path centrality must be zero
- No Strava rider counts

Using this criteria about 41 miles or 10% of the local street network, become eligible for having a zero count assigned to it. Of these local streets, 30 links are randomly selected and those 30 locations are added to the counts data and the machine learning algorithms are retrained with the

inclusion of the simulated zero counts data. The remainder of this section will detail the BMT results of the modeling with the inclusion of these randomly selected zero count locations.

With the introduction of the zero counts the distribution of the data is altered and the negative binomial model is no longer appropriate and instead a Poisson model is used to estimate the model using the simulated zero counts. Future research should explore the use of zero inflated hurdle models to see if that specification changes the final BMT results. With about 25% of the counts now being zeros it's likely this would be a more proper specification. Table 8.25Table 8.25below details the results for the new PMT estimate scenario where 30 zero count locations were inserted into the model training data. On an aggregate basis, the total PMT decreases to 67% of initial estimate for the 2017/2018 estimation period, and 55% for the 2018/2019 estimation period when estimated using the XgBoost machine learning algorithm with the *All + Strava* specification. Using the Poisson regression approach but including the simulated zeros the estimated BMT drops to 62% of initial estimate for the 2017+2018 estimation period and 58% for the 2018+2019 estimation period. These inclusion of these zero counts significantly reduces the total PMT estimated for the entire system.

Table 8.25: Total Pedestrian Miles Traveled Comparison with Simulated Zero Counts Scenario

Model Type and Specification	Estimation Periods	Total Annual Bicycle Miles Traveled		Percent Difference
		No Zero Counts	Simulated Zero Counts	
Machine Learning: All + Strava	2017+2018	17,300,492	11,585,489	67%
	2018+2019	18,338,312	10,040,018	55%
Regression: Lowest MAPE	2017+2018	11,393,385	7,100,915	62%
	2018+2019	13,758,836	7,975,721	58%

Figure 8.36 below details the aggregate PMT by functional classification for both modeling approaches (machine learning vs. regression) and shows the PMT estimate without simulated zero counts and with those simulated zero counts. The insertion of zero counts into the machine learning training data depress the estimated PMT for the local streets, as designed, reducing the estimated BMT on those facilities from 9.45 million PMT to 4.99 million PMT for the 2018/2019 estimation period, a reduction of roughly 48 percent. When the zero counts are included in the regression model approach the PMT on local streets goes from 6.86 million PMT to 2.05 million for the 2018/2019, a change of about 70% percent. Most facility types have a diminished PMT estimate in both periods.

Figure 8.36: Pedestrian miles traveled estimates comparison of zero counts scenario by bicycle facility type and functional classification

The insertion of zero counts at locations with low density and low network connectivity appear to have the desired effect of moderating the overall PMT estimates. Figure 8.37and Figure 8.38 below shows the results of the network wide application of both model approaches and the scenarios using counts data and counts data with simulated zeros. The left panel shows the results of the model applied to the network with all observed data while the right panel shows the model with simulated zero counts at low density locations. Where as in the left panel there are no locations where zero counts are estimated (denoted by grey) while the right hand panel shows a small number of links in far flung parts of the network with no estimated pedestrian activity. Additionally, the simulated zero counts scenario moderates pedestrian volumes throughout the low density areas surrounding the core of the study region, with many more links in the 1-5 AADT volume bin. In fact there no links in the No Zero Counts scenario with 1-5 pedestrian AADT while in the Simulated Zero Counts scenario there are 2,112 links with volume in this range for the XgBoost based model.

Aggregate measure of PMT between the two model approaches in the 2018+2019 period are different by about 25% with the machine learning model estimating more pedestrian activity. These differences are most stark in the core of the study region. The XgBoost model appears to spread the activity out in the downtown area while the regression model targets the activity to a discrete corridors. Those corridors are more pronounced in the scenarios where the zero counts were injected into the training data. The XgBoost results do create about 800 links where the estimate is a negative value which are then converted to a zero for the purposes of aggregation and network visualization.

The table below presents some summary statistics of the estimated bicycle volumes on the 14,000 links that make up the study region network and which were presented below in Figure 8.37 and Figure 8.38 via map visualization. As expected the mean estimated summary statistics all decrease with the injection of simulated zero counts with the XgBoost model estimating a negative values on about 800 links, which are converted to zero.

Table 8.26: Summary Statistics of Estimated Counts for Total Network Application of Pedestrian Fusion Models

Model Specification	Scenario	Estimated AADT Summary Statistics				
		Minimum	Maximum	Mean	Median	Std. Dev.
All + Strava	No Zero Counts	8.62	601	69.5	49	84.1
Lowest MAPE	No Zero Counts	6.86	995	52.2	37	46.8
All + Strava	Simulated Zero Counts	0*	578	43.3	26	60.2
Lowest MAPE	Simulated Zero Counts	3.06	1057	25.8	11	47

*811 links given an estimated AADT of between -2.55 & -0.0003

The regression model maximum estimate is larger than the machine learning model maximum estimated values and exceeds the maximum range of the observed counts data.

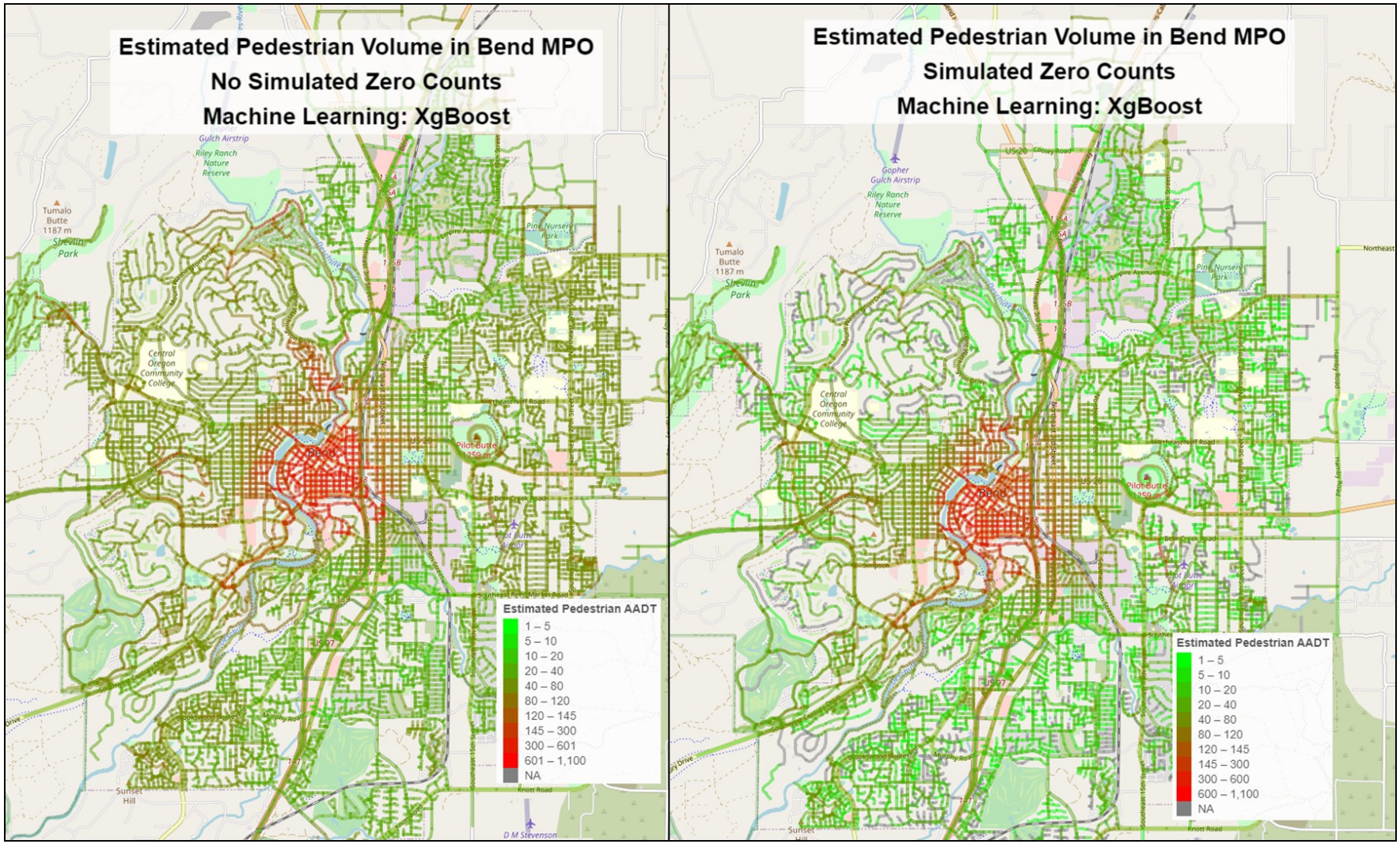

Figure 8.37: XgBoost - comparison of bicycle miles traveled scenarios – network level estimates

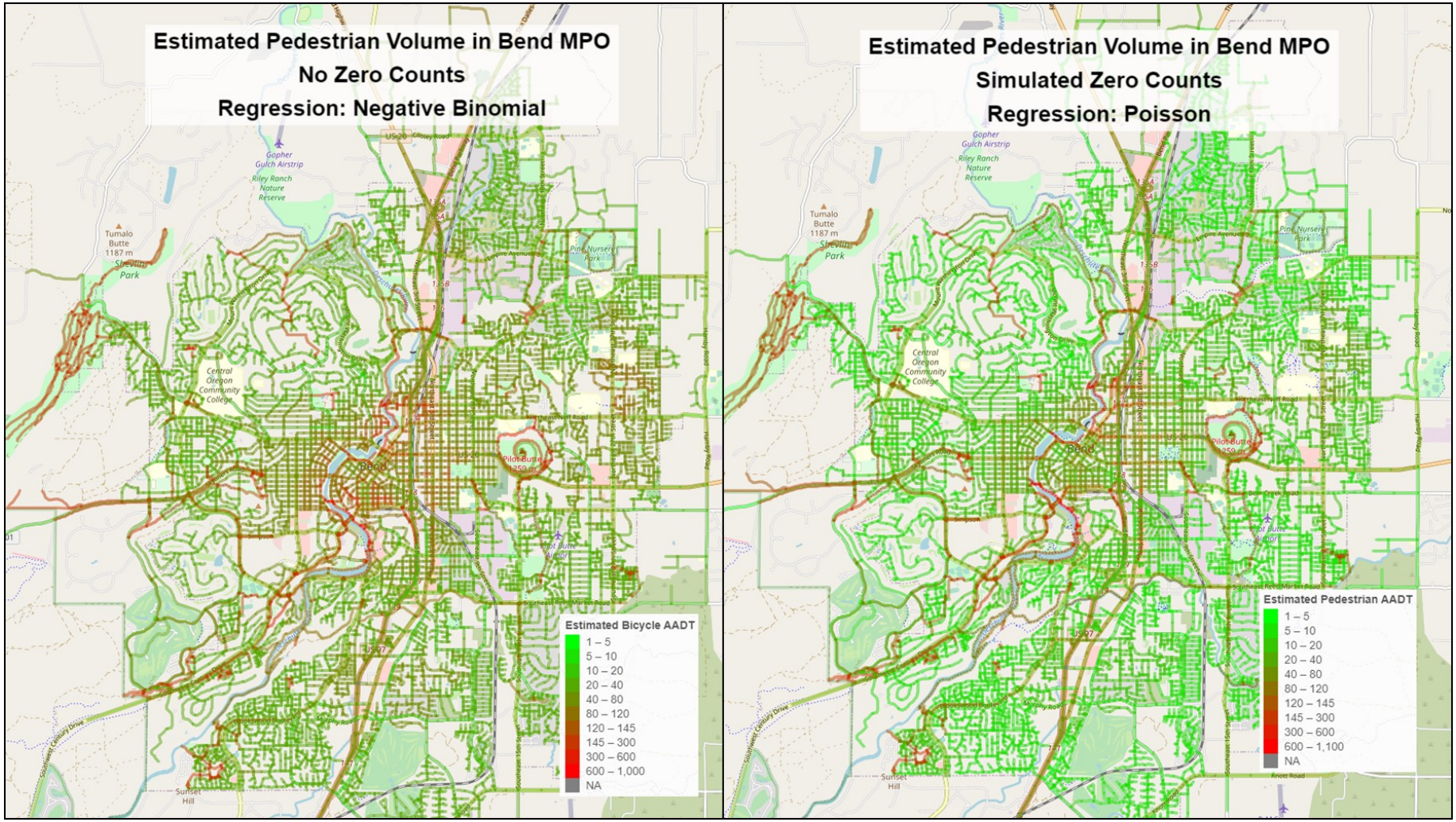

Figure 8.38: Regression - comparison of bicycle miles traveled scenarios – network level estimates

8.8.4 Pedestrian Data Fusion Discussion

The above section detailed the data, estimation procedures, validation, and results of data fusion models for pedestrian traffic volumes in the study region. The validation results for the machine learning models showed that the XgBoost machine learning algorithm worked better than random forest across three separate cross-validation procedures though for some volume bins in the 8-fold and LOO cross validations random forest performed better in terms of median APE. These validation tests also showed that including the Strava data only improved APE marginally in the 8-fold cross validation and decreased performance marginally in the LOO cross-validation.

The *All + Strava* model was chosen for additional exploration comparing results of the applied model with the regression mode. The network wide application of the data fusion models PMT results seemed showed that Strava data increased total activity estimates, which was the opposite effect of the inclusion of Strava data in the bicycle models where Strava data moderated the overall estimate. However, using just the observed data in the data fusion model is likely biasing the PMT estimate upward, due to the selection of count locations where pedestrians are expected. To handle this bias, an approach is suggested whereby zero counts are injected into the training data at locations where zero pedestrians would be expected. The results of this approach present the expected outcomes, further moderating estimated pedestrian activity across the network, especially at locations where pedestrian activity is likely to be low. Continued discussions are necessary with potential model users about an application ready pedestrian data fusion model so model users completely understand the advantages and limitations of using either of the models examined in this research as tradeoffs exists.

The use of machine learning in estimating network wide pedestrian activity is novel, based on the current status of the literature. Machine learning offers significant advantages for predicting important quantities such as pedestrian volumes where inferential data is less important for model users. Additionally, the selected machine learning algorithms offer powerful mechanisms for accounting for the interaction of many complicated relationships between network variables and are likely important tools for monitoring the system and understanding network wide activity. These models will only improve as more data is collected and the data collected and fed into the model estimation process. However based on the features currently being used in the machine learning algorithms, results appear less reasonable than the regression models with activity estimates being spread out across the network instead of being concentrated on select corridors.

