3-2022

Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes

Sirisha Kothuri
Portland State University, skothuri@pdx.edu

Joseph Broach
Portland State University

Nathan McNeil
Portland State University, nmcneil@pdx.edu

Kate Hyun
University of Texas at Arlington

Stephen Mattingly
University of Texas at Arlington

See next page for additional authors

Recommended Citation

Sirisha Kothuri, Joseph Broach, Nathan McNeil, Kate Hyun, Stephen Mattingly, Md. Mintu Miah, Krista Nordback, and Frank Proulx. Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes. NITC-RR-1269. Portland, OR: Transportation Research and Education Center (TREC), 2022 https://doi.org/10.15760/trec.273

Authors

Sirisha Kothuri, Joseph Broach, Nathan McNeil, Kate Hyun, Stephen Mattingly, Md. Mintu Miah, Krista Nordback, and Frank Proulx

NITC
NATIONAL INSTITUTE for TRANSPORTATION and COMMUNITIES

Final Report 1269
March 2022

Photo by Lacey Friedly

Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes

Sirisha Kothuri, Ph.D.
Joseph Broach, Ph.D.
Nathan McNeil, MURP
Kate Hyun, Ph.D.
Stephen Mattingly, Ph.D.
Md. Mintu Miah
Krista Nordback, Ph.D.
Frank Proulx, Ph.D.

Portland State UNIVERSITY
UNIVERSITY OF TEXAS ARLINGTON

EXPLORING DATA FUSION TECHNIQUES TO ESTIMATE NETWORK-WIDE BICYCLE VOLUMES

Final Report

NITC-RR-1269

by

Sirisha Kothuri
Joseph Broach
Nathan McNeil
Portland State University

Kate Hyun
Stephen Mattingly
Md. Mintu Miah
University of Texas, Arlington

Krista Nordback
Highway Safety Research Center, University of North Carolina

Frank Proulx
Frank Proulx Consulting LLC

for National Institute for Transportation and Communities (NITC)
P.O. Box 751
Portland, OR 97207

March 2022

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NITC-RR-1269		

4. Title and Subtitle	5. Report Date
Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes	March 2022
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Sirisha Kothuri 0000-0002-2952-169X; Joseph Broach 0000-0001-7753-501X; Nathan McNeil 0000-0002-0490-9794; Kate Hyun 0000-0001-7432-8058; Stephen Mattingly 0000-0001-6515-6813; Md Mintu Miah 0000-0001-6073-3896; Krista Nordback 0000-0001-7967-1998; Frank Proulx 0000-0001-5157-0300	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
National Institute for Transportation and Communities (NITC) P.O. Box 751 Portland, OR 97207	11. Contract or Grant No. USDOT Grant 69A3551747112

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590	Final Report, 1/1/2019 – 03/09/22
	14. Sponsoring Agency Code USDOT OST-R

15. Supplementary Notes

16. Abstract

This research developed a method for evaluating and integrating emerging sources (Strava, StreetLight, and Bikeshare) of bicycle activity data with conventional demand data (permanent counts, short-duration counts) using traditional (Poisson) and advanced machine learning techniques. First, a literature review was conducted, along with cataloging and evaluating available third-party data sources and existing applications. Next, six sites (Boulder, Charlotte, Dallas, Portland, Bend, and Eugene) that represented a variety of contexts (urban, suburban) and geographical diversity were selected. Of these, Boulder, Charlotte and Dallas constituted the basic sites, where one year of data (i.e., 2019) was used for modeling. Portland, Bend, and Eugene in Oregon were considered enhanced sites, where three years of data (2017-2019) were used for model estimation. Demographic, network, count and emerging data were gathered for these sites. Using these data, Poisson and Random Forest models were estimated. The model estimation process was designed to allow for comparison of the relative accuracy and value added by different data sources and modeling techniques. Three sets of models were specified – All City Pooled, Oregon Pooled and city-specific models. In general, the three data sources (static, Strava, and StreetLight) appeared to be complementary to one another; that is, adding any two data sources together tended to outperform each data source on its own. Low-volume sites proved challenging, with the best-performing models still demonstrating considerable prediction error. City-specific models generally displayed better model fit and prediction performance. Using Strava or StreetLight counts to predict annual average daily bicycle traffic (AADBT) without static adjustment variables increased expected prediction error by a factor of about 1.4 (i.e., a 40% increase in %RMSE). That rule of thumb figure of 1.4 times was only slightly lower when combining Strava plus StreetLight without static variables (1.3x). Tests of transferability showed that transferring the model specifications without reestimating the model parameters resulted in 10-50% increase in error rate across models. Performance of machine learning models was comparable to count models. The findings from this study indicate that rather than replacing conventional bike data sources and count programs, big data sources like Strava and StreetLight actually make the old "small" data even more important.

17. Key Words	18. Distribution Statement
Bicycle volume, machine learning, crowdsource data	No restrictions. Copies available from NITC: www.nitc-utc.net

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	139	

ACKNOWLEDGEMENTS

This project was funded by a Pooled Fund Grant from the National Institute for Transportation and Communities (NITC-RR-1269), a U.S. DOT University Transportation Center, and by Pooled Fund contributors including the Oregon DOT, Virginia DOT, Colorado DOT, Central Lane MPO, Portland Bureau of Transportation, District DOT, and Utah DOT.

We would also like to recognize the contributions of the Technical Advisory Committee for this project, who offered their time and expertise. They include Josh Roll, Oregon DOT, who was the chair of the committee; Peter Ohlms, Virginia DOT; In-Kyu Lim, Virginia DOT; Betsy Jacobsen, Colorado DOT; Ellen Currier, Central Lane MPO; Roger Gellar, Portland Bureau of Transportation; Jovi Anderson, Bend MPO; Stephanie Dock, Washington, D.C. DOT; Will Handfield, Washington, D.C. DOT; Ted Knowlton, Wasatch Front MPO; Bert Granberg, Wasatch Front MPO; Chad Worthen, Utah DOT/Wasatch Front MPO; and Heidi Goedhart, Utah DOT. We would also like to thank Erik Sunde, Strava Metro; Matt Petit, StreetLight Data; Steve Ferguson, StreetLight Data, and Naveen Juvva, StreetLight Data for providing access to data sources that were used in this study.

DISCLAIMER

RECOMMENDED CITATION

Sirisha Kothuri, Joseph Broach, Nathan McNeil, Kate Hyun, Stephen Mattingly, Md. Mintu Miah, Krista Nordback, and Frank Proulx. *Exploring Data Fusion Techniques to Estimate Network-Wide Bicycle Volumes.* NITC-RR-1269. Portland, OR: Transportation Research and Education Center (TREC), 2022.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

Planners and decision makers have increasingly voiced a need for network-wide estimates of bicycling activity. Such volume estimates have for decades informed motor vehicle planning and analysis, but have only recently become feasible for non-motorized travel modes. To date, the bulk of our information on bicycling activity has come from national and regional household travel surveys or observed counts of cyclists – either short-duration manual or longer-term automated counts – in a limited set of locations. Based on these datasets, models must be developed to assess network-wide conditions. Direct demand models and regional travel demand models have been developed, but in practice such models that include bicycling at a useful level of detail remain extremely rare. Recently, new sources of bicycling activity data have emerged. These derive primarily from GPS-based smartphone apps (e.g., Strava Metro Data), GPS-enabled devices which provide location (e.g., StreetLight) and GPS-enabled public bicycle sharing systems. Strava Metro Data (hereafter referred to as just Strava) is a mode-specified dataset, in which the user identifies a specific trip and mode of travel, while StreetLight is a mode-unspecified dataset, in which a user has not specified a mode and may not be aware trips are being tracked. These emerging data sources have potential advantages as a complement to traditional count data, and have even been proposed as replacements for such data since they are collected continuously and for larger portions of local bicycle networks. However, the representativeness of these new data sources has been questioned, and their suitability for producing bicycle volume estimates has yet to be rigorously explored. Evidence has been mixed on their ability to represent the full spectrum of cycling activity, and about their accuracy and reliability. To date, studies have been confined almost entirely to single areas and data sources, often with limited comparison data from reliable counts. Most have used only relatively weak assessments of validity and accuracy, such as linear correlation coefficients. Data fusion and other advanced modeling techniques have shown promise in other areas of transportation to get more value out of disparate data sources. There appears to be similar potential with bicycle activity, given the increasing range of data available.

This research developed a method for evaluating and integrating emerging sources (Strava, StreetLight, and bikeshare) of bicycle activity data with conventional demand data (permanent counts, short-duration counts) and methods using traditional (Poisson) and advanced machine learning techniques. First, a literature review was conducted, along with cataloging and evaluating available third-party data sources and existing applications. Next, six cities (Boulder, Charlotte, Dallas, Portland, Bend, and Eugene) that represented a variety of contexts (urban, suburban) and geographical diversity were selected. Of these, Boulder, Charlotte and Dallas constituted the basic sites, where one year of data (2019) was used for modeling. Portland, Bend, and Eugene in Oregon were considered enhanced sites, where three years of data (2017-2019) were

used for model estimation. At each of these locations, counts (minimum 24 hours duration), Strava, StreetLight, and bikeshare data were obtained, extracted and processed. In addition to these count and volume data, other data sources utilized in the modeling efforts include transportation network and other built environment data and sociodemographic variables (termed “static” variables in this report), and weather-related data. Using these data, Poisson and Random Forest models were estimated. The model estimation process was designed to allow for comparison of the relative accuracy and value added by different data sources and modeling techniques. The scripts developed for the data processing and model estimation will be openly available in GitHub (https://gitlab.com/joebroach/bike-data-fusion) to help others evaluate, process, and apply emerging data sources for network-wide bicycle volume estimation.

We developed a range of models to better understand the likely feasibility and accuracy of predicting bicycle counts on a network. Three sets of models were specified – All City Pooled, Oregon Pooled and city-specific models. This allowed us to benefit from a larger sample size and range of contexts while also considering the benefit of region- or city-specific models. We also considered the transferability of models across time and location, as well as the potential for more flexible machine learning modeling techniques. Across all the modeling streams, the crowdsourced data (Strava, StreetLight), bikeshare, and static location variables were added systematically to test their impacts on predicting annual average daily bicycle traffic (AADBT) . Two sets of models – count models and advanced models using machine learning were developed. All models (count and machine learning) were developed using 10-fold cross-validation with five repeats. Count model specifications were developed by drawing theoretically likely explanatory variables from the dataset, and examining estimated coefficients for expected sign and statistical significance. Three different machine learning algorithms– Classification and Regression Tree, Random Forest, and XGboost –were tested to compare their capacity of AADBT estimations. Additionally, the models’ hyperparameters were tuned using Hyper-opt TPE (Tree-structured Parzen Estimator Approach) algorithms.

We separately modeled sites meeting a standard definition of continuous, permanent counts (at least 10 months and every day of the week captured for every month). The primary motivation for separately modeling sites with 10-plus months of valid count data was to rigorously assess the contribution of each of our data sources. These sites did not require factor group matching and expansion, and served as our closest measure to ground truth cycling activity. Static variables representing the count location context were estimated as a baseline reflecting typical direct demand modeling practice. In general, the three available data sources (static, Strava, and StreetLight; bikeshare data were not available outside of the Oregon cities) appeared to be complementary to one another; that is, adding any two data sources together tended to outperform each data source on its own. In the All City Pooled model, the combination of all three variable types was clearly the best-performing model by most measures. While the All City Pooled model fits the data relatively well, as shown by the high pseudo-R^2 value (>0.8), prediction success varied considerably by volume. Low-volume sites proved challenging, with the best-performing model still demonstrating considerable prediction

error (>100% MAPE, mean absolute percent error), while higher-volume sites (150 AADBT or more) had much lower error rates of around 30% MAPE. Prediction at low-volume sites is also made more challenging by the lack of variety in count locations. Where cycling volumes are low, permanent counts tend to be conducted only at off-street locations. The Dallas-only model, unsurprisingly, displayed generally better model fit and prediction performance than the All City Pooled model using all cities' data. In Dallas, Strava and StreetLight were particularly complementary since most sites were primarily used for recreational cycling, with Strava performing better at mid- to high-volume spots, while StreetLight was a better option at low-volume locations. Keeping in mind the limited samples, these results are interesting and mostly in keeping with expectations that each source is providing unique and valuable information about bicycling activity. While the best-fitting model still included all three sets of variables (static, Strava, and StreetLight), a model combining just Strava and StreetLight data performed about as well in terms of predictive performance. In terms of MAPE, expected performance was better than in the All City Pooled model, with best MAPE less than 20% at mid- and high-volume sites, and low-volume MAPE as low as 55%.

For 2019, All City Pooled and Oregon Pooled models with the full count dataset are, for the most part, consistent with the full-year PC pooled model results: combinations of data sources outperform single sources, and the best-fitting models combine all three. An exception is the Oregon Pooled model, where adding StreetLight to Strava data does not improve the model performance. In fact, StreetLight appears to provide the least information of the three, significantly underperforming Strava data whether individually or in combination with static variables. One possibility is that additional static variables are needed to adjust StreetLight, which is unique due to its need to impute travel mode. Variables capturing different aspects of the count location context might be needed to complement StreetLight. The general patterns also hold, for the most part, in city-specific models, with a couple of specific results worth noting. In Dallas, no combination of static variables was found that improved on the combination of Strava and StreetLight data. It was interesting to note that while StreetLight on its own performed poorly at Dallas locations, it significantly improved the Strava estimates there. In Bend, the smallest community in our study (2020 city population just under 100,000) with a maximum site AADBT of 344 and mean AADBT of 78, performance was the worst among study sites by most measures, and different variable combinations had little impact on a model's predictive ability. Bikeshare data showed potential on their own, but their modeling impact faded to insignificance with the inclusion of static, Strava, StreetLight variables. With systems expanding and more detailed data increasingly available, bikeshare should continue to be considered as a predictor.

We also estimated the increase in error observed over all the 2019 pooled and city-specific models when using Strava or StreetLight data without adjustment factors (either static variables, or the other third-party user data). For example, using Strava or StreetLight counts to predict AADBT without static adjustment variables increased expected prediction error by a factor of about 1.4 (i.e., a 40% increase in %RMSE). That rule of thumb figure of 1.4 times was only slightly lower for Strava plus StreetLight without static variables (1.3x). The case of Strava alone versus Strava plus StreetLight

was the only mixed result. In some cases, combining the two third-party user data sources greatly improved results versus using Strava only (Dallas, Boulder), while for the rest the addition of StreetLight only modestly improved or even reduced performance.

Additional data from 2018 (all Oregon locations, Strava, and StreetLight) and 2017 (Portland, no StreetLight) provided additional tests of the model and data's stability and transferability over time and space. First, 2019 model specifications for Oregon Pooled and city-specific models were transferred to 2018 data, and parameters were reestimated. Second, 2019 models were used to predict 2018 data without reestimating parameters. With reestimation, the familiar pattern held, and the best-fitting models included all sets of variables, followed by models including both static contextual variables and Strava counts, and then the rest. Comparing error rates with the corresponding 2019 models resulted in performance within expectations (+/- <10 %RMSE) despite the time shift and sample change. Without reestimation, however, results changed significantly. Error rates in general were higher, and the Strava plus static models were clearly preferred in most cases over the fully specified models. On average, reusing model estimates resulted in a 10-50% increase in the error rate across models.

Machine learning algorithms have been used to understand how complex and flexible modeling forms could handle data variability and bias to provide a better prediction. This study used Random Forest regression to predict AADBT for All City Pooled, Oregon Pooled, and city-specific models (Portland and Eugene only) using various data and buffer fusion methods to evaluate the value of third-party user data for modeling bicycle activity. The results indicated that Strava and StreetLight played a supplementary role. When each model's performance was broken down by volume bin, the best RMSE was 14% and 8% for high- and medium-volume bins, respectively. The All City Pooled and Oregon Pooled models show similar medium-volume bin RMSEs, at 8-9%, while individual city-specific models show varying levels of error; for instance, Eugene and Portland show 11% and 19% of RMSE, respectively. For low volume, StreetLight alone overestimates AADBT; however, the full data fusion with static, Strava and StreetLight significantly improved the model performance. Even though the machine learning model is more computationally complex with respect to model development, the performance of the machine learning algorithm was comparable with the count models, perhaps due to limited data samples and variations within data. With more data collection capturing additional local contexts, machine learning is expected to improve model performance for a network-wide prediction. It is very difficult to identify the optimal number of sites or data collection duration that makes the machine learning model reliable and accurate since performance levels depend on the algorithm used and the complexity of data. One previous study (El Esawey, 2015) found that at least 1,950 data points are required to obtain 10%-15% MAPE from neural network-based bicycle volume estimation models.

2 INTRODUCTION

Planners and decision makers have increasingly voiced a need for network-wide estimates of bicycling activity. Such volume estimates have for decades informed planning and analysis for motorized travel modes, but have only recently become feasible for non-motorized travel modes. To date, the bulk of our information on bicycling activity has come from national and regional household travel surveys or observed counts of cyclists – either short-duration manual or longer-term automated counts – in a limited set of locations. Based on these datasets, models must be developed to assess network-wide conditions. Direct demand models and regional travel demand models have been developed, but in practice such models that include bicycling at a useful level of detail remain extremely rare. Recently, new sources of bicycling activity data have emerged. These derive primarily from GPS-based smartphone apps (e.g., Strava), GPS-enabled devices which provide location (e.g., StreetLight) and GPS-enabled public bicycle-sharing systems. These emerging data sources have potential advantages as a complement to traditional count data and have even been proposed as replacements for such data, since they are collected continuously and for larger portions of local bicycle networks. However, the representativeness of these new data sources has been questioned, and their suitability for producing bicycle volume estimates has yet to be rigorously explored.

This research developed a method for evaluating and integrating emerging sources (Strava, StreetLight, and Bikeshare) of bicycle activity data with conventional demand data (permanent counts, short-duration counts) and methods using traditional (Poisson) and advanced machine learning techniques. First, a literature review was conducted, along with cataloging and evaluating available third-party data sources and existing applications. Next, six sites (Boulder, Charlotte, Dallas, Portland, Bend, and Eugene) that represented a variety of contexts (urban, suburban) and geographical diversity were selected. Of these, Boulder, Charlotte and Dallas constituted the basic sites, where one year of data (2019) was used for modeling. Portland, Bend, and Eugene in Oregon were considered enhanced sites, where three years of data (2017-2019) were used for model estimation. Demographic, network, count and emerging data were gathered for these sites. Using these data, Poisson and Random Forest models were estimated. The model estimation process was designed to allow for comparison of the relative accuracy and value added by different data sources and modeling techniques. The scripts developed for the data processing and model estimation will be openly available in GitHub (https://gitlab.com/joebroach/bike-data-fusion) to help others evaluate, process, and apply emerging data sources for network-wide bicycle volume estimation. A description of the scripts developed as well as the knowledge and technical skills recommended to apply the models is provided in Appendix 8.3.

The remainder of the report is organized as follows. Chapter 3 contains a review of the relevant literature. A description of the data is presented in Chapter 4. Model formulation and results are described in Chapter 5, while a discussion and recommendations are presented in Chapter 6.

3 LITERATURE REVIEW

This chapter includes a detailed review of published literature to summarize the state of bicycle volume estimation using counters, emerging third-party applications, and methods that merge or fuse these approaches.

3.1 BICYCLE COUNT HISTORY, STATUS, AND MOTIVATION

As bicycling has increased in popularity nationwide in recent years, efforts to better measure riding activity, along with how to interpret available data and to assess the impacts of cycling, are developing along several tracks. Bicycle volumes (usually in the form of annual average daily bicycle traffic, AADBT) are useful for measuring trends, prioritizing infrastructure investments, and as exposure/activity measures in safety, public health and other studies (Roll, 2018; Ryus et al., 2014). To date, observed counts of cyclists at limited sets of locations have continued to provide most of our information on bicycling activity at the facility level (Romanillos et al., 2016; Ryus et al., 2014). In addition to fixed-location counts, emerging sources – such as public bikeshare systems, smartphone apps like Strava and other sources –provide potentially useful data for estimating activity on each link of a network (Romanillos et al., 2016). This project proposes to demonstrate and rigorously evaluate their value for that purpose. As a first step, this document reviews the history and current state of the practice regarding bicycle counting and volume estimation, and also provides a scan of evaluations and applications of emerging data sources.

FHWA's *Traffic Monitoring Guide* (TMG) was first published in 1985; however, that edition did not cover non-motorized count data. The 2013 update to the TMG was the first edition to provide guidance on collecting non-motorized count data, including bicycle data (FHWA, 2013), while the 2016 TMG provided updates to the format (FHWA, 2016). In the period in between the publication of the 2001 TMG and subsequent updates, participation and interest in bicycling increased substantially. The percentage of commuters who bike nationwide increased from 0.4% in 2007 to 0.6% in 2018, while those numbers in the 50 largest cities increased from 0.7% to 1.2% – an increase of around 70% (League of American Bicyclists, 2018). During this period, the facilities available increased as well. For example, between 2007 and 2010, the average reported miles of bicycle facilities per square mile increased from 1.2 to 1.6 (a 33% increase) across the largest 50 U.S. cities plus New Orleans (Steele and Altmaier, 2010).

The growing interest in bicycling and bicycle facilities prompted the need to better document and understand how people were using these facilities, whether to understand the impact of building bicycle facilities, trying to gauge potential demand for new facilities, or for other purposes. The National Bicycle & Pedestrian Documentation Project (NBPD) was launched in 2004 by Alta Planning + Design and the Institute of Transportation Engineers (ITE) Pedestrian and Bicycle Council, with the goal of creating a consistent count data collection model (O'Toole and Piper, 2016). The NBPD

encouraged agencies to collect manual short-duration bicycle and pedestrian count data following a consistent methodology and to submit it to a central repository.

Following the inclusion of non-motorized count data in the TMG, FHWA produced several resources designed to provide guidance to state and local agencies on collecting bicycle and pedestrian count data. These include the *Guidebook for Developing Pedestrian and Bicycle Performance Measures* (Semler et al., 2016); *Coding Nonmotorized Station Location Information in the 2016 Traffic Monitoring Guide Format* (Laustsen et al., 2016); *Exploring Pedestrian Counting Procedures* (Nordback et al., 2016); and the *Bicycle-Pedestrian Count Technology Pilot Project* (Baas et al., 2016). *NCHRP Report 797: Guidebook on Pedestrian and Bicycle Volume Data Collection* and its two associated web-only documents grew out of NCHRP Project 07-19 (Ryus et al., 2016, 2014).

Bicycle count data can be important for a number of planning purposes. A Pedestrian and Bicycle Information Center (PBIC) brief lists a number of such purposes, including measuring change over time, prioritizing projects, planning and designing future facilities, calibrating regional models, assessing and marketing commercial real estate, identifying and assessing the value of locations for advertising, safety performance studies, adjusting signal timing, conducting before-and-after studies, and more (Nordback et al., 2018).

3.2 DATA QUALITY

Data quality is critical to providing useful information to end users (Turner et al., 2019b). Poor data quality can limit the uses of the data. Nordback et al. (2016) suggest that different levels of data quality may be acceptable depending on the purpose for which the data are intended to be used. They suggest that for safety analysis the required data quality is high, while for sketch planning and project planning, low data quality may be acceptable (Nordback et al., 2016). To ensure good data quality, it is important to establish quality control checks for both equipment and data.

3.3 EQUIPMENT

FHWA's *Traffic Monitoring Guide* outlines eight aspects of data quality – accuracy, completeness, validity, timeliness, coverage, accessibility, data usage and formats (FHWA, 2016). Prior to the data being collected, it is necessary to ensure that the counters are properly installed, and a regular and frequent maintenance program is established (Turner et al., 2019b). It is also imperative to recognize the sources of error and minimize their impacts if they do occur. The *Traffic Monitoring Guide* provides the following guidance for data quality (FHWA, 2016):

- Ensure that the equipment is tested to meet required level of accuracy prior to installation;
- Calibrate equipment periodically to ensure that it is functioning properly;
- Validate data collected;
- Conduct routine quality assurance tests;

- Analyze data collected quickly and provide data to end users, so that any errors that may have been missed previously can be identified in a timely manner; and
- Install a feedback process to respond quickly to user feedback.

For automated counters, equipment calibration and validation are critical to ensuring good data quality. The objective of the calibration process is to ensure that the equipment is functioning correctly, and a good process can identify both major and minor errors. Validation includes testing the counter both on the installation day and several days after installation (Ryus et al., 2014). TMG provides the following guidance with respect to calibration (FHWA, 2016):

- Implement software tools that can help with automating the process;
- Perform daily checks to ensure that the data are properly collected, processed and stored;
- Determine validity of the data using monthly and yearly trends;
- Conduct field calibration;
- Validate the data from the automated counters using manual counts; and
- Perform manual and electronic calibration of data and hardware annually.

NCHRP 797 lists occlusion, environmental conditions, counter bypassing, and mixed-traffic effects as potential sources of counter inaccuracy (Ryus et al., 2014). Occlusion occurs when multiple people pass a counter and one person can obscure others from the counter's field of detection, leading to an erroneous count. Environmental conditions such as extreme hot or cold temperatures, precipitation for thermal sensors, and precipitation and lighting for optical sensors may affect accuracy; however, the report did not observe significant errors associated with these conditions with the technologies that were studied. Counter bypassing causes errors when users move outside of the detection zone. Mixed-traffic errors can occur when pedestrians and bicyclists use the same path, and one mode is mistaken for another. In addition to calibration and validation of the equipment, NCHRP 797 also suggests using bias compensation factors computed based on manual validation counts to correct any systematic under- or overcounting (Ryus et al., 2014).

3.3.1 **Data**

Several methods have been proposed and recommended to detect suspect count data or to flag suspicious data for further scrutiny. Table 3-1, adapted from Nordback et al. (2016) and updated to reflect latest practices, summarizes common count error-checking techniques.

Turner and Lasley (2013) used the first and third quartiles of hourly counts per direction for weekdays and weekends to identify outliers in the data outside the interquartile range (IQR).

$$IQR = 2.5 (Q_3-Q_1) + Q_3$$

Where Q_1 and Q_3 are the first and third quartiles, respectively. They also recommend checking the data manually through visual inspection to identify any errors. Turner and Lasley (2013) also recommend technology-specific detailed diagnostics.

As part of the Travel Monitoring Analysis System (TMAS) version 2.8, FHWA lists four types of flags that are used to check the data (FHWA, 2016)

- Fatal – This is the most severe flag and indicates issues that prevent TMAS from reading the data.
- Critical – This flag indicates that TMAS database can read the data but there are potential conflicts or data quality issues with the data.
- Caution – This flag indicates that there are concerns with the data and although the data is allowed into the database, the data is flagged.
- Warning – This is the least severe flag and indicates that the database could not perform a quality control check or that there was insufficient data.

Turner et al. (2019b) recommend the following flags for non-motorized data:

- Valid/Invalid – Count appears realistic and normal for the location. If the count appears invalid, flag it.
- Inverted AM/PM – Flag if the ratio of 3:00 a.m. to 3:00 p.m. counts in a day is not less than 1. This rule could be violated at locations near entertainment districts.
- Abnormal but Valid – This flag is used for counts that appear outside the normal range of values. These unusual counts may occur due to events, weather or disasters.

Turner et al. (2019b) also list the common metadata errors which include data mislabeling, latitude/longitude errors, and erroneous entries. They also recommend visually inspecting the data.

Table 3-1 Common Non-motorized Count Data Validity Checks

Source	Upper Bound/Lower Bound	Data Gaps	Identical non-zeroes	Consecutive zeros	Directional split
FHWA TMAS	Flag if: - Total hourly count exceeds 4,000 - Total daily count exceeds 50,000 or if total minimum count is less than 100 - Difference between any zero interval and an adjacent interval is greater than 50 - 3AM count is greater than 3 PM count - Variation in the monthly average daily traffic (MADT) estimated for the same month in the previous year is greater than 20%. Historical data: Average the daily totals for the previous six weeks (min. 2 weeks) for a given day of the week at a given location. Flag if variance is ±20%.	Not addressed	- Flag intervals when there are more than three identical adjacent non-zero values	-For any count interval, flag if there are more than 7 consecutive zeros	Not addressed
CDOT	Weekly check: Flag counts with any daily total higher than three times the previous year's average daily traffic (ADT) Annual Check: Suspicious daily totals for each continuous count site are identified using the interquartile range formula: IQR= 2.5 (Q3-Q1) + Q3 (Q3 = Third quartile of quarterly data; Q1 = First quartile of quarterly data)	Weekly Check: Count sites with missing data days and flag sites with > 5 days of missing data Annual Check: A count is valid only if it has a full 24 hours of count data for each 24-hour period	Not addressed	Weekly check: During warm weather months, sites with more than two continuous days of hourly zero values flagged; this check is not applicable for cold-weather locations Annual check: Same as weekly check	Weekly Check: Flag any count site exhibiting a direction split greater than 70/30 Annual Check: Same as weekly check
MnDOT	Data greater than two standard deviations above the mean flagged Web search to identify special events related to unexpected high counts	Not addressed	Not addressed	Visual inspection after installation Check daily zero values in summer months	
NCDOT	Data flagged if the upper bound exceeds IQR = Q3 + 3 (Q3 – Q1)	Not addressed	Not addressed	Over three days of zero counts	Splits 3 std. dev. > avg.
BikePed Portal, PSU	Flagged when hourly counts exceed 1000 (low-volume sites), 2000 (medium-volume sites), 4000 (high-volume sites)	Not addressed	Suspicious if: • 7+ consecutive non-zero values;	Possibly suspicious at 12.5 hours; suspicious at 25	Not addressed

Source	Upper Bound/Lower Bound	Data Gaps	Identical non-zeroes	Consecutive zeros	Directional split
			• 6+ consecutive non-zero values, volume >2; • 5+ consecutive non-zero values, volume >5; • 4+ consecutive non-zero values, volume >16; • 3+ consecutive non-zero values, volume >100	hours	
Turner and Lasley	Upper and lower bounds for expected counts in a given time period Interquartile range (IQR) = 2.5 (Q3-Q1) + Q3 Comparisons with previous counts at a given location and other stations in the vicinity and comparison of directional counts at the same facility (less than 80% deviation between directions is recommended); Expected ratios of peak hour to daily volumes				
Turner et al 2019b	Flag the counts if the sum exceeds 5,000 of one day or 1,500 per hour. Flag the count if it lies outside the interquartile range. Adjacent Interval – If count is less than 100, flag if adjacent count is > 100 percent different; if count is > 100, flag if adjacent count is +/- 100 percent different.	Timestamp exists, but count data are missing.	Flag the data if the same count value is repeated three or more times when 15 or more counts are available at the finest level of detail.	Flag the data when there are 15 hours or sixty 15-minute periods of consecutive zero counts. If this error occurs for a short time period, consider if weather could be a factor. Also, this flag may not be applicable to low volume locations.	

3.4 BICYCLE VOLUME ESTIMATION METHODS

This section provides an overview of methods to expand short-duration counts at specific locations and methods used to estimate network-wide bicycle volumes. Typically, motorized and non-motorized counts are collected using a mixture of continuous and short-duration counts. While continuous counters provide temporal trends, they are expensive to install and maintain and, hence, are conducted in limited locations (Romanillos et al., 2016). Short-duration counts are less expensive than continuous counts and, therefore, are conducted at multiple locations on the network, thus providing spatial coverage. Ryus et al. (2014) estimated that 87% of bicycle count programs include short-duration, manual counts. Bicycle volume functions can be estimated from continuous counters using seasonal and other factors, which are used to extrapolate short-duration counts to annual average daily bicycle traffic (AADBT). This methodology has been adopted from motorized counting where it is commonly used for extrapolating short-duration motorized counts. A key difference between motorized and non-motorized counts is that variation is lower for motorized counts as compared to non-motorized counts.

Despite advances in counting technology, cost and other considerations will continue to limit direct observation to small subsets of entire networks, as is the case for motorized traffic (Proulx and Pozdnukhov, 2017; Roll, 2018; Romanillos et al., 2016). Even with temporal expansion, stationary counters can tell us only about the bicycle traffic passing directly by them, and nothing about activity on the rest of the network. For these reasons, models developed from observed counts provide the most viable option for network-wide volume estimation. An extensive and well-documented range of methods is available for motorized volume estimation (Unnikrishnan et al., 2017). Comparable options for non-motorized traffic are less well-developed but have followed two primary approaches – bicycling submodels specified within larger travel demand models, and simpler, standalone direct demand models (Turner et al., 2017).

3.4.1 Extrapolating from short-duration counts to AADBT

Several researchers have conducted studies and analyzed the errors resulting from the application of factors and grouping techniques to extrapolate short-duration counts to AADBT estimates. These studies have used different factors/models to extrapolate short-duration counts as shown below.

- Day-of-Week, Month-of-Year Factors (19 factors): (El Esawey, 2018a, 2018b, 2014; El Esawey et al., 2013; El Esawey and Mosa, 2015; Figliozzi et al., 2014; Hankey et al., 2014; Nordback et al., 2018, 2013; Nosal et al., 2014; Nordback et al., 2018).
- Month-of-Year Factors (12 factors): (Nordback et al., 2019).
- Day-of-Week-of-Month (84 factors): (Nordback et al., 2019; Nosal et al., 2014).
- Day-of-Year (365 factors): (Beitel et al., 2017; Budowski et al., 2017; El Esawey, 2016; Hankey et al., 2014; Nosal et al., 2014).
- Weekend-Weekday: (El Esawey, 2018b; El Esawey et al., 2013; El Esawey and Mosa, 2015).

- Seasonal factors: (El Esawey, 2014).
- K factors (Design factors): (Beitel et al., 2017; El Esawey and Mosa, 2015).
- Statistical models: (Figliozzi et al., 2014; Nordback, 2012; Nosal et al., 2014).

These studies have been summarized in Nordback et al. (2019). Average annual daily nonmotorized traffic (AADNT) estimation errors are minimized by using day-of-year rather than traditional factors (El Esawey, 2016; Hankey et al., 2014; Nosal et al., 2014) and by using monthly rather than seasonal factors (El Esawey, 2014; El Esawey and Mosa, 2015). Extrapolating AADNT from two-hour counts results in high error: “80% chance that the error will be within plus or minus 60%” (Johnstone et al., 2017). Collecting short-duration counts over a one-week period and during the summer as opposed to shorter counts or counts collected in other seasons also reduces AADBT estimation errors (El Esawey, 2016; Hankey et al., 2014; Nordback et al., 2013; Nosal et al., 2014).

Other researchers have grouped sites based on similar characteristics and then created factors to extrapolate short-duration counts to AADBT estimates. Researchers have employed techniques such as visual inspection, cluster analysis, grouping by index and grouping by spatial variables to determine factor groups. Commonly used factor groups include commuter/utilitarian, recreational and a mixed group. At least 14 different factor groups have been identified in various studies and are listed in Nordback et al. 2019. Commonly used indices are the Weekend/Weekday Index (WWI) and the Average Morning/Midday Index (AMI), which are defined as follows (Miranda-Moreno et al., 2013):

$$WWI = V_{we}/V_{wd}$$

where:

WWI = Weekend/Weekday Index

V_{we}=average weekend daily traffic

V_{wd}=average weekday daily traffic

and

$$AMI = \frac{\sum_{7}^{9} v_h}{\sum_{11}^{13} v_h}$$

where:

AMI = Average Morning/Midday Index

v_h = Average weekday hourly count for hour (h) where hours are given as starting time of the hour (7:00 a.m., 8:00 a.m., 11:00 a.m. and noon).

Recent research on AADNT estimation has provided further guidance on the best time of the day to count, for short-duration counts and the number of counters needed per factor group to reduce estimation errors. Nordback et al. (2018) found that error was lower for the commute factor group, bicycle-only counts, scenarios in which more peak hours are counted, and when more than one permanent counter is available to estimate adjustment factors. In another study, using continuous count data from six cities,

Nordback et al. (2019) found that four or more continuous counters per factor group reduced estimation errors for bicycle volumes.

Another issue in AADNT estimation is how to assign short-duration count sites to groups to match up continuous counter groups with short-duration count sites for extrapolation purposes. The goal is to match short-duration sites with groups of continuous counters with the same travel pattern. Strategies for such matching include matching by geographic proximity (such as within the same city or neighborhood), using facility or land use as a proxy for travel pattern, or using the AMI and WWI indexes discussed above if sufficient data are available (Nordback et al., 2019). Using AMI requires short-duration counts in the morning and noon peak hours, and using WWI requires short-duration count data on at least one weekday and one weekend day. Johnstone et al. (2017) recommends at least eight hours of short-duration counts per site for such index-based grouping (Tuesday, Wednesday, or Thursday 7-9 a.m., 11 a.m.-1 p.m., and 4-6 p.m.; Saturday 12-2 p.m.).

3.4.2 **Travel demand models**

Travel demand models (TDMs) describe decisions to travel for activities, and typically model locations, modes, and routes. Models of this type have been utilized for modeling bicycle route choice (see, for example, Broach et al. (2012)) and mode choice (Broach and Dill, 2016). However, bicycle volumes along the network are determined in a network assignment phase, which to date has been done only experimentally at select agencies including Portland Metro, San Francisco County Transportation Authority, and the Metropolitan Transportation Commission in the San Francisco Bay Area. Proulx and Pozdnukhov (2017) tested both Strava and local and regional San Francisco-area TDM volume estimates in a model of observed count data. Beyond complexity, drawbacks of existing travel demand models for AADBT estimation include their typically large aggregation scale, exclusion of recreational trips, and lack of rigorous validation against observed counts.

More recently, research has attempted to combine the strengths of the two techniques by using travel model-derived outputs as inputs to direct demand models (McDaniel et al., 2014; Proulx and Pozdnukhov, 2017). Each of these studies focused on prediction of peak bicycle travel only and did not attempt to expand to AADBT estimates. Proulx and Pozdnukhov (2017) reported RMSE (root mean squared error) for PM peak volumes (4-7 p.m. September weekdays) as low as 24 for the best predictive model, calculated using 10-fold cross-validation holdout samples. Strava data alone resulted in a best RMSE of 35.

3.4.3 **Direct demand models**

Direct demand models typically are estimated by regressing observed counts on characteristics of the surrounding built environment and transportation service characteristics (Munira and Sener, 2017; Ortúzar and Willumsen, 2002). Relatively simple to estimate, direct demand models have shown potential for estimating both point location (Chen et al., 2017; Lindsey et al., 2018; Sanders et al., 2017) and

network-wide AADBT across a variety of locations (Hankey et al., 2017; Le et al., 2017; Lu et al., 2018). However, questions remain on which variables to include and how to measure them, as well as suitability for extrapolation across entire networks (Proulx and Pozdnukhov, 2017).

Munira and Sener (2017) provide a recent review that identifies the following common factor groups for direct demand bicycle volume models:

- Sociodemographic: income, race/ethnicity, education, percent students, age
- Network measures: centrality, bridges, bicycle facility density (on-street, off-street), major road density, number of lanes, curb-lane width, bike-lane width, intersection connectivity, parking entrances, slope
- Land use: population density, employment density, commercial/retail density, industrial use, open space, low-density residential, institutional use, land use mix
- Accessibility: jobs, bike trail entrances, transit stops, schools, distance from central business district
- Temporal: weather, time of day

3.5 EMERGING THIRD-PARTY BICYCLE USER DATA

Bicycle use data from smartphone apps and other emerging sources, often referred to as crowdsourced data, promise potential improvements for existing models of AADBT. These new data sources can collect large datasets with broad coverage that potentially contain information missing from conventional sources such as trip types and locations missing from conventional sources. Mode-specified datasets such as Strava, in which the user identifies a specific trip and mode of travel, contrast with mode-unspecified datasets such as StreetLight, in which a user has not specified a mode and may not be aware trips are being tracked (Lee and Sener 2020; Harrison et al., 2020; Tsapakis et al., 2021).

Several research studies have evaluated and applied bicycle user data over the past few years. Most peer-reviewed publications to date (Table 3-2) have reported on locales outside the U.S. Table 3-2 provides an overview of other studies and reports evaluating third-party data. Several state DOTs have recently purchased Strava data, resulting in a number of reports.

All the studies in Table 3-2 and Table 3-3 use data derived from GPS-enabled smartphone apps, most notably Strava. Such data are collected by self-selected samples of users that likely vary systematically from the full population of cyclists. The resulting bias in samples also likely varies across locations and over time. Other potential sources of emerging user data may draw from more representative slices of bicycle users – for example, data drawn passively from a larger sample of smartphone users – but we were unable to identify any rigorous evaluation of such sources.

