Challenges and Opportunities of Emerging Data Sources to Estimate Network-wide Bike Counts

Md. Mintu Miah, MS
Graduate Researcher
Department of Civil Engineering
University of Texas at Arlington
425 Nedderman Hall
Arlington, TX 76019, USA
mdmintu.miah@mavs.uta.edu
808-387-4052
ORCID: 0000-0001-6073-3896

Kate Kyung Hyun, PhD
Assistant Professor
Department of Civil Engineering
University of Texas at Arlington
425 Nedderman Hall
Arlington, TX 76019, USA
kate.hyun@uta.edu
817-272-9748
ORCID: 0000-0001-7432-8058

Stephen P Mattingly, PhD
Professor
Department of Civil Engineering
University of Texas at Arlington
425 Nedderman Hall
Arlington, TX 76019, USA
mattingly@uta.edu
817-272-2859
ORCID: 0000-0001-6515-6813

Joseph Broach, PhD
Adjunct Research Associate
Toulan School of Urban Studies and Planning
Portland State University
PO Box 751
Portland, OR 97201
jbroach@pdx.edu
ORCID: 0000-0001-7753-501X

Nathan McNeil
Research Associate
Toulan School of Urban Studies and Planning
Portland State University
PO Box 751,
Portland, OR – 97201
nmcneil@pdx.edu
ORCID: 0000-0002-0490-9794

Sirisha Kothuri, PhD
Senior Research Associate
Department of Civil and Environmental Engineering
Portland State University
PO Box 751
Portland, OR – 97201
skothuri@pdx.edu
ORCID: 0000-0002-2952-169X

Paper Length
1044 in text
224 in Abstract
3 Figures
1,268 words (1,200 max)

ABSTRACT

Emerging sources of mobile location data such as Strava and other phone-based apps may provide useful information for assessing activity on each link of a network. Despite their potential to complement traditional bike count programs, the representativeness and the suitability of these emerging sources for producing bicycle volume estimates require further exploration. This study investigates the challenges and opportunities by fusing Strava data with short-term and permanent conventional count program data to produce bicycle volume estimations that could be expanded to entire networks. This study finds that the concentration of permanent counters at high bicycle volume locations presents a significant challenge to network-wide daily volume or AADBT modeling. Strava data demonstrates some potential in mitigating the resulting bias at lower-volume sites, but significant challenges remain to rely on Strava counts alone to characterize network-level activities. A non-parametric method using treed regression generally reduces the sampling bias and non-linearity between Strava and observed counts compared to a linear modeling approach; however effective network-level modeling requires more extensive data collection to avoid extrapolating the patterns present at higher volume sites to low volume sites. This study will help planners and stakeholders to discern the challenges and opportunities of using emerging data in bicycle volume estimation and to assess the potential for use in planning and decision making.

Key Words: Bicycle, Network, Volume, Strava, Treed Regression

1. INTRODUCTION

The recent increase in bicycling popularity nationwide has led to efforts to better measure riding activity, to improve the interpretation of available data, and to assess the impacts of cycling. Bicycle volumes (usually in the form of Daily Bicycle Traffic DBT or Average Annual Daily Bicycle Traffic, AADBT) are useful for measuring trends and prioritizing infrastructure investments, and as exposure/activity measures in safety and public health studies (*1*, *2*). However, observing bicycle traffic using automated counters across entire networks, or even large portions of them, is seen as impractical. Hence, several approaches have been employed to capture a more holistic picture of bicycle traffic throughout a network including factoring, extrapolation, and modeling through passively-collected smartphone or similar data, and more recently, efforts to fuse two or more of these approaches.

A handful of studies combined or "fused" count data with third-party volume and other data sources to model observed bicycle volumes. Dadashova et al. (*3*) added segment functional classifications and adjacent numbers of upper-income households (>$200k/yr). Jestico et al. (*4*) included segment slope, speed limit, on-street parking presence, and a seasonal adjustment along with Strava counts. Roll (*1*) combined Strava counts with segment functional class, bicycle facility types, local accessibility and design measures, and a measure of network centrality. Sanders et al. (*5*) included the number of bike lanes and proximity to the university in conjunction with Strava data. Roy et al. (*6*) incorporated speed, land use, and socio-demographic variables in count regressions including Strava counts. Even after controlling for facility and surrounding contextual factors, third-party (mainly Strava) data significantly improved bicycle volume and safety performance models.