Model results would be improved with updated data for certain data elements. For instance, the decrease in pedestrian miles traveled from the first estimation period to the second could be because the employment data used in training and application was a single year, representing 2017 since 2018 data has yet to be released by Census Bureau, as noted above in the bicycle data fusion model discussion section. Other data from LEHD could be harnessed, including origin-destination information that connects worker residential locations and their place of work. A major issue in the training feature data is the use of population data from 2011. These data were used because of their ease of availability but more updated data from American Community Survey could be used to better reflect the conditions when traffic counts were collected, again as noted in the section above on bicycle data fusion. Other model estimation and application improvements could be to evaluate the Strava data in more detail and correct places where

potential issues are present. Transit ridership data would likely improve the models instead of the grosser measure of transit stop access.

8.9 DATA FUSION RESULTS COMPARISON

The sections above detail data fusion models using parametric and non-parametric approaches for three modes of travel and this section will briefly compare the results from each of those exercises.

Table 8.27 below summarizes the r-squared, RMSE, and median absolute error for the optimal models for each mode. Optimal models were selected based on the algorithm and specification with the lowest absolute percent error found through various cross-validations tests performed above. The vehicle models work best compared to the other modes with lower error across validation tests as well as in the full model where a model is estimated on the full data set in order to predict the *not* out-of-sample data. Since the regression model only did a 10-fold cross validation these results are the most comparable across tests. Based on these tests the XgBoost model and regression model are comparable with the regression model exhibiting lower median APE but lower r-square values. Vehicle counts data are more numerous which helps with model training for both modeling approaches and the functional designation provides a significant clue to the training of models as to what bucket the volume is likely to fall into thus helping to improve model performance. With independent estimates of VMT from the HPMS these models can be more fully validated and show that at the network level and functional classification level both of these modeling approaches work well. With probe data the vehicle models would likely improve.

The bicycle models do not perform as well as the vehicle models, likely due to a much smaller training set which makes cross validation harder. Lower overall volumes also make reported error hard to compare with the vehicle volumes. For instance if the actual volume for a given location were 30 bikes per day (average volume for all sites in 2018/2019 period) and the model estimated 20 the error is 50 percent. These same issues exist for pedestrian counts. Bike and pedestrian traffic volume have some correlation to facility type, namely off-street paths, but no volume classification is yet defined for bike and pedestrian transportation networks and since traffic monitoring for these modes is still in the beginning stages a full enough understanding of how to develop such a classification scheme does not yet exist. Even though error for the bicycle models is not extreme, though certainly higher than the vehicle models, an aim for the future of bicycle and pedestrian data collection should be to continue collecting data in new locations to try and meet the number of locations available in the vehicle counts (n = 250) though its likely more sites will be needed to make the situation of low volumes overall for bike and pedestrian workable from an error perspective.

For the bike and pedestrian models the regression approach appears superior based on the cross validation tests. The XgBoost approach resulted in median error of 43% while the regression approach produced only 39% error. For the pedestrian model the XgBoost model resulted in 57% error while the regression model was able to reduce the median APE to just 36 percent.

Table 8.27: Model Diagnostic Information Summary All Modes and Select Specifications

Mode	Model Type	Performance Metric	Cross-Validation Method			Full Model
			Internal	External - 10-Fold*	External - LOO	
Vehicle	XgBoost	R-Squared	54%	63%	67%	0.999
		RMSE	6631	6131	6551	379
		Median APE	NA	39%	40%	0.990
	Regression	R-Squared	NA	55%	NA	50%
		RMSE	NA	7897	NA	7212
		Median APE	NA	40%	NA	39%
Bicycle	XgBoost	R-Squared	32%	15%	19%	95%
		RMSE	25.7	24.6	27.4	9.7
		Median APE	NA	43%	44%	13%
	Regression	R-Squared	NA	35%	NA	42%
		RMSE	NA	24.4	NA	22.7
		Median APE	NA	39%	NA	55%
Pedestrian	XgBoost	R-Squared	36%	10%	32%	99.0%
		RMSE	129.3	165.1	130.6	39.7
		Median APE	NA	57%	67%	15%
	Regression	R-Squared	NA	71%	NA	80%
		RMSE	NA	71.9	NA	70
		Median APE	NA	36%	NA	77%

8.10 DISCUSSION AND LIMITATIONS

The above section describes the data, development and application of data fusion models for vehicle, bicycle and pedestrian travel activity. The vehicle models show significantly better performance compared to the bicycle and pedestrian models. With additional data collected over the next few years these models may achieve better accuracy but some challenges inherent in bicycle and pedestrian volume data for this study area, such as low overall volumes, may continue to limit the overall accuracy of modeled volume data.

Other data sources and training features would likely help the nonmotorized models. The bicycle models would benefit from access to parks and trails outside the urban area while the pedestrian models would benefit from better information on transit such as ridership. Both bicycle and pedestrian models (and vehicle) would benefit from updated population data. Pedestrian models may benefit from Strava's running/walking data and both modes could benefit from the origin and destination product. Additionally, other third party data sources currently on the market should be evaluated to understand how they could impact model performance.

Other improvements might come from adjustment of the hyper parameters used in the model training which were tuned with some benefit in model performance in this work but could be explored more in any future application of these techniques.

For network scale travel monitoring it's not yet certain what level of confidence is necessary for useful bicycle and pedestrian miles traveled. The next section of this report will explore the use of these BMT and PMT measures in aggregate level crash risk analysis to see if the current imprecision in the estimates is acceptable.

9.0 NONMOTORIZED CRASH ANALYSIS

Crash risk for people that use nonmotorized travel is typically understood to be greater than for motorized travel though few studies assessing risk using exposure based crash rates have been conducted. Exposure based crash risk analyses require traffic counts and related estimates of annual activity which are not common data for nonmotorized travel. One of the primary objectives of this research project is to use the nonmotorized traffic count data and related network wide traffic estimates in crash risk analysis to better document the crash risk disparities for nonmotorized users. Additionally, this research aims to offer information for how roadway features impact disparate crash risk at the system level, modeling features like nonmotorized traffic volume, functional classification, and vehicle volume and their role increasing risk for nonmotorized users. For the Bend, Oregon study area crash modeling is limited due to small number of nonmotorized crash injuries.

A literature is presented summarizing existing literature on nonmotorized crash risk at the system level. Additional literature is provided documenting other examples of using measures of nonmotorized travel activity from direct demand models for crash analysis. This research adds to the literature by offering additional information on crash risk at both an aggregate and disaggregate level for a small urban area in Oregon.

10.0 LITERATURE REVIEW OF NONMOTORIZED CRASHES

The literature review section below examines the past research and public agency reports that examined nonmotorized crash risk at an aggregate system wide level. The literature review then documents the literature that has employed direct demand modeling for nonmotorized activity estimates to be used in crash analysis.

10.1 AGGREGATE NONMOTORIZED CRASH RISK LITERATURE REVIEW

Only minimal research has attempted to describe the bicycle crash risk on the aggregate, system wide level. Using data from the National Household Travel Survey for 2001, Pucher and Dijkstra (2003) showed that fatal bicycle crash rates are 12 times higher than vehicle occupants and pedestrians 23 times more likely to be killed in a traffic related injury crash. The authors also found that bicycle crash rates in the U.S. are double those in Germany and three times higher than kilometer based rates in the Netherlands. The authors are not able to compute non-fatal injury rates due to unreliable data. Pucher and Dijkstra point out that Germany and the Netherlands enforce much lower speed limits for vehicle traffic and also administer more widespread implementation of turn restrictions at intersections that prioritize nonmotorized user safety. Beck et al. (2007) used data from 2001 National Household Travel Survey and fatal crash information from the Fatal Accident Reporting System (FARS) and non-fatal crash data from the General Estimates System (GES) to calculate person trip crash rates for multiple modes of travel. The researchers found that fatal crash rates for people riding bicycles were more than double passenger vehicle rates and nonfatal injury rates for bicyclists were nearly double those of passenger vehicle occupant rates. Pedestrian fatal injury rates were measured to be about 49% higher than vehicle occupant fatal injury risk. McAndrews (2011) compiled travel survey data for Stockholm, Sweden and San Francisco, CA in order to estimate travel activity for motorized traffic, bicycle and pedestrian users. The authors concluded that based on person miles of travel bicycle and pedestrian fatal injury rates were as much as 85% lower compared to motorized travelers using mileage based rates but as much as four times higher using person minutes of travel. McAndrews et al. (2013) measured travel activity in Wisconsin for all modes using an add-on to the National Household Travel Survey and created exposure based rates for fatal injuries and non-fatal injuries. The authors found that the relative risk of bicycle travel compared to motor vehicle travel was 10.5 and 17.1 for fatal and non-fatal injuries respectively using the mileage based exposure measures. For pedestrian fatal and non-fatal injury rates the relative risk was 11 and 11.8 respectively. Mindell et al. (2012) calculate miles of travel based fatal and injury rates for vehicle, bicycle, and pedestrians using a national household travel survey to measure travel activity and multiple sources of crash data. They demonstrate that the relative fatal injury risk for people who bike and walk in the UK is 10 to 11 and 13 to 16 respectively, times, higher than the fatal injury rate to of people who drive. For non-fatal injury the relative risk per distance traveled compared to driving is 50-58 and 49-59 for biking and walking respectively. The authors not that the bicycle injury rates are likely over estimates because the inj8ury data over counts traffic related injuries of people biking but that the pedestrian injury rates are likely underestimated because injury data is missing for on-road

pedestrian injuries. Teschke et al. (2013) used a travel survey of British Columbia, Canada to calculate fatal and injury crash rates per kilometer for automobile users and people who ride bicycles and walk. They found that fatal crash rate for bicyclists was over two and a half times that of automobile users and injury crash rates were nearly three and a half times higher for people riding bicycles. For pedestrian injuries the mileage based fatal injury rate was measured to be 7.6 times higher than vehicle fatal injury and 2.7 times for non-fatal injury. This research relies on travel surveys for calculating travel distance which is relies on self-reported distance which can introduce error into the travel distance measures. Additionally, travel surveys do not typically account for recreational trips which can make up a large proportion of bicycle travel for a given region. These two limitations may bias the previous crash rate estimates upward since they do not fully account for the full value of the denominator.

Roll (2018) estimated bicycle fatal and non-fatal injury rates using miles traveled for the Eugene-Springfield urban area in Oregon. The results demonstrated that for the time period examined, bicycle fatal injury rates were three times higher than motorized fatal injury rates, and non-fatal injury rates were 2.1 times higher.

There is some debate about whether distance based exposure measures should be used versus time based measures. Hakkert and Brainmaister (2002) examine this debate and concluded that deciding between distances versus time based risk depends on the issue being examined. They point out but don't examine fully one contradiction where increased speed can reduce time based exposure but then inherently increase risk due to the implications of higher speed. However, it is commonly understood that speed increases risk, especially at the upper margins of vehicle speeds when the driver's ability to react is further limited. Though an interesting philosophical debate, this research will rely on distance based metrics for the crash analyses presented below.

Table 10.1 below summarizes the factor by which nonmotorized crash injury rates differ from motorized crash injury rates, further summarizing the literature review above. This summary shows that bicycle fatal injury rates are between 2.3 and 23 times fatal injury rates for motorized travel with non-fatal injury rates being between 2.1 and 3.7 times higher than motorized injury rates. Pedestrian fatal injury rates are between 1.5 and 12 times higher than motorized fatal injury rates with non-fatal injury rates at least 2.7 times higher than motorized non-fatal injury rates.

Table 10.1: Summary of Crash Risk Disparity

Reference	Pedestrian		Bicycle		Study Area
	Fatal Injury	Non-Fatal Injury	Fatal Injury	Non-Fatal Injury	
Pucher and Dijkstra (2003)	12	Not-Reported	23	Not-Reported	US
Beck et al. (2007)	1.5	Not-Reported	2.3	Not-Reported	US
McAndrews et al (2011)	0.15 - 0.28	Not-Reported	0.18 - 0.55	Not-Reported	Stockholm, Sweden & San Francisco, CA
Mindell et al. (2012)	13-16	49-59	10-11	50-58	United Kingdom
Teschke et al. (2013)	7.6	2.7		3.7	British Columbia
McAndrews et al (2013)	11	11.8	10.5	17.1	Wisconsin
Roll (2018)	Not-Reported	Not-Reported	3.0	2.1	Eugene-Springield Oregon

10.2 DIRECT DEMAND MODELS AND CRASH RISK ANALYSIS

This section of the literature review documents the previous work estimating and deploying direct demand models for use in nonmotorized crash risk analysis. Though less common just a few years ago, the approach of estimating exposure using this analytic method now has a number of examples. In 2018 the FHWA released the Guide for Scalable Risk Assessment Methods for Pedestrians and Bicyclists (SCRAM) outlining acceptable approaches for nonmotorized crash analysis. This guide discusses appropriate methods for assessing risk at various levels including at the system level where estimates of bicycle and pedestrian activity can be used to generate measures of risk for use in performance monitoring. A method for deriving exposure measures discussed in detail includes the direct demand modeling approach.

Thomas et al. (2017) develop safety performance functions (SPFs) for three types of bicycle crashes using volume measures from a direct demand model. Crash types include all intersection crashes, bicyclists opposite direction, and bicyclists, angle crashes using eight years of crash police crash data from Seattle, WA. Bicycle traffic volume data is estimated using a direct demand model through a so-called “ball park’ method that relates short-term and automated counter data at 46 intersections to factors correlated with bicycle activity. Vehicle traffic volume was unavailable and functional classification was used instead. The authors employ a Conditional Random Forest (CRF) regression analysis to uncover eligible crash predictors before specifying an SPF using negative binomial regression. The safety performance function uses the natural log of bicycle volume as well as estimates of annual average daily pedestrian traffic in conjunction with intersection variables like the presence of signals, entering segment legs, parking, lanes, and transit stops. The authors also include the amount of commercial building space within a specified buffer. Thomas et al. (2017) find that an increase in motor vehicle volumes as measured by the functional classification increases the risk for bicycle crashes for all crash types. Intersections with traffic signals increased the risk of bicycle crashes as did the

presence of parking. This research also found the presence of bicycle lane and shared markings had a positive correlation with bicycle crashes. The authors apply the estimated SPFs using three approaches including an unadjusted prediction of bicycle crashes, Empirical Bayes adjusted prediction of bicycle crashes, and a Potential for Safety Improvement (Persaud et al. 1999) where the difference between the EB expected and SPF predicted crashes is calculated. The authors conclude that the data and methods used in the analysis offer a way for cities to prioritize locations for further investigation and likely treatments.