Four major questions have emerged from research on user data to date:

1) To what extent do various user data sample demographics overlap with overall cyclist demographics?
2) What are the sampling rates of user data relative to on-the-ground counts? And, how stable are these rates across location types and over time?
3) What is the correlation between user data-derived counts and observed counts of cyclists, and how much of the variation from location to location can we explain?
4) How much overlap exists between different user data sources, and how much between user data and other bicycle activity data (e.g., bikeshare and demand estimates)?

Table 3-2 Studies Evaluating Third-party User Data: Peer-reviewed Journal Articles

Study	Location (Country)	User data	Demographic (vs. actual)	Count correspondence[1] (count data)
Boss et al. (2018)	Ottawa (CA)	Strava	78% male (68%) Noted age 25-44 over-represented	(n=11 PC, single month, two years); Strava/counts=1-30%; r=0.76-0.96
Conrow et al. (2018)	Sydney (AU)	Strava	77% male 23% 35-44 39% commuting	(n=122 1-day SC manual); r=0.79 overall; r=0.17-0.57 (low/med/high count locations); Noted less agreement where few bike lanes and higher socioeconomic status
Griffin & Jiao (2015)	Austin, TX	Cycle Tracks (4 mo.); Strava (1 wk.)	CT: 70% male S: >75% male Noted Strava skewed age 35-54	(n=5 PC); Strava/counts=3-9% Noted site ranking order disagreements & 3 sites with r<0
Heesch & Langdon (2016)	Queensland (AU)	Strava	82% male (72%) 35% 35-44 (28%)	(PC, n not provided); Strava/counts=3-7% Noted consistent patterns over 3-month period
Hochmair et al. (2019)	Miami, FL	Strava	N/A	(n=32 SC 3-day video counts, 3-month period); r=0.55 weekday Noted Strava/counts ratio higher on streets than trails
Jestico et al. (2016)	Victoria, BC (CA)	Strava	77% male	(n=18 SC manual, four 1-day); 2% SR; r^2=0.58 peak; Noted 45% of sites had error > 30%
Livingston et al. (2021)	Glasgow (UK)	Strava	N/A	(n=38 SC count sites) r=0.781 hourly; r=0.861 AM-peak /PM-peak /off-peak r=0.882 daily r=0.887 two-day totals
Musakwa & Selala (2016)	Johannesburg (ZA)	Strava	80% recreational	N/A
Sanders et al. (2017)	Seattle, WA	Strava	83% male 62% 25-44	(n=46 PC + SC, not distinguished, 2 years); Strava/counts=~2%
Strauss et al. (2015)	Montreal (CA)	*Mon RésoVélo* (phone app)	N/A	(n=600 8-hr manual counts over 2 years, n=435 additional manual, and n=30 PC); r^2=0.48-0.76 (regression adjusting for bicycle facility type and surrounding context)
Sun et al. (2017a)	Glasgow (UK)	Strava	N/A	r=0.83
Sun et al. (2017b)	Glasgow (UK)	Strava	Noted 35-44 largest age group	N/A

Table abbreviations:
r=simple correlation coefficient (assumed Pearson, unless noted)
r^2=multiple correlation coefficient (assumed ordinary least squares regression, unless noted)
PC = permanent location count
SC = short-duration count
Strava/count = Strava to total count sampling ratio, the ratio of crowdsourced counts to observed counts

Table 3-3 Studies Evaluating Third-party User Data: Reports, White Papers, and Other Documents

Study	Location	User data	Demographic (vs. actual)	Count correspondence[1]
CDM (2018)	Brisbane (AU)	Strava	N/A	(n=27 PC); Strava/counts=~5-20% Noted wide variation in r across sites
CDOT (2018)	Colorado	Strava	N/A	(n=16 PC); Strava/counts=1-30% r^2=0.8-0.99
Chen (2017)	Portland, OR	Strava	N/A	(1 PC, single bridge location); Strava/counts=1.4%; Noted SR stable over seasons; Strava significant In estimated Safety Performance Function
PacTrans/Wang et al. (2016)	Portland, OR	Strava	N/A	Noted Strava counts significant in Safety Performance Function
Proulx & Pozdnukhov (2017)	San Francisco, CA	Strava	N/A	(n=25 PC, n=69 SC manual counts, 4-7p Sep weekdays only) Noted Strava most associated w/ counts (25 perm., 69 short dur.) compared with other data sources RMSE=24, best fit, cross-validation hold out, including Strava, bikeshare, and regional travel demand model inputs
ODOT/Roll (2018)	Eugene, OR	Strava	N/A	(n=52 annualized SC); Strava/counts=~1%; Noted Strava counts significantly improved volume estimates
Strava (2014)	Seattle, WA	Strava	N/A	(n=2 PC, 11 months, two bridge locations); Strava/counts=3-5%; r^2 >0.9
StreetLight (2018)	San Francisco, CA	StreetLight	N/A	Strava/counts=6.4% mean; R^2=0.69-0.78 (higher weekdays); Noted Streetlight data seemed to underestimate weekday AM peak; also compared trip stats with other travel surveys
TTI/Dadashova et al. (2018)	Texas	Strava	N/A	(n=100 PC); Strava/counts=0-63% (mean 5%); MAPE=29% (Strava adjusted by street class and income level); Noted lower error where Strava/counts=5-15%
TTI/Turner et al. (2019)	Austin, TX; Houston, TX (Strava only)	Strava; Ride Report	S: 80-82% male S: 67-68% <44 years old	(n=12 PC); Strava/counts=3-19% Strava r=0.59-0.81; RR: 0.03-0.3%; RR: r=0.61
Watkins et al. (2016)	Atlanta, GA	Strava; Cycle Atlanta (phone app)	S: 84% male S: 31% age 35-44 S: 29% commute CA: 76% male CA: 60% commute	(78 manual, AM+PM peak, 1-day) CA: ~3% SR

Table abbreviations:
r=simple correlation coefficient (assumed Pearson, unless noted)
r^2=multiple correlation coefficient (assumed ordinary least squares regression, unless noted)
PC = permanent location count
SC = short-duration count
Strava/counts = Strava to total count sampling ratio, the ratio of crowdsourced counts to observed counts
N/A = Not Available or Not Provided

3.5.1 **Demographic overlap**

The issue of representation is a primary question for agencies interested in using third-party bicycle user data. Unfortunately, user data demographics are generally limited to age, gender, and type of cycling activity (e.g., commuting vs. recreational riding). Information on key equity-related variables such as income and race/ethnicity is generally lacking. Further complicating attempts to measure representativeness is a lack of comparison data on the composition of cyclists in a city or region. Still, research to date suggests that Strava users, the most common source of user data, are more likely to be male, age 35-44, and to log a greater share of recreational trips than the general cyclist population. For reference, recent national estimates of adult cyclist gender split range from 51% to 60% male for all cyclists and 73% male for bicycle commuters (2017 American Community Survey; Dill, 2015; PeopleForBikes, 2015; Schroeder and Wilbur, 2013). Similar estimates for share in the 35-44 age group range from 18% to 21% (PeopleForBikes, 2015; Schroeder and Wilbur, 2013).

Garber et al. (2019) compared characteristics of survey respondents who used smartphone apps to track bicycle rides with those who did not use such apps. They found that app users were more likely to ride more often, take more leisure trips, and self-classify as stronger riders (e.g., “strong and fearless”).

3.5.2 **Correspondence with observed counts**

Several studies have compared Strava and other user data with observed short- and long-duration bicycle counts (Table 3-2). While there are similarities across comparison methodologies, there are also distinct differences driven by data availability, application, and researcher preferences. Most evaluations have been made comparing annual (sometimes annualized) or monthly observed versus third-party counts at the segment or intersection level. One study compared count sources at hourly and daily AM/PM peak intervals (Jestico et al., 2016).

“Ground truth” count data are often limited, and many studies have relied on a mix of permanent and short-duration counts or even short-duration counts alone. While most correlation examples summarize results across all count locations, a few have segmented by time of day (CDM Research, 2018; Jestico et al., 2016); weekday/weekend (CDM Research, 2018; CDOT, 2018); or even by individual count location (CDOT, 2018).

App users typically make up only a small fraction of cyclists at a specific location. Strava data, on average, has usually represented somewhere between 2% to 10% of all bicycle traffic at evaluated locations. Researchers have also noted that sampling ratios can vary significantly even within the same region (Conrow et al., 2018; Griffin and Jiao, 2015; Heesch and Langdon, 2016; Hochmair et al., 2019; Jestico et al., 2016). Sampling rates appear to vary systematically by context, in terms of both bicycle facility types and surrounding context. For example, Hochmair et al. (2019) noted sampling rates more than six times higher on streets relative to trails in Miami, and Conrow et al. (2018) found Strava sampling rates higher in areas with few bike lanes and in areas of

higher socioeconomic status. Dadashova et al. (2018) also reported the need to adjust Strava sampling rates by road class and surrounding households with higher incomes.

A majority of research to date has reported simple linear correlations (r) or coefficients of determination (R^2). A problem with simple correlations as an evaluation tool is their sensitivity to outliers. It is easy to come up with scenarios in which error rates are equivalent but linear correlations are wildly different.[1] It is important to consider prediction error rates along with correlations. Measurement issues aside, most evidence points to strong or linear correlation between third-party data and observed counts. Controlling for facility and surrounding contextual factors, third-party (mainly Strava) data has consistently been found to significantly improve bicycle volume and safety performance models (Dadashova et al., 2018; Proulx and Pozdnukhov, 2017; Roll, 2018; Sanders et al., 2017; Strauss et al., 2015; Wang et al., 2016).

Despite consistent findings of association, a few details are worth noting:

- Correlation does not necessarily track sampling ratio; for example, in Austin, despite much lower sampling rates (SRs), Ride Report (r=0.61) had a slightly stronger association with observed counts than Strava (0.59) (Turner et al., 2019a).
- There is evidence that Strava correlation varies by bicycling volume; for example, Conrow et al. (2018) reported an overall r=0.79, but when segmenting into low-/medium-/high-count locations, correlations fell to 0.55/0.17/0.57, respectively.
- High overall correlation may mask poor site-specific performance such as large site ranking errors or locations with negative correlation (Griffin and Jiao, 2015).
- Areas of low correlation may be spatially correlated (Conrow et al., 2018).

Several studies combined or “fused” data with third-party counts to model observed bicycle volumes. Dadashova et al. (2018) added segment functional classifications and adjacent numbers of upper-income households (>$200k/yr.). Jestico et al. (2016) included segment slope, speed limit, on-street parking presence, and a seasonal adjustment along with Strava counts. Proulx and Pozdnukhov (2017) used TDM bicycle volume estimates and bikeshare counts alongside Strava data. Roll (2018) combined Strava counts with segment functional class, bicycle facility types, local accessibility and design measures, and a measure of network centrality. Sanders et al. (2017) included the number of bike lanes and proximity to the university alongside Strava count data. Others have included third-party data directly in safety performance functions (Strauss et al., 2015; Wang et al., 2016).

3.5.3 **Third-party user data applications**

A handful of studies applied third-party user data, either directly or as part of a statistical model, to estimate area or network volumes. Most commonly, these estimates were

[1] As an example: consider two locations A & B each with three count sites like (observed: third-party) {A: (100: 10, 150: 15, 1000: 50), B: (100:10, 150: 15, 200: 10)}. A & B have equivalent error rates per site and overall (MAPE=92%), but A has a correlation coefficient r=0.92 and B r=0.00. Site A “benefits” from the outlying high-volume location dominating the correlation calculation.

used in safety analyses to estimate injury risk and safety performance functions (Chen, 2017; Saad et al., 2019; Sanders et al., 2017; Strauss et al., 2015; Wang et al., 2016). Other studies used third-party user data to better understand cyclist exposure to air pollution (Sun et al., 2017b); correlates of recreational cycling demand (Sun et al., 2017a); and to assess changes in bicycle ridership patterns (Boss et al., 2018; Heesch and Langdon, 2016).

3.6 RELATED DATA FUSION METHODS

During the last decades, an unprecedented volume of data for both non-motorized and motorized traffic in transportation has become available through a wide range of advanced technology and conventional data collection methods. Passive detection systems such as inductive loop or vision-based detectors, which estimate traffic volumes, occupancy, and vehicle speed only for the limited sections where they are installed, represent the most commonly used data collection method for both non-motorized and motorized traffic. The increased implementation of Intelligent Transportation Systems (ITS) technologies has led to the use of advanced sensors such as ultrasonic and LIDAR to obtain traffic counts or flows at point locations (Srour and Newton, 2006).

For large spatial coverage for both motorized and non-motorized traffic, Bluetooth, WiFi, and GPS remain the most available advanced technologies. GPS provides additional information such as route choice, origin-destination (OD) and travel time beyond other passive detection systems. Even though the GPS monitors and collects vehicle (or biker) path flow, this information often can only be captured from a small portion of the population, which may result in biased estimates in flow or travel time. Bluetooth has similar capabilities and limitations in collecting information because the vehicles connected to Bluetooth devices can be treated as a sample of the overall population. Vehicle-identification sensors such as the Automatic Vehicle Identification system identify and track individual vehicle sensor locations. However, sampling bias remains the main shortcoming of these sources because only a small fraction of vehicles equipped with electronic tags supply travel information (Dion and Rakha, 2006). Studies often obtain detailed trip information or travel behavior using a travel survey such as National Household Travel Survey (NHTS). The most recent NHTS data (2017) includes information regarding travel behavior and socioeconomic data where individuals or households reported their personal, vehicle, and trip information via survey questionnaires. Table 3-4 summarizes the aforementioned data sources and their advantages and limitations.

Table 3-4 Traffic Data Sources

Type	Source (examples)	Advantage	Limitation
Survey	National Household Travel Survey National level survey (Vehicle inventory and use survey (VIUS)) State level intercept survey	Detailed information on OD, vehicle category, weight, and Vehicle Miles Traveled Extensive data collection throughout the US or by State	Obtain partial data from sampled population Inaccurate responses and potentially biased survey sampling Limited spatial and temporal data collection span Cross-sectional data
Passive sensor technology	Inductive Loop Detector Vision-based (camera)	Obtain population data Time-series data Potential in obtaining vehicle flow and type data with an additional modeling effort	Obtain point observation Measurement error from sensor calibration and sensitivity issues
Active tracking sensor technology	Automatic Vehicle Identification (AVI) Electronic tolling GPS and Bluetooth	Time-series data Flow data with detailed vehicle class information without additional modeling	Typically short duration observation for sampled population Costly and privacy concerns
Active remote sensing	Radar and LIDAR	Detecting presence and volume with higher accuracy	Security concerns Obtain point observation

Data fusion combines multiple data sources prior to or during model development. Data fusion increases the accuracy and robustness of a model because different data sources complement or supplement each other to minimize the level of uncertainty or ambiguity from the data source and to enhance spatial and temporal coverage. Previous transportation studies have used data fusion techniques to estimate travel time or speed for motorized traffic (Bachmann et al., 2013; El Faouzi et al., 2011; Han, 2012). Inductive loops and GPS represent the most common data fusion sources for speed and time estimation (Qing-Jie Kong et al., 2009) since the loops have been heavily implemented nationwide and GPS provides accurate location-based information for an individual vehicle. Other studies used automatic plate number recognition (Han, 2012); inputs from toll collection such as entry-exit times at toll gates (El Faouzi et al., 2009); vision-based systems such as video cameras (Anand et al., 2011); and floating car data (Ambühl and Menendez, 2016; Cipriani et al., 2012) for traffic state estimation.

Safety research is one of the emerging data fusion applications as vehicles or network systems are being equipped with advanced sensors. Data from conventional sources such as loops are fused with GPS, vision-based (camera) and wave-based sensors (ultrasonic, radar, and LIDAR) for applications of route guidance or driver warning systems (El Faouzi et al., 2011; Zhang et al., 2011).

Freight research benefits significantly from data fusion techniques because of the proprietary nature and limited availability of freight data sources. Researchers fused

point sensor data such as from loops and weigh-in-motion (WIM) sensors with a mobile source of GPS, and survey data (FAF and VIUS) for various applications such as classifying truck configuration (Hernandez et al., 2016) or tracking flow movement (Hyun et al., 2017). These fusion techniques capture spatially and temporally varying freight movements and improve model transferability.

For non-motorized traffic, two studies used data fusion to develop a pedestrian detection algorithm. Garcia et al. (2013) used data from GPS and laser detectors, while Premebida and Nunes (2013) used LIDAR data with a camera system. Finally, for public transit travel, Kusakabe and Asakura (2014) developed a fusion technique using travel survey and smart card data to estimate continuous long-term changes in trip behaviors and demonstrate accurate estimation of trip purposes using the trip records in the smart card. Sener et al. (2021) employed count data, static data, bikeshare, Strava and StreetLight to model AADBT in Austin, TX, and found that a fusion of these sources can be helpful to estimate volumes when adequate actual count data are available, and noted that “performance of fusion firmly depends on the fusion method coupled with the data and situation characteristics” (p. 14).

Various data fusion techniques have previously been developed, such as spatial regression (Hernandez and Hyun, 2017); Kalman filtering (Chu et al., 2005); Bayesian method (El Faouzi, 2006); artificial neural network (He et al., 2016); and multiple classifiers using machine learning (Hernandez et al., 2016). A modeling approach with Kalman filtering and an artificial neural network has higher proven accuracy in model validation, especially when estimating short-term traffic flow because the model effectively captures varying traffic conditions (Anand et al., 2011; Kumar et al., 2013). The spatial regression method considers the spatial relationship between traffic count sites such that two distant count sites along the same corridor have a stronger relationship than closer sites located on different network corridors. Hernandez and Hyun (2017) developed a spatial matrix between count sites using GPS trajectories and estimated the truck weight distribution at the count sites where weight information was not available. Machine learning techniques such as neural network and support vector may perform data fusion for various input sources. Hernandez et al. (2016) fused loop and WIM to classify truck body configuration. They introduced a multiple (five) classifier in machine learning to improve the classification performance and transferability of the models. Data fusion also proves useful to fill missing data by integrating similar data sources with an imputation approach. Zhu et al. (2018) integrated GPS and Commercial Vehicle Survey using hot-deck and k-nearest neighbor imputation methods to impute missing survey data using GPS variables.

However, some researchers have identified potential complications and limitations with data fusion. El Faouzi et al. (2011) showed that the accuracy or robustness of data fusion depends on the data sources themselves as well as the analysis techniques that capture the unique characteristics of data sources and applications. Nantes et al. (2016) compared the characteristics of different data sources including Bluetooth, GPS data, and Eulerian loop sensors, and highlighted that understanding the heterogeneous nature of a data source represents an important step prior to the data fusion. Bachmann

et al. (2012) noted that many of the popular techniques being used for data fusion have limitations. For example, Kalman filtering may only work when the information about the underlying process is known, while machine learning techniques may require large input datasets.

3.7 LITERATURE REVIEW SUMMARY

This review surveyed the past, present, and emerging future of bicycle counting and volume estimation options. There has been tremendous growth in the number of count programs and supporting resources, including numerous established methodologies for both count data collection and quality control. While automated counters have become more common, the vast majority of bike count programs rely on manual counts for some or all of their network coverage, and even an extensive program leaves most of the network uncounted. Direct demand models, and, in a select few locations, regional bicycle travel demand models, have shown some promise for network-wide volume estimates, but serious questions remain about the proper form and the validity of those estimates.

Also apparent is the surge in interest from both agencies and researchers around emerging GPS and mobile phone data. These have the advantage of wide network coverage and continuous collection. Evidence has been mixed on their ability to represent the full spectrum of cycling activity, and about their accuracy and reliability. To date, studies have been confined almost entirely to single areas and data sources, often with limited comparison data from reliable counts. Most have used only relatively simple techniques to assess validity and to model volumes using these new data.

Data fusion and other advanced modeling techniques have shown promise in other areas of transportation to get more value out of disparate data sources. There appears to be similar potential with bicycle activity, given the increasing range of data available. This study explores how the various traditional and emerging sources might complement one another and provide a more complete picture of bicycling activity for applications such as safety, demand shifts, and equity analysis. The next chapter describes the data used for this study.

4 DATA

This chapter first describes the criteria developed for selecting sites, including an overview of a survey that was conducted to identify sites, and explains the selection process. Next, it describes the process for gathering and processing count, Strava, StreetLight, bikeshare, and static (demographic and network) data (Figure 4-1). We automated and standardized data collection to make the process as general and portable as possible. All data elements, aside from the StreetLight and (at the time) Strava data, were drawn from public sources, although accessing count data in some cases required local agency requests or assistance.

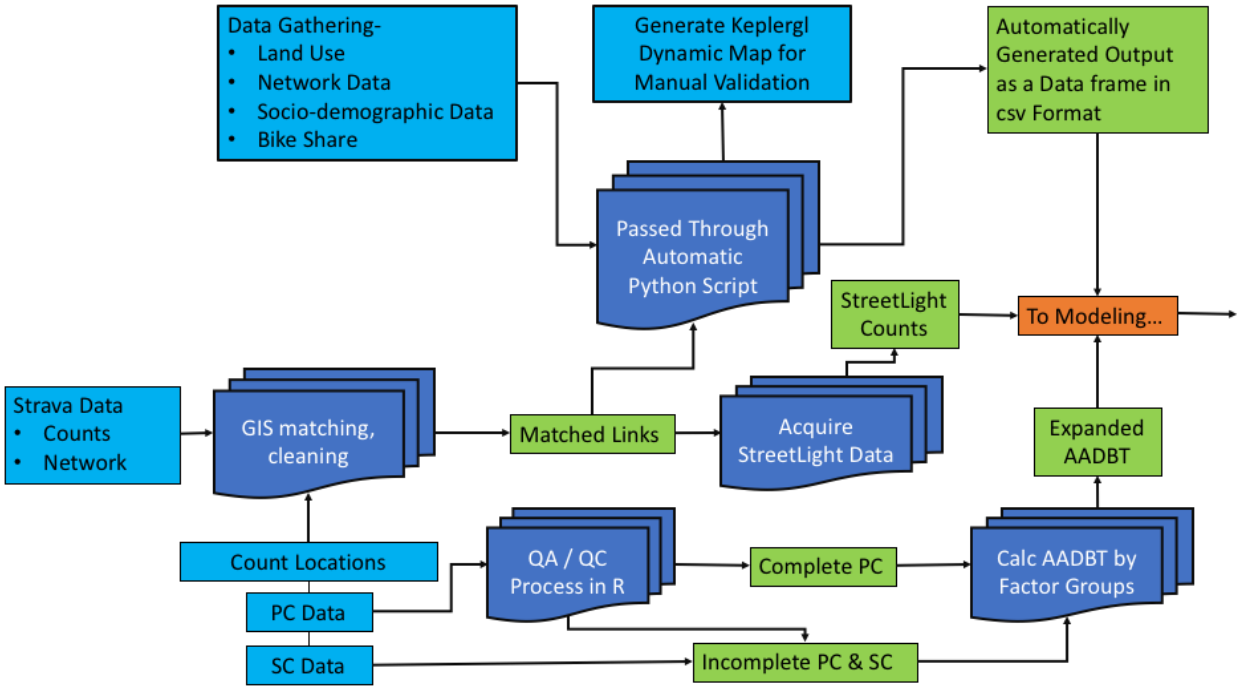


Figure 4-1 Data Process Overview (PC = Permanent Counts, SC = Short-duration Count, QA/QC = Quality Assurance/Quality Control, AADBT = Annual Average Daily Bicycle Traffic)

4.1 SITE SELECTION

The goal of the site selection process was to select sites that were geographically dispersed to provide national coverage and cover a range of contexts (large/small urban, trails/on-street). The sites also needed to have a rich array of data to fuse, including broad continuous counter coverage temporally and spatially, third-party user data (Strava, StreetLight), and bikeshare data.

To identify count sites, a survey questionnaire was created and sent out to various email listservs (APBP, ITE). The full survey questionnaire is provided in the appendix.

The questionnaire sought to identify the location of bicycle count sites and the types of counts conducted by agencies to count bicycles. Respondents were given a choice to either enter the information pertaining to the count sites in the survey or upload a file that contained the details. In the survey questionnaire, the respondents were asked to select the types of counts and counters that their agency used for counting bicycles (permanent automated, short-duration automated, or manual counts). For the permanent automated sites, the information requested included the number of counters, and, for each counter, the location (street/path, latitude/longitude); year of installation; make and model; mode being counted (bicycles only or bicycles and pedestrians combined); and directionality, if known. Specifically, for the short-duration automated counters, respondents were also asked how long the counters stayed in one location and how the locations were identified and selected. For all counts that were conducted, respondents were asked if their agencies conducted any data quality checks on count data and were asked to further describe these checks. Lastly, respondents were asked about the other non-count-related data that they have access to and/or use (Strava, StreetLight, Ride Report[2], Bikeshare, CycleTracks, Other) and to report on how they typically shared this count information.

We received 24 unique responses, including responses from 10 cities, six MPOs, three counties, two state DOTs, two universities, and one non-profit/advocacy organization. Among the 24 respondents, 22 entered information about the type of bike count data their organization collects, with 13 noting that they collect data from permanent automated bicycle counters, 13 collecting data from short-term duration bicycle counters, and 13 having manual bicycle count programs (many collected more than one type). Three also indicated that they had access to Strava data, while four had access to StreetLight data (although, in one case, the StreetLight data was noted as being exclusively for motor vehicle traffic).

Using the information obtained from the survey and known anecdotal information about count locations, a spreadsheet inventory was created. First, regions having 10 or more permanent count sites and at least two years of count data were identified. Next, for each of these regions, the location of all known continuous count sites was noted, whether they were located on a street or on a trail location, along with the extent of data availability (start and end dates) and known gaps. Finally, based on criteria described previously (range of contexts, geographical spread, availability of count and other data sources) six regions were chosen as shown in Figure 4-2 below.

[2] At the time when this survey was conducted, the RideReport app that allowed users to enter their bicycle trip data was still operational; however, it has since been discontinued.

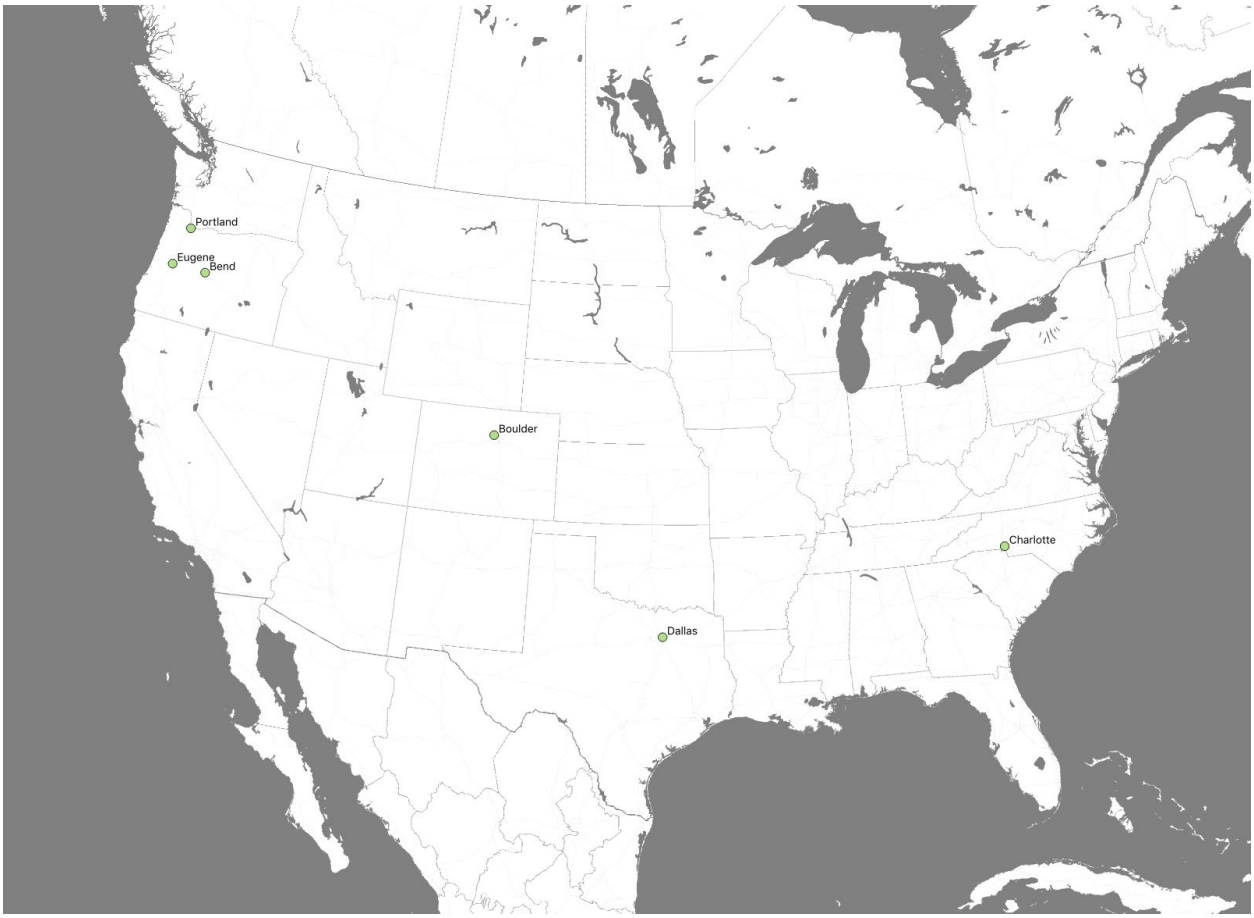


Figure 4-2 Selected Sites (Note: Each dot in a region represents one permanent counter)

Of these, Boulder, Dallas, and Charlotte formed our basic sites, where one year of data (2019) was used for modeling. The Oregon locations, Portland, Bend and Eugene. constituted our enhanced locations where three years of data (2017-2019) was used for modeling.

4.2 COUNT DATA

Bicycle count data archives were accessed with the assistance of local agencies. Both permanent (PC) and short-duration counts (SCs) were collected. PC data was collected via Eco-Counter's Eco-Visio web platform or directly via their API, and only counters able to distinguish bikes from pedestrians were considered. PC locations were further divided into "full-year" (>=10 months valid data) and "less than full-year" (<10 months valid data). The less than full-year PC locations were treated as short-duration sites. Full-year PC locations were used in an initial set of models and to create factors to expand I and SC counts into AADBT estimates. SC data were considered if they included a minimum of 24 consecutive hours counted. Overall, 603 count locations (permanent and short-duration) were available across all sites and all years, with the most count locations available in Portland and Eugene (Table 4-1). As expected, the

number of available count sites was highest in 2019 (as counts for 2017-2018 were not included for Boulder, Charlotte and Dallas).

Table 4-1 Total Usable Count Locations by Year for Six Study Regions

Region	2017	2018	2019	Grand Total
Bend	0	60	65	125
Boulder	n.c.	n.c.	41	41
Charlotte	n.c.	n.c.	16	16
Dallas	n.c.	n.c.	32	32
Eugene	0	86	78	164
Portland	104	33	88	225
Grand Total	**104**	**179**	**320**	**603**

n.c. = not considered

4.2.1 **Permanent count locations**

Hourly count data were obtained for (Eco-Counter) permanent count sites in each of the six locations. Boulder and Charlotte data were accessed via local Eco-Visio web portals. Dallas PC data were obtained via the Texas Bicycle and Pedestrian Count Exchange (BP|CX). Data for the Oregon regions were obtained from ODOT, which has developed R scripts to automate retrieval using the Eco-Counter API. For each permanent count site, a QA/QC process for filtering out the erroneous count data was created. The following are the steps undertaken in the QA/QC process:

- Obtain hourly counts at each count site.
- Remove consecutive zero counts longer than one week.
- Flag and manually inspect if any of the following conditions are met.
    - Consecutive zeros lasting at least 48 hours
    - Non-zero repeated counts lasting at least six hours
    - Daily volume is greater than 15,000
    - Hourly volume is greater than 1,500
    - Site-specific relative ADB checks are met
        - ADB < 100, flag as possibly suspicious when count is >400 and suspicious when hourly count is >1,000
        - ADB 100 to 500, flag as possibly suspicious when count is >1,000 and suspicious when hourly count is >2,000
        - ADB >500, flag as possibly suspicious when count is >4,000 and suspicious when hourly count is >8,000
- Consider daily counts valid if at least 22 valid hours are available.
- Consider months valid if at least one valid daily count for each day of the week (Sun-Sat).
- Consider counters “full-year” for AADBT calculation if they include 10 or more valid months of data.

Table 4-2 reports availability by region and year (for the Oregon regions; 2017-2018 data for the other regions were not considered due to unavailability of Strava data) after cleaning the raw permanent count data. Note that Dallas and Charlotte data had already been run through external QC processes, and we were unable to obtain raw data from those locations. We still ran the data through our own checks but very few additional data points were flagged or removed. Even at our enhanced sites, only Portland had permanent count data available for all three years (2017-2019), while Bend and Eugene did not have permanent count programs in place for 2017. Overall counter availability was disappointing, averaging less than 70% valid data over the period, and this excludes known counters that recorded no valid data over a given year.

Table 4-2 Usable Permanent Count Availability After Cleaning

	2017		2018		2019		Total – All Years	
Region	Full-year counters (>=10 mos.)	Total counters (any data)	Full-year counters (>=10 mos.)	Total counters (any data)	Full-year counters (>=10 mos.)	Total counters (any data)	Full-year counters (>=10 mos.)	Total counters (any data)
Bend	0	0	3	5	3	5	6	10
Boulder	n.c.	n.c.	n.c.	n.c.	8	11	8	11
Charlotte	n.c.	n.c.	n.c.	n.c.	9	13	9	13
Dallas	n.c.	n.c.	n.c.	n.c.	24	32	24	32
Eugene	0	0	11	15	13	15	24	31
Portland	8	13	6	10	4	14	18	37
Total	***8***	***13***	***20***	***30***	***61***	***89***	***89***	***134***

n.c. = not considered

4.2.2 **Short-duration counts**

Non-permanent, SCs were included if they were collected for a period greater than 24 hours. Except for Bend, where SCs were collected from mobile Eco-Counter devices along with the PC data, SC data had to be acquired individually from local jurisdictions in each region. The Dallas region had no SC program during 2019. SC data were not collected in Bend and Eugene for 2017 because there were no full-year PC sites available for factoring. Raw data were processed to a consistent format, but in some cases hourly counts were not available. In Boulder, we were only able to obtain daily counts, and factor groups could not be established. In Portland, we could only acquire hourly counts averaged over the days of collection (typically one to three days), and so we weighted each hour by the number of days collected. Table 4-3 shows the distribution of 471 short-duration count sites across regions and years. The short-duration data for Dallas was not available, hence, it is not included in the table. Across

all years and locations, the duration of the majority of these counts was between one week and less than one month, with most of these locations in Bend and Eugene. Bend has a number of automated counters that are part of its mobile counts program, and are rotated among different locations to gather short-duration counts. Portland had a large number of counts whose duration was greater than 25 hours and less than one week. In Boulder, all counts in 2019 were 24-hour counts.

Table 4-3 Short-Duration Count Locations and Availability

Region	24hr	25hr to <1wk	1wk to <1mo	1mo to <3mo	3mo to <1yr	Grand Total
2017						
Portland	6	85	0	0	0	91
2018						
Bend	0	0	43	9	3	55
Eugene	0	3	67	1	1	72
Portland	3	20	0	0	0	23
2019						
Bend	0	3	48	8	1	60
Boulder	30	0	0	0	0	30
Charlotte	0	0	3	0	0	3
Eugene	0	9	53	1	0	63
Portland	0	75	0	0	0	75
Grand Total	39	195	214	**18**	**5**	**471**

Note: Short-duration locations include both intentional short-term counts and continuous permanent count sites that recorded less than 10 months of complete data in a given calendar year.

4.2.3 **Factoring approaches**

Where “full-year” PCs had missing days, we calculated AADBT using the method detailed in the FHWA report *Assessing Roadway Traffic Count Duration and Frequency Impacts on Annual Average Daily Traffic (AADT) Estimation* (Krile et al., 2014). First, a monthly average daily traffic (MADT) is computed for each day of the week for each month using the following formula:

$$MADTm,y = \frac{1}{7}\sum_{j=7}^{1}\left[\frac{1}{n}\sum_{i=1}^{n} Vijmy\right]$$

Where:

V = total traffic volume for i^{th} occurrence of the j^{th} day of the week within the m^{th} month, for year y.

n = the count of the j^{th} day of the week within the m^{th} month, for which traffic volume is available (a number from 1 to 5)

The resulting MADTs are averaged to compute AADT with this formula:

$$AADT_y = \frac{1}{12}\sum_{m=1}^{12} MADT_{m,y}$$

Where m is the month of the year, y”

To expand the less than full-year PC and SC data to AADBT, we used a factoring approach based on travel pattern factor groups. Three factor groups - commute, mixed and non-commute - were identified based on travel patterns using the AMI index and a modified WWI index (Johnstone et al., 2017). WWI was modified to be based on total daily travel instead of a single peak hour.

The travel pattern identification approach used in this research is derived from the WSDOT report *Collecting Network-wide Bicycle and Pedestrian Data: A Guidebook for When and Where to Count* (Johnstone et al., 2017), except weekend ratio based on total daily travel instead of peak single hour, at the recommendation of the authors of that report. For less than full-year PC and all SC sites, the observed counts were expanded to impute AADBT using the following methodology:

First, full-year PC sites were assigned to factor groups using the approach suggested by Johnstone et al. (2017), as shown in Table 4-4.

Table 4-4 Classification of Bicycle Travel Patterns

Travel Pattern	WWI		AMI
Commute	less than 1.0	and	greater than 1.5
Mixed or Multipurpose	less than 1.0	and	less than 1.5
	-or- 1.0 - 1.8	and	greater than 1.5
Non-Commute or Noon Activity	1.0 - 1.8	and	less than 1.5
	-or- greater than 1.8	and	any

Next, generate day-of-year (DOY, n=365) factors and day-of-week-of-month (DOW, n=84) factors. Then, assign partial PC and SC sites to factor groups. Where SC site lacks weekend count data, assign based on AMI factor only. If only AMI is available, assign hourly groups by average AMI metric:

- *Recreation (Hourly Noon Activity): Average AMI <=0.7*
- *Mixed (Hourly Multipurpose): 0.7< (Average AMI)<=1.4*
- *Commute (Hourly Commute): Average AMI>1.4*

If no full-year PC sites are available for a factor group OR data are insufficient to assign factor group, use average of DOY or DOW factors across all available factor groups.

Table 4-5 shows the factor groups for the PC sites, where 1 is commute, 2 is mixed and 3 is non-commute. The majority of Portland sites are commute sites, whereas Dallas sites are predominantly non-commute. The travel patterns at Bend, Boulder and Eugene sites are mostly mixed.

Table 4-5 Number of Permanent Count Sites Per Factor Group

Year	2017			2018			2019			Grand Total
Factor Group	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	
Bend	0	0	0	0	5	0	1	3	1	10
Boulder	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	1	8	2	11
Charlotte	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	1	2	10	13
Dallas	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	0	2	30	32
Eugene	0	0	0	1	9	5	5	7	3	30
Portland	6	3	4	6	3	1	6	4	3	36
Grand Total	**6**	**3**	**4**	**7**	**17**	**6**	**14**	**26**	**49**	**132**

n.c = not considered
Factor Groups: 1=commute; 2=mixed; 3=non-commute
Note: Numbers include all permanent counter installation sites, even if the counter did not provide a full year of data

Table 4-6 shows the factor groups for the SC sites. Note that Dallas is not listed because no short-duration counts were available. The travel pattern trends seen at the SC sites mirrored those at PC sites, with Portland sites being mainly commute type, whereas Eugene and Bend were mixed sites. Boulder County could not supply any hourly or weekday/weekend splits, so all SC sites had to use pooled factors and, thus, could not be assigned to any specific factor groups.

Table 4-6 Factor Groups for Short-Duration Count Sites

Year	2017			2018			2019				Total
Factor Group	1	2	3	1	2	3	1	2	3	NA	
Bend	0	0	0	3	34	18	5	37	18	0	115
Boulder	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	0	0	0	30	30
Charlotte	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	2	0	1	0	3
Eugene	0	0	0	15	42	14	17	36	10	0	134
Portland	56	22	13	14	5	4	41	23	11	0	189
Grand Total	**56**	**22**	**13**	**32**	**81**	**36**	**65**	**96**	**40**	**30**	**471**

n.c = not considered

Factor Groups: 1=commute; 2=mixed; 3=non-commute; NA - factor group could not be assigned.

We compared the factored AADBT results from both the DOY and DOW methods and found no major disagreement between the two for this dataset. We therefore chose to use the DOW approach, since that allowed us to retain data at locations where no full-year PC data were available on a specific day in a region.

