Transportation Research Record

Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data

--Manuscript Draft--

Full Title:	Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data
Abstract:	Traffic volumes are fundamental for evaluating transportation systems, regardless of travel mode. A lack of counts for non-motorized modes poses a challenge for practitioners developing and managing multimodal transportation facilities, whether they want to evaluate transportation safety, the potential need for infrastructure changes, or to answer other questions about how and where people bicycle and walk. In recent years, researchers and practitioners alike, have been using crowdsourced data to supplement the non-motorized counts. As such, several methods and tools have been developed. The objective of this paper is to take advantage of new data sources that provide a limited and biased sample of trips and combine them with traditional counts to develop a practical tool for estimating the annual average daily bicycle (AADB) counts. In this paper, we have developed a direct-demand model for estimating AADB in Texas. We have used data from 100 stations, installed in 12 cities across the state, together with the crowdsourced Strava, roadway inventory, and American Community Survey (ACS) data to develop the count model for estimating AADB. The results indicate that crowdsourced Strava data is an acceptable predictor of bicycle counts, and when used with the roadway function class and number of high-income household in a block group, can provide quite an accurate AADB estimate (29% prediction error).
Manuscript Classifications:	Data and Information Technology; Bicycle and Pedestrian Data ABJ35SB; Operations and Traffic Management; Highway Traffic Monitoring ABJ35; Pedestrians and Bicycles; Bicycle and Pedestrian Data ABJ35SB; Bicycle Transportation ANF20
Manuscript Number:	20-02372
Article Type:	Presentation and Publication
Order of Authors:	Bahar Dadashova, PhD
	Greg Griffin, Ph.D.
	Subasish Das, Ph.D.
	Shawn Turner
	Bonnie Sherman

Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data

Bahar Dadashova, Ph.D.[1]
Associate Transportation Researcher
Texas A&M Transportation Institute
3135 TAMU, College Station, TX 77843-3135
Tel: (979) 317-2137; Email: B-Dadashova@tti.tamu.edu

Greg Griffin, Ph.D.
Assistant Professor
The University of Texas at San Antonio
501 W.César E. Chávez Blvd, San Antonio, TX 78207
Tel: (210) 458-3090; Email: Greg.Griffin@utsa.edu

Subasish Das, Ph.D.
Associate Transportation Researcher
Texas A&M Transportation Institute
3135 TAMU, College Station, TX 77843-3135
Tel: (979) 317-2153; Email: S-Das@tti.tamu.edu

Shawn Turner, P.E.
Senior Research Engineer
Associate Transportation Researcher
Texas A&M Transportation Institute
3135 TAMU, College Station, TX 77843-3135
Tel: (979) 317-2481; Email: S-Turner@tti.tamu.edu

Bonnie Sherman
Bicycle and Pedestrian Program Manager
Texas Department of Transportation
125 E 11th St, Austin, TX
Tel: (512) 486-5972; Email: Bionnie.Sherman@txdot.gov

Submitted to TRB Standing Committee on Bicycle Transportation (ANF20)
for consideration of presentation and publication at:
Transportation Research Board
99th Annual Meeting
January 2020
Washington, D.C.

Word count: 5,613 text words + 7 tables x 250 words (each) = 7,363 words

[1]Corresponding author

ABSTRACT

Traffic volumes are fundamental for evaluating transportation systems, regardless of travel mode. A lack of counts for non-motorized modes poses a challenge for practitioners developing and managing multimodal transportation facilities, whether they want to evaluate transportation safety, the potential need for infrastructure changes, or to answer other questions about how and where people bicycle and walk. In recent years, researchers and practitioners alike, have been using crowdsourced data to supplement the non-motorized counts. As such, several methods and tools have been developed. The objective of this paper is to take advantage of new data sources that provide a limited and biased sample of trips and combine them with traditional counts to develop a practical tool for estimating the annual average daily bicycle (AADB) counts. In this paper, we have developed a direct-demand model for estimating AADB in Texas. We have used data from 100 stations, installed in 12 cities across the state, together with the crowdsourced Strava, roadway inventory, and American Community Survey (ACS) data to develop the count model for estimating AADB. The results indicate that crowdsourced Strava data is an acceptable predictor of bicycle counts, and when used with the roadway function class and number of high-income household in a block group, can provide quite an accurate AADB estimate (29% prediction error).