Griswold et al (2018) develop and apply a direct demand pedestrian model for the purposes of safety analysis for the California Department of Transportation. Using short-term pedestrian counts data from 1,270 intersections on the CalTrans system, a direct demand model is estimated using employment, population, street density, walk commute share and functional classification as independent variables. The authors specified their model using an ordinary least squares regression with log-transformation of many of the independent variables. Model performance was tested by randomly splitting the data into 90 percent training and 10 percent testing partitions. This Monte Carlo cross validation scheme was performance 300 times with the adjusted r-squared results of 0.714. The results of the model are then applied to the entire CalTrans network in order to provide estimates of pedestrian traffic for use in project prioritization. No aggregate risk measures are calculated using the exposure measures.

Hasani et all (2018) use bicycle and pedestrian volume data collected at 45 intersections in San Diego, CA to estimate a direct demand model for use in nonmotorized risk analysis. These data are collected using video and processed by computer vision algorithm, then factored to represent annual traffic conditions. The authors employ the activity estimates calculate risk at the intersection level across the study area. The authors weight injuries with different severities by using a cost of injuries method that gives higher weight to more severe injuries. This research concludes by offering priority locations for intervention based on their proposed methodology

11.0 CRASH DATA DESCRIPTIVES

This section summarizes the crash injuries for motorized, bicycle, and pedestrian crash participants in the Bend, Oregon urban area, the area under examination in this research. The crash injury data are derived from the official Oregon Department of Transportation Crash Data System (CDS) and include all injuries reported to the agency. It's possible the crash injuries for nonmotorized participants are not fully reported since research in cities outside of Oregon have demonstrated a systematic underreporting of these kinds of crashes (Shinar et al 2018; Winters & Branion-Calles 2017; Langley et al 2003). Nevertheless, the ODOT CDS is the most comprehensive and high quality database of crash injury available for this research.

Injury severity codes are defined in Table 11.1 below and include fatal, severe, moderate and minor injuries categorized into the KABC index. The figure below summarizes the annual number of injuries by injury severity for each of three modes of travel including motorized, pedestrian, and bicycle.

Table 11.1: Injury Severity Description

Code	Short Description	Long Description
K	Fatal	Fatality information includes motor vehicle traffic crashes that result in the death of an occupant of a vehicle or a non-motorist within 30 days of the crash.
A	Incapacitated/ Severe Injury	Any injury to the driver of the identified UNIT that prevents the injured party from walking, driving, or normally continuing the activities he or she was capable of performing before the injury occurred. Examples include broken or distorted limbs, skull or chest injuries, abdominal injuries, unconscious at or when taken from the crash scene, unable to leave crash scene without assistance, etc.
B	Visible Injury	Check this box to indicate any injury to the driver of the identified UNIT which is evident to observers at the scene of the crash. Examples include a visible lump, abrasions, cuts, bruises, minor lacerations, etc.
C	Complaint of Pain	Any injury claimed by the driver of the identified UNIT. Examples include momentary unconsciousness, complaint of pain, limping, nausea, etc.

Motorized transport includes passenger car, heavy and light duty truck, and motorcycle. Figure 11.1 shows that for both motorized and nonmotorized travel, fatal injuries are relatively infrequent compared to severe and all injuries but generally consistent from one year to the next, especially for nonmotorized injuries.

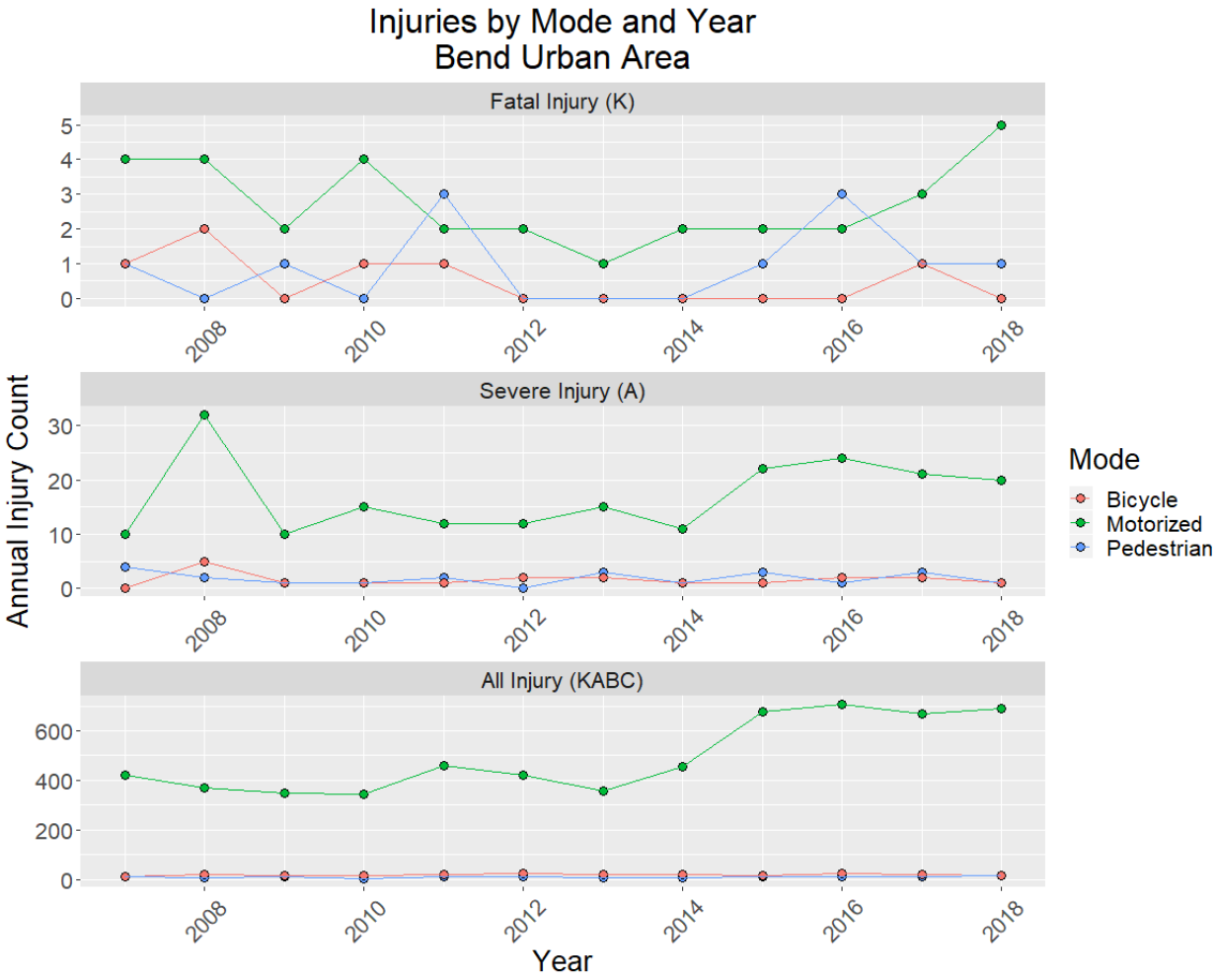

Figure 11.1: Injuries by travel mode and year in Bend urban area

Figure 11.2 below summarizes the average annual injury count for each mode, severity and aggregation period. Three aggregation periods are shown including periods that include the years 2014 through 2018 (2014+2018) 2013 through 2017 (2013+2017) and 2007 through 2012 (2007+2012). The first two periods, 2014+2018 and 2013+2019 will be used later in this report as injury data for crash rate calculation. The third period, 2007+2012, is used as a reference to compare the other two periods to assess stability of annual average injury counts.

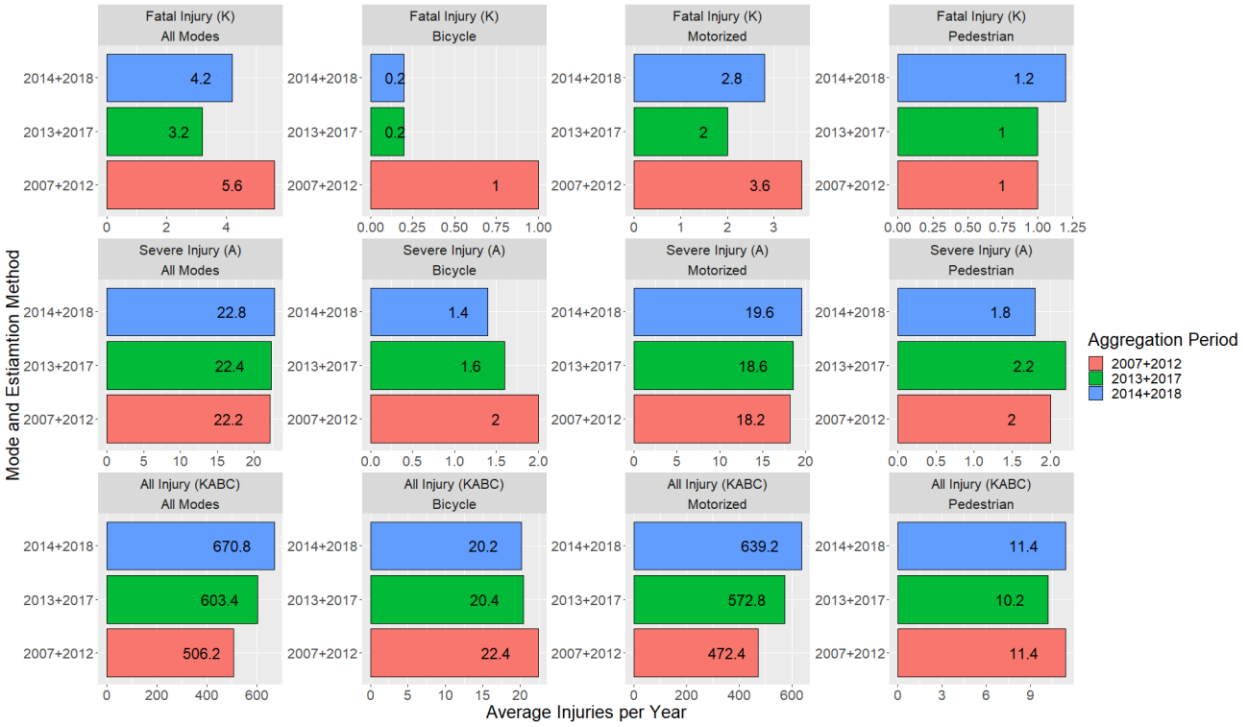

Figure 11.2: Average annual injuries by mode, and aggregation period

Figure 11.2 shows that annual average bicycle and pedestrian injury counts are relatively stable for most severity categories. There are 0.2 fatal bicycle injuries per year in the latter study periods, lower than the reference aggregation period. Average annual pedestrian fatal injuries are consistent in each aggregation period with around one of these injuries on average each year. Average annual severe injuries are also consistent from year to year with around 1.5 bicycle incidents per year and two pedestrian severe injuries per year for the 2014+2018 and 2013+2017 aggregation periods, similar to the reference period. A similar story is true for all injuries where on average for the 2014+2018 and 2013+2017 aggregation periods there are about 20 bicycle injuries and 10 pedestrian injuries which is similar to the reference aggregation period. Though it's true the bicycle and pedestrian injuries have been relatively consistent each year the total motorized injuries have increased in the latter two aggregation periods compared to the reference period while the fatal injuries are slightly down and severe injuries exhibiting little change across aggregation periods.

For additional review of these traffic injury data Table 11.2 is presented below along with information about the activity period in which the aggregation periods will be used for crash rate estimation in Chapter Five. Since 2018 is the latest available data but bicycle and pedestrian activity were estimated using counts data from 2019, there is imperfect alignment in these data. However, since bicycle and pedestrian injury counts are relatively consistent from year to year and five year averages are being used in the rate calculation, this approach should accurately represent the injury conditions during the activity estimation periods.

Table 11.2: Average Annual Injuries by Mode, Year, and Aggregation Period

Mode	Injury Severity	Average Injury Count	Standard Deviation	Crash Years	Associated Activity Estimation Period
All Modes	Fatal Injury (K)	3.2	1.8	2013-2017	NA
	Severe Injury (A)	22.4	5.9		
	All Injury (KABC)	603.4	158.4		
Motorized	Fatal Injury (K)	2	0.7		2017
	Severe Injury (A)	18.6	5.4		
	All Injury (KABC)	572.8	156.2		
Bicycle	Fatal Injury (K)	0.2	0.4		2017+2018
	Severe Injury (A)	1.6	0.5		
	All Injury (KABC)	20.4	3.8		
Pedestrian	Fatal Injury (K)	1	1.2		
	Severe Injury (A)	2.2	1.1		
	All Injury (KABC)	10.2	2.2		
All Modes	Fatal Injury (K)	4.2	1.6	2014-2018	NA
	Severe Injury (A)	22.8	5.8		
	All Injury (KABC)	670.8	105.7		
Motorized	Fatal Injury (K)	2.8	1.3		2018
	Severe Injury (A)	19.6	5.0		
	All Injury (KABC)	639.2	103.9		
Bicycle	Fatal Injury (K)	0.2	0.4		2017+2019
	Severe Injury (A)	1.4	0.5		
	All Injury (KABC)	20.2	3.9		
Pedestrian	Fatal Injury (K)	1.2	1.1		
	Severe Injury (A)	1.8	1.1		
	All Injury (KABC)	11.4	2.9		

These average annual injuries counts are further disaggregated by functional classification and featured below in Figure 11.3. Because of the low fatal injury counts for bicycle and pedestrian injuries, further disaggregation by functional classification reveals some facilities have zero average annual injuries for these modes. Annual average bicycle and pedestrian severe injuries are generally higher on minor and major arterials with similar trends for all injuries where the annual average injury count is four to five times higher than local and collector streets. There is a similarly low number of fatal motorized injuries which makes the disaggregation by functional classification produce low numbers though major arterials have consistently higher average annual fatal injuries. Severe injuries for motorized users also are higher for major and minor arterials with similar results shown for all injuries.

Higher injury counts for motorized users would be expected on arterials considering these facilities move more vehicle and higher speeds. It's also not surprising that these facility types

are locations with higher nonmotorized injury counts considering the literature review has documented these facility types typically presenting higher risk for nonmotorized users.