4.3 STRAVA

Strava relies on self-selected users of the Strava app who record and upload their rides. Raw GPS data are post-processed by joining trips to an OpenStreetMap (OSM)-based network. Utilitarian ("commute" in Strava's terminology) and recreational trip purposes are imputed based on route directness. Data are then provided as trip ("activity") and user ("athlete") counts on each matched OSM network link. We were provided data in hourly and monthly aggregations. The updated Strava Metro product, now freely available to agencies, is downloadable in hourly and daily aggregations.

We noted and confirmed with Strava a few important characteristics of the data to keep in mind when using for analysis:

- Counts are rounded in an unusual way (note: these apply to all time intervals—hourly, daily, monthly)
    - trip counts < 3 are not included (these were set to zero for each period missing in the raw Strava data file)
    - trip counts >= 3 always round up to the nearest 5 (e.g., 3 rounds to 5, 6 rounds to 10, and so on)
        - A network-wide investigation in Portland comparing monthly totals to aggregated hourly counts revealed significant bias owing to the masking and rounding process.
        - We avoided this issue by switching to monthly totals to calculate Strava average daily bicycle traffic (ADBT).
- Map matching of GPS points to the OSM network does not enforce complete routes, such that shorter links are sometimes "skipped over." We compared

volumes of adjacent links by length, and found 0.2km to be a safe threshold to avoid this issue and removed links below the threshold from our analysis.
- Strava uses the full OSM network, including sidewalks, footways, and transit-only facilities. This, combined with map matching errors, creates a problem of Strava volumes spreading across parallel links in the network (e.g., on bridges with side paths and auto/transit travel lanes or where separated cycle tracks run adjacent to streets). Since this posed a particular problem at many permanent counter locations, we decided to manually merge the volumes where multiple adjacent links paralleled a count site. We noted the issue was much less common with short-duration count sites, which tended to be conducted in simpler parts of the network. This issue would need to be addressed in an automated way—or ignored—for network-wide volume estimation with Strava data.

Count locations were joined to the nearest Strava/OSM network link (>0.2km in length only) and manually reviewed. We noted frequent inaccuracies in provided permanent counter spatial coordinates when compared with photos and Google Street View imagery. We adjusted count locations to improve accuracy, where doing so mattered for joining to the correct Strava/OSM data. After adjusting, there were nine SC sites and three PC (2% of total count sites), where Strava data were not available.

4.4 STREETLIGHT

StreetLight Data takes anonymized location records derived from mobile phones/devices (along with in-vehicle navigation systems). Through an internal process, StreetLight imputes travel mode for all trips. Bicycle trips can be converted into counts through several processes, including using user-created “gates,” which count all imputed bicycle trips that pass through a specific gate or contiguous set of gates. Counts are available as either a StreetLight Index value or as raw data; the latter data type was only available on request. The StreetLight Index is a normalized format designed to reduce variation, primarily from monthly sample size variations, and does not represent an actual estimated count, but rather a relative value that can be compared to other locations for the same time period. The raw data output is merely the number of trips recorded through the StreetLight data collection and imputation process and would represent some fraction of actual trips. After testing both the StreetLight Index and raw data and discussing the relative value of each format with StreetLight Data and the project team members, the project team opted to focus on the raw data as a model input.

The process of downloading StreetLight data started with creating gates, through which bicycle trips would be counted. To create gates, the project team uploaded sets of OSM segments associated with count locations. The segments were identified first by proximity, and then visually inspected to confirm that they appropriately represented the expected bicycle traffic route being counted. When uploaded into the StreetLight system, three gates are associated with each segment, reflecting a pathway through the segment along which bicyclists would be counted (see Figure 4-3). In most cases, the matched segments and auto-created gates were appropriate for use.

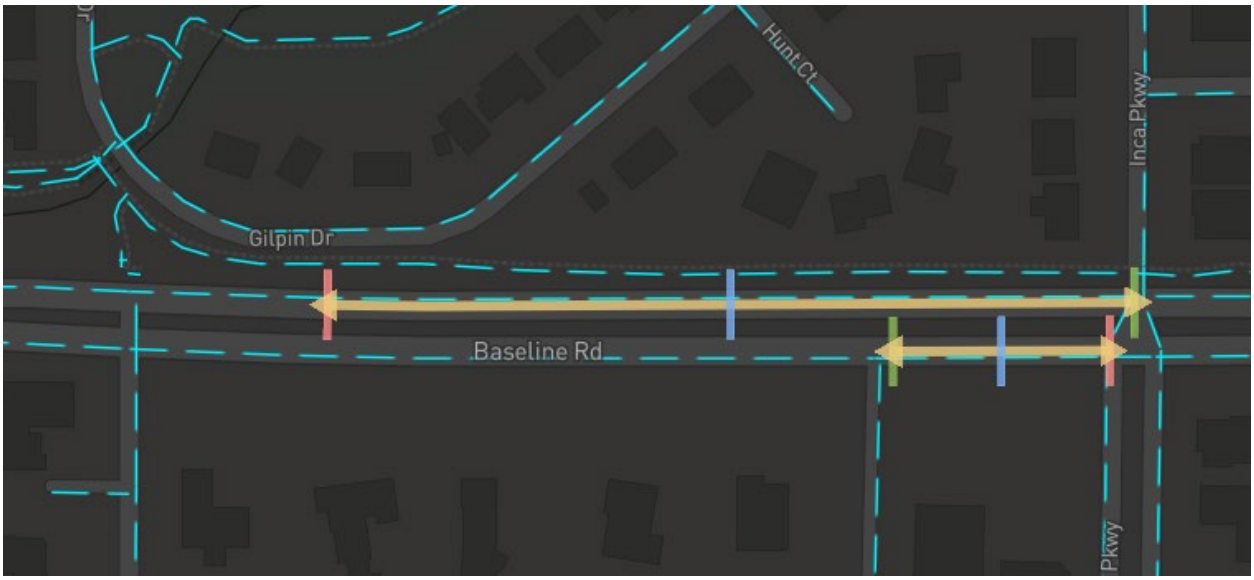


Figure 4-3 Two Example Segments, Each with a Set of Three Gates

In some cases, particularly for very long segments, as can exist for paths or for curved segments, auto-created gates were deemed to be imprecise. In those cases, gates could be moved or redrawn to reflect more accurately the expected paths of bicyclists crossing the counter locations.

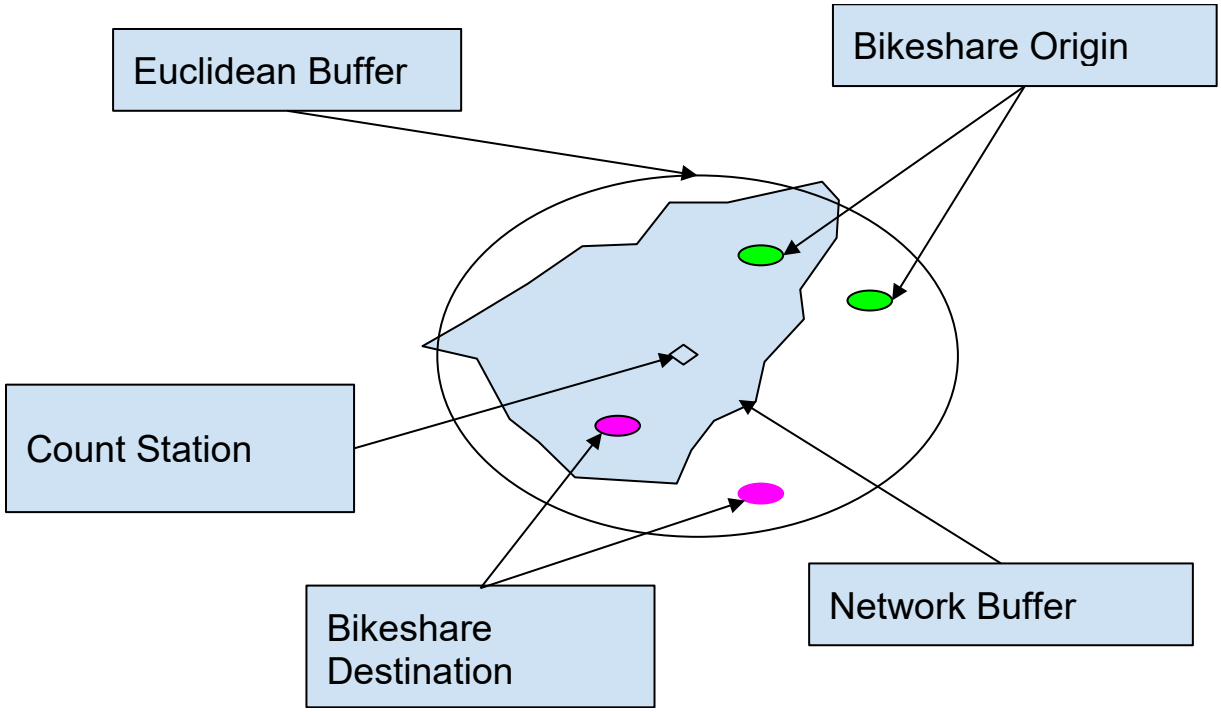
Once the research team was satisfied with the segment and gate placement, analyses were initiated in the StreetLight system to output the raw data values. Analyses could be run for a specific month (in 2018 or 2019), or for a selection of months (including a full year or more). Annual data were pulled for each location for 2018 and 2019 (2017 was not available). StreetLight output data included total trips (raw counts) and the average daily zone traffic, which is the StreetLight Index. For the count sites in the sample for 2018-2019, StreetLight data was not available at four locations (0.8%).

4.5 BIKESHARE

Bikeshare data were available as either continuous GPS latitude and longitude points for the entire trip (Bend, Eugene), or only the latitude and longitude at the start and end of each individual trip (Portland). Each trip has a unique route id, start-to-end date, time, and duration. Of the six study sites, bikeshare data were only available for the Oregon locations. Programs in other cities had either ceased operations (Charlotte, Dallas), or were too limited in coverage to use for this work (Boulder).

This study used bikeshare as an independent variable to estimate the AADBT. Four independent variables were generated from bikeshare data as follows: bikeshare trip origin count (production); bikeshare trip destination count (attraction); bikeshare crossings (straight line between origin and destination); and bikeshare original route crossing (for Eugene and Bend only). All these four variables were estimated for each count location separately for Euclidean or network buffers. Figure 4-4 shows the extraction of bikeshare origin or destination using the buffer around the count station.

Buffers around the count station were created, and the number of bikeshare origins or destinations within each buffer for each count location were counted.



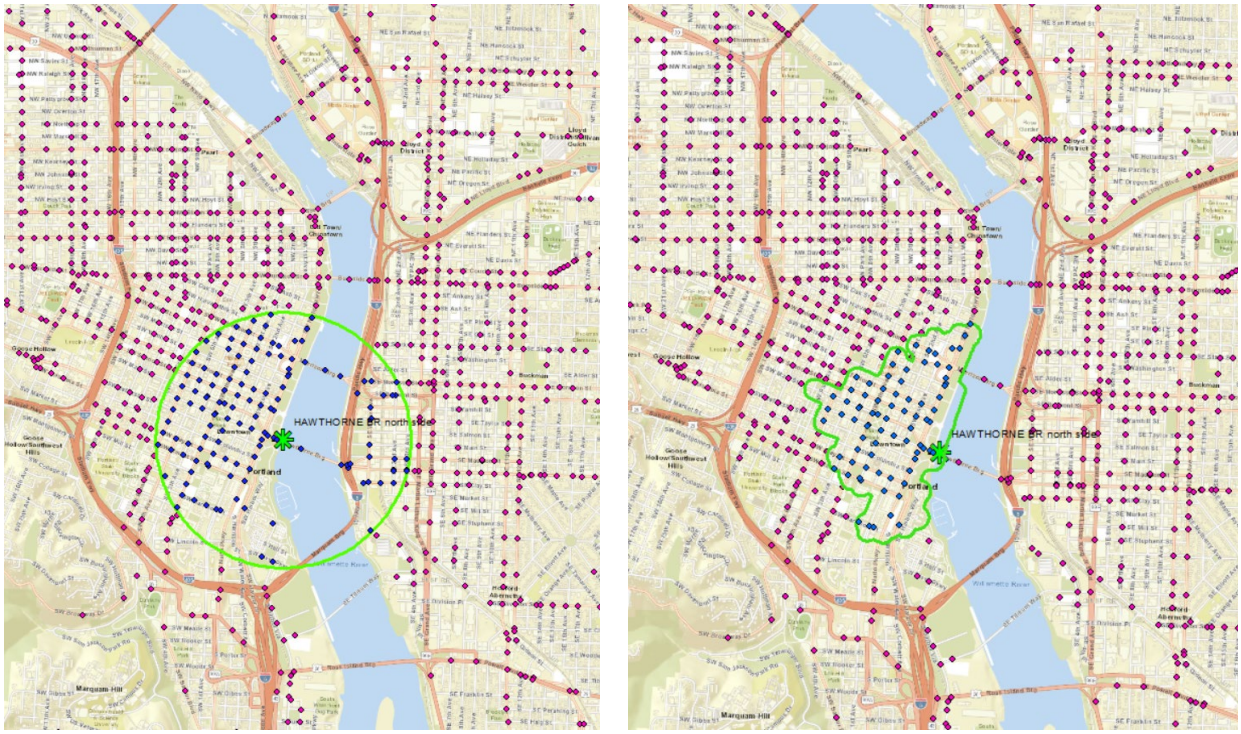


Figure 4-4 Counting of Origins or Destinations of Bike Trip

Figure 4-5 shows the bikeshare crossing count where the origin and destination of each trip were connected by a straight line and with the original route using continuous

latitude and longitude, then the number of lines passing over the buffer for each count station were counted. Please see 4.6.1 for details on Euclidean and network buffers.

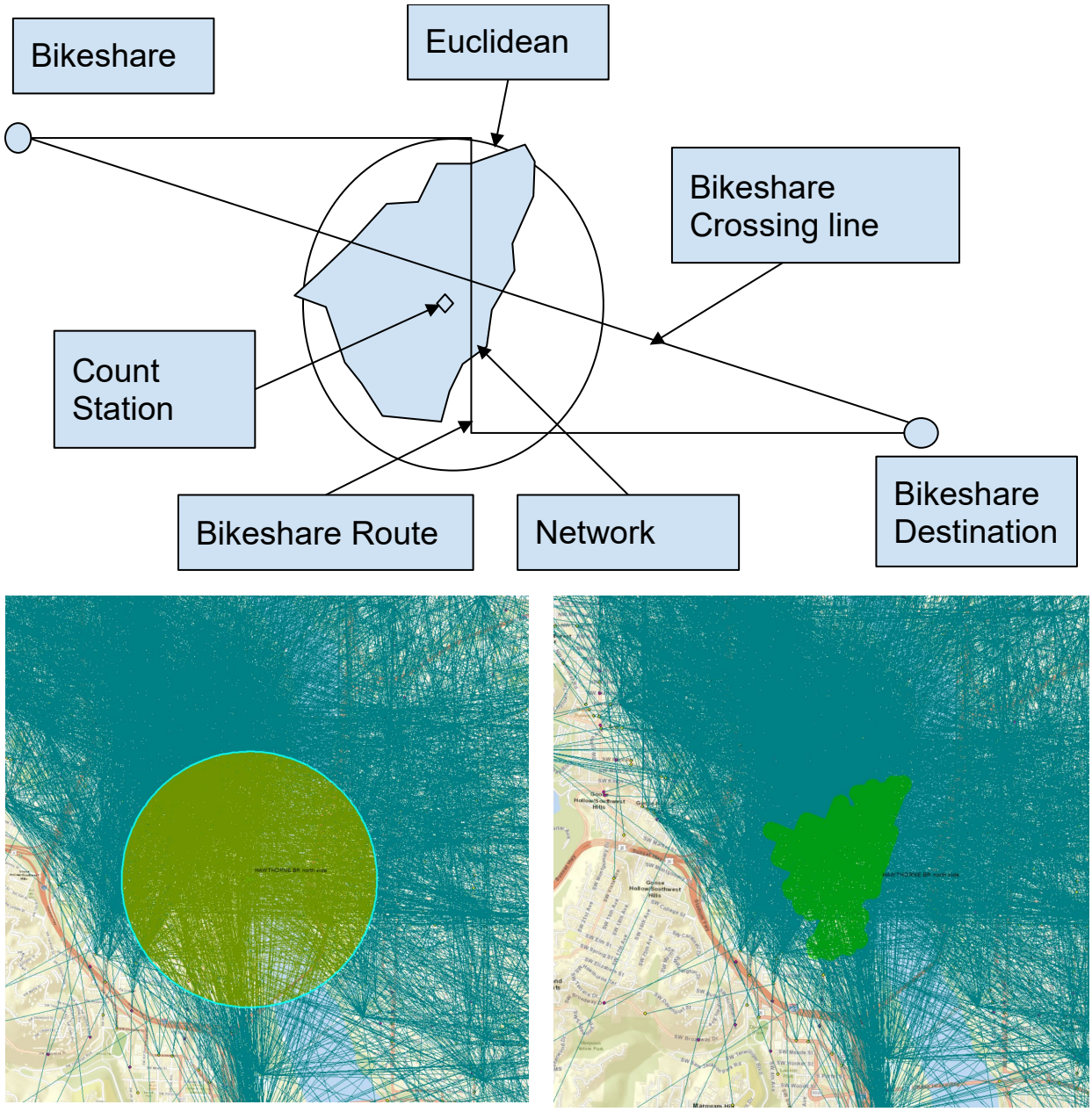


Figure 4-5 Bikeshare Crossing Count using Buffers

4.6 STATIC DATA

For the modeling process, the first step was the extraction of the variable feature set for all the selected sites using two buffer systems. Variables within the various categories - counts, network, land use, sociodemographic, and weather/temporal - were all selected based on their use and significance in prior research studies. The following sections describe the buffer systems and variable extraction within each of the buffers.

4.6.1 **Types of buffers**

Buffers show the area that is within some distance of the input features. Buffering plays an important role in many geoprocessing workflows to identify the features that are within a certain distance of other features. Two types of buffers were used in this study - air and network.

4.6.1.1 Air buffer

The air buffer is also known as the Euclidean buffer. It measures distance on a two-dimensional Cartesian plane. With Euclidean buffers, straight-line distances are calculated between two points on a plane. Euclidean buffers appear as perfect circles when drawn on a projected flat map. All the required shape files from 3D or existing coordinate reference systems were projected to 2D using the local coordinate reference system (CRS) (i.e., city of Portland local CRS: 2838, Dallas local CRS:2845). The dot buffer function on geodata frame was used to create a buffer around each count station.

4.6.1.2 Network buffer

Network buffers are created by solving from a given location outward along a travel network until some distance (or cost) threshold is reached. Typically, a second step applies some method to create a reachable area from all the links and/or nodes within the threshold. In this case, we solved from the nearest node to each count location along all OSM links open to bikes. The routable OSM network was acquired using the osmnx Python package (Boeing, 2017). We then extended buffers 60m (~200ft) outward from each reachable street centerline to create a network buffer area. The 60m buffer distance was chosen such that in a dense street grid (<100m blocks), adjacent buffers would cover the entire enclosed block in between.

4.6.1.3 Buffer size

Buffers around the count stations are needed to extract the independent variables for bicycle volume estimation. However, no consensus exists with respect to buffer size based on past research. Several researchers have used different buffer sizes to test the consistency of the results, with the buffer size ranging from 0.031 mile (50m) to 1.864 mile (3,000m) (Hankey et al., 2017a; Hankey and Lindsey, 2016; Lu et al., 2018; Strauss et al., 2013; Strauss and Miranda-Moreno, 2013; Griswold et al., 2011; Tabeshian and Kattan, 2014; Roll, 2018; Chen et al., 2017; Dill and Voros, 2007). Different buffer sizes may impact the outputs from the model. As the site characteristics vary, different sizes of buffer should be considered to get an accurate estimation of AADBT.

Four different sizes of buffer (0.5, 1.0, 1.5 and 2.0 miles) based on existing literature were considered in this study. The research team extracted all the listed variables in Table 4-7 through Table 4-10 for the four different buffer sizes for both Euclidean and network buffers. Some of the variables such as bike counts; bicycle functional class

(primary, secondary, residential, tertiary, path, cycleway, footway, cycleway lane, cycleway track); and speed limit were extracted at the link level. The research team developed an automatic python script to extract a wide range of variables for different buffer sizes. The Euclidean buffer was created using the dot buffer function on the geodata frame, while the network buffer was created using the GIS tool and passed through a python script to choose different sizes of buffers for different study areas. Data extraction time varied between 1.400 to 8.700 seconds depending on buffer size and number of count locations.

4.7 VARIABLE DESCRIPTIONS

This section describes the variables considered and/or used in the modeling process, broken down into count-related variables, network-related variables, sociodemographic variables, and weather-related variables. For each type of variable, a table is provided that includes descriptive information about the variable (including the different ways each variable was considered: air buffer, network buffer or link level), and information about what prior research has utilized similar variables for bicycle volume estimation purposes.

4.7.1 **Count-related variables**

Bicycle counts at specific counter locations, including permanent counts and short-duration counts, are key variables for estimated network volumes. Table 4-7 shows the count-related variables. Both the permanent and the short-term counts were run through the QA/QC process described previously to remove erroneous data and outliers. The Strava counts and the percentage of commuter Strava trips were obtained from Strava for each region. The percentage of commuter Strava trips for each segment is the ratio between the Strava commuter AADBT to the Strava AADBT. These variables were extracted as described previously. The StreetLight raw counts were obtained from the online portal. The bikeshare origins and destinations within each buffer were obtained by summing the number of productions and attractions. Lastly, the bikeshare crossing OD line was estimated as the number of straight lines between the origins and destinations crossing through the buffer around each count station.

Table 4-7 Count-Related Variables

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
no	no	yes	Permanent counts	City	Permanent count station for continuous counting throughout the year	4	(Wang *et al.*, 2016; Lu *et al.*, 2018; Dadashova *et al.*, 2020; Nelson *et al.*, 2021)
no	no	yes	Short-duration count	City	Temporary count station to collect data for certain time period	10	(Lindsey, 2011; Hankey and Lindsey, 2016; Jestico, Nelson and Winters, 2016; Wang *et al.*, 2016; Lu *et al.*, 2018; Roll and Proulx, 2018; Kwigizile, Oh and Kwayu, 2019; Roy *et al.*, 2019; Lin and Fan, 2020; Nelson *et al.*, 2021)
no	no	yes	Strava count	Strava Metro roll-ups	Strava AADBT or Strava commute or Strava non-commute count at street segments level	10	(Griffin and Jiao, 2015; Jestico, Nelson and Winters, 2016; Roll, 2018; Hochmair, Bardin and Ahmouda, 2019a; Kwigizile, Oh and Kwayu, 2019; Roy *et al.*, 2019; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
no	no	yes	Commuter or non-commuter Strava trips[1,2]	Strava	Percentage of Strava commuters or non-commuter trips in segment level	1	(Nelson et al., 2021)
no	no	yes	StreetLight count	StreetLight	StreetLight AADBT count at street segment levels	0	N/A
yes	yes	no	Bikeshare origin	City	Number of origins (O) of bikeshare trips (productions) within the buffer around the count station	0	N/A
yes	yes	no	Bikeshare destination	City	Number of destinations (D) of bikeshare trips (attractions) within the buffer around the count station	0	N/A
yes	yes	no	Bikeshare crossing	City	Number of bikeshare trips crossing over the buffer as a straight line between origin and destination	0	N/A
yes	yes	no	Bikeshare route	City	Number of bikeshare routes crossing its corresponding buffer	0	N/A

[1] straight-line, Euclidean buffers, varying radius 1/8 - 2 mi
[2] area based on max shortest-path distance around count/link location, varying distance 1/8 – 2 mi
[3] single value for specific link at location
N/A – no existing peer-reviewed studies identified in our review

4.7.2 **Built environment-related variables**

Characteristics of the built environment, particularly related to the transportation network, are important considerations in how and where people will travel. Table 4-8 shows the built environment-related variables that were extracted to be used in volume estimations. To keep the process standardized between the various regions, the BBBike network based on OpenStreetMap (OSM) was used for extracting these variables. First, the total length of the various facility types within the buffer were extracted. These included primary, secondary, tertiary, residential and path segments. Next, the cycle way, cycle track and footway within the buffer were extracted. Other variables extracted include speed limit, number of bicycle parking spots, number of bus stops, intersection density, bridges, roadway slope, water body area, park area, forest area, grass area, commercial area, industrial area, residential area, retail area, number of schools/colleges/universities and distance from the central business district.

Table 4-8 Built Environment Variables

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
yes	yes	yes	Primary	OSM for all cities (BBBike)	Two types of variables are used: (i) Total length of primary road segments within the buffer around each count station (ii) link type (0=absence and 1= presence).	1	(Dadashova et al., 2020)
yes	yes	yes	Secondary	OSM for all cities (BBBike)	Total length of secondary road segments within the buffer around each count station or link type (0=absence and 1= presence).	1	(Dadashova et al., 2020)
yes	yes	yes	Tertiary	OSM for all cities (BBBike)	Total length of tertiary road segments within the buffer around each count station or link type (0=absence and 1= presence).	1	(Dadashova et al., 2020)
yes	yes	yes	Residential	OSM for all cities (BBBike)	Total length of residential road segments within the buffer around each count station or link type (0=absence and 1= presence).	1	(Dadashova et al., 2020)
yes	yes	yes	Path	OSM for all cities (BBBike)	Total length of path segments within the buffer around each count station or link type (0=absence and 1= presence)	1	(Dadashova et al., 2020)
yes	yes	yes	Cycleway	OSM for all cities (BBBike)	Total length of cycle way segments within the buffer around each count station or link type (0=absence and 1= presence).	2	(Dadashova et al., 2020) (Griffin and Jiao, 2015)
yes	yes	yes	Cycleway_lane_all	OSM for all cities (osmnx)	Total length of cycleway lane, left and right segments within the buffer around each count station or link type (0=absence and 1= presence)	0	N/A
yes	yes	yes	Cycleway_track_all	OSM for all cities (osmnx)	Total length of cycleway track, left and right segments within the buffer around each count station or link type (0=absence and 1= presence)	0	N/A
yes	yes	yes	Footway	OSM for all cities (BBBike)	Total length of footway segments within the buffer around each count station or link type (0=absence and 1= presence)	1	(Dadashova et al., 2020)

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
yes	yes	yes	Speed limit	OSM for all cities (BBBike)	Speed limit on the link where the counter is situated; if unavailable the speed of the nearest link with the same functional class is extracted. For average speed within the buffer, the mode of speed for the functional class was obtained.	6	(Fagnant;, 2016; Jestico, Nelson and Winters, 2016; Roy *et al.*, 2019; Dadashova *et al.*, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	no	Number of Bicycle Parking Spaces	OSM for all cities (BBBike)	Count of bicycle parking spots within the buffer around the count station	3	(Jestico, Nelson and Winters, 2016; Hochmair, Bardin and Ahmouda, 2019a; Nelson *et al.*, 2021)
yes	yes	no	Number of Bus stops	OSM for all cities (BBBike)	Count of bus or rail stops within the buffer around the count station	5	(Strauss and Miranda-Moreno, 2013; Strauss, Miranda-Moreno and Morency, 2013; Tabeshian and Kattan, 2014; Hankey and Lindsey, 2016; Lu *et al.*, 2018)
yes	yes	no	Intersection Density	OSM for all cities (osmnx)	Number of intersections per square mile	4	(Hankey and Lindsey, 2016; Wang *et al.*, 2016; Lu *et al.*, 2018; Hochmair, Bardin and Ahmouda, 2019a)
yes	yes	yes	Number of lanes	OSM for all cities (BBBike)	Number of traffic lanes along corresponding count station street segment	6	(Tabeshian and Kattan, 2014; Fagnant;, 2016; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	yes	Presence of Bridges	OSM for all cities (BBBike)	Binary variable: 1=presence and 2=absence of bridges within the buffer around the count station	2	(Fagnant, 2016; Hochmair, Bardin and Ahmouda, 2019a)
yes	yes	yes	Roadway Slope	National Elevation Dataset	average absolute % slope along link	0	N/A
yes	yes	no	Water Body area or Distance to water body	OSM for all cities (BBBike)	Two variables were created: (i) Water body area within the buffer around the count station (ii) nearest distance to water body from the count station	5	(Hankey *et al.*, 2012; Wang *et al.*, 2016; Chen, Zhou and Sun, 2017; Ermagun, Lindsey and Hadden Loh, 2018; Nelson *et al.*, 2021)
yes	yes	no	Park Area or Distance to park	OSM for all cities (BBBike)	Two variables were created: (i) Park or open space area within the buffer around the count station, (ii) nearest distance to park area from count station	1	(Hankey and Lindsey, 2016)
yes	yes	no	Forest Area	OSM for all cities	Two variables were created: (i) Forest area within the buffer (ii) nearest	0	N/A

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
				(BBBike)	distance to forest area from count station		
yes	yes	no	Grass area or Distance to grass area	OSM for all cities (BBBike)	Grass area within the buffer around the count station or nearest distance to grass space from count station.	6	(Terri Pikora,, Billie Giles-Cortia, Fiona Bulla, b, Konrad Jamrozika, c, 2003; Noland, Deka and Walia, 2011a; Hochmair, Bardin and Ahmouda, 2019a; Roy *et al.*, 2019; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	no	Commercial area or Distance to commercial	OSM for all cities (BBBike)	Commercial area within the buffer around the count station or nearest distance to commercial area from count station.	2	(Strauss, Miranda-Moreno and Morency, 2013; Tabeshian and Kattan, 2014).
yes	yes	no	Industrial Area or Distance to industrial area	OSM for all cities (BBBike)	Industrial area within the buffer around the count station or nearest distance to industrial area from count station.	3	(Hankey and Lindsey, 2016; Wang *et al.*, 2016; Lu *et al.*, 2018)
yes	yes	no	Residential Area or Distance to residential area	OSM for all cities (BBBike)	Residential area within the buffer around the count station or nearest distance to residential area from count station.	4	(Lu *et al.*, 2018; Kwigizile, Oh and Kwayu, 2019; Roy *et al.*, 2019; Nelson *et al.*, 2021)
yes	yes	no	Retail Area or Distance to Retail area	OSM for all cities (BBBike)	Retail area within the buffer around the count station or nearest distance to retail area from count station.	3	(Griswold, Medury and Schneider, 2011; Hankey and Lindsey, 2016; Roll, 2018)
yes	yes	no	Number of Schools/ Colleges/ Universities or Distance to school/ college/ University	OSM for all cities (osmnx)	Number of schools (total count) or colleges (total count) or Universities (yes=1 and no=0) within the buffer around the count station or nearest distance to school/college/University	6	(Griswold, Medury and Schneider, 2011; Strauss and Miranda-Moreno, 2013; Strauss, Miranda-Moreno and Morency, 2013; Roll, 2018; Kwigizile, Oh and Kwayu, 2019; Nelson *et al.*, 2021)
no	no	no	Distance from CBD	OSM for all cities (BBBike)	Nearest distance to count station from City Hall (CBD)	>=2	Hankey et al., 2012; Lindsey et al., 2012 (possibly others)

[1] straight-line, Euclidean buffers, varying radius 1/8 - 2 mi
[2] area based on max shortest-path distance around count/link location, varying distance 1/8 – 2 mi
[3] single value for specific link at location
N/A = no peer-reviewed studies identified in our review

4.7.3 **Sociodemographic variables**

Characteristics of the population and employment for an area have been shown to help in estimating bicycle volumes. Table 4-9 shows the list of the sociodemographic variables that were extracted for the study sites. These included population density, employment density, number of jobs, household density, number of students, median age, percentage of males and females, percentage of African American and white population, median household income, and education. These were all extracted from the National Historical Geographic Information System (NHGIS).

Table 4-9 Sociodemographic Variables

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
yes	yes	no	Population density	NHGIS for all cities	Area weighted population per square mile within the block group (s) of the count station	17	(Noland, Deka and Walia, 2011b; Griswold, Medury and Schneider, 2011; Hankey *et al.*, 2012, 2017; Strauss and Miranda-Moreno, 2013; Wang *et al.*, 2014, 2016; Fagnant;, 2016; Hankey and Lindsey, 2016; Jestico, Nelson and Winters, 2016; Lu *et al.*, 2018; Roll, 2018; Hochmair, Bardin and Ahmouda, 2019a; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	no	Employment density	NHGIS for all cities	Area weighted employment per square mile within the block group (s) of the count station	8	(Jones *et al.*, 2010; Hankey *et al.*, 2012; Strauss and Miranda-Moreno, 2013; Strauss, Miranda-Moreno and Morency, 2013; Fagnant;, 2016; Hankey and Lindsey, 2016; Wang *et al.*, 2016; Chen, Zhou and Sun, 2017)
yes	yes	no	Number of Jobs	LEHD	Area weighted number of jobs within the block group (s) of the count station	0	N/A
yes	yes	no	Household Density	NHGIS for all cities	Area weighted households per square mile within the block group (s) of the count station	3	(Hankey and Lindsey, 2016; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020)
yes	yes	no	Number of students (student access)	NHGIS for all cities	Area weighted number of students (>12 grade) within the buffer within the block group (s) of the count station	1	(Roll, 2018)
yes	yes	no	Median age	NHGIS for all cities	Median age of the population within the buffer within the block group (s) of the count station	7	(Lindsey, 2011; Hankey *et al.*, 2012; Wang *et al.*, 2014; Chen, Zhou and Sun, 2017; Dadashova and Griffin, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	no	Percentage of female	NHGIS for all cities	Area weighted % of females within the buffer within the block group (s) of the count station	3	(Dadashova and Griffin, 2020; Dadashova *et al.*, 2020; Nelson *et al.*, 2021)

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
yes	yes	no	Percentage of male	NHGIS for all cities	Area weighted % of males within the buffer within the block group (s) of the count station	8	(Jestico, Nelson and Winters, 2016; Watkins *et al.*, 2016; Sanders *et al.*, 2017; Hochmair, Bardin and Ahmouda, 2019a; Kwigizile, Oh and Kwayu, 2019; Nickkar *et al.*, 2019; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020)
yes	yes	no	Percentage of African American population	NHGIS for all cities	Area weighted % of African American population within the buffer at count station block group	6	(Lindsey, 2011; Hankey *et al.*, 2012; Lindsey *et al.*, 2012; Wang *et al.*, 2014, 2016; Hochmair, Bardin and Ahmouda, 2019b)
yes	yes	no	Percentage of white population	NHGIS for all cities	Area weighted % of white population within the buffer within the block group (s) of the count station	3	(Chen, Zhou and Sun, 2017; Kwigizile, Oh and Kwayu, 2019; Roy *et al.*, 2019)
yes	yes	no	Median Household Income	NHGIS for all cities	Area weighted median household income within the buffer within the block group (s) of the count station	16	(Griswold, Medury and Schneider, 2011; Lindsey, 2011; Hankey *et al.*, 2012, 2017; Strauss and Miranda-Moreno, 2013; Tabeshian and Kattan, 2014; Wang *et al.*, 2014, 2016; Hankey and Lindsey, 2016; Lu *et al.*, 2018; Roy *et al.*, 2019; Hochmair, Bardin and Ahmouda, 2019b; Dadashova and Griffin, 2020; Dadashova *et al.*, 2020; Lin and Fan, 2020; Nelson *et al.*, 2021)
yes	yes	no	Education (% of population having at least college degree)	NHGIS for all cities	Area weighted % of population having at least a college degree within the buffer within the block group (s) of the count station	5	(Lindsey, 2011; Hankey *et al.*, 2012; Wang *et al.*, 2014; Ermagun, Lindsey and Hadden Loh, 2018; Nelson *et al.*, 2021)
yes	yes	no	Bicycle commuting	NHGIS	Four variables were used: (i) Total commuter, (ii) bike commuter (iii) % bike commute (iv) bike commuters/sq. mi	1	(Nelson et al., 2021)

[1] straight-line, Euclidean buffers, varying radius 1/8 - 2 mi
[2] area based on max shortest-path distance around count/link location, varying distance 1/8 – 2 mi
[3] single value for specific link at location
N/A = no peer-reviewed studies identified in our review

4.7.4 **Weather-related variables**

Table 4-10 shows the weather variables such as average humidity, precipitation, and average temperature, which were extracted from the Weather Underground website (https://www.wunderground.com/history).

Table 4-10 Weather Variables

Air[1]	Net[2]	Link[3]	Variables	Data Sources	Description	Prior studies utilizing variable to estimate volumes (count and references)	
no	no	yes	Average Humidity	Weather Underground for all cities	Average humidity at count station	2	(Strauss and Miranda-Moreno, 2013; Kwigizile, Oh and Kwayu, 2019)
no	no	yes	Precipitation	Weather Underground for all cities	Average precipitation at count station	9	(Niemeier, 1996; Hankey et al., 2012; Strauss and Miranda-Moreno, 2013; Wang et al., 2014; Fagnant;, 2016; Hankey and Lindsey, 2016; Ermagun, Lindsey and Hadden Loh, 2018; Esawey, 2018; Dadashova and Griffin, 2020)
no	no	yes	Average Temperature	Weather Underground for all cities	Average temperature at count station	9	(Niemeier, 1996; Lindsey, 2011; Hankey et al., 2012; Strauss and Miranda-Moreno, 2013; Wang et al., 2014; Fagnant;, 2016; Hankey and Lindsey, 2016; Esawey, 2018; Dadashova and Griffin, 2020)

[1] straight-line, Euclidean buffers, varying radius 1/8 - 2 mi
[2] area based on max shortest-path distance around count/link location, varying distance 1/8 – 2 mi
[3] single value for specific link at location

4.8 DATA SUMMARY

This chapter described the criteria developed for selecting sites, including an overview of a survey that was conducted to identify sites. Based on availability of count, Strava and StreetLight data, geographical variation in sites and contexts, six regions were chosen for this study. These included Bend, Boulder, Charlotte, Dallas, Eugene and Portland. At each of these locations, count, Strava, StreetLight, bikeshare, and static data were obtained, and the process for extracting and processing each of these data sources was described.

5 MODELING

We developed a range of models to better understand the likely feasibility and accuracy of predicting bicycle counts on a network. The models also allowed us to explore the relative value of static, third-party user, and other variables for modeling bicycle activity across a range of contexts. We also considered the transferability of models across time and location, as well as the potential for more flexible machine learning modeling techniques. The following sections describe the development of two types of models: traditional count and Random Forest (machine learning) models. Results are presented in each section, and an overall summary is also provided.

5.1 MODEL ROADMAP

Table 5-1 shows the overall modeling framework that was developed for this study. Three sets of models were specified – All City Pooled, Oregon Pooled and city-specific models. This allowed us to benefit from a larger sample size and range of contexts while also considering the benefit of region or city-specific models. We decided to model full-year PC locations in a separate model. This was primarily to evaluate accuracy and value added by different data sources against something close to “ground truth” counts of actual bicycling volumes. Sample size limitations restricted estimation of city-specific and machine learning models to select subsets of the data.

Table 5-1 Overall Modeling Framework

	Full-year Permanent	All Permanent + Short-duration
All City Pooled	Count - 2019 Machine Learning - N/A	Count - 2019 Machine Learning - 2019
Oregon Pooled	Count - N/A Machine Learning - N/A	Count - 2018 -2019 Machine Learning - 2019
City-specific	Count Model - Dallas only Machine Learning - N/A	Count - All Cities Machine Learning - Portland-2019, Eugene-2019

Across all the modeling streams, the crowdsourced data (Strava, StreetLight), and static location variables were added systematically to test their impacts on predicting AADBT, as shown in Table 5-2. Additional sources of user data including bikeshare were also evaluated where available.

Table 5-2 Count Model Specifications

Model	Specification
PM0:	AADBT=f(Static)
PM1:	AADBT=f(Strava)
PM2:	AADBT=f(StreetLight)
PM3:	AADBT=f(Strava + StreetLight)
PM4:	AADBT=f(Strava + Static)
PM5:	AADBT=f(StreetLight + Static)
PM6:	AADBT=f(Strava + StreetLight+ Static)

All models (count and machine learning) were developed using K-fold cross-validation with tenfolds and five repeats. K-fold cross-validation allows out-of-sample assessment

of model performance by holding back sets of data, and this external performance assessment helps avoid overfitting, providing a better sense of how the models might adapt to new data. In addition, we employed stratified resampling to ensure that test sets are reasonably balanced across AADBT ranges. All count modeling was implemented using the caret package in R (Kuhn, 2008).