Keywords: *Bicycle Counts, Crowdsourced Data, Mobile App, Strava, Travel Demand*

1. INTRODUCTION

Traffic volumes are fundamental for evaluating transportation systems, regardless of travel mode. A lack of counts for non-motorized modes poses a challenge for practitioners developing and managing multimodal transportation facilities, whether they want to evaluate transportation safety, the potential need for infrastructure changes, or to answer other questions about how and where people bicycle and walk. Bicyclist and pedestrian counts that are not feasible to collect with field equipment might be estimated through smartphone apps and other online methods to leverage the knowledge of networked communities, known as crowdsourcing. Crowdsourcing apps, such as Strava and Ride Report, have the potential of collecting data at any time and location that the apps are used. However, they are limited by the number of users and the target market for the apps. Crowdsourcing uses a broad pool of individuals through an online platform that aggregates and formats the information for a specific use. The company aggregates these trips onto a transportation system network, processes them for privacy, and then re-sells the information as a crowdsourced traffic data product, available in many places around the globe.

The objective of this paper is to take advantage of new data sources that provide a limited and biased sample of trips and combine them with traditional counts to develop a practical approach for estimating the annual average daily bicycle (AADB) counts. Crowdsourced data can provide important insights for both the agencies in terms of planning and policy decision and road users in terms of travel decisions; these data sources can be used to reveal quantitative insights into the behavior of non-motorized road users, such as route choice. Although crowdsourced data has a much more extensive coverage compared to non-motorized count stations, nevertheless the data still represent a small percentage of non-motorized users. For instance, researchers found that 3–9 percent of bicycle trips counted on trails in Austin used Strava at the time of the count (*1*). This percentage, on the other hand, can change based on the location, land use, non-motorized facility type, socioeconomic, demographic, and meteorological factors (*2–5*).

Moreover, the travel behavior of app users may be different than the population hence affecting the functional form of the underlying process that generates the crowdsourced data; for example, users of activity-based smartphone apps are more consistent, hence the temporal data produced by these users are stationary, whereas observed non-motorized user counts are highly volatile and non-stationary. Other challenges of using crowdsourced data include quality control, data redundancy, sampling biases, data conflation, etc. In this paper, we have used crowdsourced Strava, roadway characteristics, household income, and population demographics data to develop direct demand models for estimating the AADB counts in Texas.

The rest of this paper is organized as follows. A literature review in section 2 discusses the previous research on this subject. Section 3 describes data used in the study and modeling approach for developing the direct-demand models. Section 4 presents the results of data mining and data analysis. The paper ends with conclusions, acknowledgments, author contribution, and references.

2. PREVIOUS RESEARCH

Growth in research on bicycling and planning for the travel mode, in general, supported a range of methods to assess bicycle traffic, including sparse data from reference counts using automatic traffic recorders (*6*, *7*), and big data sourced from active crowdsourcing applications or passively-sensed activities (*8*, *9*). Continuous recorder data provides a key reference 24-hour dataset at a known location, but is relatively sparse because of the resources necessary for developing and operating the sensors, often requiring pavement cuts for inductive-loop wires, equipment cost, and maintenance (*10*). These permanent recorders are few in number, and sometimes miss data due to equipment failures such as battery problems, insect activity or vandalism (*11*, *12*), but state-level efforts are increasing availability and predictability of these reference counters (*13*, *14*). Conversely, big data sensed through smartphones or vehicles is spatially extensive, but represents only a portion of trips at any given location, representing users in markets rather than populations (*2*, *15*, *16*). Whether called data fusion, expansion, or weighting (*4*, *17*, *18*), this research suggests opportunities for leveraging the relative advantages of sparse and big data to improve understanding of bicycle traffic for planning and safety.