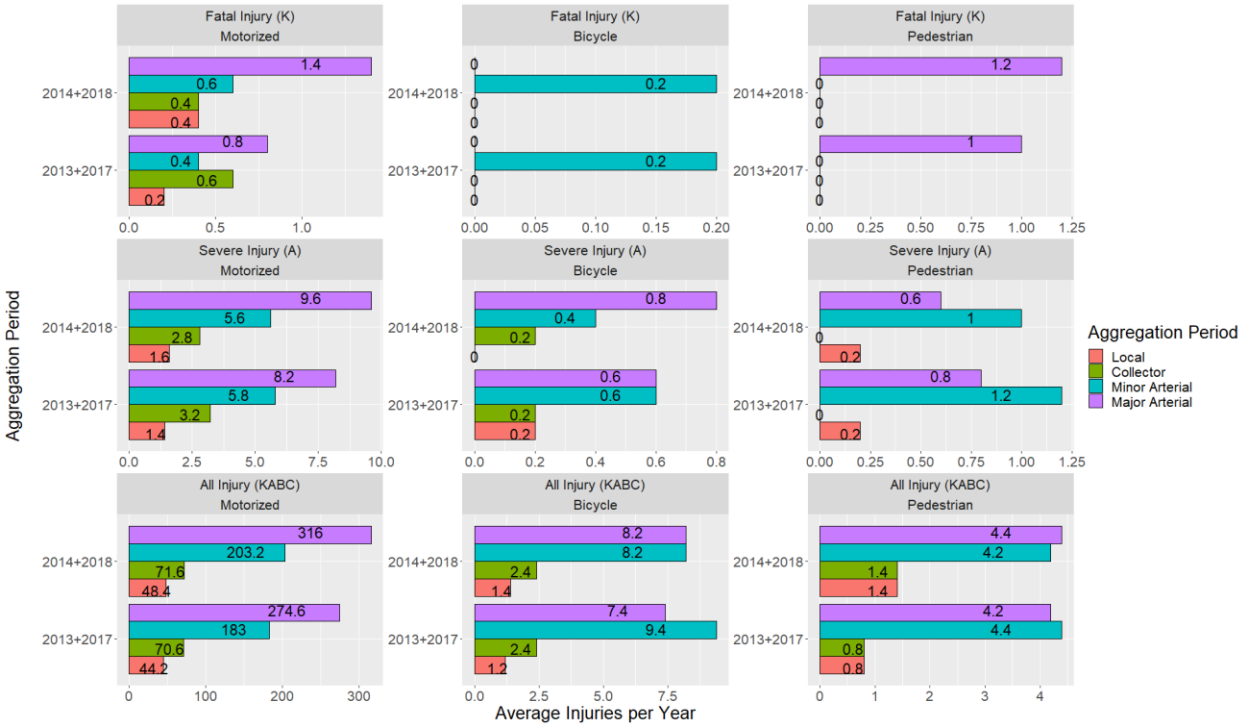

Figure 11.3: Average annual injuries by mode, aggregation period and functional classification

This section summarized the traffic injuries in the study area comparing two aggregation periods with a reference period in order to show stability in injuries across periods. Showing stability is important to prevent any perception of cherry picking injury data that is not representative of longer term conditions. Based on this review the motorized crash injuries in the 2013+2017 and 2014+2018 periods are higher than the reference period while the nonmotorized crash injuries are slightly down.

12.0 AGGREGATE CRASH RATE ANALYSIS

This section computes and reports injury rates for bicycle, pedestrian and motorized transport in the study area. These rates will be calculated using the bicycle and pedestrian activity estimates from the earlier modeling section using the equation below:

$$\boldsymbol{Injury\ Rate_{severity}} = \frac{\boldsymbol{Injury\ Count_{\ severity}}}{\boldsymbol{Miles\ of\ Travel_{mode}}} \quad \textbf{(12-1)}$$

Where:

$Injury\ Rate_{severity}$ = Annual average injury rate for severity category K, A, B, and C

$Injury\ Count_{severity}$ = Annual average injury count for severity category K, A, B, and/or C

$Miles\ of\ Travel_{mode}$ = Annual miles of travel for each mode including motorized, bicycle, and pedestrian

These rates are expressed in injuries per 100 million miles of travel, a common standard when reporting (NHTSA 2018).

12.1 REGIONAL TRAFFIC INJURY RATES

This section will summarize the traffic injuries for each mode of travel where crash data and estimates of travel activity are available including motorized, bicycle and pedestrian traffic. Motorized traffic estimates are derived from the Highway Monitoring Performance System (HPMS) while the bicycle and pedestrian traffic are derived from two modeling approaches including a machine learning algorithm and a regression approach. In addition to the two modeling approaches, the bicycle and pedestrian travel activity estimates also have a scenario in which zero counts data were injected into the observed data at sites in low density, low connectivity areas of the study region in an attempt to moderate the overall modeling estimates. This was done because it is likely that many places in these parts of the network do not have bicycle or pedestrian traffic but because the structure of the counts program nonmotorized traffic are not collected in these areas and so no zero traffic observations are actually collected. Rates are calculated using estimates from each approach and scenario to see how rates vary and measure certainty in the injury disparity between modes.

Figure 12.1 below summarizes the injury crash rates for each injury severity, for each mode, modeling approach and zero counts inclusion scenario for just the 2017+2018 estimation period. The two modeling approaches produce estimates of nonmotorized travel that differ enough to impact the injury crash rates but do not typically change the outcome that nonmotorized injury

rates are significantly higher than motorized injury rates. Only in the pedestrian All Injury rate does the rate switch from being higher than the motorized crash rate to lower. In all other results nonmotorized injury rates are higher than motorized injury rates. The scenarios defined as *No Zero Counts* are lower than results for scenarios defined as *Simulated Zero Counts* because the former scenario has higher overall estimates of nonmotorized travel activity. These would likely represent an over estimate of bicycle and pedestrian activity and so with a larger denominator the crash rates decrease relative to other scenarios. The rates from these scenarios should be considered conservative and the true rate is probably somewhere in the middle between the two scenarios. However, for the purposes of the remaining results the *No Zero Counts* scenario will be used since the objective of this chapter is to demonstrate the disparity in injury risk between travel modes so using the most conservative estimate of nonmotorized traffic hopefully reduces uncertainty in the final conclusions regarding disparity in crash risk between modes.

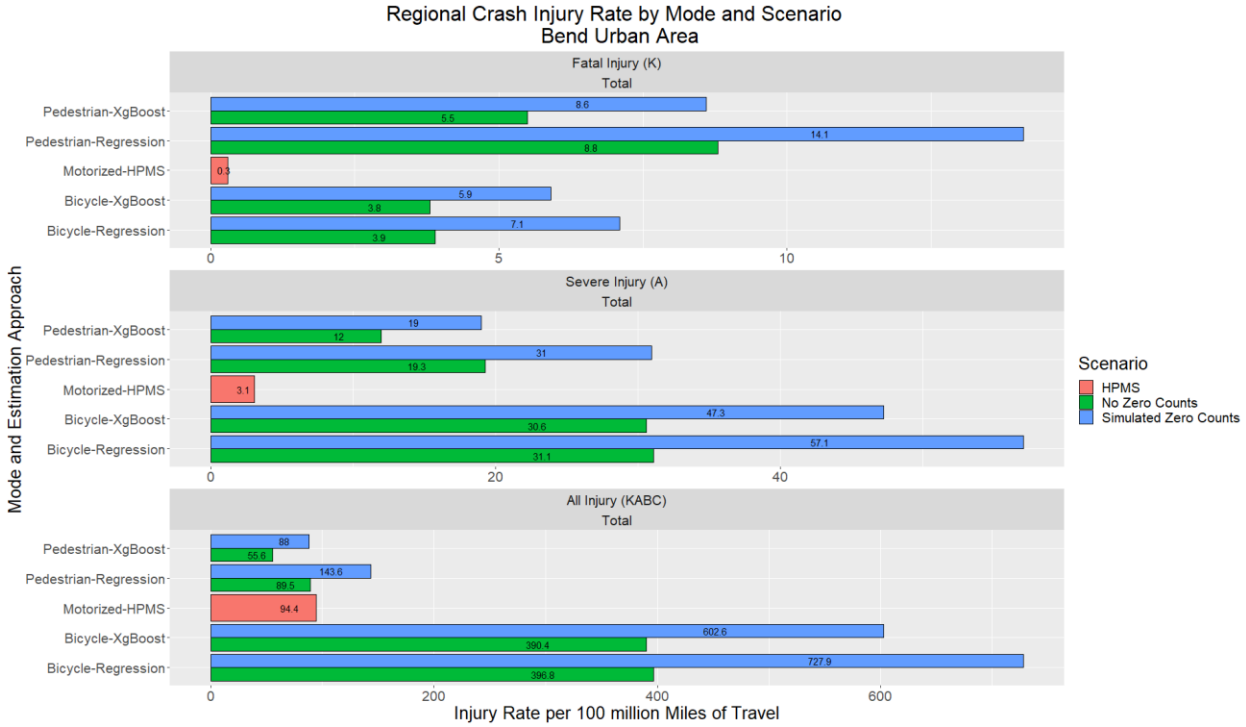

Figure 12.1: Regional crash injury rate by mode and scenario (2017+2018 estimation period)

The chart in Figure 12.2 shows the *No Zero Count* scenario still showing crash rates calculated using both modeling approaches XgBoost and Regression) for estimating bicycle and pedestrian travel activity but includes both the 2017+2018 and 2018+2019 estimation periods to give a sense of the stability in the rates from period to period. Because estimates of nonmotorized travel decrease from the 2017+2018 period to the 2018+2019 period for most of the modeling approaches the rates are generally higher in the latter period. Rates for both periods are generally many times higher for the nonmotorized users compared to the motorized users. For instance, in the 2017+2018 period the bicycle fatal injury rate is about 9 times higher (0.4 for MV compared to 3.8 for bicycle) than the motorized injury rate while the pedestrian fatal injury rate is 12 times higher (0.45 for MV compared to 5.5 for pedestrian). For severe injury rates

bicycle users face risk 10 times higher (3.1 for MV compared to 30.6 for bicycle) than motorized users while the pedestrian severe injury rate is about 4 times greater (3.1 for MV and 12 for pedestrian) than motorized crash risk. The total injury rate for bicycle users is about 4 times higher (101.9 for MV compared to 390.3 for bicycle) while the pedestrian total injury risk is lower than the motorized total injury rate by about 38 percent (89.5 for MV compared to 55.6 for pedestrian. The described crash rates between nonmotorized users and motorized users are the most conservatively derived rates and should be considered a floor for rate comparison but rates may actually be higher for nonmotorized users and thus, disparities greater.

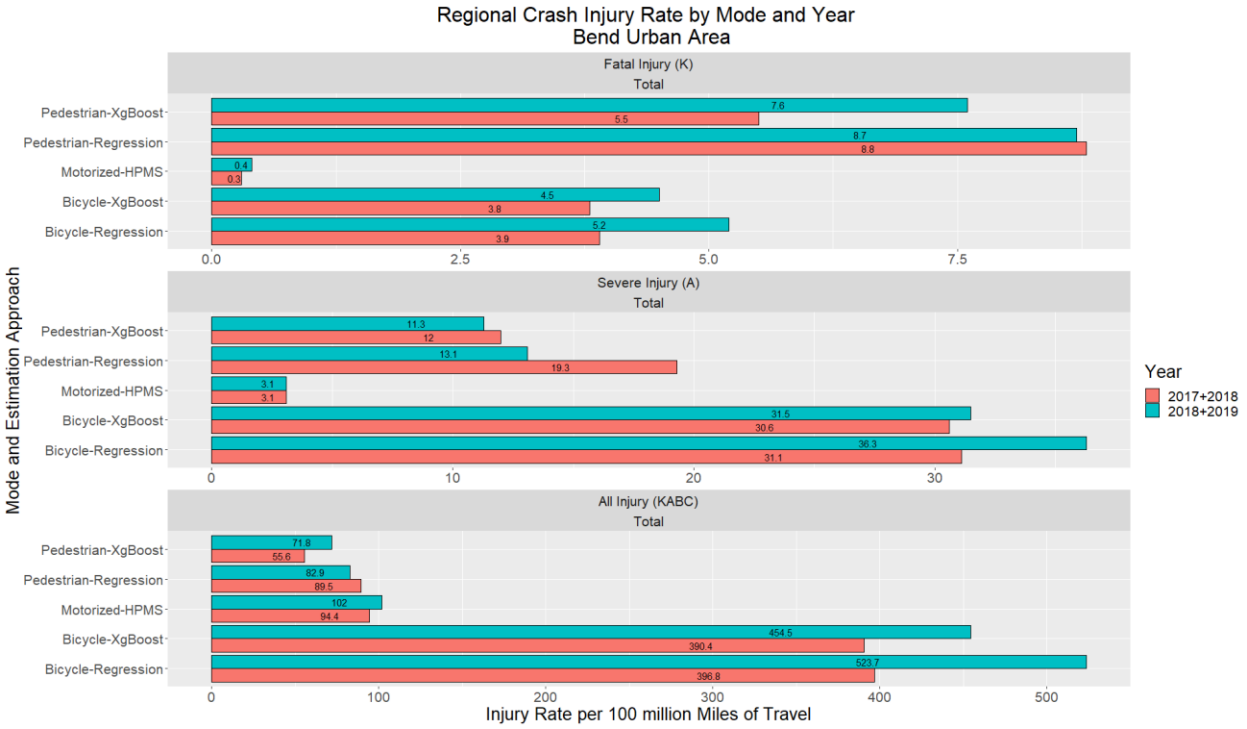

Figure 12.2: Regional crash injury rate by mode and year

Since the rates using the machine learning (XgBoost) modeling approach are generally lower than the rates using activity estimates from the regression approach these results will be used below to highlight the nonmotorized risk by functional classification. Comparisons will be made for each functional classification and injury severity for each mode. In Figure 12.3, fatal and severe injuries have been combined to simplify the number of panels shown and to reduce problems of small injury counts when disaggregated by functional classification.