5.2 COUNT MODELS

5.2.1 Estimation framework

For initial modeling, we tested two generalized linear model forms commonly used to model count-based data: the Poisson and negative binomial models. Both take the same basic form with differences related to assumptions about the error structure:

$AADBT_i = \exp\left(\sum_{n=1}^{n} \beta_n X_{i,n} + \varepsilon_i\right)$

Where, $AADBT_i$ is the average annual daily bike traffic at counter n; β_n is the co − efficient estimate; $X_{i,n}$ is the matrix of explanatory variables at site n and ε_i *is the error term*

Poisson models assume a fixed variance-to-mean ratio, while the negative binomial model relaxes that assumption by estimating a dispersion term from the data. When the Poisson error variance assumption is not met, the data are said to be “overdispersed,” and while coefficient estimates themselves remain consistent, standard errors will be biased and affect parameter significance testing. One solution is to specify a negative binomial model; another is to estimate the Poisson but calculate standard errors that are robust to overdispersion. The latter avoids the need to estimate an additional model parameter and, in this case, appeared to perform slightly better across a range of model selection metrics, as well as showing greater stability in parameters. For these reasons, we present results here from the Poisson model with robust standard errors, but it should be noted that the choice of model form here did not meaningfully change any of the results or conclusions.

This study collected a wide range of data (described in the Data Sources section), which was passed through automated scripts to process and generate data outputs for different buffer sizes and at the individual link level, as well as to standardize the permanent count data. After generating the estimation dataset, one additional Dallas counter (a loop path in a park) was removed due to having no nearby link in the OSM network that we could substitute. We retained another two count locations where an OSM link existed, but Strava reported no data, either due to privacy masking of low volumes or lack of user data and assumed a zero value for Strava volumes at those sites. StreetLight was able to provide data at all valid sites in the dataset.

Count model specifications were developed by drawing theoretically likely explanatory variables from the dataset and examining estimated coefficients for expected sign and statistical significance. In addition, all models were evaluated by a mix of standard metrics on the model’s in-sample fit as well as out-of-sample prediction performance (using statistics from the cross-validation described previously): primarily the Akaike

Information Criterion (AIC) and Root Mean Square Error (RMSE). We also examined prediction performance segmented by volume bin and region (for pooled models).

5.2.2 Model data overview

Most count sites were retained for modeling, with a few exceptions to ensure reasonable AADBT estimates and comparability across models:

- Locations that could not be matched to the Strava/OSM network (n=9) [Note: of the nine, four were also unavailable from StreetLight]
- One short-duration location (Boulder) with zero total counts, since AADBT could not be calculated properly (n=1)

To reduce the search set of explanatory variables to something meaningful, based on bivariate correlations with AADBT, we retained variables measured within the buffers at the following distances only: half-mile air, one-mile air, half-mile network, one-mile network, and two-mile network.

5.2.3 Full-year permanent count models

We developed models with full-year 2019 PC data only, primarily as a test of the relative value of the third-party user data sources. Full-year PCs served as our closest measure to ground truth cycling activity.

5.2.3.1 Data

Table 5-3 and Figure 5-1 summarize the valid PC data used in the models. The number of counters ranged from three to 23 per city. Only Dallas, with 23 PCs, had sufficient data to reasonably specify a city-specific model. Overall, AADBT for the PC sites ranged from 12 to 1,775. With the exception of Bend, it was not particularly surprising that sites selected for PCs tended to be in moderate- to high-volume locations.

Table 5-3 Full-year 2019 PC Data Used for Modeling by Region

	Total Number of Count Stations	Mean AADBT ±SD	Min AADBT	Max AADBT
Portland, OR	4	1375 ±396	842	1775
Bend, OR	3	78 ±44	46	128
Eugene, OR	13	308 ±197	59	683
Charlotte, NC	9	349 ±286	33	727
Boulder, CO	8	393 ±174	180	607
Dallas, TX	23	312 ±320	12	1091
All Sites	*60*	*387* ±379	*12*	*1775*

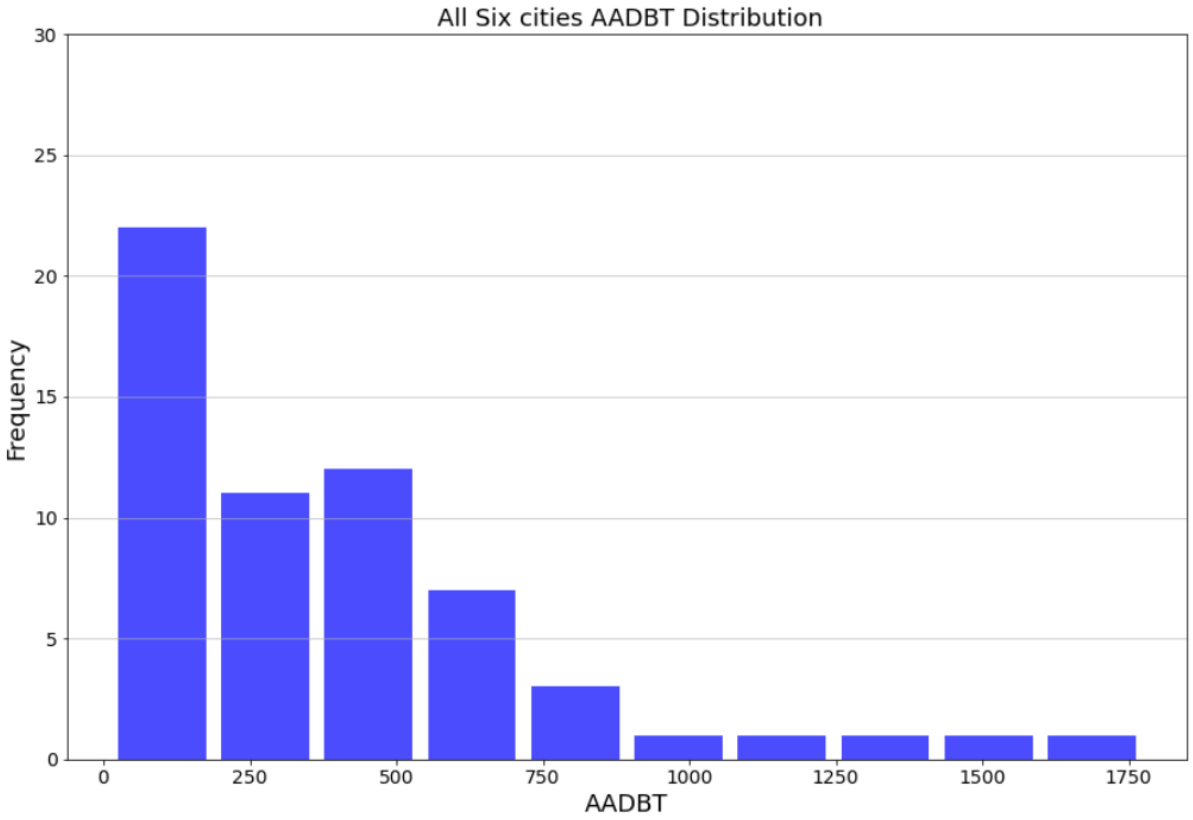


Figure 5-1 Histogram of All City Pooled Model Data

We considered a broad range of explanatory variables when specifying the full-year PC models (see Chapter 4). In some cases, variables were modified from their raw form. Table 5-4 provides a reference of variables retained in the final, best-performing All City Pooled and Dallas models. Bike and road facility type were measured at the count location itself. Other variables were measured within a half- or one-mile radius (miles of designated bike facilities, street intersection density, number of K-12 schools and colleges, and acres of park land). Proximity to a major university was measured as whether or not one was within two miles along the street and bike network (excluding facilities closed to biking).

Table 5-4 Explanatory Variable Reference for Full-year PC Count Models

Variable	mean (All Cities)	mean (Dallas)	def
log(stv_adb + 1)	-	2.73	natural log of average daily total Strava counts
log(stv_c_adb + 1)	1.69	-	natural log of average daily “commute” Strava counts
log(stv_nc_adb + 1)	2.31	-	natural log of average daily “non-commute” Strava counts
log(stl_raw + 1)	6.72	5.94	natural log of total annual StreetLight bike trips
log_stv_stl	9.38	8.68	ln(stv_adb +1) + ln(stl_raw + 1)
sep_bikeway_binary	0.58	*0.52	Location on path or cycleway
cycleway_lane_binary	0.08	-	Location in bike lane

Variable	mean (All Cities)	mean (Dallas)	def
arterial_binary	0.13	-	Location on primary or secondary arterial
BikeFac_hm	5.10	1.53	Miles of dedicated bike facility (on or off-street) in half-mile radius (does NOT include just signed bike routes or shared lanes)
intersection_density_om	147.97	114.80	street intersections per sq. mi. w/in 1-mi radius
University_tm_net_binary	0.50	0.22	Location w/in 2 network miles of a University
Schools_hm + Colleges_hm	1.25	0.43	Number of schools and colleges w/in 0.5-mi radius
Park_acres_om	192.50	379.05	Acres of park w/in 1-mile radius
Distance_to_Water_mi	-	0.49	Miles to nearest water body

* An additional 8 (35%) locations were on shared footways

5.2.3.2 Results

Table 5-5 summarizes prediction performance of the All City Pooled and Dallas count models at PC locations with a full year (10 or more months) of 2019 data. Test percent RMSE (Root Mean Square Error, presented as % of mean AADBT for comparability) was the primary prediction summary statistic considered. Test MAPE (Mean Absolut Percent Error) by volume tertile is also presented. All statistics were calculated as the mean value among hold-out sites (i.e., those sites not included in estimation for a given iteration) across the tenfolds and five repetitions. This provides a stronger test of model performance and one more in line with what we might expect when applying the model to new data. Standard errors associated with the mean values presented are omitted here for brevity and were generally relatively small (<15% of the mean for all volume bins). Full model specifications and model fit are presented in Table 5-5 and Table 5-6.

Table 5-5 Full-year PC Count Models Prediction Performance Summary

	PM0 Static	PM1 STV	PM2 SL	PM3 STV+ SL	PM4 STV+ Static	PM5 SL+ Static	PM6 All
All City Pooled (n=60)							
%RMSE	52%	52%	72%	49%	47%	51%	**42%**
MAPE (low vol. <150, n=20)	345%	310%	324%	176%	264%	186%	**167%**
MAPE (mid vol. <468, n=20)	39%	64%	50%	53%	34%	37%	**31%**
MAPE (high vol. <=1775, n=20)	29%	**27%**	38%	30%	29%	31%	29%
Dallas (n=23)							
%RMSE	102%	54%	64%	29%	39%	67%	**26%**
MAPE (low vol. <64, n=8)	369%	213%	117%	**55%**	226%	117%	86%
MAPE (mid vol. <376, n=8)	86%	46%	54%	31%	32%	49%	**14%**
MAPE (high vol. <=1091, n=7)	28%	30%	45%	15%	**13%**	47%	18%

Notes: All statistics shown are mean of cross-validation test (out of sample) performance, **best-performing model specification for each measure is shown in bold**. *n* refers to number of count sites, not the number of iterations for statistic calculation, which is always 10 folds times 5 repeats equals 50.

The primary motivation for the full-year PC models was to assess the contribution of each of our data sources. Static variables representing the count location context were estimated as a baseline reflecting typical direct demand modeling practice. In general, the three data sources (static, Strava, and StreetLight) appeared to be complementary to one another; that is, adding any two data sources together tended to outperform each data source on its own. In the All City Pooled model, the combination of all three variable types (PM6) was the best-performing model by most measures. In Dallas, where PCs were almost entirely at off-street locations, the combination of Strava and StreetLight data (PM3) performed about as well as, and by some measures slightly better than, the combination of those data sources and the static context variables (PM6). In Dallas, Strava and StreetLight were particularly complementary, with Strava performing better at mid- to high-volume spots, while StreetLight was a better option at low-volume locations. This result seems consistent with the known potential biases in Strava data (rounding and user base) being magnified at low-volume locations. Keeping in mind the limited sample size, these results are interesting and mostly in keeping with expectations that each source is providing unique and valuable information about bicycling activity.

The All City Pooled baseline static model fit the data well, and all coefficients had the expected signs (Table 5-6). We chose to retain the full set of static variables, even in models where they failed to reach standard significance levels. This was partly to provide a consistent test of the performance of emerging variables. It was also done with expanding predictions to the network in mind, where the joint performance and effects of related variables (like facility types) are more important than any individual variable's contribution to model fit. The signs and relative size of the parameters remained relatively constant, suggesting a stable model structure.

Although the All City Pooled models fit the data relatively well, as shown by the high pseudo-R^2 values, prediction success varied considerably by volume. Low-volume sites proved challenging, with the best-performing model still demonstrating considerable prediction error (167% MAPE). It is important to remember that even relatively small errors in predicting the number of cyclists can lead to large relative errors at low-volume sites (e.g., at the lowest-volume site included here) being off by 10 cyclists per day results in an 83% MAPE. Prediction at low-volume sites is also made more challenging by the lack of variety in count locations. Where cycling volumes are low, permanent counts tend to be conducted only at off-street locations, and that was true here. Among the lowest-volume third of full-year PC sites, 75% were off-street locations. Prediction success at mid- (>149) and high-volume (>467) sites was more promising at 31% and 29% MAPE, respectively, using the best-fitting overall model.

The best-performing Dallas-only model (Table 5-7, PM6), unsurprisingly, displayed generally better model fit and prediction performance than the respective All City Pooled model (and this remained true even when looking at performance of the All City Pooled

model specifically in Dallas). Establishing a good static baseline model was challenging due to a lack of variation in count site facilities. All but two of the 23 sites were on off-street trails (and one of the remaining included a sidewalk count), 17 were in parks, and 16 were within half a mile of a water feature. All of these suggest predominately recreational use patterns. The only significant variables were nearby park acres and presence of a designated, separate bike facility. A handful of others were retained from the All City Pooled model based on model fit and reasonable coefficient signs and magnitudes. While the best-fitting model still included all three sets of variables (static, Strava, and StreetLight), a model combining just Strava and StreetLight data performed about as well in terms of predictive performance. In terms of MAPE, expected performance was better than in the All City Pooled model, with best MAPE less than 20% at mid- and high-volume sites, and low-volume MAPE as low as 55%.

Figure 5-2 (All City Pooled model) and Figure 5-3 (Dallas model) show (test) prediction performance for the best-fitting models. Unfortunately, due to low numbers of full-year PC counters in other locations, comparable city-specific models could not be estimated reliably. Because of this, we cannot comment on whether city-specific model performance in other locations would show similar improvements over the All City Pooled model. We explore this question further with the full set of count data in the next section.

Table 5-6 AADBT Poisson All City Pooled Model Full-year PC Location Results (10-fold cross-validation with five repeats and robust SEs)

parameter estimate w/ significance level: .<=0.1, * <=0.05 , **<=0.01 , ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	3.4627***	4.4602***	2.0671**	2.9164***	3.4197***	1.7455***	2.2601***
log(stv_adb +1)					0.3362***		
log(stv_c_adb +1)		0.5514***		0.3308***			
log(stv_nc_adb +1)		0.1215*		0.1715***			
log(stl_raw + 1)			0.5383***	0.2583***		0.3890***	
log(stv_stl)[a]							0.2612***
sep_bikeway_binary	0.5047**				0.326.	0.6065***	0.4185**
cycleway_lane_binary	0.5826*				0.1895	0.3528	0.1014
arterial_binary	-0.852***				-0.6529***	-0.8371***	-0.6731***
BikeFac_hm	0.08**				0.0773**	0.0406.	0.0553**
intersection_density_om	0.0061***				0.0028*	0.0018	0.0006
School_hm + college_hm	0.1311*				0.0949*	0.1178***	0.0899**
Park_acres_om	0.0023***				0.0012***	0.0015***	0.0009**
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	60	60	60	60	60	60	60
AIC[b]	6535	7438	10443	6378	4692	4925	4033
Pseudo-R^2[c]	0.692	0.646	0.494	0.670	0.786	0.774	0.819
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	192 (+/- 10)	194 (+/- 8)	275 (+/- 14)	166 (+/- 8)	165 (+/- 9)	190 (+/- 9)	153 (+/- 7)
MAPE (+/- SE)	138% (+/- 10)	128% (+/- 8)	129% (+/- 11)	91% (+/- 6)	110% (+/-11)	87% (+/- 5)	77% (+/- 6)
MAE (+/- SE)[d]	157 (+/- 9)	152 (+/- 6)	200 (+/- 11)	128 (+/- 6)	127 (+/- 7)	149 (+/- 7)	121 (+/- 5)

[a] combined variable formulated as (log(stv_adb +1) + log(stl_raw + 1))
[b] Akaike Information (Loss) Criterion (lower is better)
[c] McFadden's R^2 = 1 – (Deviance in Final Model / deviance in Intercept-only Model, higher is better)
[d] Mean absolute error (unit=AADBT)

Table 5-7 AADBT Poisson Dallas Model Full-year PC Location Results (10-fold cross-validation with five repeats and robust SEs)

parameter estimate w/ significance level . <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	2.9443***	3.3860***	1.1457	1.0088**	3.0208***	2.0726***	1.4502***
log(stv_adb +1)					0.4502***		
log(stv_c_adb +1)		0.6500***					
log(stv_nc_adb +1)							
log(stl_raw + 1)			0.6803***			0.3276*	
log(stv_stl)[a]				0.4503***			0.3613***
sep_bikeway_binary	0.5994*						
cycleway_lane_binary							
arterial_binary							
BikeFac_hm	0.1299				0.2061***		0.1179***
intersection_density_om	0.0029				0.0021.	0.0075*	
University_tm_net	0.9867				0.4345		
School_hm + college_hm	0.2011						
Park_acres_om	0.0030***				0.0012***	0.0014**	0.0006***
Distance_to_Water_mi						-0.5472***	
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	23	23	23	23	23	23	23
AIC[b]	1430	1225	2281	579	572	1166	415
Pseudo-R^2[c]	0.825	0.853	0.706	0.942	0.944	0.862	0.965
Cross-validation test performance (mean of 10 folds x 5 repeats)							
RMSE (+/- SE)	194 (+/- 10)	115 (+/- 8)	179 (+/- 22)	73 (+/- 10)	74 (+/- 6)	173 (+/- 19)	70 (+/- 10)
MAPE (+/- SE)	127% (+/- 10)	90% (+/- 12)	71% (+/- 5)	34% (+/- 2)	83% (+/- 17)	70% (+/- 6)	38% (+/- 5)
MAE (+/- SE)	157 (+/- 9)	102 (+/- 8)	147 (+/- 18)	62 (+/- 8)	64 (+/- 6)	141 (+/- 15)	58 (+/- 8)

[a] combined variable formulated as (log(stv_adb +1) + log(stl_raw + 1))
[b] Akaike Information (Loss) Criterion (lower is better)
[c] McFadden's R^2 = 1 – (Deviance in Final Model / deviance in Intercept-only Model, higher is better)

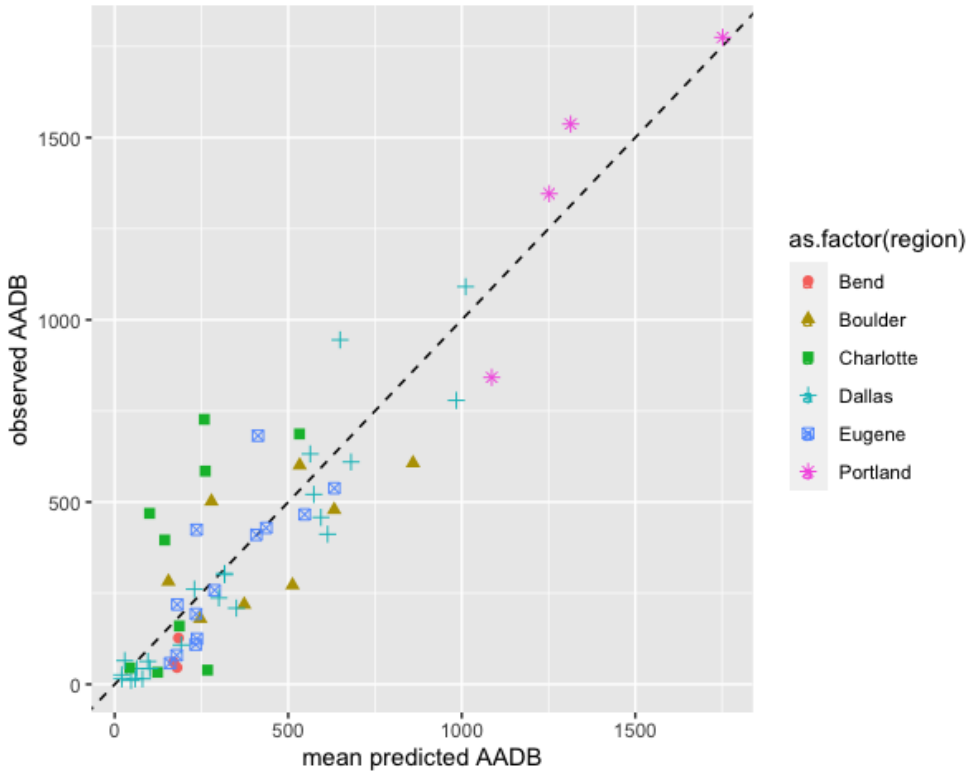


Figure 5-2 All City Pooled Model PM6 Prediction Plot (Full-year PC Data)

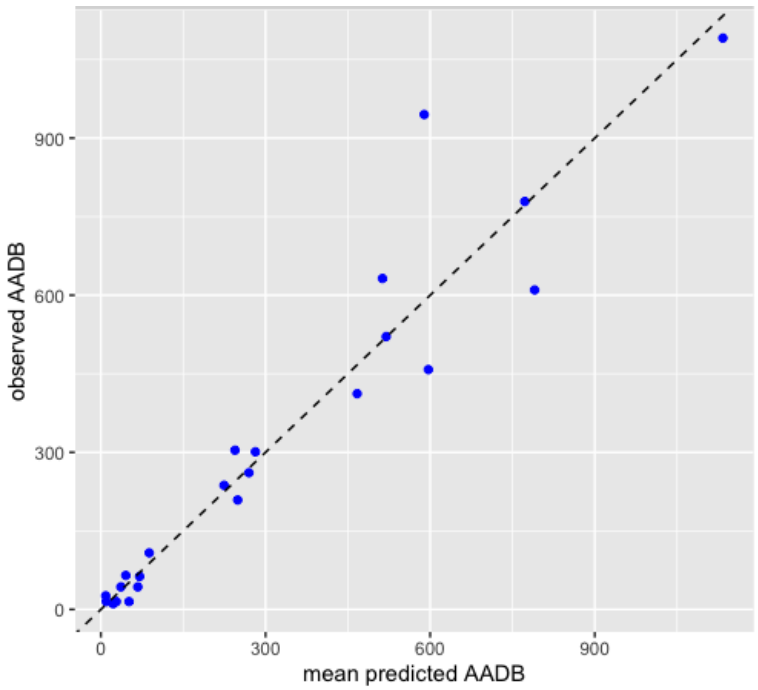


Figure 5-3 Dallas Model PM6 Prediction Plot (Full-year PC Data)

Comparison to Recent Research on Motor Vehicle AADT Estimation Using StreetLight Data

Unfortunately, existing models of bicycle volume estimation using third-party user data had too many differences in design, data, and reporting to directly compare with results presented here. However, a comparable effort was recently completed testing StreetLight's AADT product as a predictor of motor vehicle AADT under the Transportation Pooled Fund program (TPF-5(384), https://www.pooledfund.org/Details/Study/636).

StreetLight motorized AADT estimates differ from the raw bicycle counts used here in that they include sampling and contextual adjustment factors, based on sociodemographic, weather, and extensive traffic monitor calibration data (StreetLight Data, 2019). The StreetLight AADT estimates most closely resemble our static plus Streetlight (or Strava) data.

Comparison of overall results makes little sense, since motor vehicle counts tend to concentrate on facilities with much higher volumes than almost any bicycle site, and it is generally accepted that estimation errors are inversely related to volume. Two separate validations were performed for the Pooled Fund study (Fish et al. [NREL], 2021; Tsapakis et al. [TTI], 2021). The NREL study helpfully provided their validation data for download, and we re-binned their results to line up with our All-City model volume tertiles (https://github.com/NREL/fhwa-streetlight-aadt-validation/blob/master/inpt_data/nrel_aadt_results_06022021.csv). The TTI reporting bins quite a bit different but are included with that caveat. The table below compares the AADT results with our models PM4-PM6 predicting AADBT.

Comparing results of bicycle and motor vehicle AADT using third-party user data

	All City Pooled AADBT				NREL AADT		TTI AADT[a]	
	PM4 STV +Static	PM5 SL +Static	PM6 STV +SL +Static	n	SL AADT	n	SL AADT	N
%RMSE (volume 150-467)	38%	42%	36%	20	39%	9	58%	5
%RMSE (volume 468-1775)	30%	32%	28%	20	37%	38	63%	65
Combined %RMSE	34%	37%	32%	40	38%	47	63%	70
MAPE (volume 150-467)	34%	37%	31%	20	31%	9	52%	5
MAPE (volume 468-1775)	29%	31%	29%	20	28%	38	23%	65
Combined MAPE	32%	34%	30%	40	29%	47	25%	70

[a] Reporting bins for TTI/Cambridge (bidirectional) results differ from others: 237-499 & 500-4,999

Most of the AADT locations had higher volumes than any of our bike locations, and those tended to be predicted with more accuracy (NREL: overall %RMSE=9%, MAPE=19%; TTI/Cambridge: overall bidirectional %RMSE=25%, MAPE=15%). When comparing similar volume ranges, AADBT estimation using static plus third-party user data appears to be on par with motorized AADT estimates. The authors of both AADT reports rightly note that AADT imputed from short-duration counts is itself not without significant error. If a goal is to use third-party data as an equivalent alternative to SC counts, the performance target is 0% error, but maybe something on the order of 13% RMSE and 10% MAPE at lower-volume (AADT) locations may be acceptable (Hallenbeck et al., 2021).

5.2.4 **Full-year permanent plus short-duration count models**

To expand our sample, we also developed models using all available counters across the six regions. The larger sample allowed us to estimate additional count model specifications and test more subsets of the data, including an Oregon Pooled model and city-specific models beyond Dallas. We also tested more flexible machine learning modeling techniques, which are presented in a separate section. This section presents count model results from the full count dataset.

5.2.4.1 Data

Table 5-8 and Figure 5-4 summarize the count data used in the models. For less than full-year PC and all SC locations, AADBT was calculated using a factor group approach (see Factoring Approaches, Section 4.2.3). The number of counters ranged from 14 to 104 per city. Although there is some imbalance in the sample across regions, we decided not to weight model estimation by region. Since the goal was to maximize the range of contexts considered and, hopefully, improve transferability to other settings, we felt this was the best tradeoff. Overall, AADBT for the PC sites ranged from 1 to 2,647 (note: counters with zero counts could not be properly expanded and were dropped; see Chapter 4). The inclusion of SC and less than full-year PC sites resulted in a higher share of low-volume sites in the sample than for the full-year PC-only models.

Table 5-8 Combined Usable PC and SC Locations Used for Modeling by Region and Year

	Total Count Locations			Mean AADBT +-SD			AADBT Range		
	2017	2018	2019	2017	2018	2019	2017	2018	2019
Portland, Oregon	104	33	88	*334± 470*	499± 634	427 ±496	9-2647	43-2594	6 - 1951
Bend, Oregon	0	58	63	0	97 ±96	62 ±62	0	5-393	1 - 344
Eugene, Oregon	0	86	76	0	224 ±294	205 ±270	0	13-1930	5 - 1590
Charlotte, North Carolina	n.c.	n.c.	14	n.c.	n.c.	271 ±264	n.c.	n.c.	22 - 727
Boulder, Colorado	n.c.	n.c.	39	n.c.	n.c.	162 ±230	n.c.	n.c.	1 - 1086
Dallas, Texas	n.c.	n.c.	31	n.c.	n.c.	320 ±336	n.c.	n.c.	12 - 1135
All Sites	*104*	*197*	*311*	*334± 470*	*234 ±371*	*248 ±355*	*9-2647*	*5-2594*	*1 - 1951*

n.c. = not considered

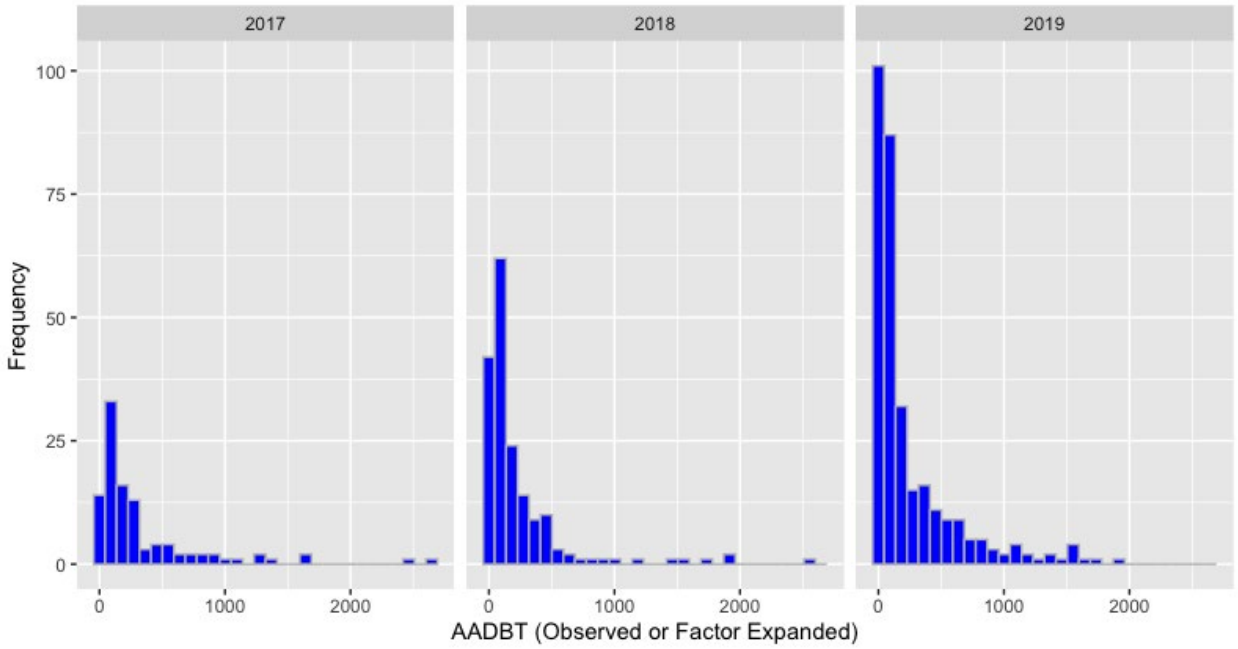


Figure 5-4 Histograms of Full Count Dataset by Year

We considered a broad range of explanatory variables when specifying models (see Chapter 4). In some cases, variables were modified from their raw form. Table 5-9 provides a reference of variables retained in the final, best-performing models. Bike and road facility type were measured at the count location itself. Other variables were measured within a half- or one-mile radius (miles of designated bike facilities, street intersection density, number of K-12 schools and colleges, and acres of park land), as noted in Table 5-9. We tested variables measured within shortest-path network buffers, but those specifications did not perform noticeably better than the straight-line buffers in terms of predictive performance and model fit. We retained the straight-line distance-based variables for their simplicity and ease of expansion to network-wide application, where the computational burden of solving paths would be much greater.

Several variations of bikeshare activity variables were tested as part of model development for the Oregon locations. While none were retained in the final models, bikeshare variables on their own had considerable explanatory power. It appears that bikeshare (or, perhaps, where bikeshare operates) might be a proxy for a range of contextual variables and might serve as a potential simple adjustment factor for count modeling where it exists. The significance of bikeshare faded away after specifying the full range of static variables.

Finally, we should note that the 2018 and (Portland-only) 2017 models were minimally re-specified from their corresponding 2019 versions. We felt this would result in a more revealing test of transferability than attempting to optimize the models to each year. We only summarize the modeling for the earlier years in the main text, but the Appendix provides full details.

Table 5-9 Explanatory Variable Reference for Combined PC and SC Data Count Models (2019 Only)

Variable	def	mean							
		All	Oregon	Bend	Boulder	Charlotte	Dallas	Eugene	Portland
log(stv_adb + 1)	natural log of average daily total Strava counts	2.17	2.02	1.26	2.64	2.09	2.78	1.48	3.02
log(stv_c_adb + 1)	natural log of average daily "commute" Strava counts	1.47	1.49	0.68	1.44	1.20	1.50	0.92	2.56
log(stv_nc_adb + 1)	natural log of average daily "non-commute" Strava counts	1.75	1.51	0.98	2.41	1.77	2.64	1.11	2.24
log(stl_raw + 1)	natural log of total annual StreetLight bike trips	6.47	6.77	6.34	5.16	6.53	5.89	6.37	7.43
log_stv_stl	Combined Strava & Streetlight: ln(stv_adb +1) + ln(stl_raw + 1)[a]	8.64	8.79	7.60	7.80	8.62	8.67	7.85	10.45
arterial_binary	Location on primary or secondary arterial	0.22	0.24	0.44	0.23	0.29	0.03	0.16	0.18
Bike.Commuter_hm	Number of bike commuters (ACS) within ½-mi radius	187.12	234.37	46.24	116.48	13.63	8.33	175.77	418.18
Bike.Commuter_om	see above, w/in 1-mi radius	668.48	843.70	171.25	373.53	62.75	30.05	623.90	1509.79
BikeFac_hm	Miles of dedicated bike facility (on or off-street) in half-mile radius (does NOT include just signed bike routes or shared lanes)	5.03	5.16	5.04	6.00	7.22	1.85	6.81	3.80
BikeFac_om	see above, w/in 1-mile radius	18.63	19.37	19.54	20.30	31.19	5.46	24.04	15.16
cycleway_binary	location on dedicated OSM cycleway	0.19	0.18	0.11	0.03	0.36	0.42	0.39	0.05
cycleway_lane_binary	Location in bike lane	0.13	0.13	0.13	0.26	0.00	0.00	0.16	0.11
Distance_to_CBD_mi	straight-line distance to City Hall	3.74	2.69	2.50	8.63	3.28	5.51	2.41	3.07
Log(Distance_to_CBD_mi+1)	natural log of above	9.40	9.18	8.77	10.23	8.97	10.13	9.17	9.49
Distance_to_Water_mi	Straight-line distance to nearest water body	0.55	0.58	0.46	0.53	0.29	0.52	0.62	0.63

Variable	def	mean							
		All	Oregon	Bend	Boulder	Charlotte	Dallas	Eugene	Portland
footway_binary	Location on footway (primarily pedestrian)	0.04	0.01	0.00	0.00	0.00	0.35	0.00	0.02
intersection_density_hm	street intersections per sq. mi. w/in 1/2-mi radius	175.57	205.71	177.32	56.19	136.96	122.52	153.17	271.40
intersection_density_hm^2	square of above	39698	48798	35394	10106	20490	18966	26719	77463
intersection_density_om	street intersections per sq. mi. w/in 1-mi radius	165.77	193.10	163.81	51.97	140.67	120.15	144.19	256.31
intersection_density_om^2	square of above	34658	42426	29290	8231	20924	17229	22510	69030
Median_HH_income_om	Area weighted block group median household income w/in 1-mi radius	53731	54068	54531	86736	24155	23097	38787	66935
log(min_dist_to_university)	natural log of the straight-line distance to nearest major university	9.00	8.80	8.95	9.99	8.44	9.47	8.82	8.67
Park_acres_hm	Acres of park w/in 1/2-mile radius	26.95	14.66	13.06	4.64	40.74	138.77	19.91	11.19
path_binary	Location on off-street bike path	0.04	0.03	0.00	0.03	0.00	0.13	0.00	0.07
pct_at_least_college_education_hm	Area weighted block group % of population having at least a college degree w/in 1-mi radius	78.35	78.30	76.91	84.84	84.68	67.65	72.33	84.45
primary_binary	Location on primary arterial	0.03	0.03	0.10	0.03	0.00	0.03	0.01	0.00
secondary_binary	Location on secondary arterial	0.19	0.21	0.34	0.21	0.29	0.00	0.14	0.18
(secondary_binary + tertiary_binary)	Location on secondary or tertiary road	0.39	0.39	0.52	0.69	0.36	0.03	0.39	0.31
sep_bikeway_binary	Location on path or cycleway	0.24	0.21	0.11	0.08	0.50	0.55	0.39	0.11
slope_hm	Avg. absolute roadway grade (%) w/in ½-mi radius	2.02	1.67	2.09	2.49	1.80	4.08	1.47	1.54
slope_hm^2	Square of above	6.34	3.75	5.57	9.96	5.19	21.53	3.03	3.07

		mean							
Variable	def	All	Oregon	Bend	Boulder	Charlotte	Dallas	Eugene	Portland
slope_om	Avg. absolute roadway grade (%) w/in 1-mi radius	1.90	1.76	2.13	2.74	1.76	1.90	1.56	1.67
tertiary_binary	Location on tertiary road	0.20	0.18	0.18	0.49	0.07	0.03	0.25	0.13

[a] The standard count model form used here causes individual terms to be essentially multiplicative, but multiplying two count variables doesn't make a lot of sense, logically (i.e., Strava *times* Streetlight). This formulation pre-combines the variables to get around the issue and allow them to behave in something like an additive fashion.

5.2.4.2 Results

Table 5-10 summarizes prediction performance of 2019 pooled and city models using all available count locations. Test RMSE (presented as % of mean AADBT for comparability) was the primary prediction summary statistic considered. Test MAPE by volume tertile is also presented. Table 5-11 and Table 5-12, respectively, examine the benefit of city-specific models and of adjusting third-party Strava and Streetlight counts using static context variables. Full model specifications and model fit for 2019 data are presented in Table 5-13 through Table 5-20. Full results for 2017 (Portland only) and 2018 (Oregon locations only) are provided in the Appendix.

Table 5-10 2019 Combined PC and SC Data Count Models Prediction Performance Summary

	PM0	PM1	PM2	PM3	PM4	PM5	PM6
All City Pooled (n=311)							
%RMSE	105%	93%	115%	87%	72%	91%	**71%**
MAPE (low vol. <49, n=104)	537%	669%	667%	415%	288%	383%	**271%**
MAPE (mid vol. <193, n=104)	123%	89%	147%	82%	60%	105%	**61%**
MAPE (high vol. <=1951, n=103)	50%	54%	53%	45%	**41%**	49%	42%
Oregon Pooled (n=227)							
%RMSE	103%	91%	118%	92%	72%	94%	**69%**
MAPE (low vol. <52, n=76)	385%	510%	591%	367%	**204%**	236%	218%
MAPE (mid vol. <178, n=76)	117%	85%	136%	86%	**65%**	101%	**65%**
MAPE (high vol. <=1951, n=75)	57%	53%	57%	52%	42%	56%	**40%**
Portland (n=88)							
%RMSE	79%	66%	91%	65%	51%	66%	**48%**
MAPE (low vol. <102, n=30)	345%	200%	528%	190%	**121%**	295%	146%
MAPE (mid vol. <392, n=29)	105%	90%	122%	84%	69%	95%	**66%**
MAPE (high vol. <=1951, n=29)	38%	30%	40%	32%	26%	30%	**23%**
Eugene (n=76)							
%RMSE	98%	93%	103%	93%	**64%**	84%	**64%**
MAPE (low vol. <64, n=26)	217%	269%	357%	189%	162%	222%	**138%**
MAPE (mid vol. <188, n=25)	64%	62%	76%	110%	45%	65%	**40%**
MAPE (high vol. <=1590, n=25)	58%	42%	49%	45%	**37%**	53%	38%
Bend (n=63)							
%RMSE	85%	80%	89%	84%	79%	**74%**	**74%**
MAPE (low vol. <33, n=21)	285%	418%	594%	568%	306%	281%	**238%**

	PM0	PM1	PM2	PM3	PM4	PM5	PM6
MAPE (mid vol. <57, n=21)	57%	**47%**	60%	60%	56%	63%	52%
MAPE (high vol. <=344, n=21)	51%	**38%**	43%	40%	41%	47%	43%
Boulder (n=39)							
%RMSE	60%	109%	89%	81%	**46%**	67%	49%
MAPE (low vol. <10, n=13)	835%	1035%	1456%	501%	321%	747%	**298**%
MAPE (mid vol. <108, n=13)	53%	112%	99%	114%	76%	**52**%	73%
MAPE (high vol. <=1086, n=13)	36%	60%	63%	48%	31%	43%	**30**%
Charlotte (n=14)							
%RMSE	139%	77%	95%	92%	**51%**	56%	**51%**
MAPE (low vol. <46, n=5)	362%	722%	838%	798%	**307**%	378%	408%
MAPE (mid vol. <396, n=5)	77%	76%	258%	85%	53%	**39**%	44%
MAPE (high vol. <=728, n=4)	46%	**44**%	56%	54%	46%	50%	56%
Dallas (n=31)							
%RMSE	60%	39%	81%	**24%**	36%	52%	30%
MAPE (low vol. <65, n=11)	494%	199%	149%	**60**%	203%	106%	74%
MAPE (mid vol. <517, n=11)	71%	47%	57%	31%	43%	45%	**27**%
MAPE (high vol. <=1135, n=10)	31%	28%	47%	**15**%	24%	43%	19%

Note: All statistics shown are mean of cross-validation test (out-of-sample) performance, **best-performing model specification for each measure is shown in bold**. *n* refers to number of count sites, not the number of iterations for statistic calculation, which is always tenfolds times five repeats equals 50.