Several platforms offer high levels of use in different countries, but Strava Metro is the only provider of aggregated and anonymized bicycling data at the global scale, to date. Previous research and evidence from planning practice show that the Strava Metro dataset is biased toward recreation and fitness-oriented bicycling, but its widespread use enables a range of applications including understanding where bicyclists ride for fitness (*19*), risk exposure to crashes (*5*, *20*), and change in ridership over time (*21*). Planners with the Oregon Department of Transportation suggested Strava Metro data could be useful despite its biases, but additional research could support broader applications. “While there currently is not a method to expand this information up to total bike riders, the relative amount of use from Strava users from one path to another does provide ODOT with more guidance than has existed previously on which routes are used more than others” (*22*). An extensive review of big data for bicycling research suggested a research agenda exploring combinations of crowdsourced and traditional information to develop new insights on travel and analysis methods that scale beyond current approaches (*23*). To date, published approaches for scaling crowdsourced data include a focus on the use of population and traffic counters (*20*), and multi-factor Poisson regression in Maricopa county, Arizona (US) (*24*). Though some studies have combined crowdsourced data with traffic counts and environmental data to better understand bicycling contexts, we suggest that both practitioners and researchers could benefit from a clear approach to expand crowdsourced data to meaningfully estimate the total volume of bicycling trips. The following method section details our approach in the state of Texas (US)

3. METHODOLOGICAL APPROACH

3.1 Data Overview

3.1.1 Field Data Collection and Quality Assurance

We collected data from 155 locations across 12 cities in Texas: Austin, Brownsville, Corpus Christi, Dallas, Houston, League City, Lubbock, Midland, Odessa, Plano, San Antonio, and Wichita Falls. FIGURE 1 shows the number of data collection stations installed in every city. Bicycle counts were collected from a wide variety of facility types, including shared-use paths, bike lanes, shoulders, sidewalks, other paths, unpaved facilities, and shared roadways. Urban areas generally attempt to count recurring locations on an annual basis; however, many have noted that due to resource constraints, counts are either sporadic or occur every other year.

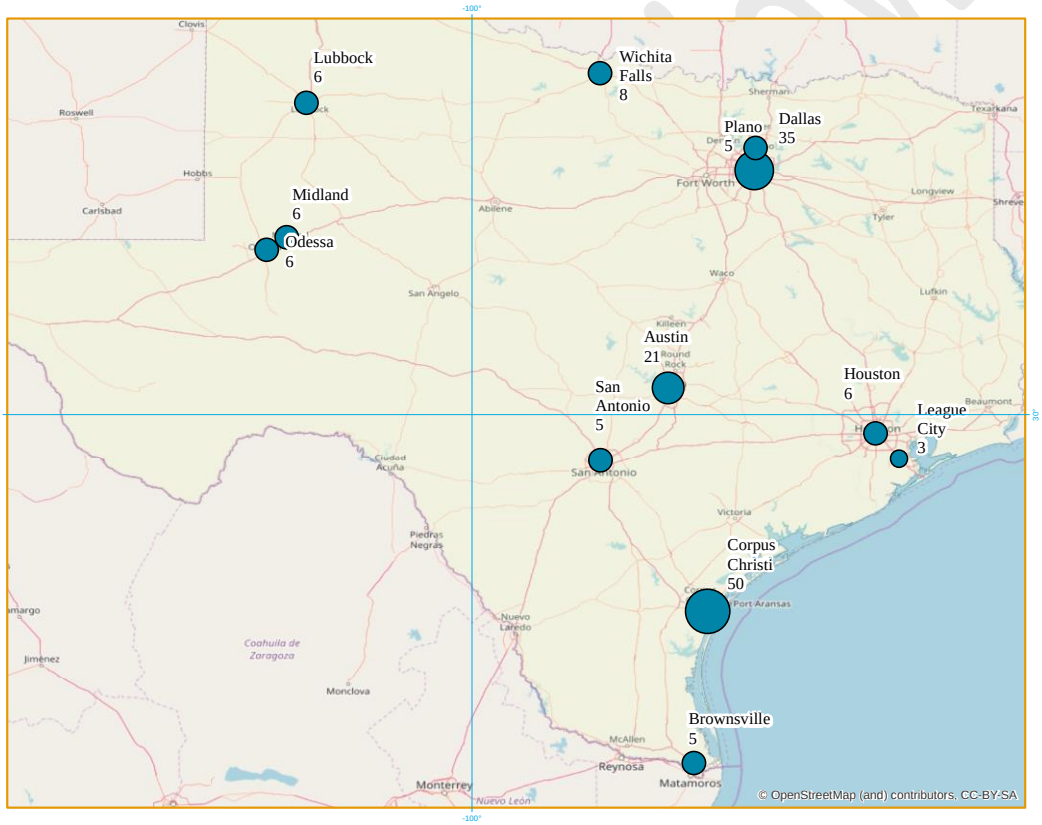


FIGURE 1. Number of Permanent and Temporary Counters per City.