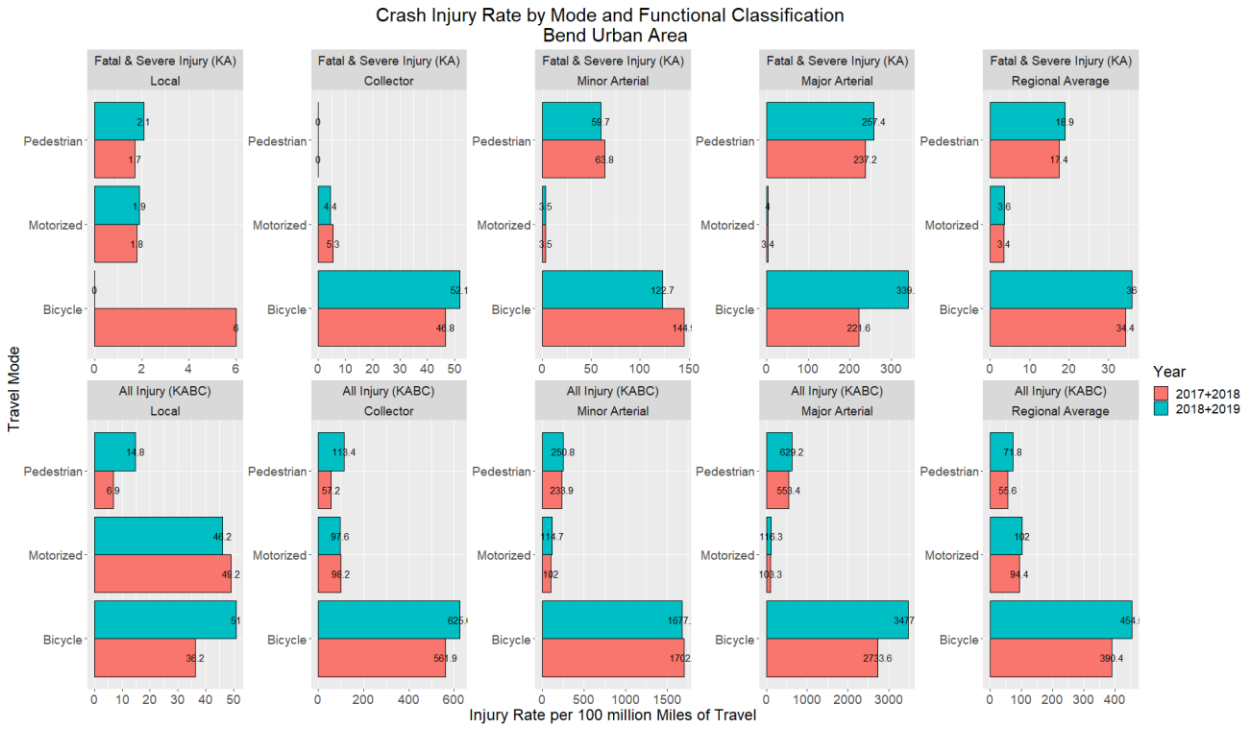

Figure 12.3: Crash injury rate by mode and functional classification

The above figure shows that risk increases for all users when vehicle traffic and speed increase with higher functional classification, i.e. local streets have lower injury rates compared to collectors streets, which in turn are lower than the two arterial classifications. Injury rates are relatively stable from one estimation period to another, especially for motorized injury rates where no change is detected on some functional classifications. Significant disparities between nonmotorized and motorized injury rates exist for almost all functional classifications with the worst disparities existing on arterials for bicycle users. Disparities on these facilities for fatal and severe injuries are 35 to 99 times higher for bicycle users compared to motorized users. For pedestrians the fatal and severe injury rate disparity is 17 to 75 times higher compared to motorized users. On some functional classifications the nonmotorized injury rates are lower including on local streets for both bicycle and pedestrian users and on collectors for just pedestrian users.

The last chart featured in Figure 12.4 shows similar information to Figure 12.3 but now only shows the nonmotorized crash injury rates in order to highlight the disparate risk across functional classification. Local and collector streets has much lower risk than arterial roads. In the case of all injury (KABC) rates for bicycle users, major arterials present about five times greater risk compared to collector streets at least 68 times greater risk compared to local streets. For pedestrian users, the total injury rates are also at least five times higher on major arterials compared to collectors and at least 42 times higher compared to local streets. If injury rates from less conservative estimates of nonmotorized travel activity were used these disparities would be even larger.

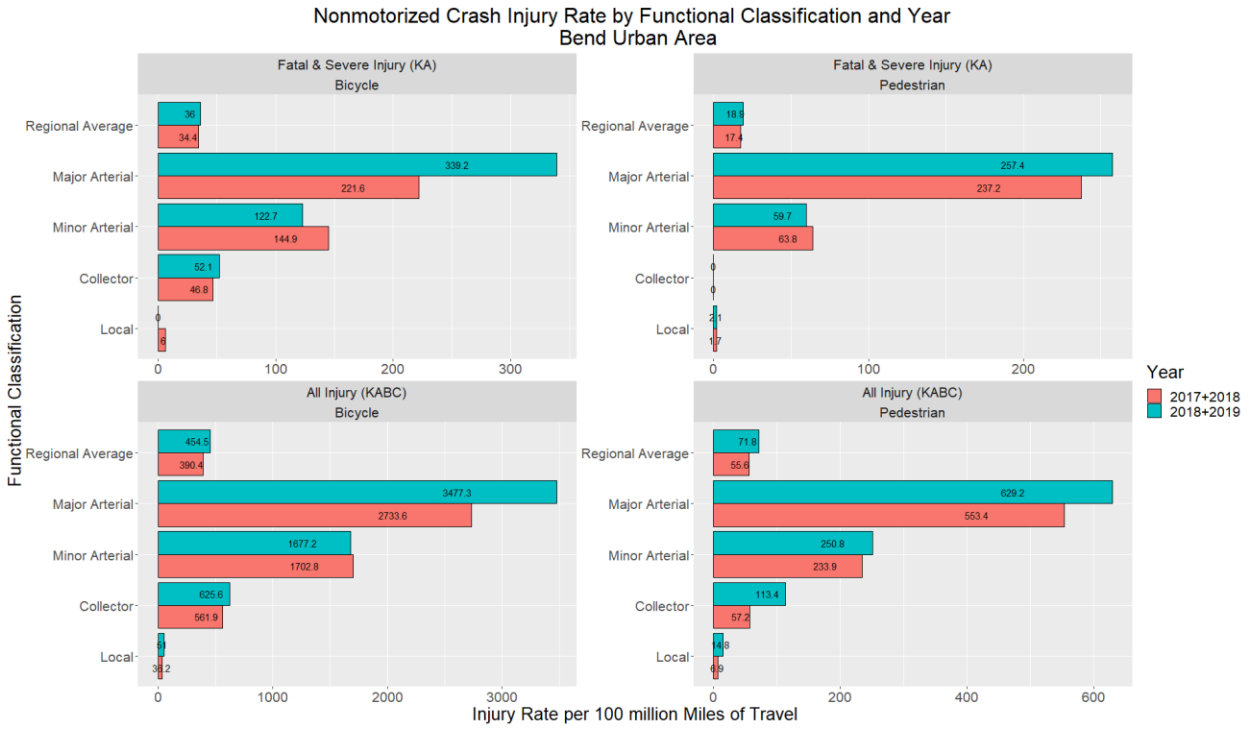

Figure 12.4: Nonmotorized traffic injury rate by mode and scenario

12.1.1 Regional Traffic Injury Rates Discussion

This section summarized crash injury rates for motorized and nonmotorized users using available miles of travel data from official HPMS sources and a novel approach to estimating nonmotorized activity. The results confirm past research that has demonstrated disparities in crash injury risk between user types, with bicycles facing significantly more risk for all injury severities compared to motorized users. Pedestrian crash injury risk is higher for fatal and severe injuries but about the same when compared to total crash injury rates for motorized users. Crash risk is relatively stable across estimation periods. Nonmotorized crash injury rates are not homogenous across the system and increase as the functional classification changes from streets with lower vehicle volumes and lower travel speeds. The next section will perform statistical modeling using the network wide estimates of nonmotorized traffic activity to determine what other factors are associated with increased crash injury risk for nonmotorized users.

13.0 CRASH MODELING

This section will use the available estimates of nonmotorized activity at the network level to perform relatively simple crash injury modeling to further understand factors associated with injuries and injury risk. Statistical crash models are recommended when data is available at the link level which the activity modeling performed in the above chapters provides for these purposes. Even though this research has measures of activity across the network and a fully attributed network for certain data elements, the number of observed injuries is so low that model stability is an issue, especially for the pedestrian crash models. Because of this only one segment model for both bicycle and pedestrian injury is presented, representing the best attempt to apply statistical methods to modeling the crash injuries. Guidance is issued in the discussion section about how to improve the functionality and confidence in these models.

13.1 BICYCLE CRASH MODELING

Crash data for the bicycle crash modeling includes all crash injuries recorded between 2014 and 2018 which occurred on the on-street network where estimates of bicycle activity are available within the Bend MPO study area. The nonmotorized activity represents average values for both estimation periods as since the crash data represents multiple years it was thought that a general representation of the bicycle activity across years was sensible. In Table 13.1 below the bicycle injury counts are summarized by functional classification and presence of bicycle lane.

Table 13.1: Bicycle Injuries by Functional Classification 2014-2018

Functional Classification	Bicycle Injury Count		
	Bike Lane	No Bike Lane	Total
Local	1	28	29
Collector	6	8	14
Minor Arterial	26	4	30
Major Arterial	27	0	27
Total	60	40	100

Figure 13.1 below shows the spatial distribution of the bicycle injuries. From the map of injury locations it's observable that many injuries occur on or near arterials that transect the study area showing spatially what the table shows in tabular format.

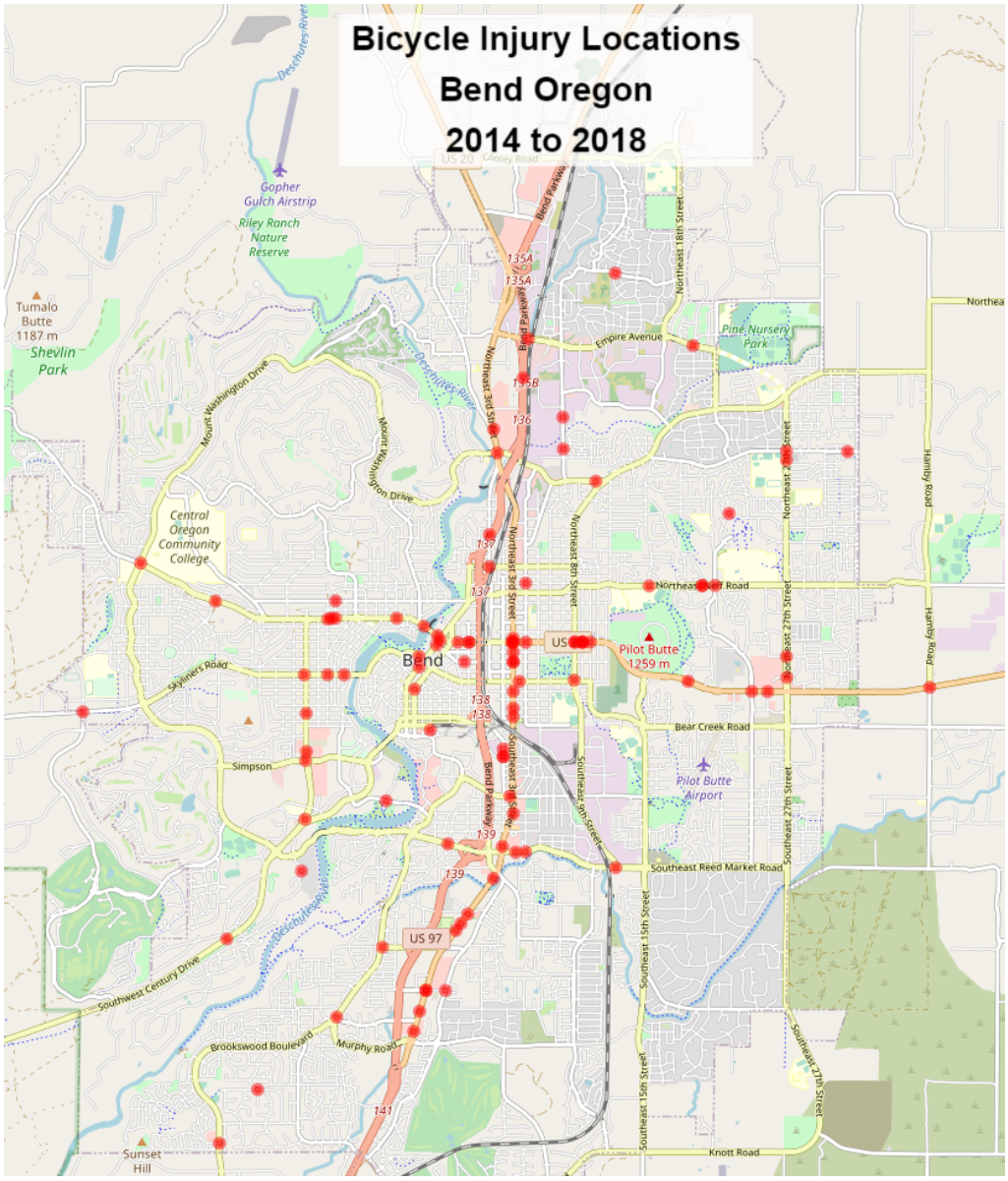

Figure 13.1: Bicycle injury locations for years 2014 to 2018

These data are combined into a statistical model to better understand the role that functional classification and presence of bicycle lane play in predicting bicycle crash injury while controlling for the effect of differences in bicycle traffic. These statistical models are commonly referred to as Safety Performance Functions (SPFs) and typically use a negative binomial regression model specification because the crash data distribution feature over dispersion, a condition when the variance exceeds the mean (HSM 2010). However, for the bicycle crash model data over dispersion is not detected, likely because the entire network is being used and the vast majority of link segments have not experienced a bicycle injury crash within the analysis timeframe resulting in excessive number of zeros. This condition requires the use of a hurdle regression model or zero-inflated regression model which combines a truncated Poisson model with a logit model. The Poisson element of the hurdle model estimates the non-zero values

while the logit model estimates the zero values. The probability distribution can be described as follows:

$$Pr(yi = j) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & \text{if } j = 0 \\ (1-\pi_i)\frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!} & \text{if } j > 0 \end{cases}$$

(12-2)

Where:

y_i is the number of injuries

π_i is the logistic link function for predicting the links with zero injuries

A few model specifications in the zero inflated regression model using covariates such as posted speed, estimated vehicle volume, presence of bicycle facility, and functional classification to predict bicycle injuries. Other treatment measures would be ideal but the kinds of treatments currently installed in the study area is limited and other treatments are not currently tracked in the available network data.

The table below summarizes a zero inflated regression model where the non-zero bicycle injury segments are a function of bicycle AADT, roadway length, and the presence of a bicycle lane while the zero bicycle injury segments are a function of functional classification. These results show that as bicycle AADT increases bicycle injury counts, as would be expected. Length and lack of bike lane is also positively correlated with bike injuries though the presence of bike lane variable is not significant at the 0.10 level.