2019 All City Pooled and Oregon Pooled models with the full count dataset are for the most part consistent with the full-year PC pooled model results: combinations of data sources outperform single sources, and the best-fitting models combine all three. An exception is the Oregon Pooled model, where adding StreetLight to Strava data did not improve the model performance. In fact, StreetLight appeared to provide the least information of the three data sources, significantly underperforming Strava data whether individually or in combination with static variables. One possibility is that additional static variables are needed to adjust StreetLight, which is unique due to its need to impute travel mode. Variables capturing different aspects of the count location context might be needed to complement StreetLight.

The general patterns also held, for the most part, in city-specific models, with a couple of specific results worth noting. In Dallas, no combination of static variables was found that improved on the combination of Strava and StreetLight data. It was interesting to note that while StreetLight performed poorly at Dallas locations on its own, it significantly improved the Strava estimates there. In Bend, the smallest community in our study with a maximum AADBT of 344, performance was the worst among study sites by most measures, and different variable combinations had little impact on a model's predictive ability.

As was the case with the full-year PC All City Pooled model, prediction at low-volume (<49) sites proved challenging, with the best-performing model still demonstrating considerable prediction error (271% MAPE). Prediction success at mid- (>48) and high-volume (>192) sites was again more promising at 61% and 42%, respectively, using the best-fitting overall model (PM6). Note that the volume bins are based on the count data, so all bins are lower volume compared with the full-year PC sample.

City-specific MAPE varied considerably from place to place, but the same general trends held: highest-volume predictions were fairly good for best-performing city models (worst: 44% MAPE, best: 15%); mid-volume locations were more variable but generally fairly good (worst: 66% MAPE, best: 27%); and low-volume sites typically had MAPE well over 100%, with the exception of Dallas (60%).

For the best overall model in each case, Table 5-11 provides results for the locations in each city by model type. As expected, city-specific models, with static variables and parameters tuned to local contexts, performed best in all cases. Results from pooling across multiple cities were mixed, with performance in Eugene nearly as good as a city-specific model but equal to or marginally worse than the All City model in the other two cities.

Table 5-11 City-specific Versus Pooled Model Performance by City (%RMSE, Best-fit 2019 Models)

	City-Specific	Oregon Pooled	All City Pooled
Portland	48%	54%	54%
Eugene	64%	67%	80%
Bend	74%	91%	86%
Boulder	46%	N/A	95%
Charlotte	51%	N/A	139%
Dallas	24%	N/A	43%

N/A = Not Applicable; no regional pooled model available

With the expanded list of locations, we found it no longer worked well to hold the base static model relatively fixed when combining with the Strava and StreetLight data. We instead worked to find the set of static variables that best “adjusted” the inherent biases or limitations in the third-party user data. As described in the literature review (see 3.5.1), Strava users are known to differ from the general cycling population, and the static variables are consistent with that finding. The significance and sign of the static variables in model PM4 suggest Strava users might be overrepresented in higher-income areas, on arterial routes, farther out of town, and where roads are steeper. Conversely, model results suggest Strava counts might need an upward adjustment where bike commuting activity and intersection densities are higher, and on separated bikeways and near parks. While the latter two variables might initially seem counterintuitive, they might be proxies for the types of low-stress locations that appeal to more casual bicycle users than the typical Strava app user.

StreetLight data, derived from a broad sample of smartphone users, should be more representative of the population at large, but interestingly the best-fitting adjustment factors were fairly similar to Strava, with a couple of notable differences. Income and terrain were not significant factors in the StreetLight model, and the distance from downtown adjustment factor was greatly reduced, consistent with the idea of a more typical set of cycling trips. However, the StreetLight model also included several significant positive factors (area bike commuting, intersection density, park acres, and separated bikeways) similar in magnitude to those same factors in the Strava model. Another interpretation is that the static variables here are just adjusting for imprecision inherent in StreetLight's bicycling mode imputation. The significant, negative coefficients on higher-order streets (where transit or even driving trips might be more likely mistaken for cycling) would support that possibility.

Table 5-12 summarizes the increase in error observed over all the 2019 pooled and city-specific models when using Strava or StreetLight data without adjustment factors (either static variables, or the other third-party user data). For example, using Strava counts to predict AADBT without static adjustment variables increased expected prediction error by a factor of about 1.4 (i.e., a 40% increase in %RMSE). That rule of thumb figure of 1.4 times the error held for StreetLight alone (vs. with static variables), and was only slightly lower for Strava plus StreetLight without static variables (1.3x). The case of Strava alone versus Strava plus StreetLight was the only mixed result. In some cases, combining the two third-party user data sources greatly improved results versus using Strava only (Dallas, Boulder), while in other cases, the addition of StreetLight only modestly improved or even reduced performance.

Table 5-12 Increase in Error When Using Strava or StreetLight Data Alone

	Strava Only vs. Strava + StreetLight	Strava Only vs. Strava + Static	Strava Only vs. All	StreetLight Only vs. Strava + StreetLight	StreetLight Only vs. StreetLight + Static	StreetLight Only vs. All	Strava + StreetLight vs. All
Mean[a]	1.1x	1.4x	1.4x	1.5x	1.4x	1.8x	1.3x
Min	0.8x	1.0x	1.1x	1.0x	1.2x	1.2x	0.8x
Max	1.6x	2.4x	2.2x	3.4x	1.7x	2.7x	1.8x

[a] Calculated as ratio of %RMSE

We found that the static explanatory variables had to be modified considerably from the full-year PC pooled model, and from place to place. Variation across the cities themselves as well as the types of locations counted likely both contributed to the lack of a universal model form. In addition, while OSM-derived variables should have consistent definitions, they are likely interpreted and prioritized differently from city to city.

The All City Pooled baseline static model fit the data moderately well and all coefficients had the expected signs. Model fit was not as good as observed for the full-year PC locations, but this is not so surprising given the more complex range of contexts

represented especially by the SC sites. Cycleways (designated bicycle facilities that can be either separated on-street or off-street) and off-street bike paths both increased the expected number of cyclists. Striped on-street bike lanes were found only on higher-order roads, and in the model they fully offset the negative effect of tertiary roads, and partially offset the impact of primary and secondary streets on cycling activity, consistent with other findings (Broach et al., 2012). The number of regular bike commuters (measured via ACS data) in the vicinity increased expected counts, as did proximity to downtown, and acres of nearby parks, while distance to the nearest water body—which seemed to be a strong attractor in at least some study regions—decreased expected cycling rates. Intersection density had a more complex connection to bike counts, initially increasing expected counts of cyclists, but at a declining rate as densities increased. This is consistent with the idea that beyond some density threshold, bicycling might actually become slower or less comfortable and. Therefore, less competitive with walking or other modes.

The best-fitting Oregon Pooled model specification was similar to the All City Model (Oregon count locations made up about 73% of the 2019 data), but city-specific models were much more varied in terms of best-fit variable sets. This was likely partly due to low sample sizes in some regions, but some locally specific factors are worth highlighting. In Eugene, distance to the University of Oregon was an important variable in all model combinations. In Boulder, only distance from downtown was found to be a significant static explanatory variable. While that obviously would not serve as a useful model for most applications on its own, the combination of proximity to downtown and Strava counts actually resulted in good model fit and performance. There might be other cases where even a relatively simple combination of third-party user data and contextual adjustment can meet basic planning needs. Already mentioned was the Dallas example, perhaps the lone example we found where third-party user data (when combined) did not benefit much, if at all, from static variable adjustment and performed well on their own.

Table 5-13 AADBT Poisson All City Pooled Model All Count Locations 2019 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	4.8064***	3.9872***	1.0019**	1.7130***	4.6863***	2.5369***	3.5488***
log(stv_adb +1)				0.4787***	0.6098***		0.5407***
log(stv_c_adb +1)		0.5729***					
log(stv_nc_adb +1)		0.1492*					
log(stl_raw + 1)			0.6366***	0.3464***			0.1835***
cycleway_binary	0.4739**						
path_binary	1.0604***						
cycleway_lane_binary	0.7425***					0.5495**	
Median_HH_income_om					-0.00001***		-0.00002***
arterial_binary					-0.3781**		-0.481***
primary_binary	-0.9821.					-1.4389**	
secondary_binary	-0.8825***					-1.0525***	
tertiary_binary	-0.7080**					-0.7058**	
Bike.Commuter_om	0.0007***				0.0004****	0.0004***	0.0004***
Distance_to_CBD_mi	-0.1058***						
Log(Distance_to_CBD_mi+1)					-0.4974***		-0.2627*
Intersection_Density_om	0.0021***						
Intersection_Density_om^2	-0.0000009*						
Intersection_Density_hm						-0.0073**	
`I(Intersection_Density_hm^2)`						0.00002***	
Park_acres_hm	0.0063***				0.0018.	0.0038***	
sep_bikeway_binary					0.2515.	0.5157**	0.1827.
Distance_to_Water_Body_mi	-0.5364**						
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	311	311	311	311	311	311	311
AIC[a]	53,289	47,961	68,099	42,332	27,858	39,885	26,622
Pseudo-R^2[b]	0.548	0.595	0.417	0.645	0.772	0.666	0.783
Cross-validation test performance (mean of 10 folds x 5 repeats)							
RMSE (+/- SE)	260(+/- 9)	230 (+/- 8)	284 (+/- 9)	216 (+/- 9)	179 (+/-8)	226 (+/-7)	175 (+/-8)
MAPE (+/- SE)	236% (+/- 8)	270% (+/-16)	287% (+/- 12)	181% (+/- 7)	129% (+/-5)	178% (+/-8)	124% (+/-5)
MAE (+/- SE)	168 (+/- 4)	148 (+/- 4)	185 (+/- 5)	131 (+/- 4)	107 (+/-4)	141 (+/-4)	105 (+/-4)

[a] Akaike Information (Loss) Criterion (lower is better); [b] McFadden's R^2 = 1 – (Deviance in Final Model / deviance in Intercept-only Model, higher is better)

Table 5-14 AADBT Poisson Oregon Pooled Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	-0.0879	4.1467***	0.1967	2.0364***	5.0274***	1.4320**	4.2627***
log(stv_adb +1)				0.5265***	0.7202***		
log(stv_c_adb +1)		0.7887***					0.6247***
log(stv_nc_adb +1)		-0.1664.					
log(stl_raw + 1)			0.7729***	0.2842***		0.5325***	0.1408
footway_binary	-1.8576*						
sep_bikeway_binary	0.6089**					0.3625.	0.1095
cycleway_lane_binary	0.5435*					0.5720**	
Median_HH_income_om					-0.00003***		-0.00002***
arterial_binary	-0.5182*				-0.407***		
primary_binary						-2.0331***	-1.0393*
secondary_binary						-1.1927***	
tertiary_binary						-0.9548***	
I(secondary_binary + tertiary_binary)							-0.4560**
Bike.Commuter_om	0.0005***				0.0003***	0.0006***	0.0002*
Distance_to_CBD_mi					-0.1473***	-0.1792**	-0.0959*
BikeFac_om	0.0450***						
Intersection_Density_om	0.0070***						
pct_at_least_college_education_hm	0.0285*						
Park_acres_hm	0.0072.					0.0047.	
slope_hm							0.2609*
Slope_hm^2							-0.0464
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	227	227	227	227	227	227	227
AIC[b]	35,561	33,022	52,317	32,982	18,600	19,357	17,183
Pseudo-R^2[c]	0.612	0.640	0.420	0.641	0.805	0.796	0.821
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	254 (+/- 11)	230 (+/- 10)	296 (+/- 11)	232 (+/- 10)	178 (+/- 7)	229 (+/- 10)	170 (+/- 7)
MAPE (+/- SE)	190% (+/- 11)	214% (+/- 18)	261% (+/- 11)	168% (+/- 7)	103% (+/- 4)	132% (+/- 6)	107% (+/- 5)
MAE (+/- SE)	159 (+/- 6)	143 (+/- 5)	190 (+/- 6)	142 (+/- 4)	106 (+/- 3)	137 (+/- 5)	101 (+/- 3)

Table 5-15 AADBT Poisson Portland Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	-2.8207	3.2076***	0.2326	2.0455**	-1.7390***	-7.0047**	-4.25***
log(stv_adb +1)					0.8055***		0.6332***
log(stv_c_adb +1)		0.8756***		0.7498***			
log(stl_raw + 1)			0.7448***	0.2010*		0.6542***	0.2128***
footway_binary						1.2148**	
sep_bikeway_binary	1.8412***				0.7312***		0.6723***
cycleway_lane_binary	1.5489**				-1.277***	1.3716*	-0.2674***
arterial_binary	-0.9606*				1.159***	-1.4174**	
Median_HH_income_om					-0.00001***		-0.00001***
Bike.Commuter_om	0.0005*				0.0002***	0.0004*	0.0002***
Distance_to_CBD_mi	-0.1525.						
BikeFac_om							
intersection_density_hm	0.0078**				0.0046***		0.0046***
intersection_density_om						0.0075***	
pct_at_least_college_education_hm	0.0677*				0.0438***	0.0534*	0.0537***
Park_acres_hm	0.025*				0.0172***	0.0268**	0.0227***
Slope_hm	-0.1893						
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	88	88	88	88	88	88	88
AIC[b]	15,811	12,718	25,695	12,177	7,044	12,349	7,347
Pseudo-R^2[c]	0.657	0.727	0.434	0.7390	0.855	0.721	0.849
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	332 (+/- 13)	280 (+/- 12)	387 (+/- 17)	277 (+/- 12)	214 (+/- 8)	273 (+/-11)	202 (+/- 8)
MAPE (+/- SE)	163% (+/- 13)	105% (+/- 8)	225% (+/- 12)	100% (+/- 7)	71% (+/-5)	139% (+/-12)	77% (+/-5)
MAE (+/- SE)	245 (+/- 10)	191 (+/- 8)	276 (+/- 12)	192 (+/- 8)	151 (+/- 6)	202 (+/-10)	143 (+/-6)

Table 5-16 AADBT Poisson Eugene Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	7.7066**	4.0453***	1.1567	2.4735***	6.8221***	4.6527***	5.4580***
log(stv_adb +1)					0.7485***		0.6509***
log(stv_c_adb +1)		2.0452***		0.8949***			
log(stv_nc_adb +1)		-0.8583*					
log(stl_raw + 1)			0.6187***	0.2684*		0.4876***	0.1832*
sep_bikeway_binary	0.3440						
Median_HH_income_om							
arterial_binary	-0.1544				-0.6493**	-0.8021**	-0.7241***
Bike_Commuter_om	0.0318				0.0005*	0.0003	0.0005**
Distance_to_CBD_mi					-0.0520		
log(min_dist_to_university)	-0.4261.				-0.3577***	-0.3317***	-0.3352***
Park_acres_hm	0.0085						
slope_om	-0.0253						
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	76	76	76	76	76	76	76
AIC[b]	7,604	8,196	12,543	9,023	4,274	6,731	4,020
Pseudo-R^2[c]	0.606	0.575	0.334	0.529	0.791	0.655	0.805
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	203 (+/- 17)	189 (+/- 17)	210 (+/- 18)	189 (+/- 18)	134 (+/- 14)	162 (+/- 12)	134 (+/- 13)
MAPE (+/- SE)	113% (+/- 6)	123% (+/- 10)	158% (+/- 7)	110% (+/- 6)	78% (+/- 5)	107% (+/- 5)	72% (+/- 4)
MAE (+/- SE)	129 (+/- 8)	118 (+/- 8)	145 (+/- 9)	118 (+/- 8)	86 (+/- 6)	115 (+/- 6)	83 (+/- 6)

Table 5-17 AADBT Poisson Bend Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	0.8151	3.6002***	3.1285***	2.9138***	4.1543***	-1.3451	3.0681***
log(stv_adb +1)					0.3235*		0.3103.
log(stv_c_adb +1)		0.6993***		0.6363***			
log(stl_raw + 1)			0.71525*	0.1126.		0.2711**	0.1399
Log(Bikeshare Crossing_hm+1)							
sep_bikeway_binary	0.3165				0.7611**	0.6864***	0.9630**
Bike.Commuter_om	0.0020					0.0028.	
Distance_to_CBD_mi	-0.0445						
Log(Distance_to_CBD_mi)					-0.7771**		
cycleway_lane_binary							-0.6028.
intersection_density_om	0.0032						
pct_at_least_college_education_hm	0.0344					0.0423***	
Park_acres_hm	0.0049						
slope_om	-0.1705					-0.1180	
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	63	63	63	63	63	63	63
AIC[b]	2,179	2,930	3,143	2,839	2,162	1,989	1,947
Pseudo-R^2[c]	0.391	0.132	0.060	0.163	0.394	0.454	0.468
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	53 (+/- 2)	52 (+/- 4)	55 (+/- 4)	52 (+/- 4)	50 (+/- 3)	47 (+/- 3)	47 (+/- 3)
MAPE (+/- SE)	144% (+/- 12)	243% (+/- 41)	216% (+/- 25)	207% (+/- 28)	135% (+/- 11)	124% (+/- 10)	131% (+/- 13)
MAE (+/- SE)	39 (+/- 1)	38 (+/- 2)	42% (+/- 2)	39 (+/- 2)	36 (+/- 2)	35 (+/- 2)	35 (+/- 2)

Table 5-18 AADBT Poisson Boulder Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	6.8581***	3.8999***	2.2417 ***	1.7176***	5.2415***	6.0857***	4.7252***
log(stv_adb +1)					0.4817***		0.4625***
log(stv_c_adb +1)		1.5774***		0.7457***			
log(stv_nc_adb +1)		-0.6511**					
log(stl_raw + 1)			0.4737***	0.3195***		0.09934***	0.0748
log(Distance_to_CBD_mi + 1)	-1.2115***				-1.1612***	-1.1030***	-1.0829***
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	39	39	39	39	39	39	39
AIC[b]	1,919	3,972	4,887	3,348	1220	1,799	1,162
Pseudo-R^2[c]	0.826	0.614	0.520	0.679	0.898	0.838	0.904
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	98 (+/- 10)	170 (+/- 17)	149 (+/- 17)	136 (+/- 16)	74 (+/- 6)	109 (+/- 12)	79 (+/- 9)
MAPE (+/- SE)	336% (+/- 45)	451% (+/- 74	597% (+/- 104)	245% (+/- 34)	152% (+/- 16)	307% (+/- 43)	144% (+/- 15)
MAE (+/- SE)	71 (+/- 6)	116 (+/- 10)	109 (+/- 10)	94 (+/- 9)	54 (+/-4)	77 (+/-7)	56(+/-5)

Table 5-19 AADBT Poisson Charlotte Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	4.9178***	4.2415***	4.3813	1.0849	2.9857*	0.8474	1.0323
log(stv_adb +1)					0.8786		
log(stv_c_adb +1)		-0.9615.		-1.2431*			0.1551
log(stv_nc_adb +1)		1.3141.		1.5142**			
log(stl_raw + 1)			0.1852	0.4704		0.6120	0.5469
sep_bikeway_binary	0.8905				0.8507*	0.7902	0.7617
cycleway_lane_binary					-0.00003		
arterial_binary	-1.5340						
BikeFac_hm					-0.6669	-1.5688*	-1.5057*
Distance_to_CBD_mi	-0.0468						
Park_acres_hm	0.0099*				0.0129*	0.0086.	0.00977*
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
N	14	14	14	14	14	14	14
AIC[b]	929	2398	3,575	2,230	660	797	671
Pseudo-R^2[c]	0.767	0.350	0.017	0.398	0.843	0.804	0.810
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	154 (+/- 16)	206 (+/- 28)	271 (+/- 14)	240 (+/- 33)	138 (+/- 17)	151 (+/- 22)	165(+/- 24)
MAPE (+/- SE)	167% (+/- 33)	269% (+/- 59)	409% (+/- 55)	301% (+/- 73)	141% (+/- 31)	163% (+/- 44)	180% (+/- 50)
MAE (+/- SE)	149 (+/- 16)	197 (+/- 27)	269 (+/- 13)	227 (+/- 31)	126(+/- 14)	139(+/- 20)	152(+/- 21)

Table 5-20 AADBT Poisson Dallas Model All Counters 2019 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	3.0939***	3.3634***	0.8926	1.0355***	2.3044**	1.7503*	0.7928
Log(stv_adb + 1)							
log(stv_c_adb +1)		0.6301**					
log(stv_nc_adb +1)		0.0940					
log(stl_raw + 1)			0.7187***			0.4366**	
log(stv_stl)[a]				0.4487***			0.3878***
intersection_density_om	0.0127***				0.0060.	0.0062*	0.0035.
Distance_to_Water_Body_mi	-1.4731***				-0.2401	-0.8204*	-0.1243
Distance_to_CBD_mi	-0.0684				-0.0447		-0.0163
Park_acres_hm	0.0049***				0.0019*	0.0028	0.0012
Slope_om	0.5769***				0.3320*	0.0053	0.1892
Model fit statistics (presented for final, best-fitting model for each variable set, including all observations)							
AIC[b]	2,706	1,574	2,938	744	1,170	1,699	654
Pseudo-R^2[c]	0.765	0.872	0.743	0.950	0.911	0.860	0.959
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	178 (+/- 14)	124 (+/- 7)	195 (+/- 17)	76 (+/- 8)	111 (+/- 8)	167 (+/- 15)	94 (+/- 8)
MAPE (+/- SE)	225% (+/- 53)	96% (+/- 10)	78% (+/- 11)	34% (+/- 2)	100% (+/- 18)	66% (+/- 7)	40% (+/- 4)
MAE (+/- SE)	144 (+/- 10)	102 (+/- 6)	146 (+/- 13)	59 (+/- 6)	92 (+/- 8)	125 (+/- 11)	69 (+/-6)

Figure 5-5 through Figure 5-7 provide selected model prediction plots for the pooled models. Additional plots are provided in the Appendix. Note that predictions shown represent the mean of each observation for the out-of-sample test sets in the cross-validation procedure. It was encouraging to see that even without any weighting or region-specific variables, pooled models fit each regional subsample reasonably well, without extreme bias or excessive outliers. The two most apparent outliers are two underpredicted high-volume locations in Eugene. Both sites are on the University of Oregon campus, along an internal street that lacked specific facility attributes in the OSM data.

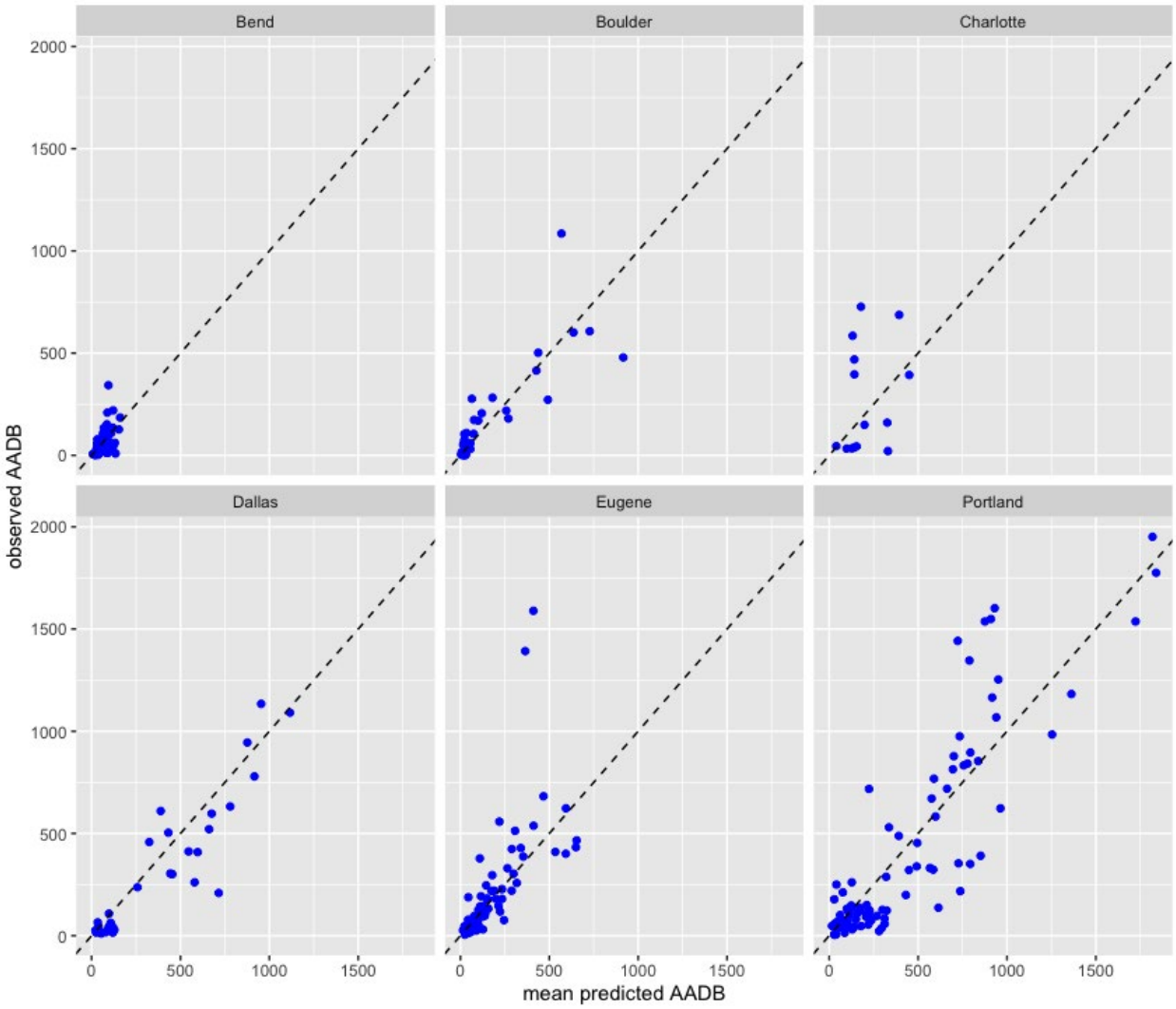


Figure 5-5 PM6 Pooled Count 2019 Model Prediction Plots by Region for All Count Locations

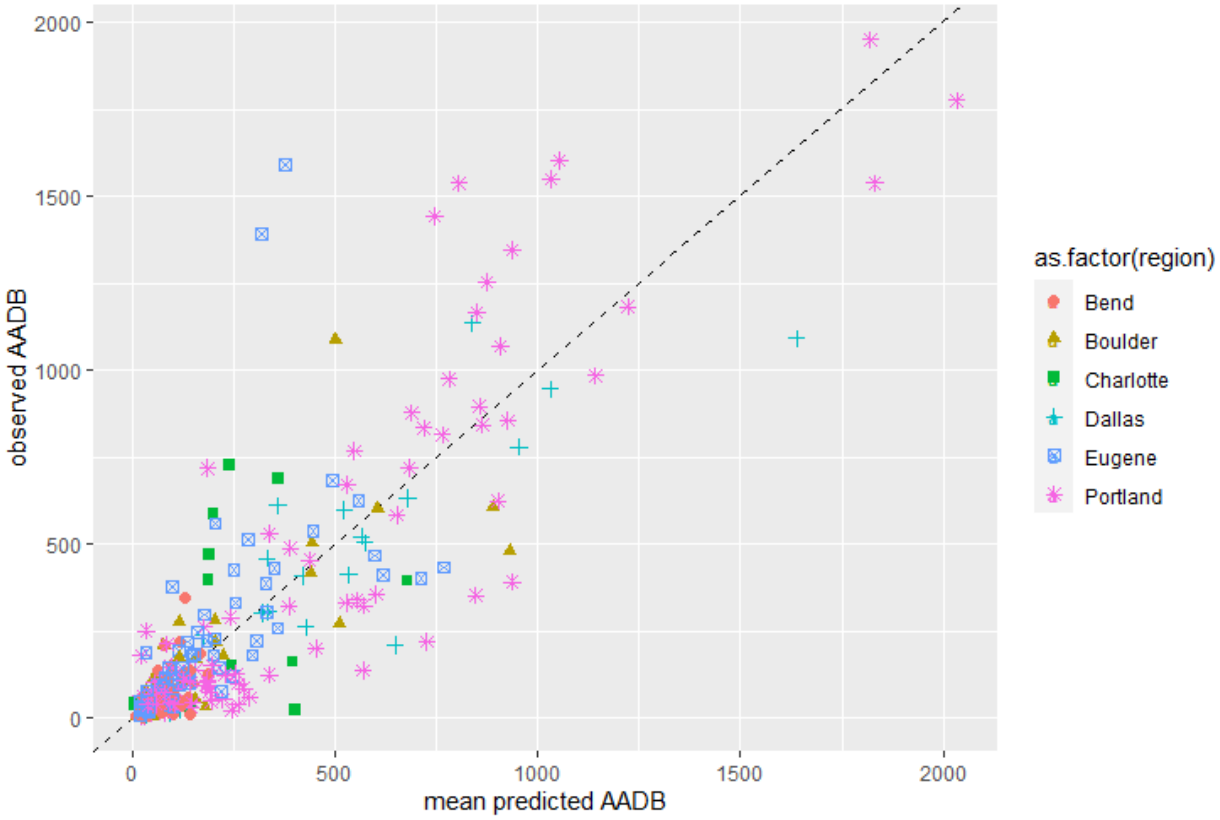


Figure 5-6 PM6 Pooled Count 2019 Model Prediction Plot for All Count Locations

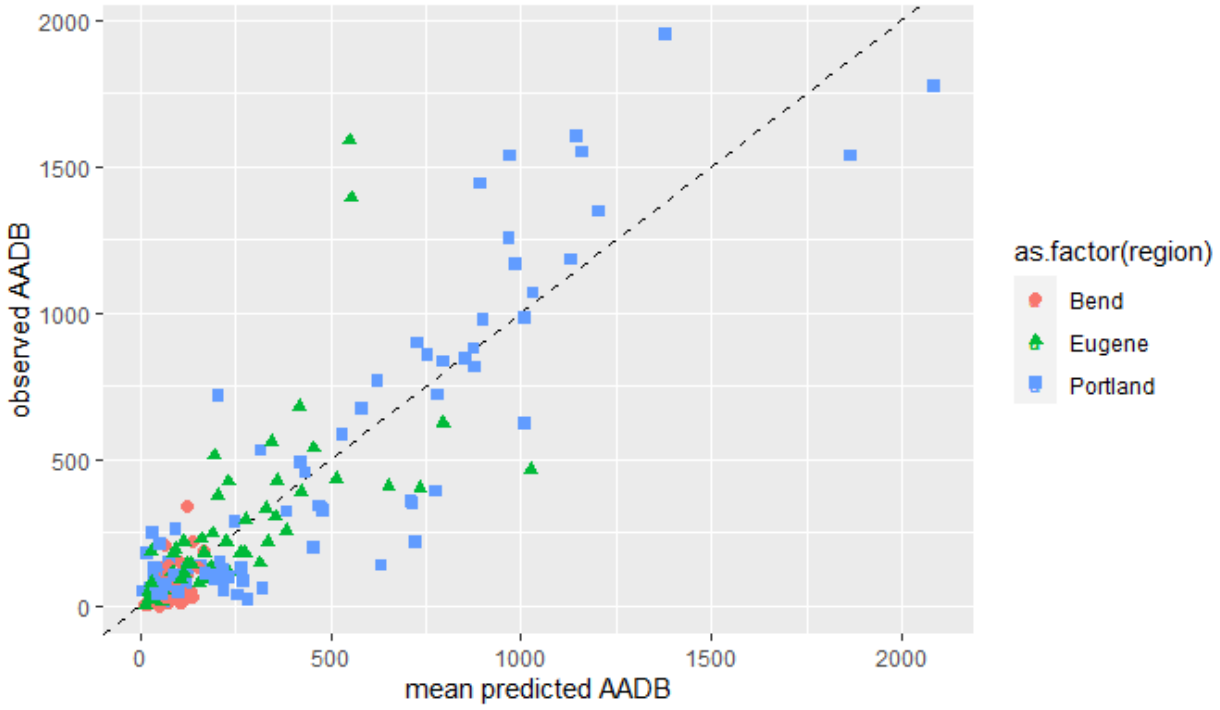


Figure 5-7 PM6 Oregon Pooled Count 2019 Model Prediction Plot for All Count Locations

Additional data from 2018 (all Oregon locations, Strava, and StreetLight) and 2017 (Portland, no StreetLight) provided further tests of the model and data's stability and transferability over time and space. Note that while there was some repetition, many count locations differed between years, making these tests more like a joint test of temporal and spatial transferability. Table 5-21 summarizes the results of two types of model transfer. First, 2019 model specifications for Oregon Pooled and city-specific models were transferred to 2018 data, and parameters were re-estimated. Second, 2019 models were used to predict 2018 data without re-estimating parameters. With re-estimation, the familiar pattern held (i.e., the best-fitting models included all sets of variables (PM6), followed by the static-adjusted Strava models (PM4), and then the rest. Comparing error rates with the corresponding models in Table 5-10 shows performance within expectations (+/- <10 %RMSE) despite the time shift and sample change. Without re-estimation, however, results changed significantly. Error rates, in general, are higher, and the Strava plus static models (PM4) were clearly preferred in all city-specific models and tied for best fit in the pooled model. On average, reusing model estimates resulted in a 10-50% increase in the error rate across models.

Table 5-21 Applying 2019 Models to 2018 Data with and Without Reestimation

	%RMSE						
	PM0	PM1	PM2	PM3	PM4	PM5	PM6
2019 Model Specifications Applied to 2018 Data WITH re-estimated coefficients							
Portland (n=33)	50%	47%	54%	46%	43%	48%	**33%**
Eugene (n=86)	99%	83%	94%	92%	73%	88%	**65%**
Bend (n=57)	98%	88%	90%	88%	81%	86%	**75%**
Oregon Pooled (n=176)	106%	99%	113%	97%	91%	101%	**71%**
2019 Model Specifications Applied to 2018 Data WITHOUT re-estimating coefficients							
Portland (n=33)	82%	76%	107%	75%	**51%**	79%	63%
Eugene (n=86)	106%	107%	127%	110%	**77%**	102%	94%
Bend (n=57)	96%	100%	106%	101%	**89%**	100%	101%
Oregon Pooled (n=176)	113%	115%	138%	115%	**95%**	121%	**95%**
Error multiple when NOT re-estimating							
Portland (n=33)	1.6	1.6	2.0	1.6	**1.2**	1.7	1.9
Eugene (n=86)	**1.1**	1.3	1.3	1.2	**1.1**	1.2	1.4
Bend (n=57)	**1.0**	1.1	1.2	1.2	1.1	1.2	1.3
Oregon Pooled (n=176)	1.1	1.2	1.2	1.2	**1.0**	1.2	1.3

The relatively better performance of the Strava plus static models (PM4) over the full model (PM6) might reflect a number of underlying reasons. Despite our efforts to avoid overfitting, these results are consistent with the more complex model potentially being less adaptable to new data. A contributing factor might also be technical change in StreetLight data from 2018-2019, during which time its base sample of trip data was

expanded. This would explain why models including StreetLight fared worse than others when holding parameters constant.

The following section expands the modeling effort further to consider more flexible model forms estimated using machine learning techniques. An overall summary of the modeling exercises follows.

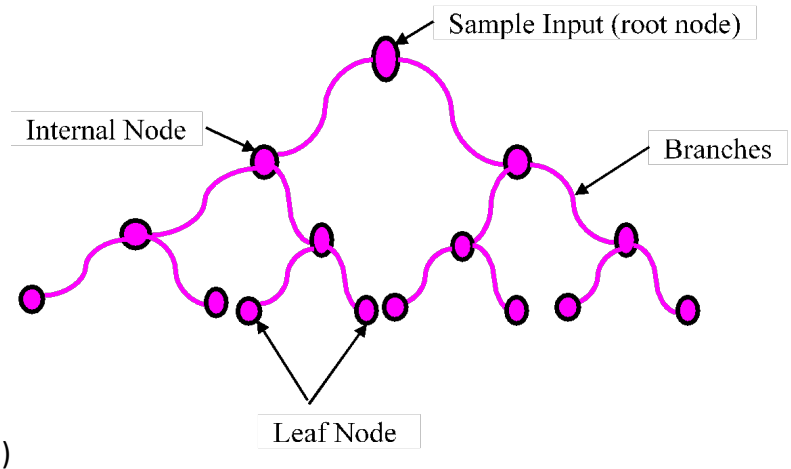
5.3 ADVANCED MODELING USING MACHINE LEARNING

We developed a range of models using various machine learning techniques to understand the feasibility of using more complex and flexible model forms in predicting bicycle counts on a network. The models also use different data fusion and variable extraction techniques to compare and evaluate the values of third-party user data for modeling bicycle activity with machine learning techniques.

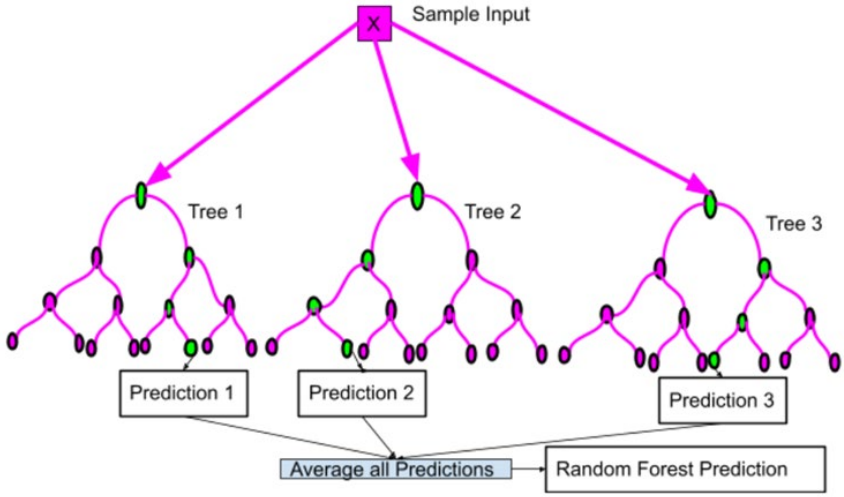
As shown in Table 5-1, this section used three sets of models specified – All City Pooled, Oregon Pooled, and city-specific models (Portland and Eugene only) to ensure a larger sample size to develop machine learning models. Like the count models, the modeling sequentially adds the crowdsourced data and static location variables to understand their roles in AADBT predictions. All models use 10-fold cross-validation with five repeats to avoid overfitting while maintaining sufficient data for model development.

5.3.1 Machine learning modeling description

This study uses Random Forest (RF) models to estimate AADBT. A decision tree (Figure 5-9 (a)) repeatedly partitions the data into multiple subspaces (i.e., smaller trees) until the outcomes in each final subspace become as homogeneous as possible. RF (Figure 5-9 (b)) builds multiple decision trees and merges them (forest) to obtain a more accurate and stable prediction. Therefore, a parent node splits into exactly two child nodes, and a child node will act again as a parent node until the branches cannot be grown any further without gaining benefits of growing out the tree. RF considers additional randomness within the model by adding more variables in smaller trees within a forest, which likely produces a wide diversity that generally results in a better model. RF aims to estimate modeling coefficients to fit the observed AADBT based on an information gain theory to select the most important features for the forest. Information gain detects the features that provide maximum information about the regression outputs. Gain indicates the relative contribution of the corresponding features to the model and is calculated based on each feature's contribution for each tree in the model. A higher value of importance compared to another feature implies that it is a better feature for prediction.



(a)



(b)

Figure 5-8 (a) A Decision Tree, (b) Random Forest

We tuned the models' parameters that control the model building process (referred to as hyperparameters) using Hyper-opt TPE (Tree-structured Parzen Estimator Approach) algorithms. TPE is a sequential model-based optimization (SMBO) approach to tune the hyperparameters. SMBO methods sequentially construct models to approximate the performance of hyperparameters based on historical measurements and subsequently choose new hyperparameters to test based on this model. The TPE approach models P(x|y) and P(y) where x represents hyperparameters and y indicates an associated quality score.