We used Federal Highway Adminstrations' Traffic Monitoring Guidelines (TMG) to develop the bicycle count database. Each location was assigned two unique station IDs to indicate the direction of travel. TABLE 1 shows the list of available information per count station. The count locations indicate the city and street names (i.e. station name), the stations ID per travel direction, and latitude and longitude among other variables.

TABLE 1. Example of Count Location Information

TMG Variables	Available Information
Location ID	10
City	Austin
Station Name	Guadalupe St N of W 21st St
Latitude	-97.74187
Longitude	30.28419
Station ID Travel Direction 1	453-1-2-60-000354
Station ID Travel Direction 2	453-5-2-60-000355
Travel Direction 1	NB
Travel Direction 2	SB

There were several inconsistencies regarding both the bicycle count location information and the bicycle counts. The common errors concerning the count locations include:

- No information about the type of counter.
- No information about user type or direction.
- Multiple entries (Names) for functionally the same location.
- Truncated latitude and longitude of location.
- The errors concerning the bicycle counts can happen due to several reasons:
- Installation of the counter
- Miscoding of the metadata
- Poor or sloppy maintenance of the metadata
- Actual counting errors

These inconsistencies were identified and removed from the count database during the data reduction and consolidation activities. The bicycle and pedestrian count data used in this study is readily available through Texas Bicycle and Pedestrian Count Exchange Program website (*25*). The count data collection methods are described in details in Turner et al. (*26*).

3.1.2 Crowdsourced Data

Strava Metro is a crowdsourced database that shows bicycle or pedestrian activity for a given edge (segments) or node (intersection). Strava Metro is the oldest and largest source of crowdsourced bicycle volumes currently available. Strava uses Open Street Map (OSM) for developing the geospatial count files. This service is a business unit of Strava, which is a smartphone app and website that seeks to "enhance the experience of sport and connect millions

of athletes from around the world. However, previous research has shown that Strava represents a sample of health-oriented contributors and may not represent the broader bicyclist population. Strava includes walking, running, and hiking trips, in addition to bicycling trips.

TABLE 2 shows the list of variables available in Strava.

TABLE 2. Strava Bicycle Count Database

Strava Data	Definition
Edge/Node ID	Numeric value indicating the segment or intersection ID
From X/Y & To X/Y	Beginning and Ending latitude and longitude of a Strava segment
Node X/Y	Latitude and longitude of a Strava intersection
Street Name	Street name of a Strava segment
Year, Day, Hour, and Minute	The timeframe of bicycle and pedestrian counts
Athlete	Number of athletes travelling on the default direction of travel
Reverse Athlete	Number of athletes travelling on the opposite direction of travel
Activity	Number of bicyclists/pedestrians travelling on the default direction of travel
Reverse Activity	Number of bicyclists/pedestrians travelling on the opposite direction of travel
Total Activity	Number of total bicyclists/pedestrians on a given Strava segment/intersection

Note that in Strava the roadway segments are labeled as "edges" while the intersections and segment end points (e.g. cul-de-sac) are labelled as "nodes". The number of athletes and activities show the number of bicyclists and pedestrians on a given segment/intersection at the given year, day, hour and minute. Strava shows the number of bicyclists and pedestrians for both directions of travel however it does not indicate the default direction of travel. To identify the default direction of travel, we used the following equation:

$$A = 180 + \arctan(Y2 - Y1, X2 - X1) \times \frac{180}{\pi} \quad (1)$$

$$Cardinal\ Direction = \begin{cases} WB & if\ 1 \le A < 90 \\ SB & if\ 90 \le A < 180 \\ EB & if\ 180 \le A < 270 \\ NB & if\ 270 \le A \le 360 \end{cases} \quad (2)$$

We matched Strava data from 2016 to 2018 with the bicycle counts collected from the aforementioned count stations. Strava assigns several edges (i.e., segments) to the same road segment based on direction of travel, and non-motorized facility (i.e., bike lane, sidewalk, etc.). To match the count stations with the correct Strava edge, we compared the name of the street and direction of travel.