Table 13.2: Bicycle Injury Model – Zero-Inflated Regression

Parameter	Regression Based Bike AADT Estimates			XgBoost Based Bike Based AADT Estimates		
	Estimate	**Standard Error**	**P-value**	**Estimate**	**Standard Error**	**P-value**
(Intercept)	-4.7455103	0.709924	0.00	-5.16503	0.7377129	0.00
log(Bike AADT)	0.7861149	0.1610009	0.00	0.89128	0.1675135	0.00
Roadway Length (ft.)	0.0008817	0.0002889	0.00	0.000756	0.0002986	0.01
No Bicycle Facility (Reference Bike Lane)	0.366269	0.3431637	0.29	0.090931	0.3371497	0.79
Zero-inflation model coefficients (binomial with logit link):						
	Estimate	**Standard Error**	**P-value**	**Estimate**	**Standard Error**	**P-value**
(Intercept)	3.9774	0.4013	0.00	3.7072	0.3957	< 2e-16
Collector Street (Refernce Local)	-1.4515	0.4191	0.00	-1.0656	0.4224	0.01
Arterial Street	-2.2725	0.4193	0.00	-2.0065	0.4033	0.00

Parameter result tables can often be better interpreted when applied through a sensitivity test holding some covariates constant while altering others to see how the model responds with new data. Figure 13.2 below shows the results of applying the model in a sensitivity test holding segment length constant but varying bicycle AADT and functional classification. Also, instead of showing predicted crash count the chart below shows the crash rate or risk to show how the bicycle injury risk decreases across the varying covariates. For instance the sensitivity test below shows bicycle injury crash risk is mitigated by the presence of a bicycle lane by 23%, or that the presence of a bicycle lane reduces crash risk by 22 percent, all else being equal. Functional classification, a proxy measure for vehicle speed and volume, increase bicycle crash injury risk as functional classification increases. For instance, local streets present 90% less injury risk compared to arterials, with collectors decreasing risk by 69 percent. These findings align with the injury rates presented in the above chapter. Additionally, this model and the accompanying sensitive test demonstrates that the safety in numbers effect is at work in the study region. Crash risk decreases as the number of daily bicycle riders on a given corridor increases. For instance, as the daily average (AADT) bicycle traffic increases from a 25 AADT to 100 AADT, the crash risk decreases by 21 percent. Increasing bicycle AADT from 10 AADT to the maximum observed bicycle AADT of 390 bicycle injury crash risk decreases by 38 percent.

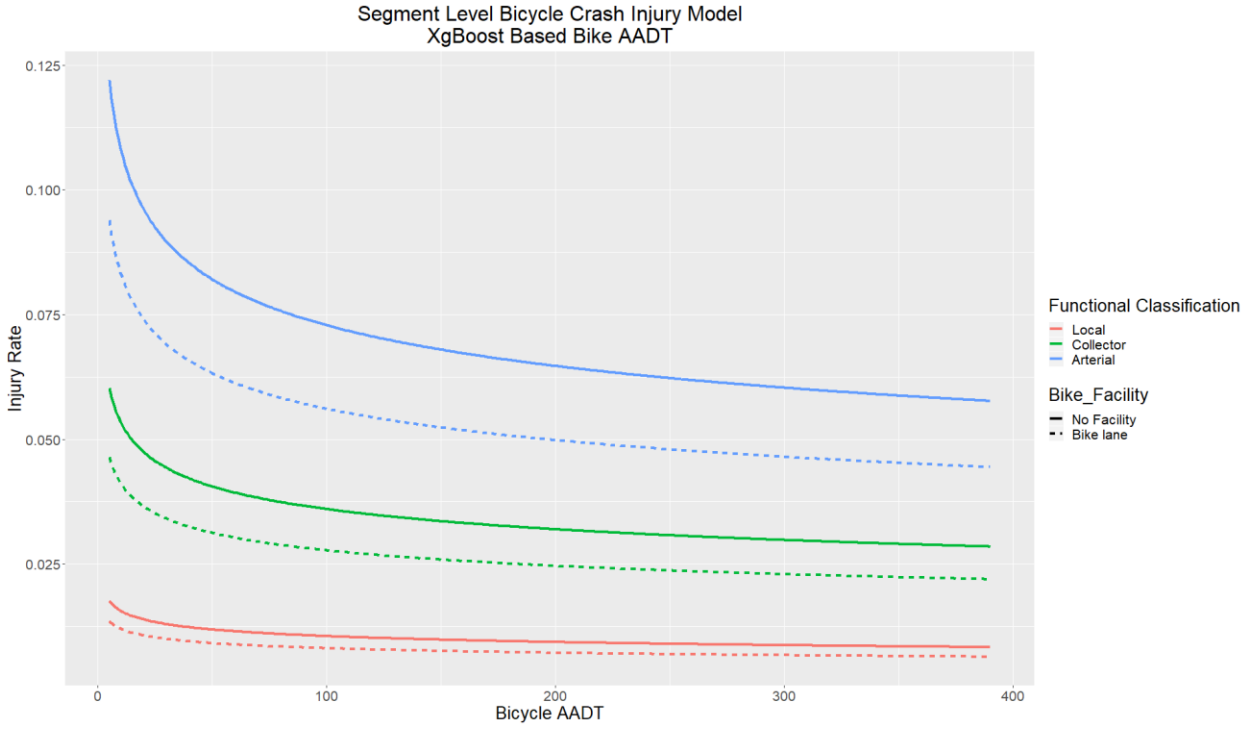

Figure 13.2: Bicycle injury crash model sensitivity test for segments

13.1.1 Bicycle Crash Modeling Discussion

The above section estimates a bicycle injury crash model using segment level analysis in the Bend study area. Though the bicycle crash injury data is sparse, the model affirms the aggregate crash risk results presented in an earlier chapter, showing that crash risk for bicycle users is higher on arterial streets compared to collector streets. The model tested the effect of the

presence of a bike lane, showing that the presence of bike lanes decreased bicycle injury risk, though the variable in the zero inflated regression model was not statistically significant.

13.2 PEDESTRIAN CRASH MODELING

Crash data for the pedestrian injury modeling includes all crash injuries recorded between 2014 and 2018 which occurred on the on-street network where estimates of bicycle activity are available within the Bend MPO study area. The nonmotorized activity represents average values for both estimation periods since the crash data represents multiple years it was though that a general representation of the pedestrian activity across years was sensible. In Table 13.3 below the pedestrian injury counts are summarized by functional. Information

Table 13.3: Pedestrian Injuries by Functional Classification 2014-2018

Functional Classification	Pedestrian Injury Count
Collector	7
Local	7
Major Arterial	22
Minor Arterial	21
Total	57

Figure 13.1 below shows the spatial distribution of the pedestrian injuries. From the map of injury locations it's observable that many injuries occur on or near arterials that transect the study area showing spatially what the table shows in tabular format. As presented in the table, a majority of pedestrian injuries occur on arterials. These facilities typically have higher vehicle speeds and volume.

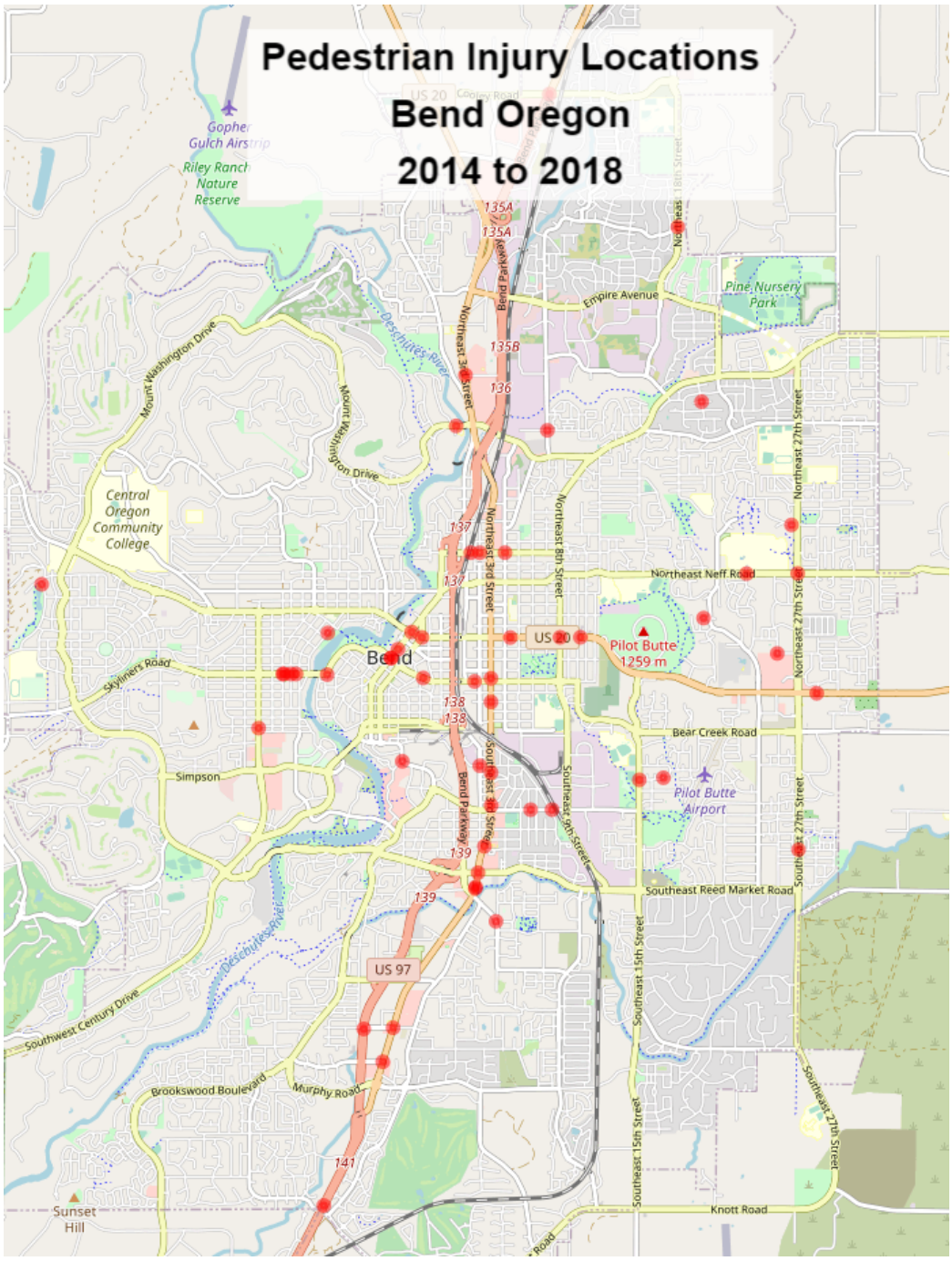

Figure 13.3: Bicycle injury locations for years 2014 to 2018

Because the distribution of the pedestrian injuries follows a different negative binomial form an alternative model specification was used to estimate the pedestrian injury model. The results of the negative binomial regression model for estimating pedestrian injuries in below in Table 13.4. The results show that there is a significant positive relationship between pedestrian injuries and pedestrian traffic volumes, as would be expected. The other covariate in the model includes vehicle volume, which is also correlated with an increase in pedestrian injuries. Both of these

variables are significant at the 0.05 level. Other variables such as segment distance were tried but models were unstable. The low number of pedestrian injury counts makes estimating pedestrian injury models using only a single urban area difficult.

Table 13.4: Pedestrian Injury Model – Zero-Inflated Regression

Parameter	Regression Based Pedestrian AADT Estimates			XgBoost Based Pedestrian Based AADT Estimates		
	Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
(Intercept)	-9.41777	0.65253	< 2e-16	-8.53464	0.51308	< 2e-16
log(Ped_AADT_Rg)	0.97517	0.17693	0.00	0.89242	0.14380	0.00
Estimated Vehicle AADT	0.000092	0.00003	0.00	0.000102	0.00002	0.00

Parameter result tables can often be better interpreted when applied through a sensitivity test holding some covariates constant while altering others to see how the model responds with new data. Figure 13.4Figure 13.2 below shows the results of applying the model in a sensitivity test holding segment length constant but varying pedestrian AADT and functional classification. Also, instead of showing predicted crash count the chart below shows the crash rate or risk to show how the pedestrian injury risk decreases across the varying covariates. For instance the sensitivity test below shows pedestrian injury crash decreases as pedestrian volume increases.

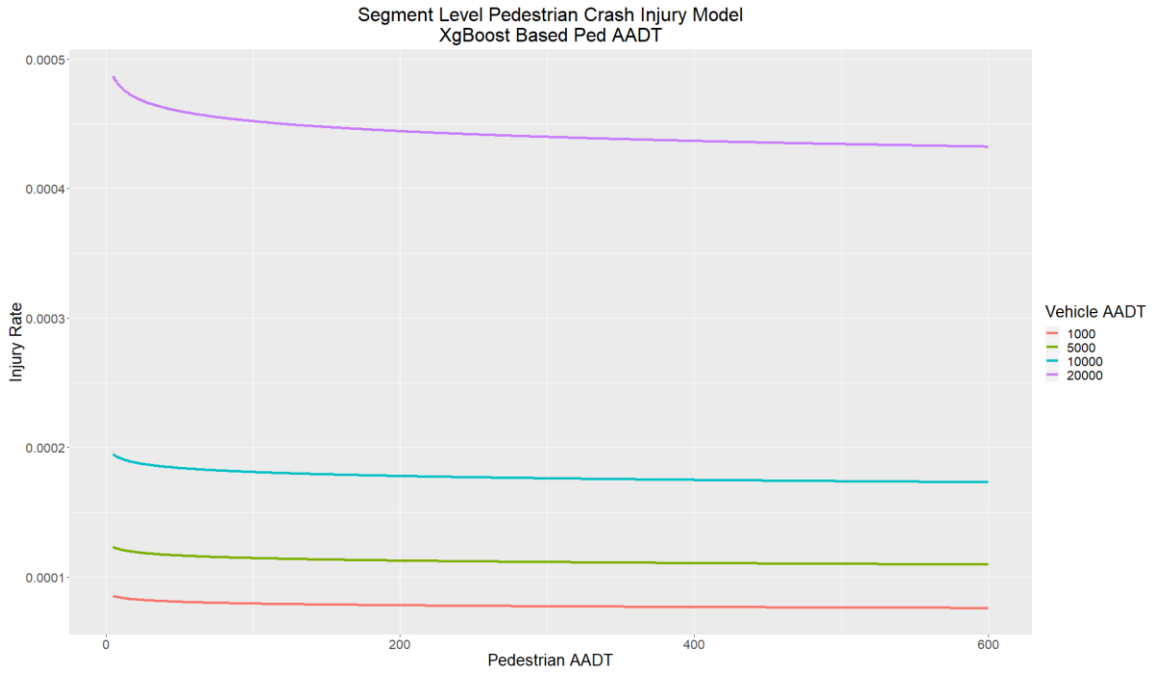

Figure 13.4: Pedestrian injury crash model sensitivity test for segments

More specifically, if pedestrian volume increases from 25 to 100 pedestrians per day the risk decreases by nearly 9% while going from 25 to the maximum observed 800 pedestrian AADT the injury risk drops by 35 percent, all else being equal. The figure shows how increased vehicle

volumes increase pedestrian crash risk. When the vehicle volume increases from 1000 AADT to 20,000 AADT the pedestrian crash risk increases by 83 percent, all else being equal.