This research begins to address several gaps in the existing literature on using third-party mobile location data for network-wide DBT/AADBT estimation. First, we explicitly allow for Strava prediction coefficients to vary by location context. Second, we use short term (low volume) sites to enhance modeling capabilities networkwide, which has been rare in work to date. Third, we employ non-parametric models alongside more traditional count models to understand whether these estimation techniques which are increasingly common for motorized volume estimation might also improve performance in non-motorized applications.

2. METHODOLOGY

This study investigates strategies to estimate network-wide bike volumes by fusing emerging data sources such as Strava with permanent count (PC) and short-term count (SC) data. The study uses a deliberate methodology that consists of three steps - preparation, modeling and assessment as shown in Figure 1. Estimating network-wide counts requires temporal (e.g. weather or weekday) and static (e.g. population, network, infrastructure and land use) variables to characterize factors that may impact the bicycle counts. During the first step, bicycle counts and supplemental data are gathered, followed by the second step where data fusion incorporates Strava data with the existing data. The initial modeling stage estimates daily volumes based on temporal and static variables using a linear structure. To handle different bicycle activity representations from various data sources, this paper uses a staged approach, which creates a model based on temporal and Strava count variables (M1) and then augments the model with static variables (M1a). This study enhances the initial models based on user profiles created from Strava trip data using a K-means clustering. Separated linear modeling by clusters creates a temporal (M2) and a static model (M2a). In the advanced modeling stage (M3), the relationship between daily count data from either

permanent or short-term sites and Strava is further investigated using a classification and regression tree approach. This non-parametric model overcomes the issues captured from the previous modeling step such as a sampling bias or non-linearity between variables.