TABLE 3 depicts the descriptive statistics of the percentage of bicyclists using the Strava app per OSM functional class. As can be observed, the percentage of Strava users vary from 6 to 16 percent. However, this percentage is different for each OSM functional class

TABLE 3 Proportion of Strava to Field Counts per OSM Functional Class

OSM Functional System	Strava User Percentage			
	Min	**Max**	**Mean**	**St. D.**
Primary	1%	35%	8%	0.04
Secondary	0%	19%	6%	0.02
Tertiary	0%	70%	16%	0.13
Residential	0%	75%	8%	0.06
Cycleway	0%	100%	7%	0.09
Footway	0%	100%	6%	0.12

3.1.3 Other Relevant Data

We compiled a list of potentially important variables that can help to explain the relationships between the observed bicycle counts and Strava activity. For this purpose, we used the American Community Survey (ACS) and the Texas Department of Transportation (TxDOT) roadway inventory database (RHiNO).

The U.S. Census Bureau's American Community Survey (ACS) is a nationwide survey that delivers information on social, economic, household, and other relevant demographic characteristics about the U.S. population every year. In general, the Census Bureau contacts over 3.5 million U.S. households to participate in the ACS every year. One of the uniqueness of using ACS is its ability to produce estimates on a wide range of geographies, including low geographic levels such as block groups. We collected block group level ACS data for Texas. As ACS contains a wide list of variables, the variable selection was conducted by using random forests (discussed below).

TxDOT maintains a database that includes a variety of roadway characteristics. This database, known as the Roadway Highway Inventory Network Offload (RHiNO), can be used to supplement information from the crash database. This database primarily provides road characteristic information, including the estimated traffic volume and corridor length, for every known road in Texas.

We conflated the acquired databases on the Strava network using ArcMap 10.5.1. It is important to note that field data is a point data, Strava and RHiNO are polynomial and ACS is a polygon data. We follow the following steps to conflate the data:

1. From the ACS block group geodatabase, select tables with population, housing unit, and income data.
2. Assign block group level information to the Strava segments
3. Conflate RHiNO roadway level data to the Strava segments

TABLE 4 depicts the descriptive statistics of all the variables and data sources considered for the analysis.

TABLE 4. Descriptive Statistics of Variables.

Variable Name	Source	Unit of Analysis	Min	Max	Mean	St. D.
Quantitative Variables						
Land Area (km square)	ACS	Polygon	200,328	9,917,652	1,749,294	1,889,419
Total Population	ACS	Polygon	486	8,977	1,992.69	2,071.78
Population Density	ACS	Polygon	532.05	23,989.05	4,680.44	4,771.44
Total Female Population	ACS	Polygon	242	4622	957.86	929.77
Female, Age 15-20	ACS	Polygon	0	3970	183.91	744.48
Female, Age 21-34	ACS	Polygon	29	1543	314.05	354.4
Female, Age 35-49	ACS	Polygon	0	868	151.36	187.58
Female, Age 5-14	ACS	Polygon	0	310	95.54	84.89
Female, Age 50-64	ACS	Polygon	0	309	130.32	90.21
Female, Age 65-85	ACS	Polygon	0	471	82.67	91.12
Total Male Population	ACS	Polygon	180	6,230	1,034.82	1,241.49
Male, Age 15-20	ACS	Polygon	0	2,996	164.94	564.29
Male, Age 21-34	ACS	Polygon	0	3,016	364.94	577.47
Male, Age 35-49	ACS	Polygon	18	1677	209.65	312.03
Male, Age 5-14	ACS	Polygon	0	362	94.79	78.07
Male, Age 50-64	ACS	Polygon	11	883	145.57	167.81
Male, Age 65-85	ACS	Polygon	0	246	54.94	54.34
Total Number of Households	ACS	Polygon	24	2317	689.78	512.43
Household Density	ACS	Polygon	0.00026	0.0031	0.00073	0.00074
Household Income (HHI) 10K	ACS	Polygon	0	134	47.83	41.78
HHI 15K	ACS	Polygon	0	101	23.39	28.8
HHI 20K	ACS	Polygon	0	148	26.99	37.42
HHI 25K	ACS	Polygon	0	85	22.57	27.49
HHI 30K	ACS	Polygon	0	113	20.57	25.99
HHI 35K	ACS	Polygon	0	63	14.4	18.08
HHI 40K	ACS	Polygon	0	146	22.97	26.36
HHI 45K	ACS	Polygon	0	160	25.35	26.87
HHI 50K	ACS	Polygon	0	65	21.31	20.04
HHI 60K	ACS	Polygon	0	275	68.97	72.9
HHI 75K	ACS	Polygon	0	261	69.25	71.19