13.2.1 Pedestrian Crash Modeling Discussion

The above section constructs a simplistic pedestrian crash injury model using estimates of pedestrian volume from an earlier chapter and available pedestrian crash injury data, which is quite sparse. Using a small set of covariates, the model highlighting the role that vehicle volume and pedestrian volume play in predicting crash risk demonstrating that vehicle volume increases pedestrian crash injury risk and increasing pedestrian traffic decreases pedestrian crash risk. Important covariates like pedestrian safety treatments were not tested because the network data used in this analysis was not attributed with those features. Future work should develop those data so that the impacts on crash risk can be captured in the models. However, because of small sample size, it's likely that the current study area would need to be combined with data from other regions in order to make the models more reliable.

13.3 CRASH MODELING DISCUSSION

The chapter above on crash modeling demonstrates that simplistic crash injury models for nonmotorized users can be developed and generate useful insights on local conditions. Though many of the risk factors used in the model like vehicle volume have been documented elsewhere, confirming the existence of and magnitude of the effect is important to help align local decision makers understanding of the role vehicle traffic play in nonmotorized injury outcomes.

13.4 DISCUSSION

Using estimates of bicycle and pedestrian activity form direct demand models is becoming a more common approach to quantifying traffic volumes for nonmotorized users across the system. At the aggregate level such as functional classification these estimates of activity likely reduce error compared to the link level estimates. With more traffic count data and some additional tuning of the direct demand models these estimates could be further refined but currently provide useful information in crash risk analysis.

The results presented in this research demonstrate what past research has shown, the major crash injury risk disparities exist currently on the system in the study area. The risk is not homogenous across the system and streets with higher vehicle traffic volumes and speed significantly increase crash risk for nonmotorized users. Risk appears to be further mitigated by design, with the bicycle crash model demonstrating risk reduction on facilities with bicycle lanes. Further, just having the presence of more nonmotorized users can reduce risk, likely utilizing the safety in number effect. Though limited in scope, this research show that for states and cities interested in getting more people to use the nonmotorized to bike and walk, interventions exist that will make people feel safer and deliver objective reduction in risk.

This research was not able to examine other treatment types but research summarized in DiGioia et al. (2017) found many common treatments exist that reduce risk for users.

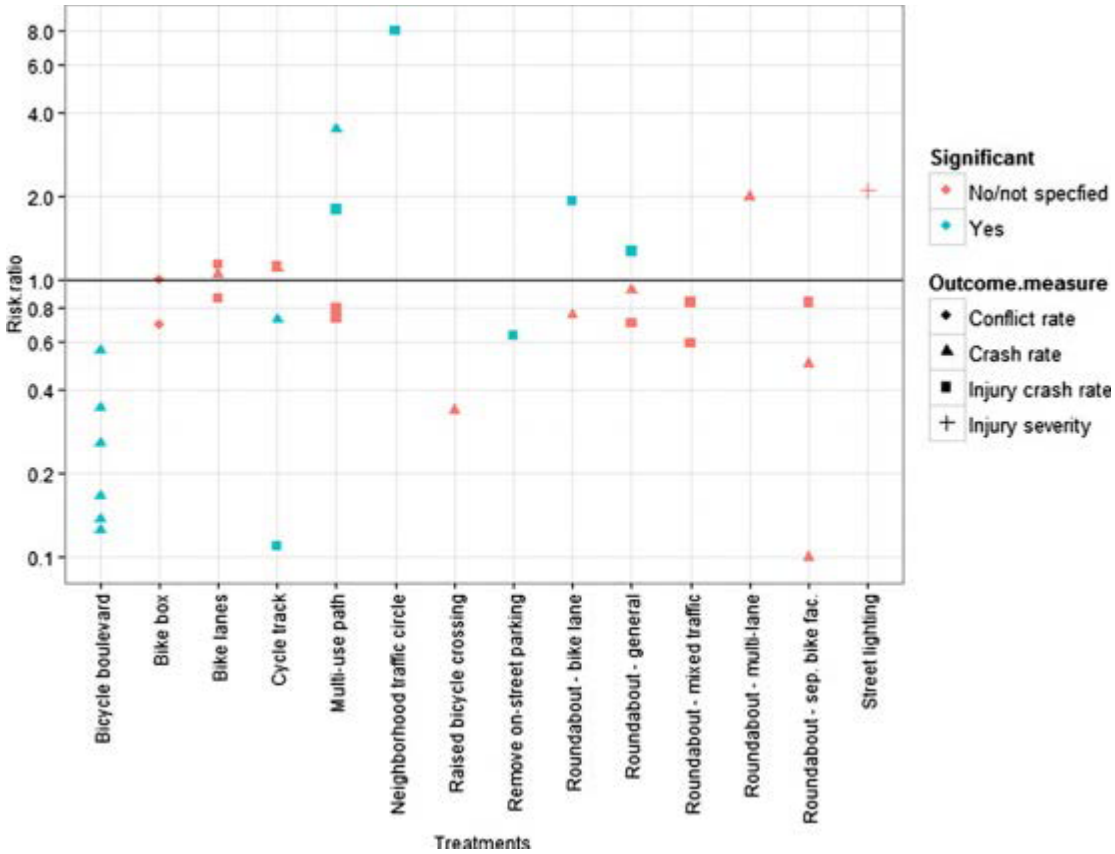

Figure 13.5: Summary of risk ratios for bicycle infrastructure treatments from DiGioia et al. (2017)

For pedestrian crash injury many treatments have been documented to reduce crash frequency and overall risk. The National Cooperative Highway Research Program released a guide to performing systemic pedestrian crash analysis and compiled treatment options and associated crash reduction factors ((NCHRP 2018). Treatments can range from low cost and easy to install to relatively more complicated interventions but these treatments have been demonstrated to reduce pedestrian injury crash outcomes including risk.

Taken together it should not be concluded that nonmotorized injury outcomes are inevitable and somehow poor behavior alone is responsible for these injury outcomes. Injuries for nonmotorists, like all crash injuries, are preventable and urban areas that lack facilities with evidence based treatment options will likely struggle to attract more users due to the existence of high risk conditions. The Oregon Department of Transportation formerly recognizes the risk many of the system elements analyzed in this research present to nonmotorized users. Bergh et al. (2015) reviewed the process developed for the agency using Oregon data to establish risk factors for people that walk and bicycle concluding that in addition to vehicle volumes and posted speed limit, the presence of traffic signals, number of lanes, lack of a bicycle facility and driveway density also increased risk for these users. The authors also note that presence of mid-block pedestrian crossing and transit stops *increased* crash risk for pedestrians highlighting the need for nonmotorized exposure data since it's likely the safety issue with these factors is actually the presence of more nonmotorized users, not the features themselves.

In addition to nonmotorized crash injuries being avoidable through design, active transportation benefits to population health outweigh the costs associated with injuries and air pollution. These benefits have been well documented In Mueller et al. (2015) where 30 studies looking at the health impact of shifting driving trips to walking and bicycle trips finding that in 27 of those studies the benefits of increased physical activity outweighed the increased risks of traffic safety and air pollution exposure. In a another study of over 250,000 people in the United Kingdom, researchers followed participants for up to five years and found that people who bicycled or bicycled and walked to work had lower risks of cardiovascular disease and cancer (Celis-Morales et al. 2017). Under current conditions, these health benefits however, will never be fully realized when potential nonmotorized road users experience the elevated crash risk documented in this and other studies. System managers at the state and local level have a significant role to play to build a complete system that allow users the freedom of movement by for the mode they choose. This freedom is currently limited by the outsized risk nonmotorized users' face on Bend's roadways.

14.0 REFERENCES

Al-Deek, H. M., Venkata, C., & Chandra, S. R. (2004). New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record: Journal of the Transportation Research Board, 1867*(1), 116-126. doi:10.3141/1867-14

ASTM International. (2018). Standard Practice for Highway Traffic Monitoring Truth-in-Data, ASTM International (Standard No. E2759-10). West Conshohocken, PA. Retrieved from www.astm.org

Albright, D. (1990). *1990 survey of traffic monitoring practices among state transportation agencies of the United States* (Publication No. FHWA/HPR/NM-90-05). Santa Fe, NM: New Mexico State Highway and Transportation Department.

Albright, D. (1991). History of estimating and evaluating annual traffic volume statistics. *Transportation Research Board,* (1305), 103-107.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning, 36*(1), 105-139. doi:10.1023/A:1007515423169

Beitel, D., Mcnee, S., Mclaughlin, F., & Miranda-Moreno, L. F. (2018). Automated validation and interpolation of long-duration bicycle counting data. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(43), 75-86. doi:10.1177/0361198118783123

Beck, L. F., Dellinger, A. M., & O'Neil, M. E. (2007). Motor Vehicle Crash Injury Rates by Mode Of Travel, United States: Using Exposure-Based Methods To Quantify Differences. *American Journal of Epidemiology, 166*(2), 212-218. doi:10.1093/aje/kwm064

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. doi:10.1007/bf00058655

Breiman, L. (1998). Arcing Classifiers. *Annals of Statistics, 26*(3), 801-849. doi:10.1214/aos/1024691079

Broach, J., Gliebe, J., & Dill, J. (2009). Development of a Multi-class Bicyclist Route Choice Model Using Revealed Preference Data. In *12th International Conference on Travel Behavior Research*. Portland, OR: Portland State University.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics, 30*(4), 927-961. doi:10.1214/aos/1031689014

Castro-Neto, M., Jeong, Y., Jeong, M. K., & Han, L. D. (2009). AADT prediction using support vector regression with data-dependent parameters. *Expert Systems with Applications, 36*(2), 2979-2986. doi:10.1016/j.eswa.2008.01.073

Celis-Morales, C. A., Lyall, D. M., Welsh, P., Anderson, J., Steell, L., Guo, Y., . . . Gill, J. M. (2017). Association between active commuting and incident cardiovascular disease, cancer, and mortality: Prospective cohort study. *Bmj*. doi:10.1136/bmj.j1456

Chamberlain, S. (2019). 'NOAA' Weather Data from R [R package rnoaa version 1.2.0]. Retrieved from https://CRAN.R-project.org/package=rnoaa

Chatfield, C. (1989). *The analysis of time series: An introduction* (4th ed.). London: Chapman & Hall/CRC.

Csardi G, & Nepusz T (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695. Retrieved from https://igraph.org.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random fores. *BMC Bioinformatics, 7*(3). doi:10.1186/1471-2105-7-3

Dietterich, T. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning, 40*(2), 139-157. doi:10.1023/A:1007607513941

Digioia, J., Watkins, K. E., Xu, Y., Rodgers, M., & Guensler, R. (2017). Safety impacts of bicycle infrastructure: A critical review. *Journal of Safety Research, 61*, 105-119. doi:10.1016/j.jsr.2017.02.015

Ermagun, A., Lindsey, G., & Loh, T. H. (2018). Bicycle, pedestrian, and mixed-mode trail traffic: A performance assessment of demand models. *Landscape and Urban Planning, 177*, 92-102. doi:10.1016/j.landurbplan.2018.05.006

Esawey, M. E. (2014). Estimation of Annual Average Daily Bicycle Traffic with Adjustment Factors. *Transportation Research Record: Journal of the Transportation Research Board, 2443*(1), 106-114. doi:10.3141/2443-12

Esawey, M.E., (2018a) Impact of data gaps on the accuracy of annual and monthly average daily bicycle volume calculation at permanent count stations, *Computers, Environment and Urban Systems,* Volume 70, Pages 125-137, ISSN 0198-9715, doi: 10.1016/j.compenvurbsys.2018.03.002.

Fagnant, D. J., & Kockelman, K. (2015). A direct-demand model for bicycle counts: The impacts of level of service and other factors. *Environment and Planning B: Planning and Design*, *43*(1), 93–107. https://doi.org/10.1177/0265813515602568

Federal Highway Administration. (2013). *Traffic Monitoring Guide* (FHWA PL-13-015). Federal Highway Administration, US Department of Transportation, Washington, D.C. Retrieved from https://www.fhwa.dot.gov/policyinformation/tmguide/tmg_fhwa_pl_13_015.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

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: Journal of the Transportation Research Board*, *2673*(4), 941–950. https://doi.org/10.1177/0361198119837235

Heidema, A. G., Boer, J. M. A., Nagelkerke, N., Mariman, E. C. M., van der A, D. L., & Feskens, E. J. M. (2006). The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, *7*(1), 23. https://doi.org/10.1186/1471-2156-7-23

Kittelson & Associates. (2016).*Technical memorandum 5.3: Count program development*. [Memorandum] Bend Transportation Planning Strategy. https://www.bendoregon.gov/home/showdocument?id=28903

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2020). CRAN - package caret. Retrieved from https://cran.r-project.org/web/packages/caret/index.html

Langley, J. D., Dow, N., Stephenson, S., & Kypri, K. (2003). Missing cyclists. *Injury Prevention*, *9*(4), 376–379. https://doi.org/10.1136/ip.9.4.376

Lewin, A. (2011). Temporal and weather impacts on bicycle volumes. In *Transportation Research Board 90th Annual Meeting*. Washington, D.C.: Transportation Research Board.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importance in forests of randomized trees. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 1). Red Hook, NY: Curran Associates.

Lu, L., & Zhang, M. (2013). Edge Betweenness Centrality. In *Encyclopedia of Systems Biology*. New York, NY: Springer. doi:10.1007/978-1-4419-9863-7_874

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2013). Cluster: Cluster analysis basics and extensions (Version 1.14.4) [R package].