5.3.2 **Machine learning (ML) modeling data fusion**

We used random forest algorithms with different data fusions techniques as shown below:

Model 1: AADBT=f(StreetLight)
Model 2: AADBT=f(Strava)
Model 3: AADBT=f(Strava +StreetLight)

Model 4: AADBT=f(Strava + StreetLight+Static)
Model 5: AADBT=f(Strava + StreetLight+Static+Bikeshare)

Using the aforementioned data fusion structures, the research team used 2019 data to develop four models by geographic regions: (i) all six cities (All City Pooled model, N=311), (ii) three cities from Oregon (Oregon Pooled model, N=227), (iii) Portland only (N=88), and (iv) Eugene only (N=77). The four remaining cities - Dallas, Charlotte, Boulder, and Bend were not considered for city-specific machine learning models due to limited sample sizes.

Four different additional variable sets were created that were based on buffer types and sizes to consider for random forest models (See Table 5-22).

Table 5-22 Buffer Fusion Structure for Machine Learning

	Buffer type	Size	# of variables
Full Data	Air and Network	0.5, 1, and 2 miles	200
Selected Data	Air	0.5 and 1 miles	99
	Network	0.5, 1, and 2 miles	133
	Network	0.5 and 1 miles	99
	Network	1 and 2 miles	100

5.3.3 Random Forest modeling results

5.3.3.1 Random Forest data fusion performance

Table 5-23 shows the overall performance of each RF model using its percentage of root mean square error (RMSE) by data fusion structure, and Table 5-24 further breaks down the results with three different volume bins - high, mid, and low volumes. Each of these bins consists of one-third of the total data points for each region. The specific AADBT range for each bin is indicated in the Table 5-24. For the four geographical levels (All City Pooled, Oregon Pooled, Portland and Eugene), there are five data fusion models (StreetLight only, Strava only, StreetLight + Strava, static + StreetLight + Strava, and static + StreetLight + Strava + bikeshare). The RF model performs best with full data fusion (Strava and StreetLight with static variables) while bikeshare data does not further improve the model performance. The modeling results indicate that Strava plays a more important role than StreetLight. When each model's performance is broken down by volume bin, the best RMSE is 14% for a high-volume bin. The All City Pooled and Oregon Pooled models show similar RMSEs at 9% while individual city-specific models show varying levels of error. For instance, Eugene and Portland show 11% and 19% of RMSE, respectively. Finally, for the low-volume bin, StreetLight alone significantly overestimates AADBT (except for Eugene), likely due to its limited coverage for the low-volume locations. The Strava-only model performs well than the Streetlight-only model; however, the full data fusion with static, Strava and StreetLight significantly improves the model performance.

Table 5-23 Overall RF Fusion Model Results (%RMSE)

Study Area	SL	Strava	Sl+Strava	Static+SL+Strava	Static+SL+Strava+BS
ALL City Pooled (n=312)	122%	107%	94%	72%	N/A
Oregon Pooled (n=228)	125%	95%	92%	71%	74%
Portland (n=88)	93%	71%	66%	54%	55%
Eugene (n=77)	101%	85%	83%	67%	68%

Table 5-24 RF Fusion Model Results (%RMSE) for Different Geographic Regions by Volume Breakdown

Volume Bin	Study Area (Volume Range)	SL	Strava	Sl+Strava	Static+SL+Strava	Static+Strava+SL+BS
Low Volume	ALL City Pooled (0-48)	121%	85%	63%	50%	N/A
	Oregon Pooled (0-51)	143%	68%	68%	48%	47%
	Portland (5-101)	152%	46%	76%	59%	60%
	Eugene (5-63)	61%	61%	49%	33%	34%
Mid Volume	ALL City Pooled (49-190)	19%	16%	14%	9%	N/A
	Oregon Pooled (52-177)	18%	14%	11%	9%	9%
	Portland (102-391)	35%	27%	22%	19%	19%
	Eugene (64-187)	20%	11%	11%	11%	12%
High Volume	ALL City Pooled (191-1951)	29%	17%	23%	18%	N/A
	Oregon Pooled (178-1951)	29%	24%	23%	20%	20%
	Portland (392-1951)	28%	26%	25%	20%	20%
	Eugene (188-1590)	20%	17%	17%	14%	14%

Another measure of performance, MAPE (Mean Absolute Percent Error) is presented in Table 5-25 and Table 5-26. MAPE has reduced significantly when static, StreetLight and Strava fused together in the model (Table 5-25). For example, the All City Pooled model's overall MAPE has reduced from 325 to 135, Oregon Pooled model's MAPE from 287 to 157, and Eugene model's MAPE from 152 to 83 when Strava and static variables were combined with Streetlight. Bikeshare did not improve the prediction. Strava-only model provides better prediction compared to StreetLight-only model as Strava may represent more close value of ground truth compared to StreetLight. The breakdown of the model performance by volume bin in Table 5-26 shows that the results are consistent with count modeling findings. High-volume sites show lower MAPE compare to mid- and low-volume sites.

Table 5-25 Overall MAPE Results

Study Area	SL	Strava	Sl+Strava	Static+SL+Strava	Static+SL+Strava+BS
ALL City Pooled (n=311)	325	272	170	135	N/A
Oregon Pooled (n = 227)	287	190	179	157	150
Portland. (n = 88)	209	112	156	105	104
Eugene (n = 77)	152	144	123	83	85

Table 5-26 RF Fusion Model Results (MAPE) for Different Geographic Regions by Volume Breakdown

Volume Bin	Study Area (Volume Range)	SL	Strava	Sl+Strava	Static+SL +Strava	Static+Strava +SL+BS
Low Volume	ALL City Pooled (0-48)	769	657	365	307	N/A
	Oregon Pooled (0-51)	677	424	409	382	357
	Portland (5-101)	464	223	363	222	219
	Eugene (5-63)	329	355	293	183	186
Mid Volume	ALL City Pooled (49-190)	133	104	95	58	N/A
	Oregon Pooled (52-177)	118	96	74	59	61
	Portland (102-391)	136	89	82	68	69
	Eugene (64-187)	87	54	50	50	53
High Volume	ALL City Pooled (191-1951)	60	56	49	40	N/A
	Oregon Pooled (178-1951)	65	54	53	43	44
	Portland (392-1951)	37	36	33	27	28
	Eugene (188-1590)	48	37	40	34	34

Figure 5-9 compares the distribution of observed and estimated AADBTs from the RF All City Pooled model using the three data sources – static, Strava and StreetLight. The distribution patterns confirm that the majority of the sites had volumes below 250, particularly the cities of Bend, Boulder, and Charlotte. In Eugene and Boulder (Figure 5-10), the model appears to estimate AADBT more closely than other locations, whereas in Portland, the model tends to overestimate volumes at low- and medium-volume locations while underestimating at high-volume sites.

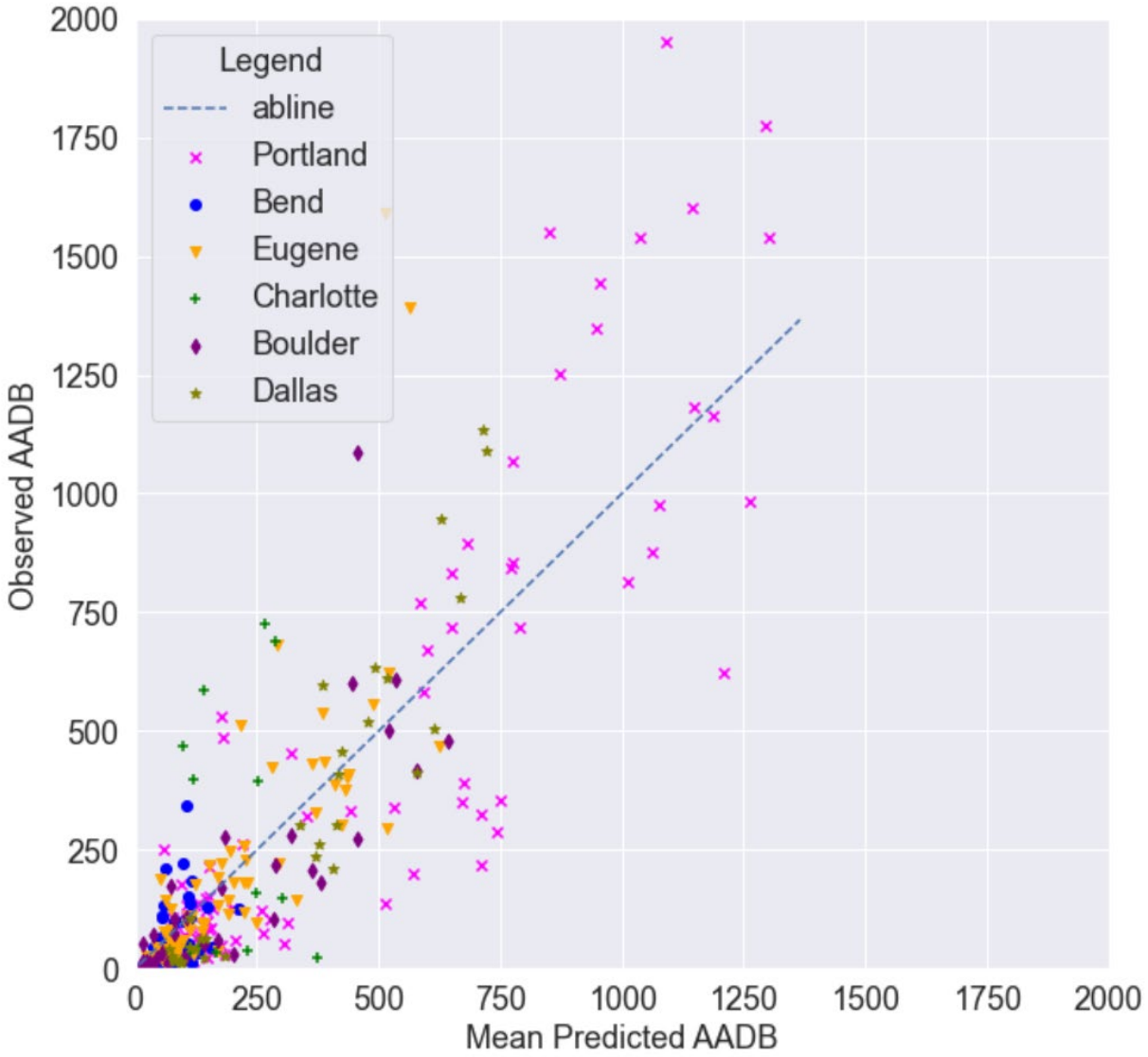


Figure 5-9 All City Pooled RF Model [Static+Strava+SL] Observed Vs Prediction

Table 5-27 and Figure 5-10 compare the performances of the pooled model by region. Portland and Dallas show lower RMSEs (better fit) than other regions. This may be because Dallas mostly collects counts at locations that share similar user/geographic characteristics. For example, Dallas collects counts along bike paths and Portland captures more commuters, and this reduces variations in both Strava and StreetLight in capturing the bicycle volumes. Mixed recreational and utilitarian activity patterns of Bend and Charlotte compared to remaining cities could contribute to their higher error rates (% RMSE). Boulder and Eugene are also significantly underestimated by the model. Due to various volume ranges and AADBT variations within each region, the All City Pooled model with combined data from six different regions still struggles to provide an effective generalized model for all regions.

Table 5-27 Pooled RF Model [Static+Strava+SL] Performance Breakdown by Region

Study Area	% RMSE	MAPE	MAE
Portland (n=88)	57	109	169
Eugene (n=77)	87	70	82
Bend (n=63)	95	163	42
Boulder (n=39)	83	216	79
Charlotte (n=14)	100	258	232
Dallas (n=31)	47	155	117

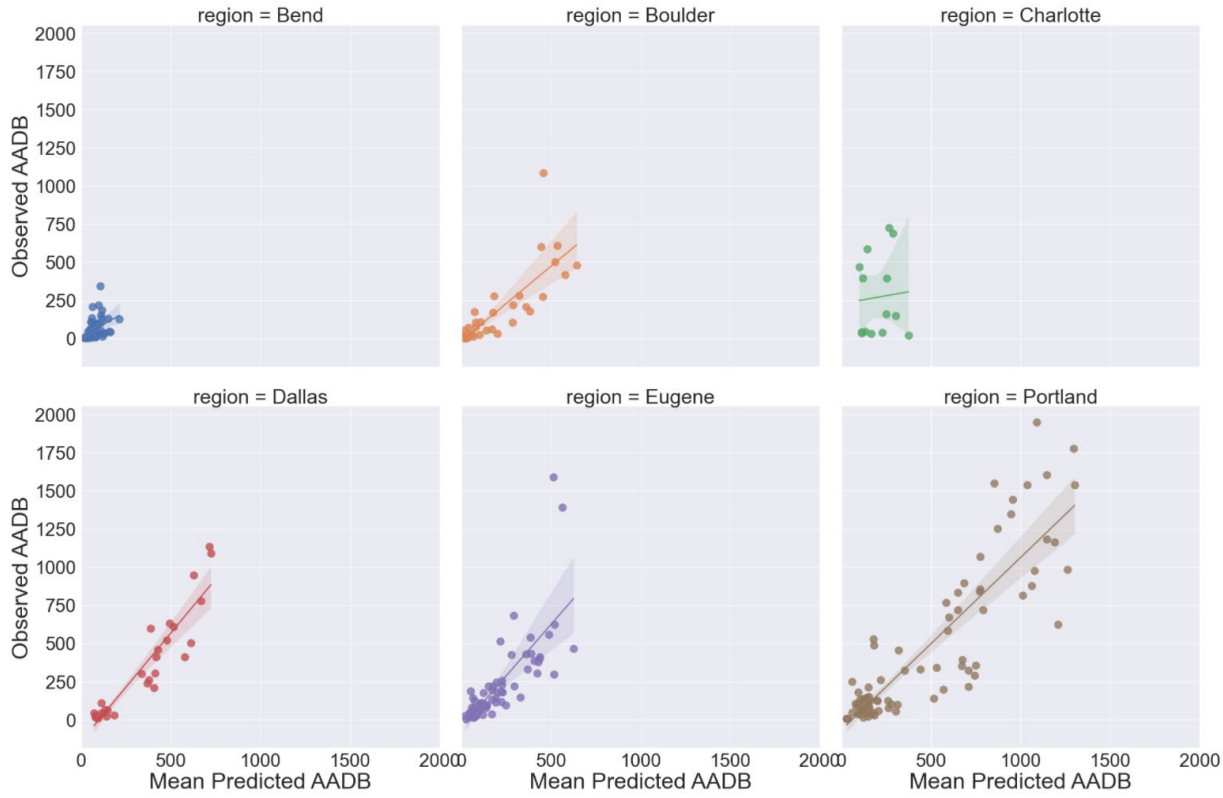


Figure 5-10 Pooled RF Model [Static+Strava+SL] Observed Vs Prediction Fit by Region

5.3.3.2 Buffer fusion results

Although RF models select optimal variables that ensure the best performance, a substantial number of features selected by RF models could hamper this variable selection process to achieve optimality. We found that a reduced number of variables that are pre-selected prior to the model development could help the model to select the best-performing variables. We created four subsets of data features based on different buffer types and sizes to test the optimal variable selection processes of the RF models. Table 5-28 shows the comparison of RF model performance in terms of MAPE, RMSE,

and MAE for different data and buffer sets. The results indicate that variables extracted at a half-mile and one-mile air buffer show the best performance. This could be because air (Euclidean) buffers create more homogenous sets of variables at static locations regardless of network or geographical characteristics.

Table 5-28 RF Model Performance for Different Set of Variable Sets

Datasets	MAPE	RMSE	MAE
Full variable set [312 data pointsx200 variables]	153	180	112
Half mile + One-mile air buffers + point variables [312 data points x 99 variables]	135	179	109
Half mile + One mile + Two-mile network buffers + point variables [312 data points x 133 variables]	153	179	111
Half miles + One-mile network buffers + point variables[312 data points x 99 variables]	146	180	110
One mile + Two miles network buffers + point variables [312 data points x 100 variables]	147	175	108

5.3.3.3 Variable importance

The RF models select the optimal variables and Figure 5-11 shows the most important 40 variables of the All City Pooled model. The results show that Strava- and StreetLight-related variables are some of the most important factors used at the root level of the forest. Intersection density within the one-mile buffer; demographics (median age, median household income, number of jobs, employment density, percentage African American, and population density); distance metrics (distance to water body, industrial area, park, CBD, and forest); land use (park area, industrial area); and bicycle facilities are also selected as important predictors to estimate AADBT in the pool model. The rest of the variables' importance values and variable definitions appear in the Appendix.

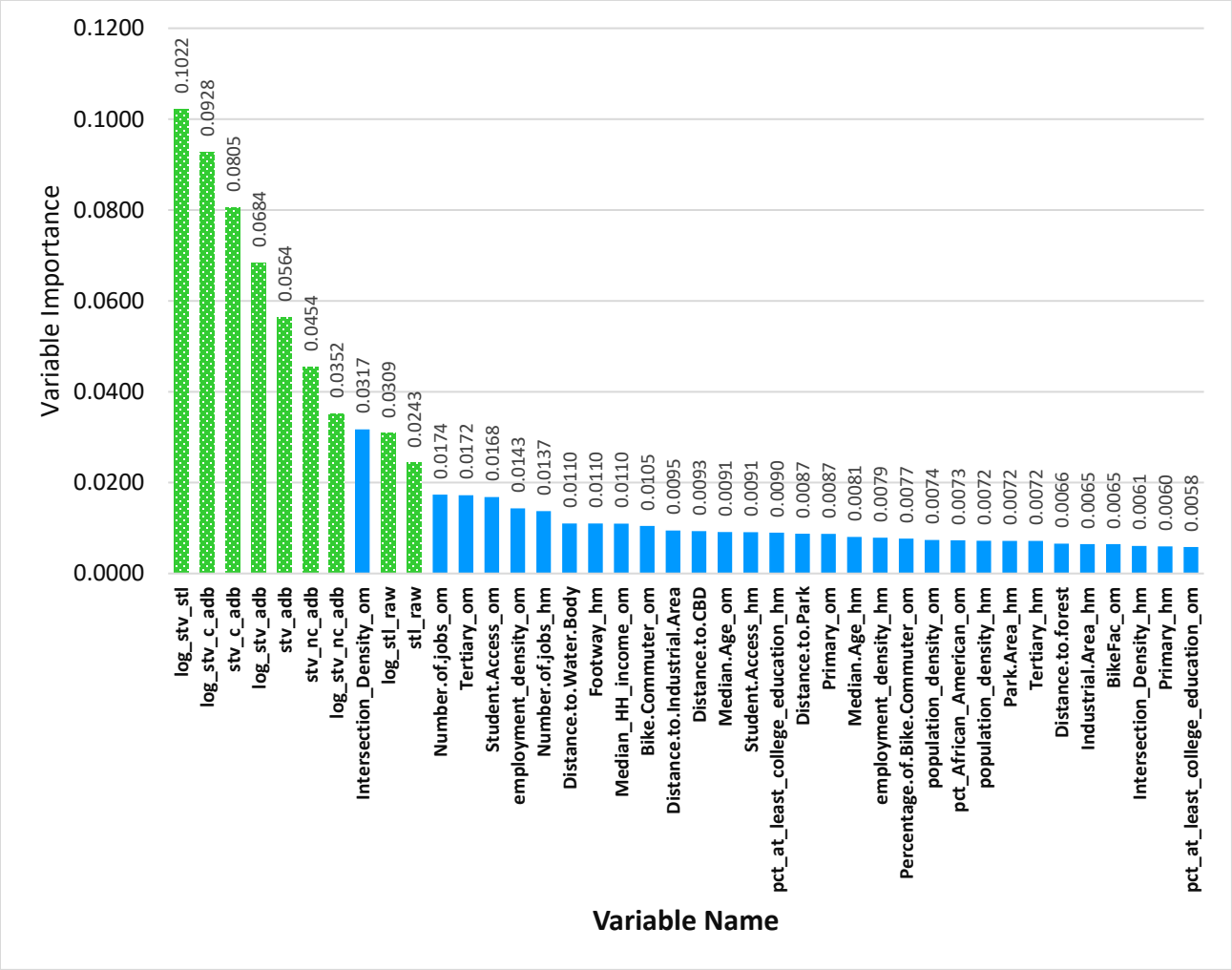


Figure 5-11 All City Pooled RF Model [Static+Strava+SL] Top 40 Variables

*Note: om=one-mile Euclidean buffer, hm=half-mile Euclidean buffer, stv=Strava, c=commute, nc=non-commute, stl=StreetLight

5.4 OVERALL MODELING RESULTS AND SUMMARY

We developed a range of model forms over multiple subsets of permanent and short-duration count, third-party user, and supporting contextual "static" data across six regions. Conventional multivariate count models as well as more flexible machine learning exercises were conducted. The general consistency of the models and key findings makes us fairly confident of some of the findings in this emerging area of research.

Strava and StreetLight data provide valuable information on bicycling activity, but they reach their full potential only when combined with more traditional contextual variables measuring the immediate context and the general neighborhood of a given network location. In almost every scenario considered, the best-performing models used a combination of all three available data sources: Strava, StreetLight, and static variables that adjust or augment the third-party user data. In most cases, the combination of

Strava and static variables performed about as well as the best-fitting models. StreetLight data tended to provide only marginal improvements, but there were specific instances where StreetLight was a valuable complement to Strava data, such as higher bicycle volume sites with few transit users sharing the link.

Useful static variable sets can be constructed using standardized, public data sources, such as OSM and Census data, without the need to gather and normalize locally sourced data. Further, simple specifications—at point locations and within Euclidean "crow fly" buffers—proved adequate, actually outperforming more involved measures based on network shortest-path buffers or bikeshare GPS data.

Specifying conventional multivariate count models requires a fair amount of effort. If count location samples are adequate—perhaps 100 locations, although 200 is better—carefully chosen machine learning techniques like the Random Forest models presented here can reduce the effort considerably without much, if any, performance penalty. The more flexible form of these models means they are likely to improve and surpass conventional count models as sample sizes increase further.

As expected, even where local data are limited, there are gains to be had from creating more localized models—whether regional or, better yet, city-specific—versus pooled models built using data from multiple regions. At the same time, where pooling data is necessary, those models can work reasonably well—without extreme error or bias—across a range of localities, although with an expected increase in error. In this case, all locales contributed at least some count sites to the All City Pooled model; further testing would be needed to better understand transferability to locations outside the pool.

Below a certain volume threshold (somewhere below around 50-200 cyclists per day, depending on the scope of the model and number of counters), current data and modeling techniques can be expected to produce very noisy count estimates (MAPE > 100%). At medium volumes (beginning somewhere between 50-200), decent if not exacting estimates are possible (40-60% MAPE). At higher volumes (roughly > 200 AADBT), fairly accurate predicted volumes are feasible (15-40% MAPE), especially if models can be calibrated to a specific city and year. Of course, that means to understand likely error at locations without counts requires first approximating the volume level, an important need for future research and (count program) practice.

6 CONCLUSIONS AND RECOMMENDATIONS

The work presented here lays a foundation for further evaluation of emerging data sources and bicycle count modeling methods more generally. One thing that the process made clear is that rather than replacing conventional bike data sources and count programs, big data sources like Strava and StreetLight actually make the old "small" data even more important. As we move away from simple correlation analysis using convenient permanent count sites and toward more rigorous and ambitious evaluation and applications, deficiencies in count and related data become clear.

Despite enthusiastic support and help from multiple agencies and jurisdictions, accessing and standardizing bike count data was challenging. Permanent count data came in multiple formats, and access was often delayed more than a calendar year. When we were able to access the data, we found poor uptime at many sites, resulting in only about two out of every three permanent counters able to provide 10 months of valid data per year. Short-duration counts were even more challenging to access and process. We also noted a tendency of regions to locate permanent counters in clusters of similar location types, resulting in little information about bicycle activity in different contexts. Robust, organized, and accessible count programs will be essential to get the most out of emerging data.

We also noted challenges with conventional modeling methods for emerging data sources. Direct demand and related models to date have mostly worked from "zero" in the sense that little direct information about bicycle use was typically available at a given location. That has led to model forms focused on extracting correlations of bicycling with adjacent land use, street networks, and other surrogate measures of use. These model forms and variable sets may be less suited to adjusting or extrapolating from data on actual use of a facility. We feel some progress was made, but this is an area in need of further research.

Third-party user data, theoretically available anywhere at any time, also puts fresh demands on contextual data such as land use, sociodemographic, and network data. Past standard practice has been to diligently assemble a static dataset, often for just one locale and at a single point in time. We initially explored this path but the relatively high consistency and availability of data from Strava and StreetLight expose the weaknesses in relying on patchy local jurisdiction data to support analysis. Many of the problems encountered with count data are also present in supporting data: access, standardization, timeliness, and extent. Big data do not stop at city boundaries, for example. This led us to pivot toward more standard open access data from OpenStreetMap (OSM) and similar standard, open datasets. Gathering data in this way is not without its own challenges, but automated processing techniques can partly offset the added effort of assembling larger datasets.

Although the initial modeling results suggest work remains to be done, they were also encouraging. We were able to assemble a standardized, reproducible dataset spanning six regions and to incorporate and apply two emerging data sources in a standard way.

The general consistency of the models adds confidence to some key findings. Strava and StreetLight can each provide valuable information on bicycle volumes at specific locations, but they still require traditional “small” data to help fill their blind spots. In almost every scenario considered, the best-performing models used a combination of three available data sources: Strava, StreetLight, and static contextual variables. Bikeshare, while showing potential alone, did not improve performance when added to the other data, at least as we specified it here. For our set of available count locations in six cities, Strava counts provided the most information about total bicycling activity. Adjusting Strava counts with contextual static variables noticeably improved model performance. Adding Streetlight counts to the Strava and static variables produced only marginal improvement, with some exceptions noted in the text.

We also tested various constructions of the data over time and space and compared more flexible machine learning methods with conventional multivariate count models. With the number of counter locations falling well short of “big” data, machine learning models did not, by and large, make better predictions than count models. That said, the Random Forest method we explored did more or less match the performance of carefully hand-specified models using a much more automated—and potentially more portable—process. The more flexible models would likely surpass conventional models given a larger sample of counts. We also found, unsurprisingly, that city-specific models outperformed a broader national pooled mode in terms of local performance. A regional pooled model (Oregon) produced mixed results. Local models required specifically tailored sets of contextual static variables to maximize performance and for that reason would be unlikely to transfer well to another locale, although we did not test that specifically. Pooled models might be a safer choice, although increased error rates should be expected. Finally, the models we developed transferred reasonably well to a prior year’s data in the same regions—even where count locations changed considerably—but we found it was important to reestimate model parameters from year to year for reliable performance, likely due at least in part to changes in third-party user data samples and methodology.

At the current time, data and methods appear capable of decent prediction performance at medium- to high-volume sites, at least at the sorts of locations where counts are available. Emerging third-party data are helpful. Future research could evaluate if these findings apply to pedestrian volume estimation. A key ongoing challenge is identifying which locations throughout a network are likely to be estimated reliably, or at least to be able to provide a likely error range along with each predicted count. A related challenge is designing programs to count cyclists where we typically have not to collect the information we need to expand our models and methods to cover more of complete networks more completely.

6.1 RECOMMENDATIONS

Recommendations are grouped into three categories to aid practitioners responsible for establishing bicycle count programs and modelers: preparing data for modeling, using results from modeling, and bicycle traffic monitoring.

Recommendations on preparing data for modeling:

- Emerging data are useful for predicting bicycle travel on individual road segments across a network and can improve model fit compared to models built on static variables alone.
    - Strava is generally a better-performing predictor of bicycle travel than StreetLight, but the two data sources tend to complement each other.
- Although it is often recommended to use more complex network buffers over simpler Euclidean buffers, they may not provide significantly better information to justify the effort required to develop them. It should be noted that we tested only shortest network paths here. More behaviorally realistic routing that considered the presence and quality of bicycle facilities might provide more useful measures and should continue to be explored in future work.
- One-mile and half-mile radius buffers are more useful for developing static variables than two-mile buffers. A two-mile radius can take up most of the study area, and thus the resulting buffer will not vary significantly from road segment to road segment. There is value in having someone familiar with both bicycle modeling and OSM data to pre-screen the variable list for model development, in order to reduce the number of potential predictor variables prior to model development. Machine learning is not (yet) able to improve upon human expert data selection, at least at likely count data sample sizes.
- Using consistent publicly available data sources from city to city (such as OpenStreetMap data) can result in easier-to-process data and substantially reduce data preparation time (using the scripts discussed in this report) and allow comparison across jurisdictions. However, these data sources may not be as predictive in one city as they are in another. For example, Portland, OR, has many “Neighborhood Greenways” (also known as bike boulevards, which encourage cycling on a particular local street) that are not specifically coded as such in the OpenStreetMaps data, but that may perform very differently than similarly coded bike routes in another city like Boulder. However, OSM does provide data across jurisdictional boundaries which might be helpful if analyzing travel across a metropolitan area or state.

Recommendations on using results from modeling:

- Don’t apply a model developed from data in one city to another. Bicyclist travel varies substantially from city to city, as does bicycle infrastructure, so city-specific models are expected to have better fit than pooled models within the city for which they were developed. A pooled model can provide insight into cycling in general, such as identifying important predictors of bicycle traffic.

- Don’t apply a model developed using one year’s data to another year without reestimating parameters. Strava data varies from year to year based on the number of users and the number of trips those users log. StreetLight’s raw data are also subject to changes. Both data sources can have data changes in algorithms internal to their platforms from year to year. Also, bicycle travel patterns, both temporally and spatially, vary from year to year.
- Predicting bicycle travel at lower-volume locations (which is most of the network) is inherently difficult, and results in high and highly variable error. However, for some applications, the error in terms of total cyclists may not be as large or concerning as it looks if the goal is simply to classify locations as having high, medium or low bicycle volumes.

Recommendations for bicycle traffic monitoring:

- Permanent continuous counter site selection:
    - More permanent continuous bicycle count stations are needed, especially at low-volume count sites. This is especially important if agencies want to use emerging data sources to estimate bicycle volume at all road segments across a network. The vast majority of permanent continuous bicycle count sites are at high-volume sites, specifically selected because they are high-volume sites. However, the vast majority of roads in the U.S. have low bicycle volumes. It is natural for count programs to begin with high-volume sites so that there is enough data to calibrate the counters. However, a count program with a few counters should start establishing counters at low-volume sites. This leaves the prediction of bicycle volumes across most of the network with little or no way to be validated and calibrated.
    - Locating permanent continuous counters at sites with a variety of bicycle travel patterns across the network is similarly important. This usually means installing counters on different facility types (not just recreational trails) and in different regions of a metropolitan area, which is especially important when estimating counts across the entire network. For example, in Dallas, the count program’s focus on off-street paths prevented us from developing adjustment factors for different on-street bicycle facility types. Based on the importance of such factors in other study cities, we would not expect the Dallas model to predict on-street cycling with similar accuracy.
- When automated counters are used, they should be validated in the field.
    - For short-duration counters, such as tube counters, this could mean a quick post-validation of 10 cyclists to determine if cyclists on both sides of the street are being counted correctly.
    - For permanent counters, this should be a more robust and ongoing process that is documented and shared with data users so that they can correctly account for site-specific under- or overcounting. These data are the foundation (the ground truth) for any bicycle volume estimates or

models and, thus, extra care is needed for their accuracy and maintenance.

- Maintain permanent continuous counting equipment. Data gaps of many months decrease the usefulness of data provided by these counters. Make the most of investment in equipment by adequately funding maintenance. This includes identifying and fixing problems promptly.
- Standardizing count data and aggregating datasets in a shared repository would greatly expedite modeling. Methods for doing this have already been proposed and should be implemented more widely (TMG 2016, BikePed Portal).
- Short-duration counts of 24 hours or more can be used, but also introduce sources of additional error. Future research can evaluate the tradeoffs of adding many more short-duration sites vs. fewer, but more robust, permanent count sites.

7 REFERENCES

Ambühl, L., Menendez, M. “Data fusion algorithm for macroscopic fundamental diagram estimation”. *Transportation Research Part C: Emerging Technologies* 71, 184–197, 2016. https://doi.org/10.1016/j.trc.2016.07.013

American Community Survey, 2017.

Anand, R.A., Vanajakshi, L., Subramanian, S.C. “Traffic density estimation under heterogeneous traffic conditions using data fusion” 2011 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2011 IEEE Intelligent Vehicles Symposium (IV), IEEE, Baden-Baden, Germany, pp. 31–36, 2011. https://doi.org/10.1109/IVS.2011.5940397

Baas, J., Galton, R., Biton, A. *FHWA Bicycle-Pedestrian Count - Technology Pilot Project* (No. FHWA-HEP-17-012), 2016. FHWA, Washington, DC.

Bachmann, C., Abdulhai, B., Roorda, M.J., Moshiri, B. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. *Transportation Research Part C: Emerging Technologies* 26, 33–48, 2013. https://doi.org/10.1016/j.trc.2012.07.00

Bachmann, C., Abdulhai, B., Roorda, M.J., Moshiri, B. Multisensor Data Integration and Fusion in Traffic Operations and Management. *Transportation Research Record: Journal of the Transportation Research Board* 2308, 27–36, 2012. https://doi.org/10.3141/2308-04

Beitel, D., McNee, S., Miranda-Moreno, L.F. Quality Measure of Short-Duration Bicycle Counts. *Transportation Research Record: Journal of the Transportation Research Board* 2644, 64–71, 2017. https://doi.org/10.3141/2644-08

Boeing, G. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *Computers, Environment and Urban Systems*, 65, 126-139, 2017. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Boss, D., Nelson, T., Winters, M., Ferster, C.J. Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. *Journal of Transport & Health* 9, 226–233, 2018. https://doi.org/10.1016/j.jth.2018.02.008

Broach, J., Dill, J. Using Predicted Bicyclist and Pedestrian Route Choice to Enhance Mode Choice Models. *Transportation Research Record: Journal of the Transportation Research Board* 2564, 52–59, 2016. https://doi.org/10.3141/2564-06

Broach, J., Dill, J., Gliebe, J. Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A: Policy and Practice 46, 1730–1740, 2012. https://doi.org/10.1016/j.tra.2012.07.005

Budowski, A., Mclaughlin, F., Nordback, K., Montufar, J. Estimating Seasonal Average Daily Bicyclists Using 2-hour Short Duration Counts, 2017. Presented at the TRB 2017 Annual Meeting, Washington, DC.

CDM Research. *How reliable is Strava?* 2018. http://cdmresearch.com.au/post/how-reliable-is-Strava/ (accessed 3.26.19).

CDOT. *Strava Metro Data Analysis Summary*. Colorado Department of Transportation, 2018.

Chen, C. *Crowdsourcing Data-driven Development of Bicycle Safety Performance Functions (SPFs): Microscopic and Macroscopic Scales*. Oregon State University, 2017.

Chen, P., Zhou, J., Sun, F. Built environment determinants of bicycle volume: A longitudinal analysis*. Journal of Transport and Land Use*, 2017. https://doi.org/10.5198/jtlu.2017.892

Chu, L., Oh, J.-S., Recker, W. Adaptive Kalman Filter Based Freeway Travel Time Estimation., 2005. Presented at the TRB 2005 Annual Meeting, Washington, DC, p. 21.

Cipriani, E., Gori, S., Mannini, L. Traffic state estimation based on data fusion techniques. Presented at the 2012 15th International IEEE Conference on Intelligent Transportation Systems - (ITSC 2012), IEEE, Anchorage, AK, USA, pp. 1477–1482, 2012. https://doi.org/10.1109/ITSC.2012.6338694

Conrow, L., Wentz, E., Nelson, T., Pettit, C. Comparing spatial patterns of crowdsourced and conventional bicycling datasets. *Applied Geography* 92, 21–30, 2018. https://doi.org/10.1016/j.apgeog.2018.01.009

Dadashova, B., Griffin, G., Das, S., Turner, S. Graham, M. *Guide for Seasonal Adjustment and Crowdsourced Data Scaling* (Cooperative Research Program Technical Report No. 0-6927-P6). Federal Highway Administration and the Texas Department of Transportation, 2018.

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

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

Dill, J. and Voros, K. Factors Affecting Bicycling Demand: Initial Survey Findings from the Portland, Oregon Region. *Transportation Research Record*, Vol. 2031(1), pp 9-17, 2007.

Dill, J. Seven quick things I learned about bicycling nationally. Transportation Research and Education Center, 2015. URL https://trec.pdx.edu/blog/seven-quick-things-i-learned-about-bicycling-nationally (accessed 3.26.19).

Dion, F., Rakha, H. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transportation Research Part B: Methodological* 40, 745–766, 2006. https://doi.org/10.1016/j.trb.2005.10.002

El Faouzi, N.-E. *Bayesian and evidential approaches for traffic data fusion:*

methodological issues and case study, 2006.

El Faouzi, N.-E., Klein, L.A., De Mouzon, O. Improving Travel Time Estimates from Inductive Loop and Toll Collection Data with Dempster–Shafer Data Fusion. *Transportation Research Record: Journal of the Transportation Research Board* 2129, 73–80, 2009. https://doi.org/10.3141/2129-09

El Faouzi, N.-E., Leung, H., Kurian, A. Data fusion in intelligent transportation systems: Progress and challenges – A survey. *Information Fusion* 12, 4–10, 2011. https://doi.org/10.1016/j.inffus.2010.06.001

El Esawey, M. 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* 70, 125–137, 2018a. https://doi.org/10.1016/j.compenvurbsys.2018.03.002

El Esawey, M. Daily Bicycle Traffic Volume Estimation: Comparison of Historical Average and Count Models. *Journal of Urban Planning and Development* 144, 04018011, 2018b. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000443

El Esawey, M. Toward a Better Estimation of Annual Average Daily Bicycle Traffic: Comparison of Methods for Calculating Daily Adjustment Factors. *Transportation Research Record: Journal of the Transportation Research Board* 2593, 28–36, 2016. https://doi.org/10.3141/2593-04

El Esawey, M. Estimation of Annual Average Daily Bicycle Traffic with Adjustment Factors. *Transportation Research Record* 2443, 106–114, 2014. https://doi.org/10.3141/2443-12

El Esawey, M., Lim, C., Sayed, T., Mosa, A.I. Development of Daily Adjustment Factors for Bicycle Traffic. *Journal of Transportation Engineering* 139, 859–871, 2013. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000565

El Esawey, M., Mosa, A.I. Determination and Application of Standard K Factors for Bicycle Traffic. *Transportation Research Record: Journal of the Transportation Research Board* 2527, 58–68, 2015. https://doi.org/10.3141/2527-07

Esawey, M. El. Daily bicycle traffic volume estimation: Comparison of historical average and count models', *Journal of Urban Planning and Development*, 144(2), pp. 1–9, 2018. doi: 10.1061/(ASCE)UP.1943-5444.0000443.

Ermagun, A., Lindsey, G. and Hadden Loh, T. Bicycle, pedestrian, and mixed-mode trail traffic: A performance assessment of demand models, *Landscape and Urban Planning*. Elsevier, 177(May), pp. 92–102, 201). doi: 10.1016/j.landurbplan.2018.05.006.

Fagnant, D.J., and Kockelman, K. 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, 2016.

FHWA. *Traffic Monitoring Guide*. U.S. Department of Transportation, Washington, D.C, 2016.

FHWA. *Traffic Monitoring Guide*. U.S. Department of Transportation, Washington, D.C,

2013.

Figliozzi, M., Johnson, P., Monsere, C., Nordback, K. Methodology to Characterize Ideal Short-Term Counting Conditions and Improve AADT Estimation Accuracy Using a Regression-Based Correcting Function. *Journal of Transportation Engineering*, 140 (5), 2014. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000663

Fish, J., Young, S., Wilson, A., and Borlaug, B. *Validation of Non-Traditional Approaches to Annual Average Daily Traffic (AADT) Volume Estimation*, FHWA-PL-21-033, Federal Highway Administration, 2021.

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

García, F., Jiménez, F., Anaya, J., Armingol, J., Naranjo, J., de la Escalera, A. Distributed Pedestrian Detection Alerts Based on Data Fusion with Accurate Localization. *Sensors* 13, 11687–11708, 2013. https://doi.org/10.3390/s130911687

Griffin, G., Jiao, J. Crowdsourcing Bicycle Volumes: Exploring the role of volunteered geographic information and established monitoring methods. *URISA* 27, 58–66, 2015. https://doi.org/10.31235/osf.io/e3hbc

Griswold, J., Medury, A. and Schneider, R. Pilot models for estimating bicycle intersection volumes, *Transportation Research Record*, (2247), pp. 1–7, 2011. doi: 10.3141/2247-01.