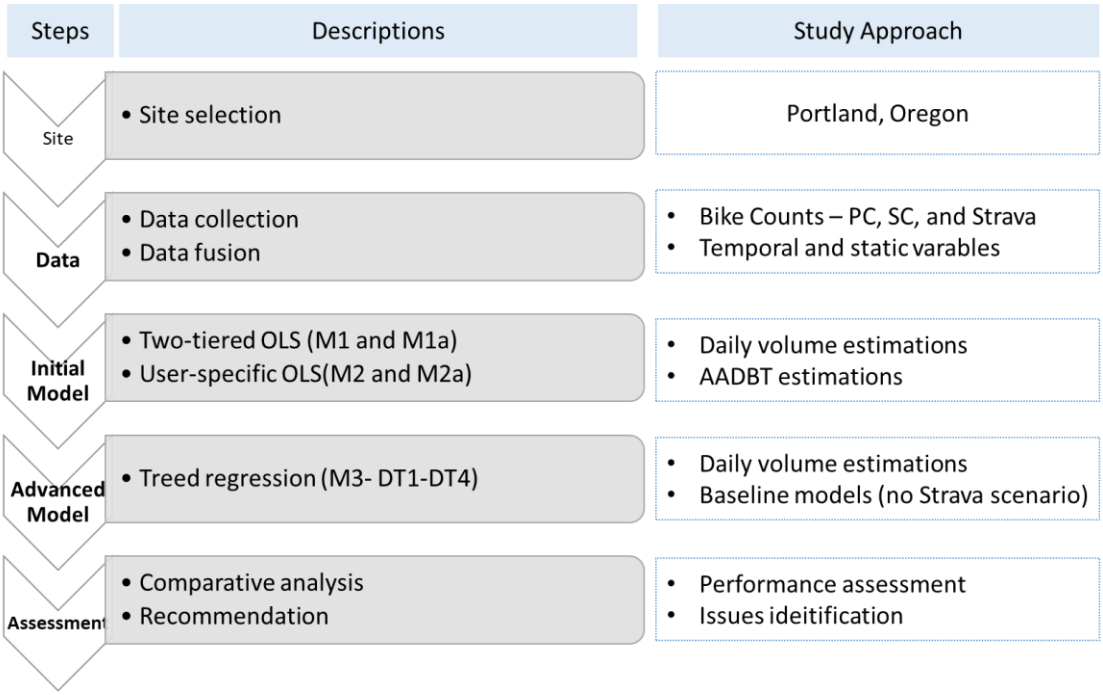


Figure 1 Methodology Overview

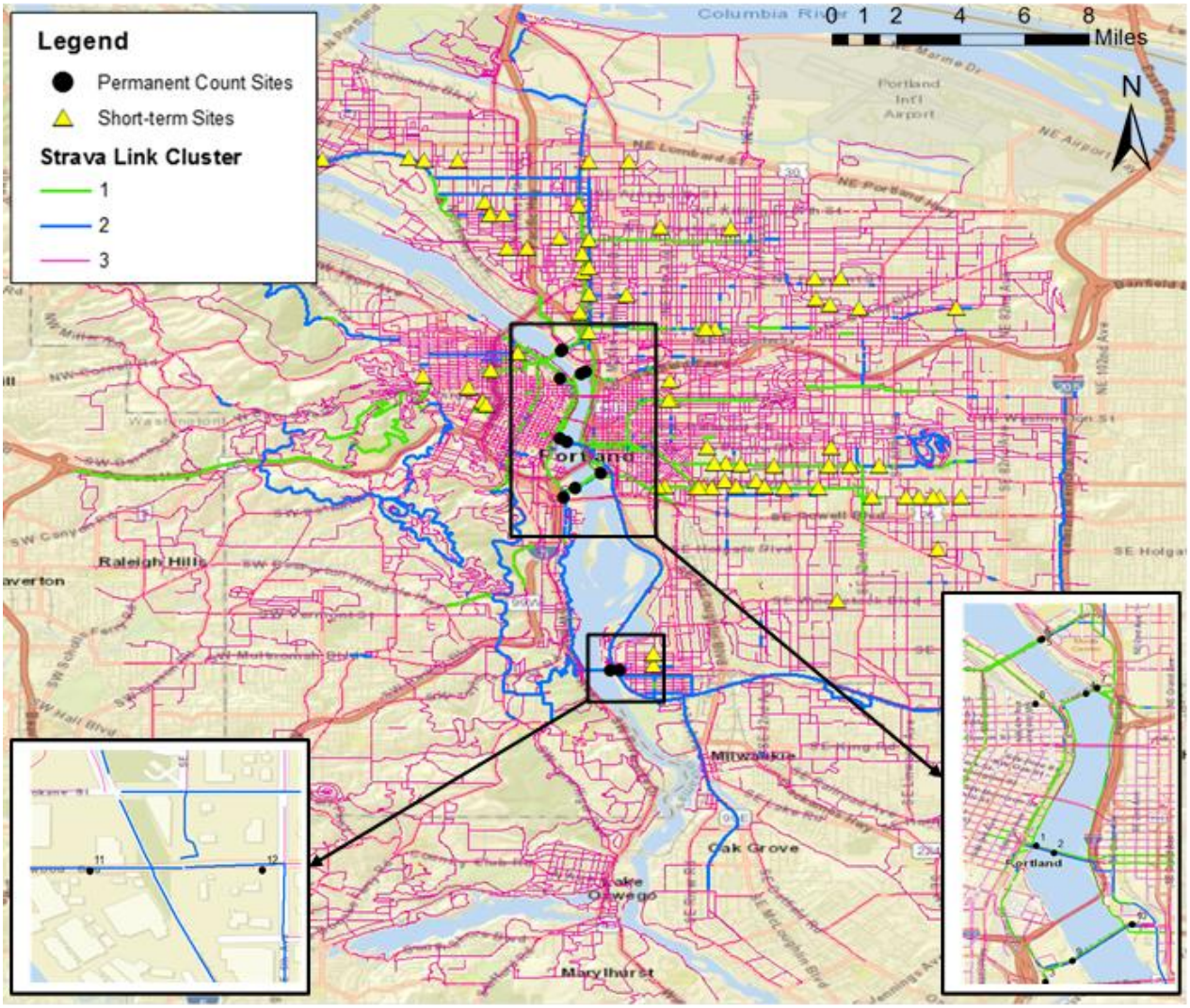


Figure 2 Study Sites and Data Sources

3. RESULTS

3.1 Findings from the Linear Models

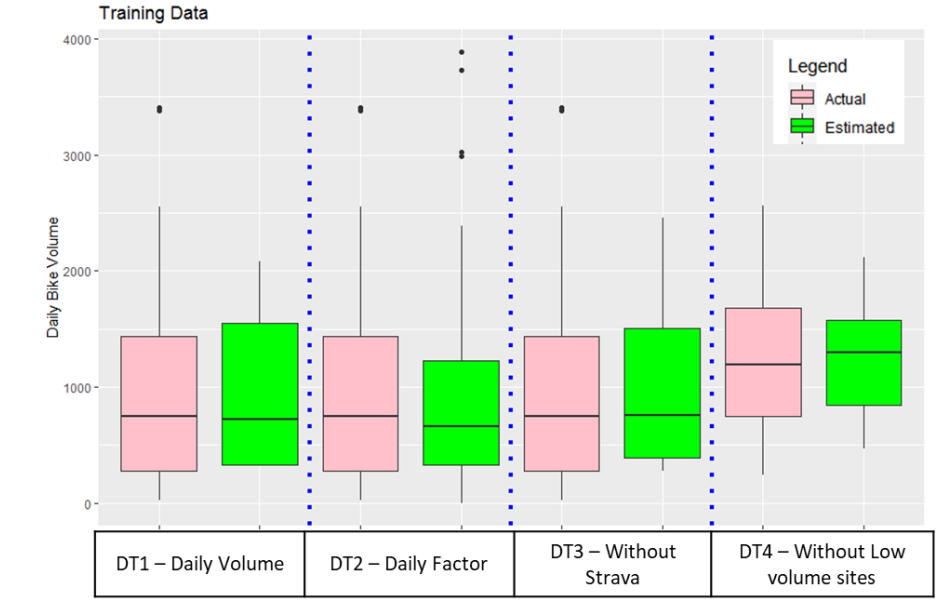
High errors in the daily counts and AADBT along with the low goodness-of-fit in the