Variable Name	Source	Unit of Analysis	Min	Max	Mean	St. D.
Quantitative Variables						
HHI 100K	ACS	Polygon	0	361	85.71	81.41
HHI 125K	ACS	Polygon	0	227	67.2	58.2
HHI 150K	ACS	Polygon	0	167	33.69	39.48
HHI 200K	ACS	Polygon	0	241	55.43	61.12
HHI > 200K	ACS	Polygon	0	755	84.15	108.02
Annual Average Daily Bicycle Counts (AADB)	Manual	Point	1	669	66.93	127.68
Non-motorized Facility Width	Manual	Line	4	25	8.47	4.13
Non-motorized Facility Buffer Width	Manual	Line	0	5	2.91	0.84
Median Width	RHiNO	Line	0	16	4.6	2.66
Number of Lanes	RHiNO	Line	0	6	2.75	1.09
Posted Speed Limit	RHiNO	Line	0	55	17.35	16.75
Inside Shoulder Width	RHiNO	Line	0	10	0.3	1.71
Outside Shoulder Width	RHiNO	Line	0	20	0.6	2.95
Surface Width	RHiNO	Line	0	76	33.85	16.33
Average Activity (AvgActivity)	Strava	Line	0	81	4.8	12.38

Variable Name	Source	Unit of Analysis	Variable Description
Qualitative Variables			
City	Manual	Polygon	Austin, Brownsville, Corpus Christi, Dallas, Houston, League City, Lubbock, Midland, Odessa, Plano, San Antonio and Wichita Falls
Non-motorized Facility Type	Manual	Line	Shared Use Path; On-street Bike Lane
Parking	Manual	Line	No on-street parking; Parallel parking
Pavement Condition	Manual	Line	Poor; Fair; Good; Excellent
Pavement Type	Manual	Line	Asphalt; Concrete; Crushed Granite/Gravel
Place of Interest (POI) within 50 Miles	Manual	Point	High School; University
Shade	Manual	Line	None; Partial; Full
Street Lighting	Manual	Line	None; One side; Both sides; Partial
Transit	Manual	Polygon	No; Yes
Functional Classification	RHiNO	Line	Principal Arterial; Minor Arterial; Collector; Local; Shared Path or Trail
OSM Functional System (CLAZZ)*	Strava	Line	15 = Primary; 21 = Secondary; 31 = Tertiary; 32 = Residential; 72 = Path; 81 = Cycleway; 91 = Footway

*Definition of OSM function class or *highway link* can be found in this link.

3.2 Data Analysis Methods

Models allow transportation planners and researchers to generalize understanding of limited data to larger geographies and explore impacts of changes in policies and infrastructure both in the present and for estimating future scenarios. These approaches help understand people's propensity to cycle in a variety of circumstances (*27*), including the roles of the built environment (*27*), time (*28*), seasonal and weather factors (*29*). Increased sophistication of modeling requires additional resources, and many of these approaches can be challenging to replicate in research or apply in practice.

Recent guidance suggests a need for improvement in balancing resource needs for modeling with accuracy, and for additional calibration data through conducting bicycle traffic counts (*30*). More resource-intensive models such as tour generation and mode split, and route choice models require substantial data and expertise, while GIS index and direct demand models may sacrifice accuracy. Methods to improve model calibration include increasing the number of count locations and times through short-term counts (*31*), and examining bicycle traffic over larger areas through crowdsourced data collected by smartphone users (*32*). Crowdsourcing was first popularized in transportation planning as a public participation method to collect ideas from a broad range of people, the approach is becoming more prevalent to monitor traffic (*33*). Regardless of input traffic data, bicycle traffic models can be assessed and improved through rigorous evaluation (*34*).

3.2.1 Variable Importance

Because the list of potential factors for including in the regression is very comprehensive, we used data mining tool, random forests (RF) to select the list of most important factors explaining the relationship between ground counts and Strava activity. RF method was proposed by Breiman (2001) and is considered to be one of the most efficient classification methods (*37*). One of the most significant byproducts of RF is of variable importance. Variable importance ranking is measured by the classification accuracy and Gini impurity. This importance measure shows how much the mean squared error or the impurity increase when the specified variable is randomly permuted. If prediction error does not change by permuting the variable, then the importance measures will not be altered significantly which in turn will change the mean squared error, MSE, of the variable only slightly (low values). This implies that the specified variable is not important. On the contrary, if the MSE significantly decreases during the permutation of the variable then the variable is deemed as important.