Hallenbeck, M., Schewel, L., Co, S., and Wergin, J. *Guidelines for Obtaining AADT Estimates from Non-Traditional Sources*, Federal Highway Administration, 2021.

Han, J. *Multi-Sensor Data Fusion for Travel Time Estimation* (Doctoral). Imperial College London, London, UK, 2012.

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

Hankey, S., Lindsey, G., Marshall, J. Day-of-Year Scaling Factors and Design Considerations for Nonmotorized Traffic Monitoring Programs. *Transportation Research Record: Journal of the Transportation Research Board* 2468, 64–73, 2014. https://doi.org/10.3141/2468-08

Hankey, S. and Lindsey, G. Facility-demand models of peak period pedestrian and bicycle traffic: Comparison of fully specified and reduced-form models, *Transportation Research Record*, 2586, pp. 48–58, 2016. doi: 10.3141/2586-06.

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

Harrison, G., Grant-Muller, S. M., & Hodgson, F. C.. New and emerging data forms in transportation planning and policy: Opportunities and challenges for “Track and Trace” data. *Transportation Research Part C: Emerging Technologies*, 117, 2020. https://doi.org/10.1016/j.trc.2020.102672

He, S., Zhang, J., Cheng, Y., Wan, X., Ran, B. Freeway Multisensor Data Fusion Approach Integrating Data from Cellphone Probes and Fixed Sensors. Journal of Sensors 2016, 1–13, 2016. https://doi.org/10.1155/2016/7269382

Heesch, K.C., Langdon, M. The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health Promotion Journal of Australia* 27, 222–229, 2016. https://doi.org/10.1071/HE16032

Hernandez, S., Hyun, K. (Kate). Truck Weight Distributions at Traffic Count Sites Using WIM and GPS Data. Presented at the TRB 2017 Annual Meeting, Washington, DC, 2017.

Hernandez, S.V., Tok, A., Ritchie, S.G. Integration of Weigh-in-Motion (WIM) and inductive signature data for truck body classification. *Transportation Research Part C: Emerging Technologies* 68, 1–21, 2016. https://doi.org/10.1016/j.trc.2016.03.003

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

Hyun, K. (Kate), Tok, A., Ritchie, S.G. Long distance truck tracking from advanced point detectors using a selective weighted Bayesian model. *Transportation Research Part C: Emerging Technologies* 82, 24–42, 2017. https://doi.org/10.1016/j.trc.2017.06.004

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

Jones, M. G., Ryan, S., Donlon, J., Ledbetter, L., Ragland, D., and Arnold, L.S. ‘Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and Its Relationship to Land Use, Transportation, Safety, and Facility Type’, *PATH Research Report*, 2010. Available at: https://merritt.cdlib.org/d/ark%3A%2F13030%2Fm54f1sft/2/producer%2FPRR-2010-12.pdf%0Ahttps://trid.trb.org/view/919880.

Johnstone, D., Nordback, K., Lowry, M. *Collecting Network-wide Bicycle and Pedestrian Data: A Guidebook for When and Where to Count*. WSDOT, 2017.

Kumar, K., Parida, M., Katiyar, V.K. Short Term Traffic Flow Prediction for a Non Urban Highway Using Artificial Neural Network. Procedia - Social and Behavioral Sciences 104, 755–764, 2013. https://doi.org/10.1016/j.sbspro.2013.11.170

Krile, R., Feng, J., and Schroeder, J. *Assessing Roadway Traffic Count Duration and Frequency Impacts on Annual Average Daily Traffic Estimation: Assessing Accuracy Issues with Current Known Methods in AADT Estimation from Continuous Traffic Monitoring Data*, FHWA-PL-015-008, Federal Highway

Administration, 2014.

Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1 – 26, 2008. doi:http://dx.doi.org/10.18637/jss.v028.i05

Kusakabe, T., Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies* 46, 179–191, 2014. https://doi.org/10.1016/j.trc.2014.05.012

Kwigizile, V., Oh, J. and Kwayu, K. *Integrating Crowdsourced Data with Traditionally Collected Data to Enhance Estimation of Bicycle Exposure Measure*, 2019. Available at: http://www.wmich.edu/transportationcenter/trclc17-3%0A https://trid.trb.org/view/1483416.

Laustsen, K., Mah, S., Semler, C., Nordback, K., Sandt, L., Sundstrom, C., Raw, J., Jessberger, S. *Coding Nonmotorized Station Location Information in the 2016 Traffic Monitoring Guide Format*. FHWA, 2016.

Le, H., Tech, V., Buehler, R., Hankey, S. Multi-City, *National Scale Direct-Demand Models of Peak-Period Bicycle and Pedestrian Traffic* (No. DTRT13- G-UTC33). MTS UTC, Blacksburg, VA, 2017.

Lee, K. and Sener, I.N. Strava Metro data for bicycle monitoring: a literature review, *Transport Reviews*, 41:1, 2747, 2021. DOI: 10.1080/01441647.2020.1798558

League of American Bicyclists. *Bicycling & Walking in the United States 2018 Benchmarking Report*. League of American Bicyclists, Washington, D.C, 2018.

Lin, Z. and Fan, W. (David) 'Modeling bicycle volume using crowdsourced data from Strava smartphone application', *International Journal of Transportation Science and Technology*. Tongji University and Tongji University Press, 2020. doi: 10.1016/j.ijtst.2020.03.003.

Lindsey, G. H. Forecasting Use of Nonmotorized Infrastructure: Models of Bicycle and Pedestrian Traffic in Minneapolis, Minnesota, Transportation Research Board Annual Meeting 2011.

Lindsey, G., Hoff, K., Hankey, S., and Wang, X. 'Understanding the Use of Non-Motorized Transportation Facilities', *Facilities*, Final Report, University of Minnesota, 2012. Available at: http://conservancy.umn.edu/bitstream/132499/1/CTS 12-24.pdf.

Lindsey, G., Wang, J., Hankey, S., Pterka, M. *Modeling Bicyclist Exposure to Risk and Crash Risk: Some Exploratory Studies* (No. DTRT13- G-UTC35). Roadway Safety Institute, 2018.

Livingston, M., McArthur, D., Hong, J., & English, K. Predicting cycling volumes using crowdsourced activity data. *Environment and Planning B: Urban Analytics and City Science*. 2021. https://doi.org/10.1177/2399808320925822

Lu, T., Mondschein, A., Buehler, R., Hankey, S. Adding temporal information to direct-demand models: Hourly estimation of bicycle and pedestrian traffic in Blacksburg, VA. *Transportation Research Part D: Transport and Environment* 63,

244–260, 2018. https://doi.org/10.1016/j.trd.2018.05.011

McDaniel, S., Lowry, M.B., Dixon, M. Using Origin–Destination Centrality to Estimate Directional Bicycle Volumes. *Transportation Research Record: Journal of the Transportation Research Board* 2430, 12–19, 2014. https://doi.org/10.3141/2430-02

Minge, E., Falero, C., Lindsey, G., Petesch, M., Vorvick, T. *Bicycle and Pedestrian Data Collection Manual*, 2017.

Miranda-Moreno, L.F., Nosal, T., Schneider, R.J., Proulx, F. Classification of Bicycle Traffic Patterns in Five North American Cities. *Transportation Research Record: Journal of the Transportation Research Board* 2339, 68–79, 2013. https://doi.org/10.3141/2339-08

Munira, S., Sener, I.N. *Use of Direct-Demand Modeling in Estimating Nonmotorized Activity: A Meta-analysis*, Texas Transportation Institute, 2017.

Musakwa, W., Selala, K.M. Mapping cycling patterns and trends using Strava Metro data in the city of Johannesburg, South Africa. *Data in Brief* 9, 898–905, 2016. https://doi.org/10.1016/j.dib.2016.11.002

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., Chung, E. Real-time traffic state estimation in urban corridors from heterogeneous data. *Transportation Research Part C: Emerging Technologies* 66, 99–118, 2016. https://doi.org/10.1016/j.trc.2015.07.005

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

Niemeier, D. A. ‘Longitudinal analysis of bicycle count variability: Results and modeling implications’, *Journal of Transportation Engineering*, 122(3), pp. 200–206, 1996. doi: 10.1061/(ASCE)0733-947X(1996)122:3(200).

Nickkar, A., Banerjee, S., Chavis, C., Bhuyan, I., Barnes, P. ‘A spatial-temporal gender and land use analysis of bikeshare ridership: The case study of Baltimore City’, *City, Culture and Society,* 2019. doi: 10.1016/j.ccs.2019.100291.

Noland, R. B., Deka, D. and Walia, R. ‘A statewide analysis of bicycling in New Jersey’, *International Journal of Sustainable Transportation*. Taylor & Francis Group, 5(5), pp. 251–269, 2011a. doi: 10.1080/15568318.2010.501482.

Noland, R. B., Deka, D. and Walia, R. ‘A statewide analysis of bicycling in New Jersey’, *International Journal of Sustainable Transportation*. Taylor & Francis Group, 5(5), pp. 251–269, 2011b. doi: 10.1080/15568318.2010.501482.

Nordback, K., Kothuri, S., Johnstone, D., Lindsey, G., Ryan, S., Raw, J. Minimizing AADNT Estimation Errors: How Many Counters are Needed per Factor Group? Presented at the Transportation Research Board Annual Meeting, p. 28, 2019.

Nordback, K., Kothuri, S., Petritsch, T., McLeod, P., Rose, E., Twaddell, H. *Exploring Pedestrian Counting Procedures: A Review and Compilation of Existing*

Procedures, Good Practices, and Recommendations (No. FHWA-HPL-16-026). FHWA, Washington, DC, 2016.

Nordback, K., Marshall, W.E., Janson, B.N., Stolz, E. Estimating Annual Average Daily Bicyclists: Error and Accuracy. *Transportation Research Record: Journal of the Transportation Research Board* 2339, 90–97, 2013. https://doi.org/10.3141/2339-10

Nordback, K., O'Brien, S., Blank, K. *Bicycle and Pedestrian Count Programs: Summary of Practice and Key Resources*. Pedestrian and Bicycle Information Center, Chapel Hill, NC, 2018.

Nordback, K.L. *Estimating annual average daily bicyclists and analyzing cyclist safety at urban intersections* (PhD Thesis). University of Colorado at Denver, 2012.

Nosal, T., Miranda-Moreno, L.F., Krstulic, Z. Incorporating Weather: Comparative Analysis of Annual Average Daily Bicyclist Traffic Estimation Methods. *Transportation Research Record: Journal of the Transportation Research Board* 2468, 100–110, 2014. https://doi.org/10.3141/2468-12

Ortúzar, J., Willumsen, L.G. *Modelling Transport*. Wiley, 2002.

O'Toole, K., Piper, S. *Innovation in Bicycle and Pedestrian Counts: A Review of Emerging Technology*. Alta Planning + Design, 2016.

PeopleForBikes,. U.S. Bicycling Participation Benchmarking Study Report, 2015.

Pikora, T., Giles-Corti, B., Bulla, F., Jamrozika, K., and Donovan, R. Developing a framework for assessment of the environmental determinants of walking and cycling, *Social Science & Medicine*, 56(3), pp. 49–57, 2003.

Portland BikeTown Data. Accessed at https://s3.amazonaws.com/biketown-tripdata-public/index.html

Premebida, C., Nunes, U. Fusing LIDAR, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research* 32, 371–384, 2013. https://doi.org/10.1177/0278364912470012

Proulx, F.R., Pozdnukhov, A. *Bicycle Traffic Volume Estimation using Geographically Weighted Data Fusion*, 2017.

Qing-Jie K., Li, Z., Chen,Y., and Liu,Y. An Approach to Urban Traffic State Estimation by Fusing Multisource Information. *IEEE Transactions on Intelligent Transportation Systems* 10, 499–511, 2009. https://doi.org/10.1109/TITS.2009.2026308

Roll, J.F. *Bicycle Count Data: What Is It Good For? A Study of Bicycle Travel Activity in Central Lane Metropolitan Planning Organization*. Oregon Department of Transportation, 2018.

Roll, J.F., Proulx, F.R. Estimating Annual Average Daily Bicycle Traffic without Permanent Counter Stations. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. https://doi.org/10.1177/0361198118798243

Romanillos, G., Austwick, M.Z., Ettema, D., Kruijf, J.D. Big Data and Cycling. *Transport Reviews* 36, 114–133, 2016. https://doi.org/10.1080/01441647.2015.1084067

Roy, A., Nelson, T., Fotheringham, A.S., Winters, M. 'Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists', *Urban Science*, 3(2), p. 62, 2019. doi: 10.3390/urbansci3020062.

Ryus, P., Butsick, A.J., Proulx, F.R., Schneider, R.J., Hull, T. *Methods and Technologies for Pedestrian and Bicycle Volume Data Collection: Phase 2*, 2016.

Ryus, P., Ferguson, E., Laustsen, K.M., Schneider, R.J., Proulx, F.R., Hull, T., Miranda-Moreno, L. *Guidebook on Pedestrian and Bicycle Volume Data Collection*. Transportation Research Board, Washington, D.C, 2014. https://doi.org/10.17226/22223

Saad, M., Abdel-Aty, M., Lee, J., & Cai, Q. Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(4), pp 1-14, 2019.

Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R., Hintze, M. Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington: Potential Option for Resource-Constrained Cities in an Age of Big Data. *Transportation Research Record: Journal of the Transportation Research Board* 2605, 32–44, 2017. https://doi.org/10.3141/2605-03

Schroeder, P., Wilbur, M. 2012 *National survey of bicyclist and pedestrian attitudes and behavior, volume 2: Findings report* (No. DOT HS 811 841 B). National Highway Traffic Safety Administration, Washington, DC, 2013.

Semler, C., Vest, A., Kingsley, K., Mah, S., Kittelson, W., Sundstrom, C., Brookshire, K. *Guidebook for developing pedestrian and bicycle performance measures. FHWA*., 2016.

Sener, I. N., Munira, S., & Zhang, Y. (2021). *Data Fusion for Nonmotorized Safety Analysis* (Final Report No. 03–049). Texas A&M Transportation Institute. https://trid.trb.org/view/1875768

Srour, F.J., Newton, D. Freight-Specific Data Derived from Intelligent Transportation Systems: Potential Uses in Planning Freight Improvement Projects. *Transportation Research Record: Journal of the Transportation Research Board* 1957, 66–74, 2006. https://doi.org/10.1177/0361198106195700110

Steele, K., Altmaier, M., 2010. *Bicycling and Walking in the U.S.: 2010 Benchmarking Report*. Alliance for Biking & Walking, Washington, DC.

Strauss, J. and Miranda-Moreno, L. F. (2013) 'Spatial modeling of bicycle activity at signalized intersections', *Journal of Transport and Land Use*, 6(2), pp. 47–58. doi: 10.5198/jtlu.v6i2.296.

Strauss, J., Miranda-Moreno, L. F. and Morency, P. (2013) 'Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach', *Accident Analysis and Prevention*. Elsevier Ltd, 59, pp. 9–17. doi: 10.1016/j.aap.2013.04.037.

Strauss, J., Miranda-Moreno, L.F., Morency, P. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. *Accident Analysis & Prevention* 83, 132–142, 2015. https://doi.org/10.1016/j.aap.2015.07.014

Strava. *Bike Counter Correlation*, 2014.

StreetLight Data. *StreetLight AADT 2018 V2 Methodology and Validation White Paper: United States + Canada*, 2019.

StreetLight Data. *StreetLight Active Mode Methodology and Validation*, 2018.

Sun, Y., Du, Y., Wang, Y., Zhuang, L. Examining Associations of Environmental Characteristics with Recreational Cycling Behaviour by Street-Level Strava Data. *International Journal of Environmental Research and Public Health* 14, 644, 2017a. https://doi.org/10.3390/ijerph14060644

Sun, Y., Moshfeghi, Y., Liu, Z. Exploiting crowdsourced geographic information and GIS for assessment of air pollution exposure during active travel. *Journal of Transport & Health* 6, 93–104, 2017b. https://doi.org/10.1016/j.jth.2017.06.004

Tabeshian, M. and Kattan, L. 'Modeling Nonmotorized Travel Demand at Intersections in Calgary, Canada: Use of Traffic Counts and Geographic Information System Data', *Transportation Research Record*, pp. 38–46, 2014. doi: 10.3141/2430-05.

Tsapakis, Y., Turner, S., Joeneman, P., Anderson, P. *Independent Evaluation of a Probe-Based Method to Estimate Annual Average Daily Traffic Volume*. FHWA-PL-21-032, Texas A&M Transportation Institute [TTI], 2021.

Turner, S., Benz, R., Hudson, J., Griffin, G., Dadashova, B., Das, S. *Improving the Amount and Availability of Pedestrian And Bicyclist Count Data in Texas* (Technical Report No. 0-6927-R1), Cooperative Research Program. Federal Highway Administration and the Texas Department of Transportation, 2019a.

Turner, S., Lasley, P. Quality Counts for Pedestrians and Bicyclists: Quality Assurance Procedures for Nonmotorized Traffic Count Data. *Transportation Research Record* 2339, 57–67, 2013. https://doi.org/10.3141/2339-07

Turner, S., Lasley, P., Hudson, J., Benz, R. *Guide for Pedestrian and Bicyclist Count Data Submittal*. Federal Highway Administration and the Texas Department of Transportation, 2019b.

Turner, S., Sener, I.N., Martin, M.E., Das, S., Hampshire, R.C., Fitzpatrick, K., Molnar, L.J., Colety, M., Robinson, S., Shipp, E. *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities*. United States. Federal Highway Administration. Office of Safety, 2017.

Unnikrishnan, A., Figliozzi, M., Moughari, M.K., Urbina, S. *A Method to Estimate Annual Average Daily Traffic for Minor Facilities for Map-21 Reporting and Statewide Safety Analysis: Literature Review and Analysis of Data Sources*. Oregon Department of Transportation & Federal Highway Administration, 2017.

Wang, X., Lindsey, L., Hankey, S., and Hoff, K. Estimating mixed-mode urban trail traffic using negative binomial regression models, *Journal of Urban Planning and Development*, 140(1), pp. 1–9, 2014. doi: 10.1061/(ASCE)UP.1943-5444.0000157.

Wang, H., Chen, C., Wang, Y., Pu, Z., Lowry, M.B. *Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions*. Final Report, Pacific Northwest Transportation Consortium, 2016.

Watkins, K., Ammanamanchi, R., LaMondia, J., Le Dantec, C.A. Comparison of smartphone-based cyclist GPS data sources. Presented at the TRB 95th Annual Meeting Compendium of Papers, 2016.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C. Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 12, 1624–1639, 2011. https://doi.org/10.1109/TITS.2011.2158001

Zhu, L., Guo, F., Polak, J.W., Krishnan, R. Urban link travel time estimation using traffic states-based data fusion. IET *Intelligent Transport Systems* 12, 651–663, 2018. https://doi.org/10.1049/iet-its.2017.0116

8 APPENDIX

8.1 STRAVA DATA NOTES

This section documents some unexpected issues that we encountered with the Strava data. Where applicable, workarounds we adopted are also described.

8.1.1 Short link undercounts

One issue that was noted was that counts tend to drop off (sometimes by 50% or more) on short links as seen in Figure 8-1 below. We confirmed w/ Strava that it's due to map matching limitations. The fix adopted was to use Strava only for links at least 20 miles long.

Figure 8-1 Short Link Undercounts

8.1.2 Junction paradox

Another issue that was noted was that funnel counts did not add up to the main line count. In Figure 8-2, 2018 Strava rider counts equal 515 and 465 (980 total) for the two-entering links, while the main line equals 780. Since the incidence of this error was low, and there was no way to automate the detection and fixing of this issue, the research team did not take further action on this issue.

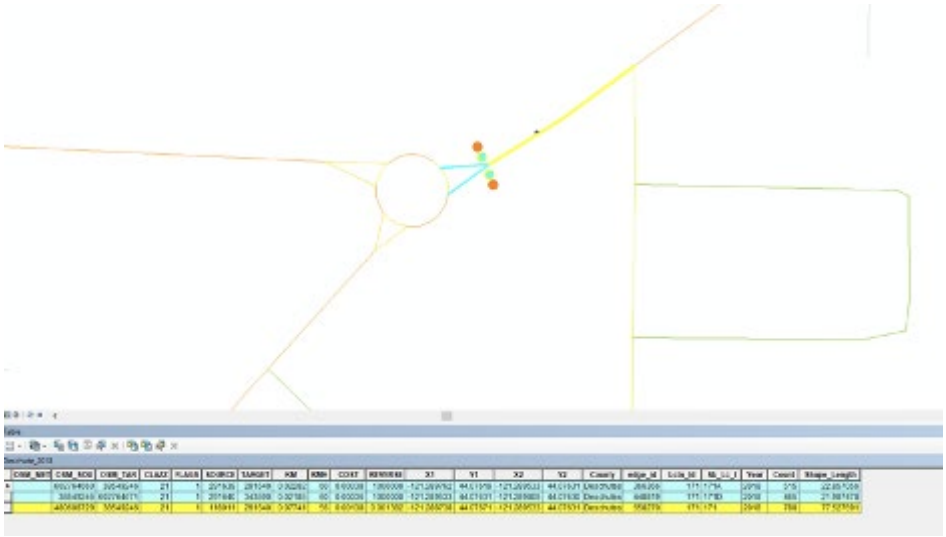


Figure 8-2 Junction Paradox

8.1.3 **Parallel links map matching problem**

Another issue that we noted was that bike counts often were wrongly assigned to the nearby parallel links. As seen in Figure 8-3, bike counts that were supposed to be largely assigned to the cycle track were also assigned to the pedestrian path and auto lanes. These instances were flagged and corrected.

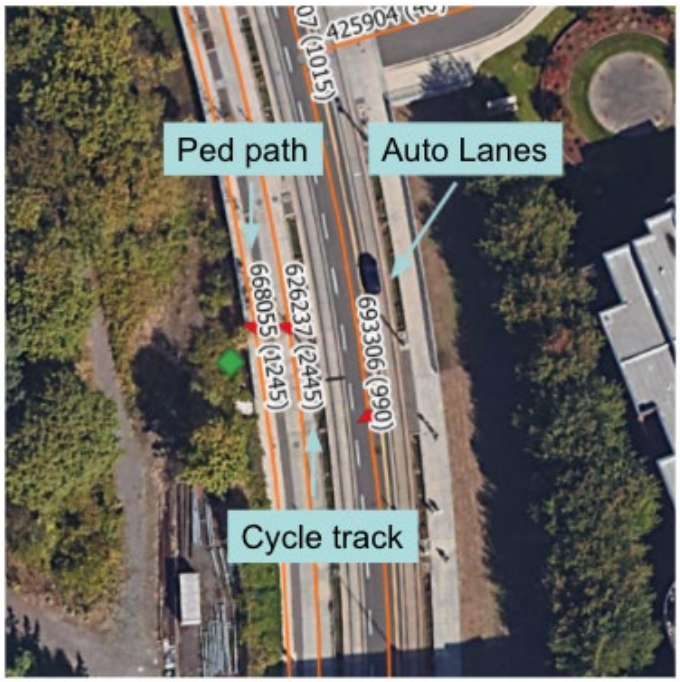


Figure 8-3 Parallel Links

8.1.4 **Rounding Error**

The Strava data that was made available to the research team included hourly and monthly rollups (but not daily). We noted that hourly counts less than 3 were masked (not reported) and hourly counts greater than 3 were rounded up to the next 5. When aggregated hourly counts were compared to the monthly rollups, undercounting was generally observed, with occasional large overcounts as seen in Figure 8-4. Low volume locations could not be rolled up from the hourly counts accurately. It is anticipated that future Strava data will include daily rollups, which will solve this issue.

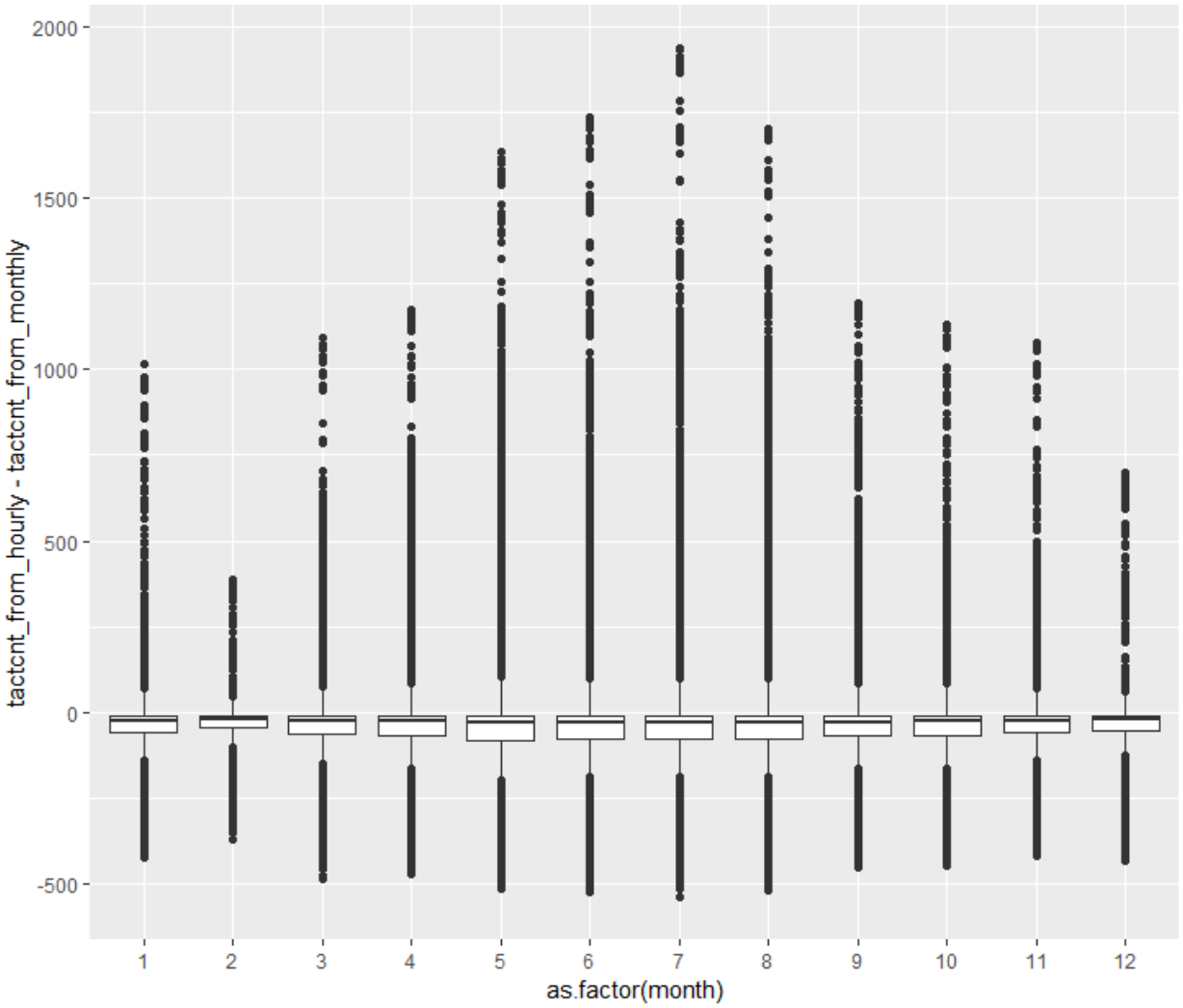


Figure 8-4 Difference between Hourly Aggregation and Monthly Rollups by Month

8.2 STREETLIGHT

Issues that we noted with the StreetLight data were overcounts at some locations which were along busy transit routes (e.g., Steel Bridge in Portland), with the assumption being that some transit riders were being considered as bicyclists, possibly due to similar speeds between the modes. Additionally, undercounts were also observed along busy bicycle corridors such as the bike boulevards in SE Portland. See Figure 8-5.

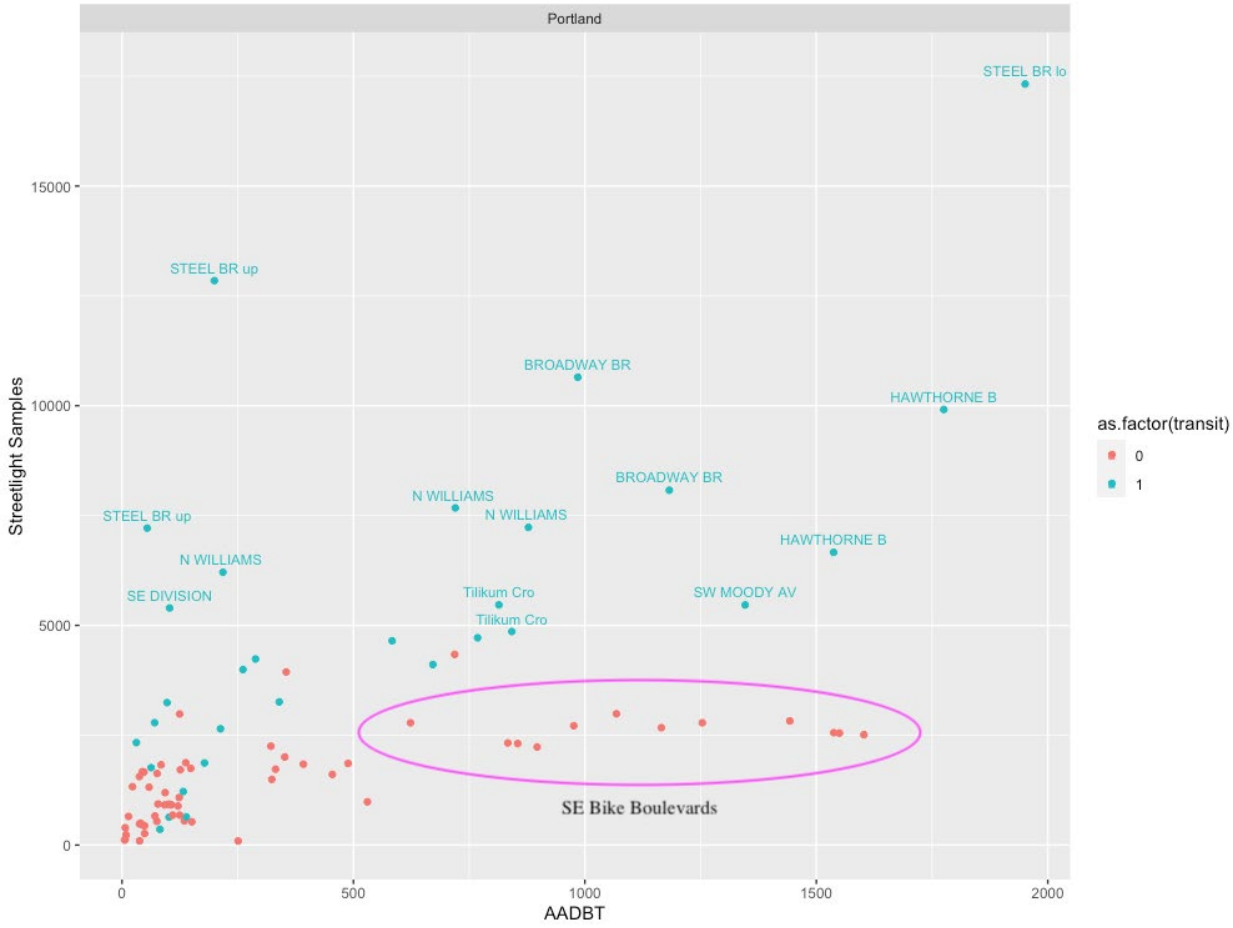


Figure 8-5 StreetLight Data on Bicycle Boulevards

8.3 DATA PROCESSING SCRIPTS

The research team developed automated python scripts to extract required static variables and uploaded them onto Gitlab (https://gitlab.com/joebroach/bike-data-fusion; see Figure 8-6). Gitlab is a web-based collaboration service that allows users to track, integrate and develop individual codes and integrate with others. GitLab enhances collaboration productivity because each team member who has access to the repository could simultaneously collaborate and share codes. The research team's workflow and data processing steps in Gitlab are described below.

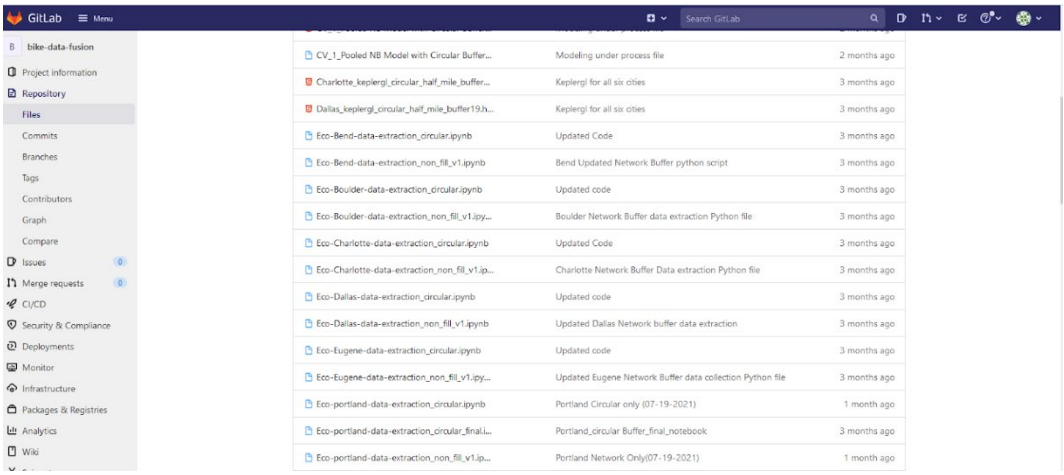


Figure 8-6 A Screenshot of Sourcetree Software Interface and Python Scripts in GitLab

Library: An initial step creates a separate environment in desktop python to install necessary libraries such as Geopanda, panda, networkx, osmnx, shapely and matplotlib. Geopanda library reads and processes geospatial data and can be downloaded from the link (https://medium.com/analytics-vidhya/fastest-way-to-install-geopandas-in-jupyter-notebook-on-windows-8f734e11fa2b). Networkx and osmnx libraries extract data from OpenStreetMap such as node density, intersection density, number of lanes, and bicycle facilities.

Data inputs: After installing Geopandas Environment in python, the developed python script reads necessary data inputs including counter locations, Strava network, Bikeshare, OSM (shape files from BBBike), LEHD job data, NED raster elevation data and NHGIS data.

Geocode type: A user needs to enter a geocode for data aggregations. In general, any geofile uses its original EPSG which could be replaced with a specific local EPSG. For example, all land use shape files for Portland, OR, are set to EPSG 4326 (default), but should be changed to EPSG 2838 because this code is specifically assigned to Portland in the GIS environment. This local EPSG ensures to extract an accurate size of buffers for data aggregation. Note that the developed script already sets this EPSG information for six study locations.

Buffer size: A user could use any given buffer size to generate and aggregate the data, as shown in Figure 8-7. This study extends one mile from the study location's boundary (e.g., Portland Metro region, Figure 8-7 (a)) to create an extended boundary (Figure 8-7(b)) to capture counters that are located close to the geographical boundary of the study area.

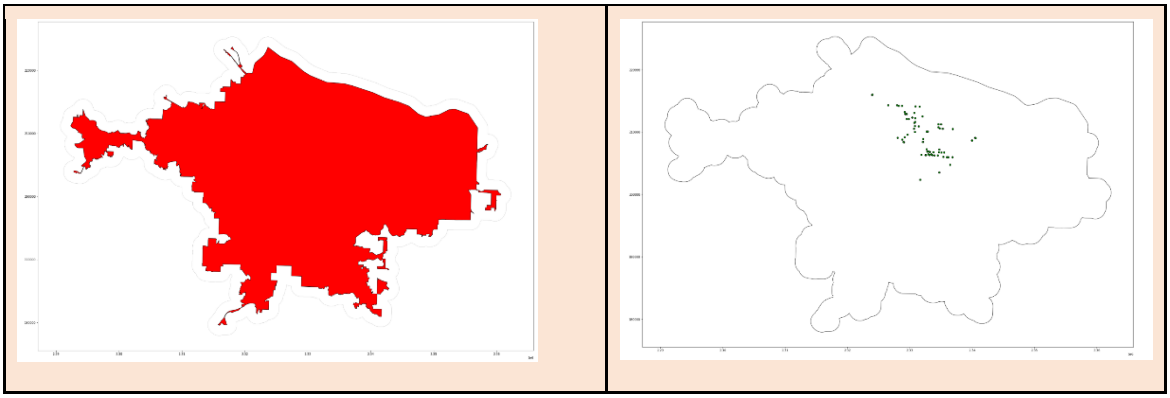


Figure 8-7 (a) Original and (b) Extended Boundary of Study Area

Variable generation - Density: Variables defining network/facility density such as the number of major generators (e.g., schools, colleges and universities) and network features (e.g., intersection density, number of lanes, and bicycle facilities) were extracted using direct overpass API through osmnx. The research team also uses BBBike source to supplement land use datasets.

Variable generation - Slope: Slope data was extracted using a National Elevation Dataset (NED) raster image (Figure 8-8). This study extracts an elevation at each node of the link and calculates its slope. The research team assumed that a bridge slope is zero.

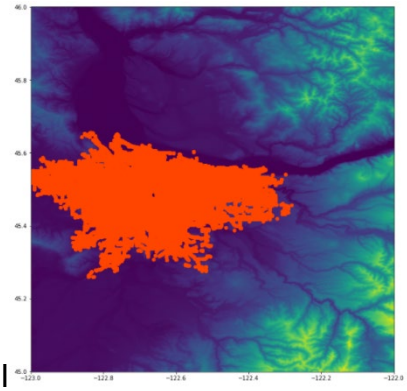


Figure 8-8 Slope Extraction from NED Raster Image

Variable generation - Weather: Weather data was collected from weather underground using direct overpass API. The collected data was automatically aggregated as a single data frame and exported in csv format for further analysis and modeling.

Variable visualization: The final data frames are passed to the KeplerGL program to create dynamic and interactive visual maps. KeplerGL is a high-performance web-based application for large-scale geo-spatial data visual exploration. KeplerGl visualizes data in 2D or 3D space. A user selects variables to visualize and saves an interactive map as HTML format to share with others. Figure 8-9 shows the heatmap of bikeshare origin data for Portland, OR.

Figure 8-9 KeplerGL Dynamic Heat Map of Bikeshare Origin at Portland, OR

Knowledge or skill required: To run the primary python script requires only basic familiarity with the Python language. As set up, users just need to set the buffer size and rename the save file as desired or needed. To extend beyond the six regions included in this study, local boundary (GIS) files also need to be supplied.

Count data cleaning and count modeling scripts: Also provided via GitLab are scripts used to process raw count data (from physical counters and third-party user data sources Strava and StreetLight). These scripts are written in the R programming language and are best classified as "semi-automated." Because bike count data (especially short-duration counts) lack a standard format, users should expect considerable tweaking of either the raw data themselves or of the processing scripts. Strava modified how their data are formatted and accessed during the course of this research, and Strava scripts will need to be updated to the new standard. Methods were described in the body of this report, such that users should be able to reproduce our methods using the software or programming language of their choice.

Count (R) and ML (Python) modeling scripts are also included. The count models are described in a way that they could be estimated in a variety of statistical software packages, and the code provided mainly serves as a record and template for future applications. While the ML scripts are more easily transferred to new data, interpreting the output likely will require specialized knowledge or external assistance.