user-specific OLS models require further investigation of the data sources and bike count patterns.

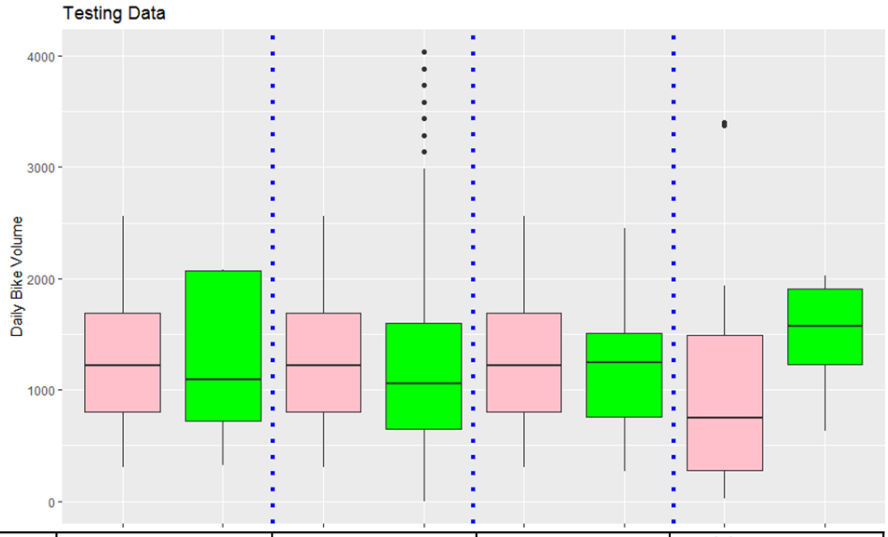
3.2 Advanced Modeling with Treed Regression

The study tests a non-parametric Classification and Regression Tree (CART) model to overcome the deficiencies observed with the linear framework.