The classification accuracy measure of the variable is averaged over the number of trees, B, used to construct the RF:

$$MDA(x_i) = \frac{\sum_{tree=1}^{B} MDA^{tree}(x_i)}{B} \quad (3)$$

where $MDA(x_i)$ is the average importance rate of the variable x_i and $MDA(x_i)$ is the importance rate of the same variable in $tree = \{tree_{b,b=1,\dots,B}\}$.

The mean decrease in Gini impurity computes the contribution of the variable to the homogeneity of the nodes and leaves in the resulting RF. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous):

$$MDG^n(x_i) = 1 - \sum_{k=1}^{K} p(k|n) \quad (4)$$

where $MDG^n(x_i)$ is the Gini impurity coefficient of the variable x_i at the node n; $p(k|n)$ is the probability of class k in node n (weights) and K is the number of classes.

A higher MDA and MDG indicate higher variable importance. In this paper, we have used the RF method to select the most important factors affecting the annual average daily bicycle demand.

3.2.2 Bicyclist Direct-Demand Model

Traditionally, bicycle demand has been estimated using several approaches such as adjustment factors (*35*), ordinary least squares (OLS) regression, or count data models (*5*, *17*, *36*). The foundational building block of count data models is Poisson regression. In this model, it is assumed that the count data follows a Poisson distribution, which is a discrete probability distribution; for example number of bicyclists traveling across a roadway segment or crossing an intersection, over a fixed time interval (e.g., every day) follows a Poisson distribution. In Poisson distribution mean and variance of count data are assumed to be equal. Although this condition may hold for a relatively big data, however, most of the data used in non-motorized data studies are relatively short. Hence most researchers use a negative binomial model which is a standard choice for a basic count data. The negative binomial regression model has the following functional form:

$$Y_i = \exp(\sum_{k=1}^{K} \beta_k \times X_{k,i} + \varepsilon_i) \quad (5)$$

where, $Y_i -$ is the vector of bicyclist counts at segment i during a given period; $\beta_k -$ is the coefficient estimates; $X_{k,i} -$ is the matrix of explanatory variables at site i and $\varepsilon_i -$ is the error term which represents the unobserved conditions of site i. The error term of NB model is the the assumed to follow a gamma distribution with mean one an variance α^2, $\exp(\varepsilon_i) \sim G(1, \alpha^2)$. α is also referred to as overdispersion parameter; lower overdispersion indicates a better model fit.

In this paper, we have used a negative binomial model to estimate the bicycle counts as the function of Strava activity and other relevant factors.

4. RESULTS

4.2 AADB Direct Demand Models

4.2.1 Selection of the Most Influential Factors

In this paper, we used ground counts from 100 stations to develop the AADB direct demand models. We removed the sites with missing data and with very-short-term counts.

As indicated earlier, we used RF methodology to select the most influential factors. FIGURE 2 shows a list of the most important factors affecting the relationship between average Strava activity and ground counts.

The initial analysis results indicate that household income and demographic variables are very influential (FIGURE 2). Because most of these variables belong to the same category, we selected the most important variables from each category and conducted RF analysis again.