McAndrews, C. (2011). Traffic risks by travel mode in the metropolitan regions of Stockholm and San Francisco: A comparison of safety indicators. *Injury Prevention*, *17*(3), 204–207. https://doi.org/10.1136/ip.2010.029306

McAndrews, C., Beyer, K., Guse, C. E., & Layde, P. (2013). Revisiting exposure: Fatal and non-fatal traffic injury risk across different populations of travelers in Wisconsin, 2001–2009. *Accident Analysis & Prevention*, *60*, 103–112. https://doi.org/10.1016/j.aap.2013.08.005

Mindell, J. S., Leslie, D., & Wardlaw, M. (2012). Exposure-Based, 'Like-for-Like' assessment of road safety by travel mode using routine health data. *PLoS ONE*, *7*(12), e50606. https://doi.org/10.1371/journal.pone.0050606

Minge, E., Falero, C., Lindsey, G., and Petesch, M.. (2015). *Bicycle and pedestrian data collection manual* (MN/RC 2015-33). Minnesota Department of Transportation. Retrieved from https://www.dot.state.mn.us/research/TS/2015/201533.pdf

Minge, E., Falero, C., Lindsey, G., Petesch, M., & Vorvick, T. (2017). *Bicycle and pedestrian data collection manual* (MN/RC 2017-03). Minnesota Department of Transportation. Retrieved from https://www.dot.state.mn.us/research/reports/2017/201703.pdf

Miranda-Moreno, L. F., & Nosal, T. (2011). Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. *Transportation Research Record: Journal of the Transportation Research Board*, *2247*(1), 42–52. https://doi.org/10.3141/2247-06

Mohamad, D., Sinha, K. C., Kuczek, T., & Scholer, C. F. (1998). Annual average daily traffic prediction model for county roads. *Transportation Research Record: Journal of the Transportation Research Board, 1617*(1), 69-77. doi:10.3141/1617-10

Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., de Nazelle, A., Dons, E., Gerike, R., … Nieuwenhuijsen, M. (2015). Health impact assessment of active transportation: A systematic review. *Preventive Medicine*, *76*, 103–114. https://doi.org/10.1016/j.ypmed.2015.04.010

Munro, C. (2013). *Evaluation of automatic cyclist counters* (0030). CDM Research. Retrieved from http://bicyclecouncil.com.au/files/research/EvaluationOfAutomatic CyclistCounters.pdf

National Highway Traffic Safety Administration (2019). *2018 Fatal Motor Vehicle Crashes: Overview* (Publication No. DOT HS 812 826). Washington, D.C. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812826

Ni, D., Leonard, J. D., Guin, A., & Feng, C. (2005). Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *Journal of Transportation Engineering*, *131*(12), 931–938. https://doi.org/10.1061/(asce)0733-947x(2005)131:12(931)

Nordback, K., Piatkowski, D., Janson, B., Marshall, W., Krizek, K., & Main, D. (2011). Using inductive loops to count bicycles in mixed traffic. *Journal of Transportation of the Institute of Transportation Engineers*, *2*(1), 39–56. Retrieved from http://www.ite.org/jot/

Nosal, T., & Miranda-Moreno, L. (2012). Cycling and weather: A multi-city and multi-facility study in North America. In *91st Annual Meeting of the Transportation Research Board*. Washington, D.C.: Transportation Research Board of the National Academics.

Proulx, F., & Pozdnukhov, A. (2017). *Bicycle traffic volume estimation using geographically weighted data fusion*.

Pucher, J., & Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: Lessons from the netherlands and germany. *American Journal of Public Health*, *93*(9), 1509–1516. https://doi.org/10.2105/ajph.93.9.1509

Redfern, E., Watson, S., Tight, M., & Clark, S. (1993). A comparative assessment of current and new techniques for detecting outliers and estimating missing values in transport related time series data. In *21st PTRC Summer Annual Meeting* (Vol. P, Ser. 366, pp. 163-174). Manchester, United Kingdom.

Roll, J. (2018). *Bicycle count data: What is it good for? A study of bicycle travel activity in central lane metropolitan planning organization* (Report No. FHWA-OR-RD-18-16). Oregon Department of Transportation. Retrieved from https://www.oregon.gov/ODOT/Programs/ResearchDocuments/304-761%20Bicycle%20Counts%20Travel%20Safety%20Health.pdf

Roll, J., & Proulx, F. R. (2018). Estimating annual average daily bicycle traffic without permanent counter stations. *Transportation Research Record Journal of the Transportation Research Board*, *2672*(43), 145–153. https://doi.org/10.1177/0361198118798243

Rose, G., Ahmed, F., Figliozzi, M., & Jakob, C. (2011). Quantifying and comparing effects of weather on bicycle demand in Melbourne, Australia, and Portland, Oregon. In *Transportation Research Board 90th Annual Meeting*. Washington, D.C.: Transportation Research Board.

Ryus, P., Ferguson, E., Laustsen, K., Schneider, R., Proulx, F., Hull, T., & Miranda-Moreno, L. (2014). *Guidebook on pedestrian and bicycle volume data collection* (Report No. 797). Washington, D.C.: National Academies of Sciences, Engineering, and Medicine. https://doi.org/10.17226/22223

Sharma, S., Lingras, P., & Zhong, M. (2004). Effect of missing values estimations on traffic parameters. *Transportation Planning and Technology, 27*(2), 119-144. doi:10.1080/0308106042000218203

Shinar, D., Valero-Mora, P., Strijp-Houtenbos, M. V., Haworth, N., Schramm, A., Bruyne, G. D., . . . Tzamalouka, G. (2018). Under-reporting bicycle accidents to police in the COST TU1101 international survey: Cross-country comparisons and associated factors. *Accident Analysis & Prevention, 110*, 177-186. doi:10.1016/j.aap.2017.09.018

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(307). https://doi.org/10.1186/1471-2105-9-307

Tang, Y. F., Lam, W. H., & Ng, P. L. (2003). Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong. *Journal of Transportation Engineering, 129*(3), 271-277. doi:10.1061/(asce)0733-947x(2003)129:3(271)

Teschke, K., Harris, M. A., Reynolds, C. C., Shen, H., Cripton, P. A., & Winters, M. (2013). Exposure-based traffic crash injury rates by mode of travel in British Columbia. *Canadian Journal of Public Health, 104*(1). doi:10.1007/bf03405659

Thomas, L., Lan, B., Sanders, R. L., Frackelton, A., Gardner, S., & Hintze, M. (2017). Changing the Future?: Development and Application of Pedestrian Safety Performance Functions to Prioritize Locations in Seattle, Washington. *Transportation Research Record: Journal of the Transportation Research Board, 2659*(1), 212-223. doi:10.3141/2659-23

Tin Tin, S., Woodward, A., Robinson, E., & Ameratunga, S. (2012). Temporal, seasonal and weather effects on cycle volume: An ecological study. *Environmental Health, 11*(1). doi:10.1186/1476-069x-11-12

Turner, S., & Park, E. S. (2008). Incomplete Archived Data of Intelligent Transportation Systems for Calculation of Annual Average Traffic Statistics. *Transportation Research Record: Journal of the Transportation Research Board, 2049*(1), 1-13. doi:10.3141/2049-01

Winters, M., & Branion-Calles, M. (2017). Cycling safety: Quantifying the under reporting of cycling incidents in Vancouver, British Columbia. *Journal of Transport & Health, 7*, 48-53. doi:10.1016/j.jth.2017.02.010

Xia, Q., Zhao, F., Chen, Z., Shen, L. D., & Ospina, D. (1999). Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County. *Transportation Research Record: Journal of the Transportation Research Board, 1660*(1), 32-40. doi:10.3141/1660-05

Zhao, F., & Park, N. (2004). Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic. *Transportation Research Record: Journal of Transportation Research Board, 1879*(1), 99-107. doi:10.3141/1879-12

Zhong, M., Sharma, S., & Liu, Z. (2005). Assessing Robustness of Imputation Models Based on Data from Different Jurisdictions: Examples of Alberta and Saskatchewan, Canada. *Transportation Research Record: Journal of Transportation Research Board, 1917*(1), 116-126. doi:10.1177/0361198105191700114

APPENDIX A:DATA DICTIONARY

Table A.1: Deployment Information

Field Name	Data Format	Description
Location_Id	numeric	Unique Identifier for count locations locations (parent 'site')
Sub_Location_Id	numeric	Unique identifer describing the sub location of the count
Deployment_Date	character	Deployment date of device
Deployment_Time	time hms	Deployment time of device
Equipment_Id	character	Serial number for hardware device
Collection_Type	character	Data collection type (e.g. Roadway, sidewalk, Combo, etc.)
Photo_Url	character	A URL link for photo(s) of the deployment
Device_Name	character	Serial Number of device. Used to link with automatically transmitted data)
Pickup_Date	character	Date when device was picked up
Pickup_Time	time hms	Time when device was picked up
Comments	character	
Email	character	
Edit_Url	character	
Time Stamp	character	Time stamp of when the record was created
Description	character	Description of device deployment

Latest version of the code links directly to Google Sheets bi-passing the need to export the file and store it on a network.

See Diagram for Data Collection graphic

Table A.2: Count Location Information

Field Name	Data Format	Description
Location_Id	integer	Unique Identifier for count locations locations (parent 'site')
Sub_Location_Id	character	Unique identifier for sites at location, e.g. sidewalk, bike lane, roadway. See Diagram tab
Vendor_Site_Id	character	Identifier assigned by device vendor. Used to link spatial data location to counts data. Only used for permanent locations.
Collection_Type	character	Data collection method, mobile vs. permanent (see Collection_Type_Code)
Collection_Type_Desc	character	Data collection method, mobile vs. permanent (see Collection_Type_Code)
Facility_Type	numeric	Value indicating the facility in which data collection occurred (See Facility_Type_Code)
Facility_Type_Desc	character	Descriptive indicating the facility in which data collection occurred (See Facility_Type_Code)
Double_Count_Location	numeric	Value indicating if two devices required to collect at site. For mobile collection only
Is_Oneway	numeric	Value indicating if site is a one-way travel direction
Oneway_Direction	character	Direction in which oneway travel is directed
Latitude	numeric	Latitude of site
Longitude	numeric	Longitude of site
Site_Name	character	Name of Sub_Location
Location_Description	character	Unique name for location
Visualize	numeric	Flag for data visualization. Applies to parent site only
User_Type	character	User type collected (See User_Type code tab)
User_Type_Desc	character	User type collected (See User_Type code tab)
Device_Type	character	Equipment type setup (see Device_Type tab)
Direction	numeric	Code to establish travel direction of traffic
Post_Needed	numeric	Indication of whether a post is needed to hang collection device
Install_Instructions	character	Information for vendor regarding how to install collection devices
Vendor_Channel_Id	numeric	Unique value for permanent sites that link spatial data to counts data. Only Used for permanent locations.
ImageFilePath	character	Relative file path for a picture of the count site
User_Updated	character	Initials of GIS user updating the record last
Street Furniture	character	Description of street furniture, temporary post, or tree used to anchor the hardware

Table A.3: Processed Count Data

Field Name	Data Format	Description
Location_Id	character	Unique Identifier for count locations (parent 'site')
Sub_Location_Id	character	Unique identifier for sites at location, e.g. sidewalk, bike lane, roadway. See Diagram tab
Date	date	Date when count was recorded
Direction	character	Direction of travel for counts
User_Type_Desc	character	User type collected (See User_Type code tab)
Facility_Type	numeric	Code indicating the facility in which data collection occurred (See Facility_Type_Code)
Counts	numeric	Traffic count
Obs_Hours	numeric	Number of hours of collected data
Weekday	character	Day of Week
Is_Weekday	character	Descriptive of weekday vs. weekend
Month	character	Month of Year
Year	numeric	Calendar Year
Device_Type_Desc	character	Description of device type collecting data. Only available and Sub_Location_Id level since Location_Id and Link_Id level aggregates Sub_Location data of whiohc multiple devices types may have been used
Is_Holiday	character	TRUE or FALSE value depending on whether date falls on the following federal US holidays(New Years Day, Inauguration Day,ML Kings Birthday,Memorial Day,Independence Day,Labor Day Veterans Day,Thanksgiving Day,ChristmasDay)
Potential_Special_Event	logical	TRUE or FALSE value based on grouping analysis of all counts sites where clusters of higher than expected counts by day are used to inform potnetial days where special events took place that may increase the traffic volumes
Error_Code	numeric	Value assigned to daily counts that indicates any error in that record. See Error_Codes tab for full description.
Ub_Conf_Bound	numeric	Upper level threshold value assigned to record to perform error flagging and understand if daily observation is within an acceptable range
Lb_Conf_Bound	numeric	Lower level threshold value assigned to record to perform error flagging and understand if daily observation is within an acceptable range
Est_Split	logical	TRUE or FALSE value based on whether the count was estimated from user only counts using an assumed split factor
Index	numeric	Index value is used to keep track of records in the error flagging process and do not persist across data process

APPENDIX B

Table B.1: Hyper parater Summary

Mode	Algorithm Specification	Algorithm	Year	Split Variable Count	Boosting Rounds	Maximum Depth	Learning Rate	Gamma	Subsample Ratio of Columns	Minimum Child Weight	Subsample Ratio
Vehicle	Federal Fc	Random Forest	2017	6	NA	NA	NA	NA	NA	NA	NA
	Federal Fc	Random Forest	2018	6	NA	NA	NA	NA	NA	NA	NA
	Local Fc	Random Forest	2017	6	NA	NA	NA	NA	NA	NA	NA
	Local Fc	Random Forest	2018	6	NA	NA	NA	NA	NA	NA	NA
	Federal Fc	XgBoost	2017	NA	100	8	0.075	0	0.5	2.5	1
	Federal Fc	XgBoost	2018	NA	75	6	0.075	0	0.5	2	1
	Local Fc	XgBoost	2017	NA	100	8	0.075	0	0.5	2	1
	Local Fc	XgBoost	2018	NA	75	6	0.1	0	0.5	2	1
Bicycle	Spec	Algorithm	2017+2018	2	NA	NA	NA	NA	NA	NA	NA
	All + Strava	RandomForest	2018+2019	2	NA	NA	NA	NA	NA	NA	NA
	All + Strava	RandomForest	2017+2018	2	NA	NA	NA	NA	NA	NA	NA
	All	RandomForest	2018+2019	2	NA	NA	NA	NA	NA	NA	NA
	All	RandomForest	2017+2018	NA	50	7	0.05	0	0.3	2.25	1
	All + Strava	XgbTree	2018+2019	NA	50	8	0.05	0	0.4	2	1
	All + Strava	XgbTree	2017+2018	NA	50	7	0.05	0	0.3	2.5	1
	All	XgbTree	2018+2019	NA	50	6	0.05	0	0.3	2	1
Pedestrian	All + Strava	RandomForest	2017+2018	2	NA	NA	NA	NA	NA	NA	NA
	All + Strava	RandomForest	2018+2019	2	NA	NA	NA	NA	NA	NA	NA
	All	RandomForest	2017+2018	2	NA	NA	NA	NA	NA	NA	NA
	All	RandomForest	2018+2019	2	NA	NA	NA	NA	NA	NA	NA
	All + Strava	XgbTree	2017+2018	NA	50	6	0.05	0	0.5	2	1
	All + Strava	XgbTree	2018+2019	NA	50	7	0.075	0	0.4	2	1
	All	XgbTree	2017+2018	NA	50	7	0.05	0	0.3	2	1
	All	XgbTree	2018+2019	NA	50	8	0.05	0	0.3	2	1