Performance Comparisons of Treed Regression Models

Figure 3(a) compares the performance of the daily MAPE from the four decision trees for cluster 1. Compared to the linear approach (M1 and M2), the CART structure improves the daily estimation accuracy by over 10% for both training and testing datasets when applying daily counts as a dependent variable. The daily factor(proportion of Strava counts to PC counts) approach (M3-DT2) does not appear to improve the overall model performance although it significantly reduces the overestimation problems for both training and testing sets. However, for cluster 1, Strava does not seem to improve the accuracy since DT3 shows a lower MAPE than DT1 or DT2. However, including low volume sites in a training data set seem to significantly enhance the spatial transferability of the model, since the MAPE of DT4 increases to 80%.



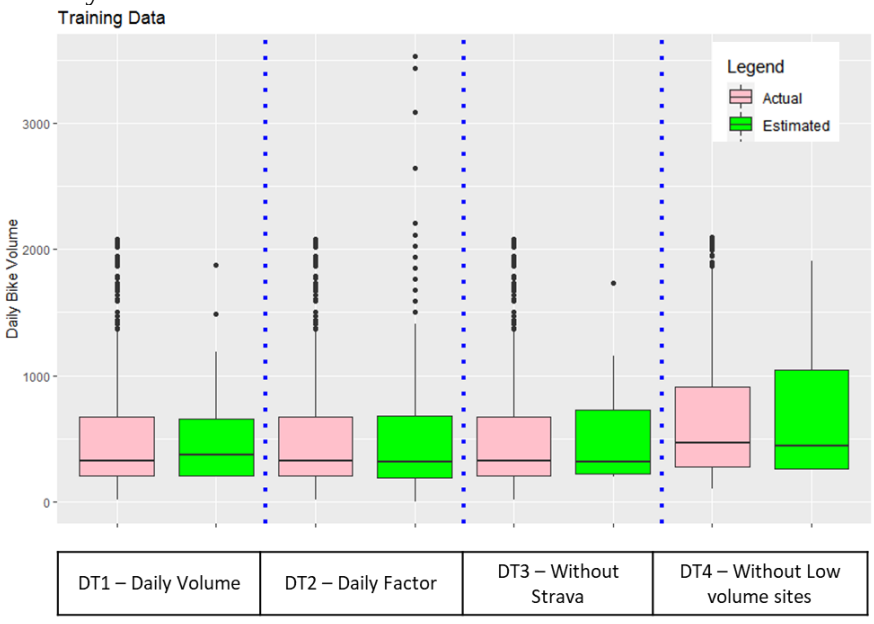


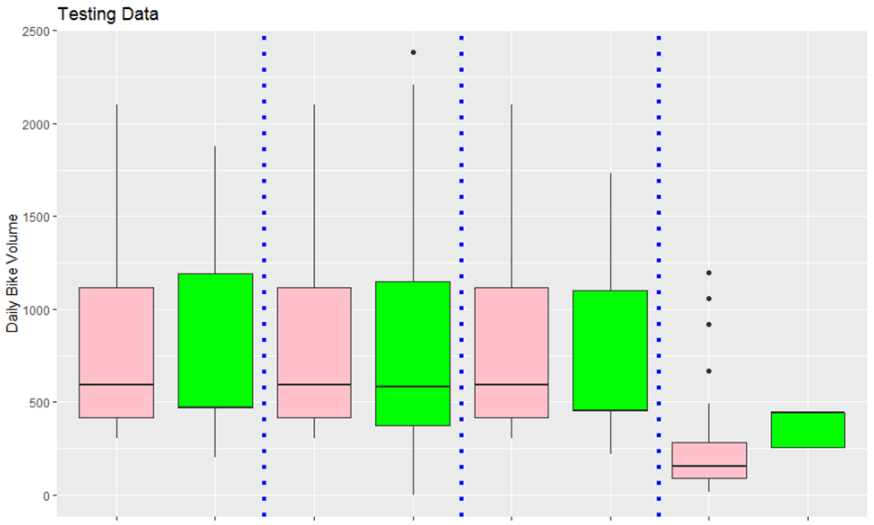
MAPE	(a)Daily Volume	(b)Daily factor	(c)Without Strava	(d) Without low volume sites
Training	20.70%	33.39%	18.05%	16.97%
Testing	23.76%	28.19%	22.42%	79.53%*