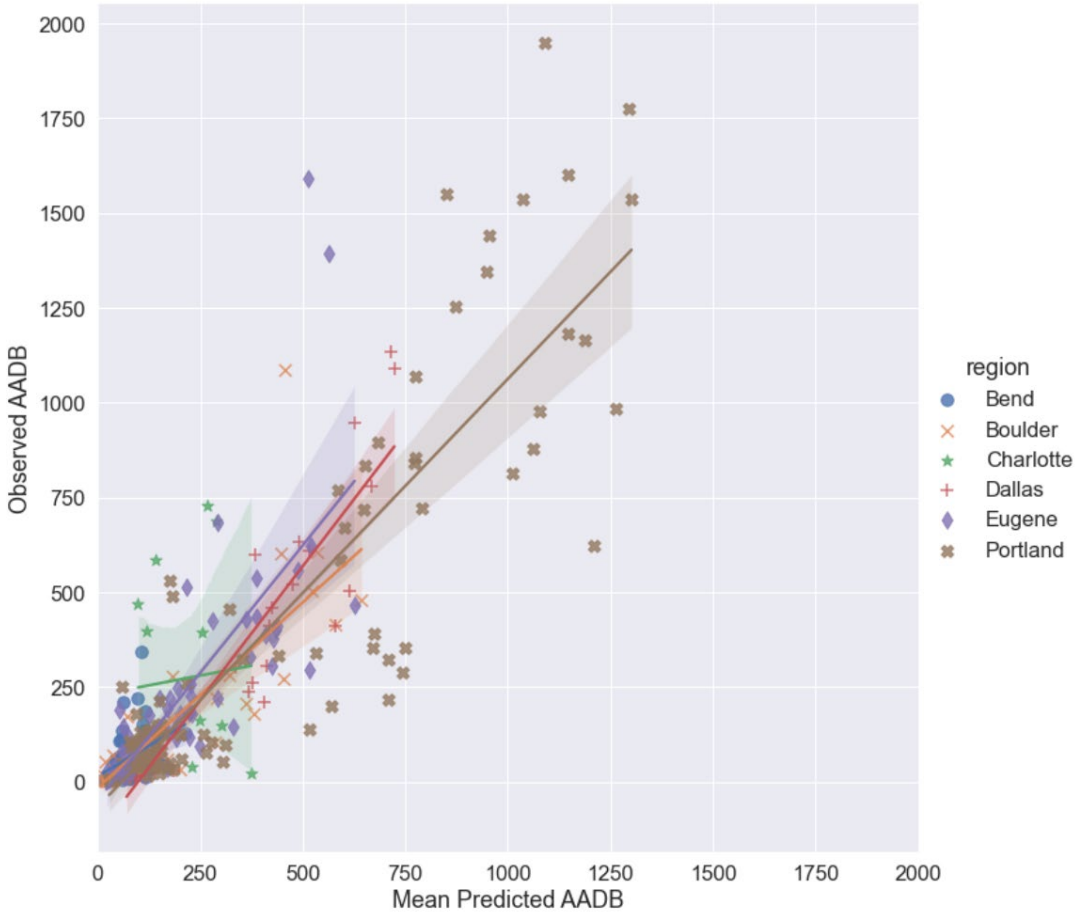


Figure 8-10 Pool RF Model [Static+Strava+SL] Observed Vs Prediction Fit

Table 8-1 Variable Formulations for Pooled Random Forest Models [Static+Strava+StreetLight]

Var	Importance	Def
log_stv_stl	0.1022	log(Strava+1)+Log(StreetLight)
log_stv_c_adb	0.0928	Natural log transformation Strava commute trips
stv_c_adb	0.0805	Natural log transformation Strava AADBT
log_stv_adb	0.0684	Strava commute AADBT
stv_adb	0.0564	Strava AADBT
stv_nc_adb	0.0454	Strava non-commute AADBT
log_stv_nc_adb	0.0352	Natural log transformation of Strava non-commute AADBT
Intersection_Density_om	0.0317	Number of intersections per square mile in one-mile air buffer
log_stl_raw	0.0309	Natural log transformation StreetLight Raw Trips
stl_raw	0.0243	StreetLight raw trips
Number.of.jobs_om	0.0174	Area weighted number of jobs at count station block groups/group for one-mile air buffer
Tertiary_om	0.0172	Total length of tertiary road segments within the one-mile air buffer around each count station
Student.Access_om	0.0168	Area weighted number of students within the one-mile air buffer at count station block groups/group

Var	Importance	Def
employment_density_om	0.0143	Area weighted employment per square mile at count station block groups/group for one-mile air buffer
Number.of.jobs_hm	0.0137	Area weighted number of jobs at count station block groups/group for half mile air buffer
Distance.to.Water.Body	0.0110	nearest distance to water body from the count station
Footway_hm	0.0110	Total length of footway segments within the half mile air buffer around each count station
Median_HH_income_om	0.0110	Area weighted Median household income within the one-mile radius buffer at count station block group
Bike.Commuter_om	0.0105	Number of Bike commuter within one-mile air buffer
Distance.to.Industrial.Area	0.0095	Nearest distance to industrial area from count station.
Distance.to.CBD	0.0093	Nearest distance to count station from City Hall (CBD)
Median.Age_om	0.0091	Median age of the population within the one-mile air buffer at count station block group
Student.Access_hm	0.0091	Area weighted number of students within the half mile air buffer at count station block groups/group
pct_at_least_college_education_hm	0.0090	Area weighted % of population having at least college degree within the one-mile air buffer at count station block group
Distance.to.Park	0.0087	nearest distance to park area from count station
Primary_om	0.0087	Total length of primary road segments within the one-mile air buffer around each count stations
Median.Age_hm	0.0081	Median age of the population within the half mile air buffer at count station block group
employment_density_hm	0.0079	Area weighted employment per square mile at count station block groups/group within half mile air buffer
Percentage.of.Bike.Commuter_om	0.0077	Percentage bike commute or bike commuters/sq mi for one-mile air buffer
population_density_om	0.0074	Area weighted population per square mile at count station block groups/group for one-mile air buffer
pct_African_American_om	0.0073	Area weighted percentage of African American population within the one mile air buffer at count station block group
population_density_hm	0.0072	Area weighted population per square mile at count station block groups/group for half mile air buffer
Park.Area_hm	0.0072	Acres of park w/in half-mile radius
Tertiary_hm	0.0072	Total length of tertiary road segments within the half mile air buffer around each count station
Distance.to.forest	0.0066	Nearest distance to forest area from count station
Industrial.Area_hm	0.0065	Industrial area within the half mile air buffer
BikeFac_om	0.0065	Miles of dedicated bike facility (on or off-street) in one-mile radius (does NOT include just signed bike routes or shared lanes)
Intersection_Density_hm	0.0061	Number of intersections per square mile in half mile air buffer
Primary_hm	0.0060	Total length of primary road segments within the half mile air buffer around each count stations
pct_at_least_college_education_om	0.0058	Area weighted % of population having at least college degree within the one-mile air buffer at count station block group
Footway_om	0.0058	Total length of footway segments within the one-mile air buffer around each count station
Path_om	0.0057	Total length of path segments within the one-mile air buffer around each count station

Var	Importance	Def
sep_bikeway_om	0.0055	length of path or cycleway within one-mile air buffer
Cycleway_om	0.0055	Total length of cycleway lane, left and right segments within the one-mile air buffer around each count station
Cycleway_hm	0.0053	Total length of cycle way segments within the half mile air buffer around each count station
Bike.Commuter_hm	0.0053	Number of Bike commuter (ACS) within ½-mi radius
Path_hm	0.0051	Total length of path segments within the half mile air buffer around each count station
cycleway_lane_all_om	0.0050	Total length of cycleway lane, left and right segments within the one-mile air buffer around each count station
sep_bikeway_hm	0.0047	Length of on path or cycleway within half mile air buffer
Median_HH_income_hm	0.0047	Area weighted Median household income within the half mile radius buffer at count station block group
Water.Area_om	0.0043	Water body area within the one-mile air buffer
pct_white_om	0.0042	Area weighted percentage of white population within the one-mile air buffer at count station block group
Park.Area_om	0.0039	Acres of park w/in one-mile radius
min_dist_to_university	0.0038	Straight distance between count station and University
Residential_Road_om	0.0038	Total length of residential road segments within the one-mile buffer around each count station
pct_African_American_hm	0.0037	Area weighted percentage of African American population within the half mile air buffer at count station block group
Grass.Area_hm	0.0036	Grass area within the half mile air buffer
Residential_Road_hm	0.0035	Total length of residential road segments within the half mile buffer around each count station
Percentage.of.Bike.Commuter_hm	0.0034	Percentage bike commute or bike commuters/sq mi for half mile air buffer
Distance.to.Grass	0.0033	nearest distance to grass space from count station.
Commercial.Area_hm	0.0031	Commercial area within the half mile air buffer
Grass.Area_om	0.0030	Grass area within the one-mile air buffer
BikeFac_hm	0.0030	Miles of dedicated bike facility (on or off-street) in half-mile radius (does NOT include just signed bike routes or shared lanes)
path_binary	0.0028	Location on path
Forest.Area_hm	0.0028	Forest area within the half mile buffer
cycleway_binary	0.0028	Location of Cycleway
pct_white_hm	0.0024	Area weighted percentage of white population within the half mile air buffer at count station block group
Secondary_om	0.0023	Total length of secondary road segments within the one-mile buffer around each count station
cycleway_lane_binary	0.0023	Location on cycleway_lane
Secondary_hm	0.0023	Total length of secondary road segments within the half mile buffer around each count station
Forest.Area_om	0.0021	Forest area within the one-mile buffer
sep_bikeway_binary	0.0020	Location on path or cycleway
Water.Area_hm	0.0020	Water body area within the half mile air buffer
Industrial.Area_om	0.0020	Industrial area within the one-mile air buffer
Distance.to.Retail.Area	0.0018	Nearest distance to retail area from count station
slope_hm	0.0018	average absolute % slope along the link within half mile air buffer
slope_om	0.0017	average absolute % slope along the link within one-mile air buffer
School_om	0.0016	Number of schools within one-mile air buffer
Retail.Area_om	0.0015	Retail area within the one-mile air buffer

Var	Importance	Def
cycleway_lane_all_hm	0.0015	Total length of cycleway lane, left and right segments within the half mile air buffer around each count station
secondary_binary	0.0015	Location on Secondary
arterial_binary	0.0015	Location on primary or secondary arterial
Retail.Area_hm	0.0014	Retail area within the half mile air buffer
Distance.to.Commercial.Area	0.0014	Nearest distance to commercial area from count station.
Commercial.Area_om	0.0012	Commercial area within the one-mile air buffer
Distance.to.Residential.Area	0.0012	Nearest distance to residential area from count station.
min_dist_to_school	0.0011	Straight distance between count station and School
residential_binary	0.0008	Location on residential
BikeFac_binary	0.0006	Location of on or off-street
maxspeed_om	0.0006	The mode of speed for the functional class was obtained within the one-mile air buffer
lanes_hm	0.0005	Number of traffic lanes along corresponding count station street segment within half mile air buffer
School_hm	0.0005	Number of schools within half mile air buffer
Point.Speed	0.0004	Speed limit on the link where the counter is situated; if unavailable the speed of the nearest link with the same functional class is extracted
maxspeed_hm	0.0003	The mode of speed for the functional class was obtained within the half mile air buffer
University_om	0.0002	Number of Universities (yes=1 and no=0) within the one-mile air buffer
University_hm	0.0001	Number of Universities (yes=1 and no=0) within the half mile air buffer

Table 8-2 Variables Used in RF Model

Study Area	Fusion RF Model	Variables Used
Pool-2019 Model	SL	'log_stl_raw'
	Strava	'log_stv_adb'
	SL+Strava	'log_stv_adb','log_stl_raw'
	SL+Strava+Static (PM4)	Please see Pool Model RF Variable Importance graph
	SL+Strava+Static+BS	PM4+ Bikeshare Best_hm+Bikeshare Best_om
Oregon-2019 Model	SL	'log_stl_raw'
	Strava	'log_stv_c_adb', 'log_stv_nc_adb'
	SL+Strava	'log_stv_adb','log_stl_raw'
	SL+Strava+Static (OM4)	Please see Oregon Pool Model RF Variable Importance graph
	SL+Strava+Static+BS	OM4+ Bikeshare Best_hm+Bikeshare Best_om
Portland-2019 Model	SL	'log_stl_raw'
	Strava	'log_stv_c_adb'
	SL+Strava	'log_stv_c_adb', 'log_stl_raw'
	SL+Strava+Static (PM4)	Please see Portalnd Pool Model RF Variable Importance graph
	SL+Strava+Static+BS	PM4+ Bikeshare Best_hm+Bikeshare Best_om
Eugene-2019 Model	SL	'log_stl_raw'
	Strava	'log_stv_c_adb','log_stv_nc_adb'
	SL+Strava	'log_stv_c_adb','log_stl_raw'
	SL+Strava+Static (EM4)	Please see Eugene Pool Model RF Variable Importance graph
	SL+Strava+Static+BS	M4+ Bikeshare Best_hm+Bikeshare Best_om

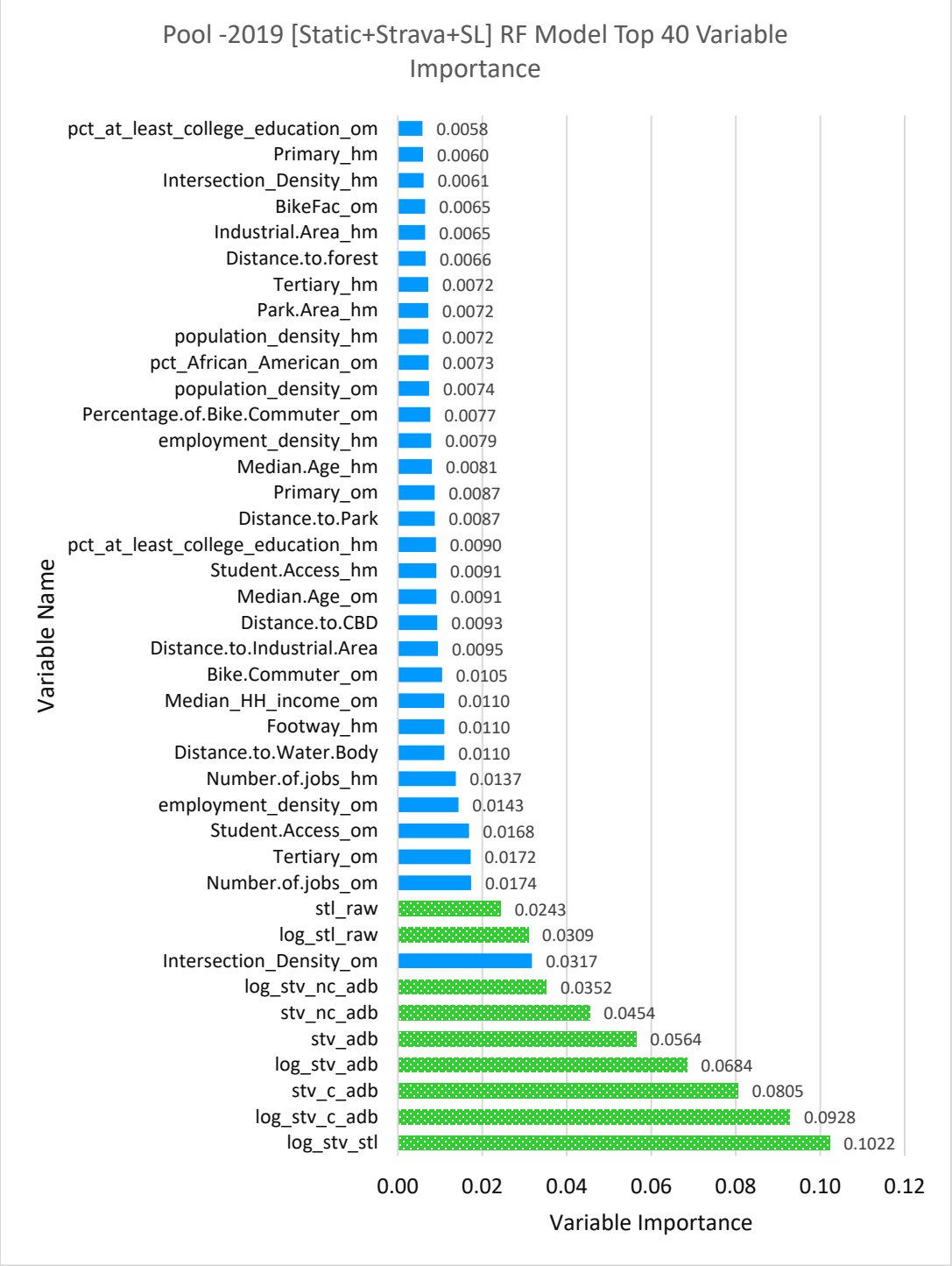


Figure 8-11 Pool -2019 [Static+Strava+SL] RF Model Top 40 Variable Importance

Note: Additional Variables (importance <0.006): Footway_om, Path,_om, sep_bikeway_om, Cycleway_om, Cycleway_hm, Bike.Commuter_hm, Path_hm, cycleway_lane_all_om, sep_bikeway_hm, Median_HH_income_hm, Water.Area_om, pct_white_om, Park.Area_om, min_dist_to_university Residential,_Road_om,

pct_African_American_hm, Grass.Area_hm, Residential_Road_hm, Percentage.of.Bike.Commuter_hm, Distance.to.Grass Commercial.Area_hm, Grass.Area_om, BikeFac_hm, path_binary, Forest.Area_hm, cycleway_binary, pct_white_hm, Secondary_om ,cycleway_lane_binary, Secondary_hm, Forest.Area_om, sep_bikeway_binary, Water.Area_hm, Industrial.Area_om, Distance.to.Retail.Area, slope_hm, slope_om, School_om, Retail.Area_om, cycleway_lane_all_hm, secondary_binary, arterial_binary, Retail.Area_hm, Distance.to.Commercial.Area, Commercial.Area_om, Distance.to.Residential.Area, min_dist_to_school, residential_binary, BikeFac_binary, maxspeed_om, lanes_hm, School_hm, Point.Speed, maxspeed_hm, University_om, University_hm.

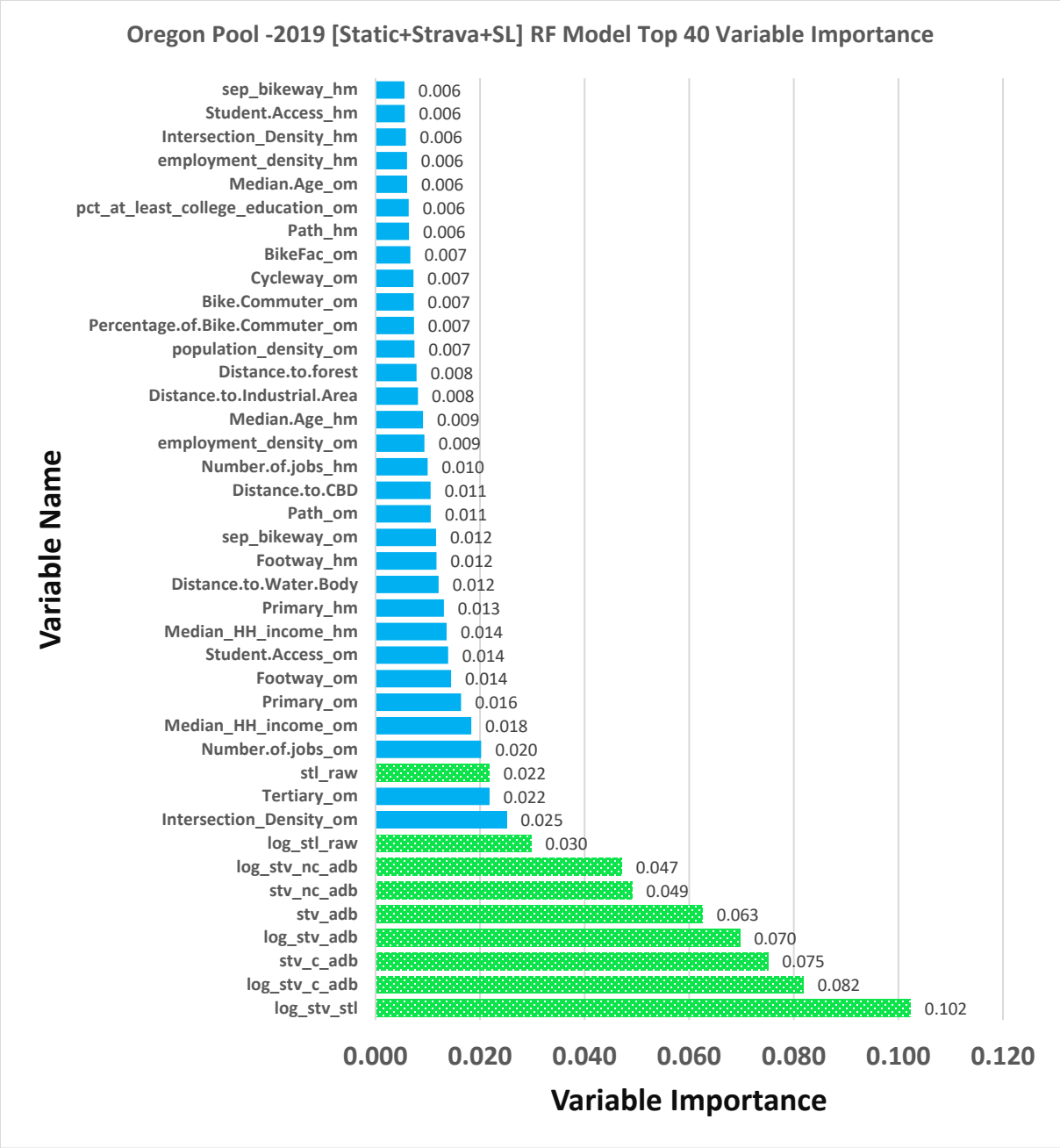


Figure 8-12 Oregon Pool -2019 [Static+Strava+SL] RF Model Top 40 Variable Importance

Note: Additional Variables (importance <0.006): Commercial.Area_hm, path_binary, Grass.Area_hm, Distance.to.Grass, pct_African_American_om,population_density_hm,Industrial.Area_hm,Residential_Road_om,Bike.Commuter_hm,Grass.Area_om,pct_at_least_college_education_hm,Cycleway_hm,pct_African_American_hm,Tertiary_hm,min_dist_to_university,pct_white_om,maxspeed_om,Percentage.of.Bike.Commuter_hm,Park.Area_hm,Residential_Road_hm,Retail.Area_hm,cycleway_lane_all_om,secondary_binary,arterial_binary,Distance.to.Park,BikeFac_hm,Forest.Area_hm,pct_white_hm,Water.Area_hm,Retail.Area_om,School_o

m,slope_om,Park.Area_om,Forest.Area_om,cycleway_lane_binary,Secondary_hm,Water.Area_om,Industrial.Area_om,slope_hm,sep_bikeway_binary,Secondary_om,cycleway_lane_all_hm,cycleway_binary,Distance.to.Retail.Area,Distance.to.Commercial.Area,Commercial.Area_om,Distance.to.Residential.Area,min_dist_to_school.

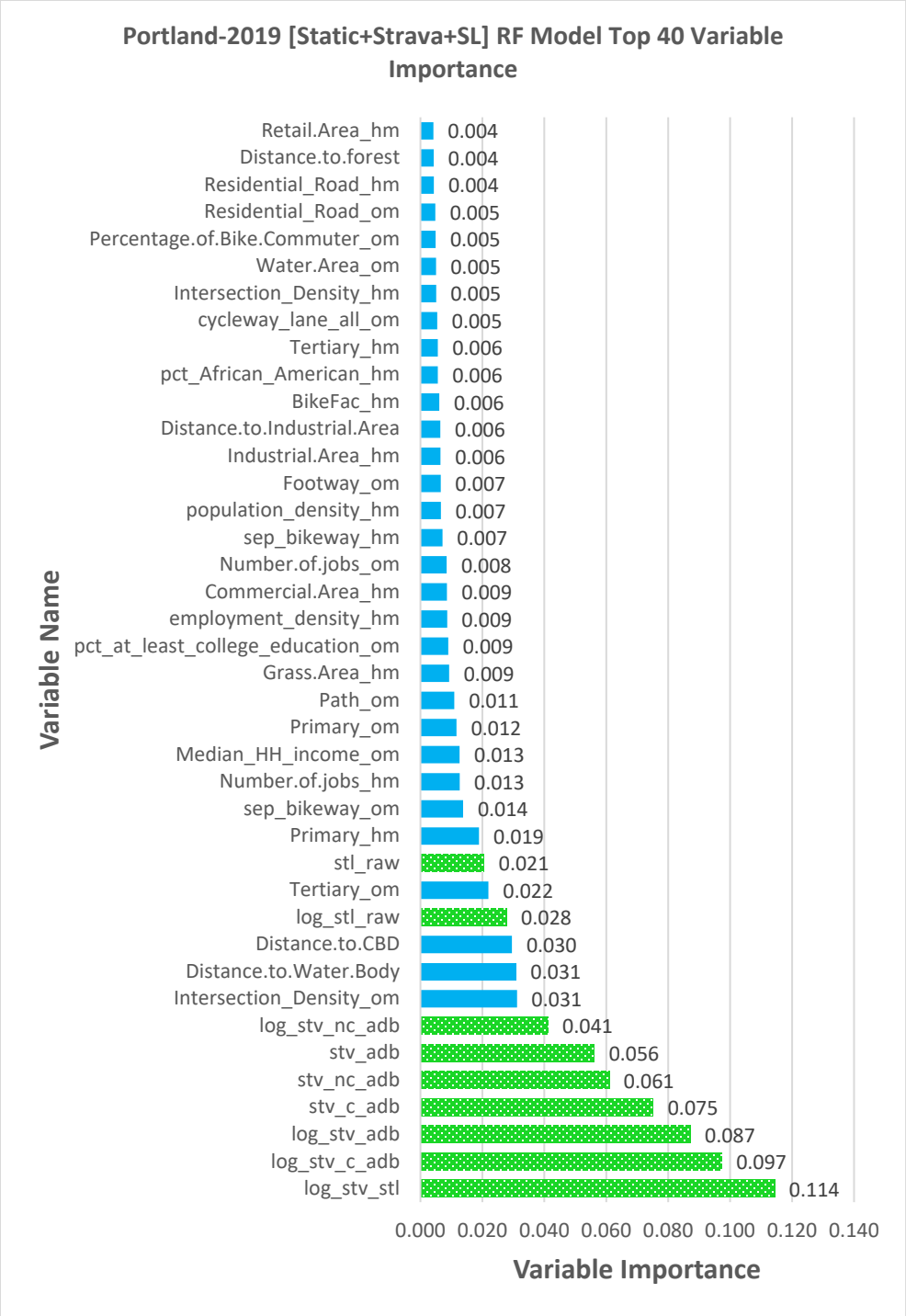


Figure 8-13 Portland-2019 [Static+Strava+SL] RF Model Top 40 Variable Importance

Note: Additional Variables (importance <0.004):
sep_bikeway_binary,Grass.Area_om,min_dist_to_university,Cycleway_om,pct_at_least_college_education_hm,cycleway_lane_binary,Path_hm,Park.Area_hm,Percentage.of.Bike.Commuter_hm,Bike.Commuter_hm,Distance.to.Park,Industrial.Area_om,slope_hm,Park.Area_om,Median_HH_income_hm,BikeFac_om,Distance.to.Grass,Water.Area_h

m,Secondary_om,Retail.Area_om,Secondary_hm,Student.Access_om,Cycleway_hm,pct_white_hm,pct_African_American_om,University_hm,School_om,arterial_binary,Median.Age_om,lanes_hm,population_density_om,cycleway_lane_all_hm,Distance.to.Retail.Area,pct_white_om,min_dist_to_school,Forest.Area_hm,Commercial.Area_om,secondary_binary,Bike.Commuter_om,School_hm,Student.Access_hm,Distance.to.Residential.Area,Median.Age_hm,Point.Speed,path_binary,Forest.Area_om,BikeFac_binary.

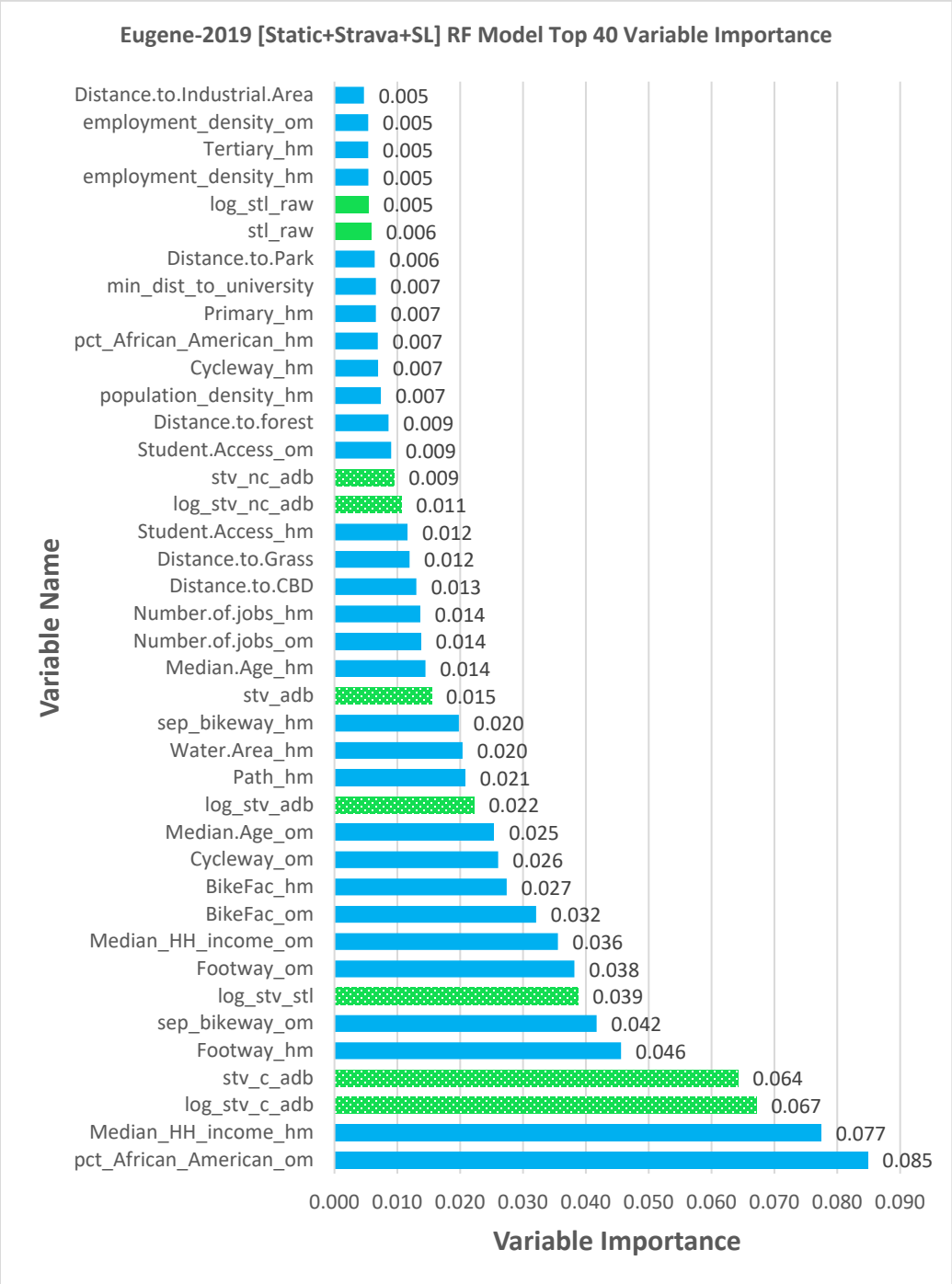


Figure 8-14 Eugene-2019 [Static+Strava+SL] RF Model Top 40 Variable Importance

Note: Additional Variables (importance <=0.005):
Distance.to.Residential.Area,BikeFac_binary,Distance.to.Water.Body,Percentage.of.Bike.Commuter_hm,Distance.to.Commercial.Area,pct_white_om,population_density_om,Tertiary_om,Grass.Area_hm,Grass.Area_om,Water.Area_om,pct_at_least_college_education_om,Secondary_om,Retail.Area_hm,Park.Area_om,Distance.to.Retail.Area,Park.A

rea_hm,Forest.Area_om,slope_om,Path_om,Forest.Area_hm,Primary_om,cycleway_la
ne_all_om,slope_hm,University_om,sep_bikeway_binary,pct_at_least_college_educatio
n_hm,Bike.Commuter_hm,Retail.Area_om,Bike.Commuter_om,cycleway_lane_all_hm,
min_dist_to_school,Residential_Road_om,School_hm,cycleway_binary,Residential_Ro
ad_hm,Industrial.Area_hm,Intersection_Density_om,Percentage.of.Bike.Commuter_om,
residential_binary,pct_white_hm,Secondary_hm,maxspeed_hm.

8.4 2017 AND 2018 MODEL RESULTS

Table 8-3 AADBT Poisson Portland Model All Counters 2018 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	7.4339	2.2518***	1.7212.	1.6019*	2.9565	0.3974	2.1165
log(stv_adb +1)					0.6365***		0.7756***
log(stv_c_adb +1)		1.2095***		0.9570***			
log(stl_raw + 1)			0.6172***	0.2026		0.1783	0.1019
sep_bikeway_binary	1.2015				0.8768	1.9776***	
Median_HH_income_om					-0.00001		-0.00001
Bike.Commuter_om	-0.0003				-0.0005	-0.0002	-0.0007*
Distance_to_CBD_mi	-0.3465						
BikeFac_om	0.0172						
intersection_density_om	-0.0020				0.0046	0.0091	0.0020
pct_at_least_college_education_hm	-0.0094				0.0017	0.0183	0.0158
Park_acres_hm	-0.0034				-0.0047	-0.0048	-0.0017
slope_om	0.0819						
Model fit statistics (presented for final model)							
N	33	33	33	33	33	33	33
AIC[b]	2490	3,034	7,955	2,689	1330	1,827	1410
Pseudo-R^2[c]	0.883	0.853	0.595	0.872	0.943	0.917	0.939
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	776 (+/- 174)	211 (+/- 28)	412 (+/- 63)	263 (+/- 39)	212 (+/- 30)	266 (+/- 30)	215 (+/- 31)
MAPE (+/- SE)	342%(+/- 155)	50% (+/- 3)	98% (+/- 15)	44 (+/- 3)	51% (+/- 4)	60% (+/- 6)	52% (+/- 4)
MAE (+/- SE)	493 (+/- 97)	161%(+/- 18)	278 (+/- 38)	188 (+/- 24)	158 (+/- 18)	187 (+/- 19)	156 (+/- 19)

Table 8-4 AADBT Poisson Portland Model All Counters 2017 Results (10-fold cross-validation with five repeats and robust SEs)

	parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001						
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	-1.7933	3.7222***			1.9694		
log(stv_adb +1)					0.5838***		
log(stv_c_adb +1)		0.8344***					
sep_bikeway_binary	0.7408				-0.00001		
cycleway_lane_binary					1.2778***		
arterial_binary	0.6239				0.1416		
Bike.Commuter_om	-0.0003				0.5178		
Distance_to_CBD_mi	-0.0866				-0.0002		
BikeFac_om	0.0014						
intersection_density_om	0.0097.				0.0087*		
pct_at_least_college_education_hm	0.0703.				-0.004		
Park_acres_hm	-0.0073				-0.0167		
slope_om	-0.0961						
Model fit statistics (presented for final model)							
N	104	104			104		
AIC[b]	19,493	17,984			10,245		
Pseudo-R^2[c]	0.586	0.619			0.790		
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	388 (+/- 33)	296 (+/- 20)			259 (+/- 20)		
MAPE (+/- SE)	158% (+/- 17)	105% (+/- 6)			84% (+/- 4)		
MAE (+/- SE)	238 (+/- 15)	191 (+/- 11)			163 (+/- 10)		

Table 8-5 AADBT Poisson Oregon Pooled Model All Counters 2018 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	2.3443**	4.4008***	1.5651**	2.2226***	5.3325***	3.0996***	4.2346***
log(stv_adb +1)				0.3681***	0.6550***		0.4572***
log(stv_c_adb +1)		0.8094***					
log(stv_nc_adb +1)		-0.1646					
log(stl_raw + 1)			0.6333***	0.3949***		0.4272***	0.2387**
sep_bikeway_binary	0.5971*					0.5319	
cycleway_lane_binary	-0.3951*					-0.0549	
Median_HH_income_om					-0.0267***		-0.00003***
arterial_binary	-0.0686				-0.5815***		
primary_binary						-1.5059*	-1.1075*
secondary_binary						-0.5177	
tertiary_binary						-0.1683	
I(secondary_binary + tertiary_binary)							-0.5840***
Bike.Commuter_om	0.00003				0.0003.	0.0002	0.0002
Distance_to_CBD_mi	-0.7155				-0.2727***	-0.2299**	-0.1262*
BikeFac_om	0.05952***						
Intersection_Density_om	0.0034						
pct_at_least_college_education_hm	0.0148						
Park_acres_hm	-0.0002					-0.0031	
slope_om	-0.0035						
Model fit statistics (presented for final model)							
N	176	176	176	176	176	176	176
AIC[b]	21,891	27,410	28,925	23,760	16,581	19,159	14468
Pseudo-R^2[c]	0.647	0.553	0.527	0.615	0.737	0.693	0.772
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	250 (+/-19)	234(+/-16)	267 (+/- 24)	229 (+/- 17)	215(+/-15)	238(+/-20)	193(+/-17)
MAPE (+/- SE)	132%(+/-7)	153%(+/-6)	162% (+/- 11)	130% (+/- 6)	100%(+/-4)	101%(+/-5)	89%(+/-4)

MAE (+/- SE)	142(+/-7)	147(+/-6)	148 (+/- 8)	135 (+/- 7)	116(+/-6)	129(+/-7)	111(+/-7)

Table 8-6 AADBT Poisson Eugene Model All Counters 2018 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	5.8157*	4.2428***	1.4809*	2.1981**	6.7478***	4.8578**	5.2751***
log(stv_adb +1)					0.7782***		0.6831***
log(stv_c_adb +1)		2.0108***		0.8353***			
log(stv_nc_adb +1)		-0.8320**					
log(stl_raw + 1)			0.6822***	0.3994**		0.4362**	0.1956*
Median_HH_income_om					0.0000005	0.000003	
sep_bikeway_binary	0.4577.						
cycleway_lane_binary							
arterial_binary	-0.2611				-0.4642*	-0.6388*	-0.5414*
Distance_to_CBD_mi					-0.0919	-0.0608	
BikeFac_om	0.0482						
Bike.Commuter_om					0.0004	0.0002	0.0005*
log(min_dist_to_university)	-0.2476				-0.3282***	-0.2295	-0.2882***
Park_acres_hm	0.0033						
slope_om	0.0105						
Model fit statistics (presented for final model)							
N	86	86	86	86	86	86	86
AIC[b]	10,168	9,417	13,252	9,768	5739	9170	5315
Pseudo-R^2[c]	0.552	0.587	0.407	0.570	0.759	0.599	0.779
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	221 (+/- 20)	185 (+/- 14)	211 (+/- 19)	207 (+/- 17)	164 (+/- 20)	197 (+/- 17)	146 (+/- 16)
MAPE (+/- SE)	105% (+/- 6)	119% (+/- 10)	138% (+/- 8)	104% (+/- 7)	69% (+/-4)	114% (+/-7)	64% (+/- 4)
MAE (+/- SE)	140 (+/- 10)	123 (+/- 8)	139 (+/- 9)	130 (+/- 8)	100 (+/- 10)	132 (+/- 9)	91 (+/- 8)

Table 8-7 AADBT Poisson Bend Model All Counters 2018 Results (10-fold cross-validation with five repeats and robust SEs)

parameters estimate w/ significance level. <=0.1. * <=0.05 **<=0.01 ***<=0.001							
	PM0: Static Model	PM1: Strava Only	PM2: StreetLight Only	PM3: Strava + StreetLight	PM4: Static + Strava	PM5: Static + StreetLight	PM6: Static + Strava + StreetLight
(Intercept)	3.8683	4.0005***	2.9805***	2.7841***	4.4834***	0.8864	0.5507
log(stv_adb +1)					0.4455**		0.3877*
log(stv_c_adb +1)		0.7123**		0.5961*			
log(stl_raw + 1)			0.2932**	0.2401*		0.7112***	0.5491**
sep_bikeway_binary	0.3680				0.6061*	1.6784***	1.4109***
Bike.Commuter_om	-0.0004					-0.0034	
Distance_to_CBD_mi	-0.2401					0.0321	
intersection_density_om	0.0036					-0.0026	
log(Distance_to_CBD_mi + 1)					-0.9221**		0.2452
pct_at_least_college_education_hm	0.0089					0.0035	
Park_acres_hm	-0.0002					0.0038	
slope_om	-0.0040					0.0195	
Model fit statistics (presented for final model)							
N	58	58	58	58	58	58	58
AIC[b]	3,807	4156	4,182	3824	3257	3016	2800
Pseudo-R^2[c]	0.210	0.128	0.122	0.204	0.334	0.392	0.440
Cross-validation test performance (mean of 10 folds)							
RMSE (+/- SE)	95(+/- 5)	85 (+/- 5)	87 (+/- 5)	85 (+/- 5)	79 (+/- 5)	83 (+/- 4)	73 (+/- 4)
MAPE (+/- SE)	201% (+/- 24)	149% (+/- 11)	166% (+/-15)	140% (+/- 11)	128% (+/- 13)	157% (+/- 16)	109% (+/- 9)
MAE (+/- SE)	78 (+/- 4)	6 (+/- 3)	69 (+/- 3)	68 (+/- 3)	63 (+/- 3)	66 (+/- 3)	57 (+/- 3)