* Short term count sites are used as testing locations for DT4

Figure 3 (a) Performance Comparisons of Cluster 1

Figure 3(b) compares the performance of the four decision trees for cluster 2. The findings appear largely the same as cluster 1 but show a higher testing MAPE (130%) for the DT4. For the daily factor case, the errors appear much more like the daily volume model which may indicate DT2 well handles lower volumes sites of cluster 2. Based on the DT3, the Strava data appears to reduce the error. This indicates that as bicycle volumes become smaller and less consistent the importance of the Strava data likely increases.





MAPE	(a)Daily Volume	(b)Daily factor	(c)Without Strava	(d) Without low volume sites
Training	19.42%	22.08%	21.33%	17.80%
Testing	16.31%	21.65%	18.14%	129.79%*

* Short term count sites are used as testing locations data for DT4

Figure 3(b) Performance Comparisons of Cluster 2

4. CONCLUSIONS

The regression models perform poorly due to non-linearities between Strava and counter data. It was found that higher errors at lower volume short-term count sites with a MAPE of over 100 percent for both cluster 1 and cluster 2. The non-parametric CART modeling improves these results significantly for low volume sites by handling low and high volume sites separately in its model structure; however, the over-representation of high volume sites still remain in the datasets and produce a MAPE of 80% for cluster 1 and a MAPE of 130% for cluster 2. Another tree modeling approach that uses factors, rather than volumes, appears to better capture low volume sites; however, the overall errors increase since the model under-estimates the counts of higher volume sites. So, Effective network-level modeling requires more extensive data collection to estimate the models, and most critically must avoid extrapolating the patterns present at higher volume sites to low volume sites.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Miah, Hyun, Mattingly; data collection: Kothuri, Broach, McNeil, Miah; analysis and interpretation of results: Miah, Hyun, Mattingly; draft manuscript preparation: All authors. All authors reviewed the results and approved the final version of the manuscript. The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

ACKNOWLEDGEMENT

This project was funded by the National Institute for Transportation and Communities (NITC-1269), a U.S. DOT University Transportation Center, though a Pooled-Fund in partnership with the following contributors: Oregon DOT, Virginia DOT, Colorado DOT, Central Lane MPO, Portland Bureau of Transportation, District DOT, and Utah DOT.

REFERENCES

1. Josh Roll. *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.
2. Ryus, P., E. Ferguson, K. M. Laustsen, R. J. Schneider, F. R. Proulx, T. Hull, and L. Miranda-Moreno. *Guidebook on Pedestrian and Bicycle Volume Data Collection*. 2014.
3. Dadashova, B., G. P. Griffin, S. Das, S. Turner, and M. Graham. *Guide for Seasonal Adjustment and Crowdsourced Data ScalingScaling (Cooperative Research Program Technical Report No. 0-6927-P6). Federal Highway Administration and the Texas Department of Transportation.* 2018.
4. Jestico, B., T. Nelson, and M. Winters. Mapping Ridership Using Crowdsourced Cycling Data. *Journal of Transport Geography*, Vol. 52, 2016, pp. 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006.
5. Sanders, R. L., A. Frackelton, S. Gardner, R. Schneider, and M. Hintze. 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*, Vol. 2605, No. 1, 2017, pp. 32–44. https://doi.org/10.3141/2605-03.
6. Roy, A., T. A. Nelson, A. S. Fotheringham, and M. Winters. Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science*, Vol. 3, No. 2, 2019, p. 62. https://doi.org/10.3390/urbansci3020062.