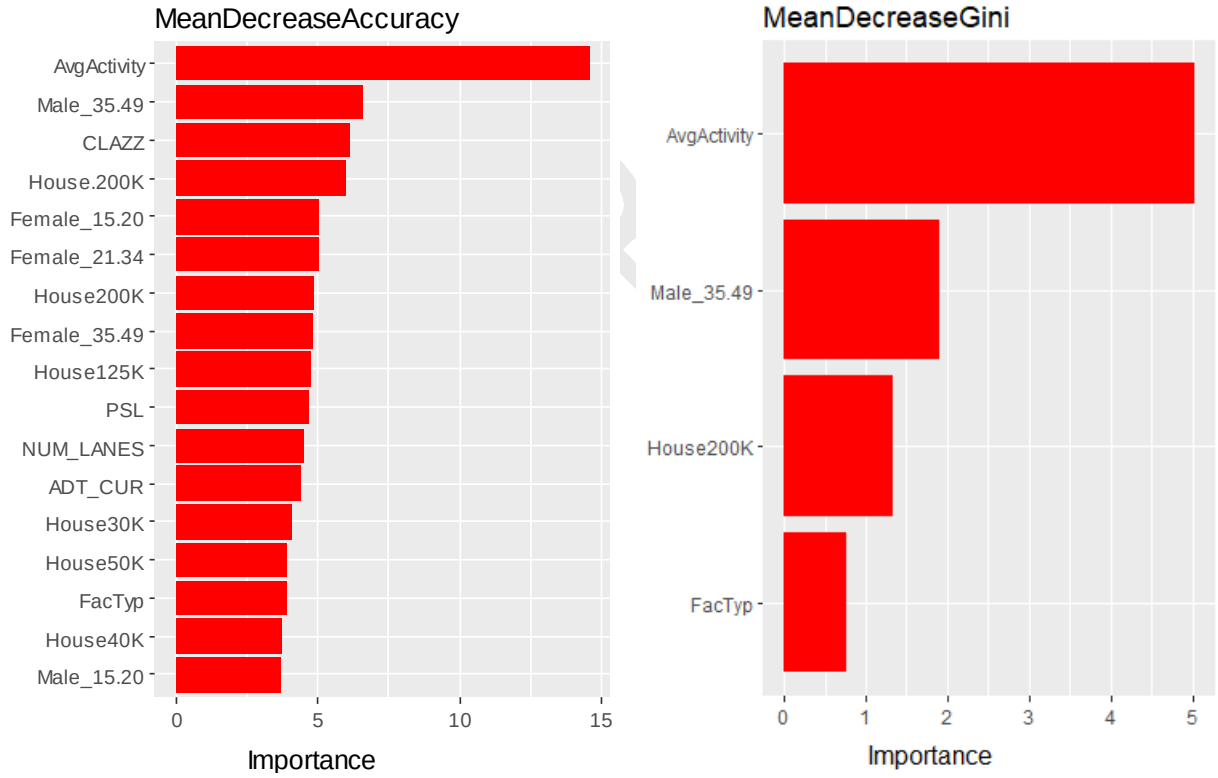


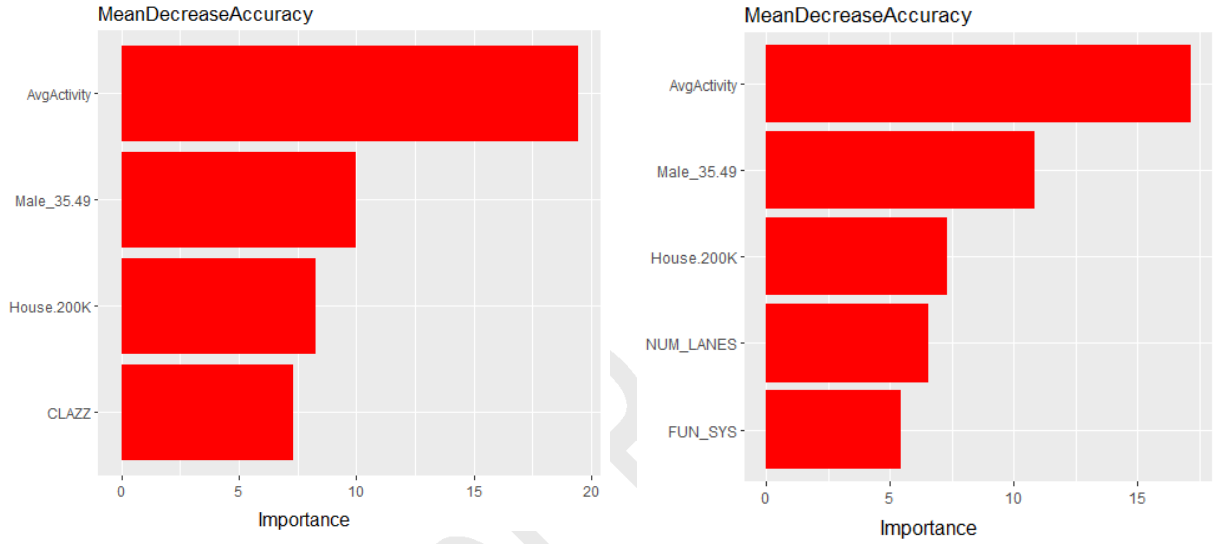
FIGURE 2. Preliminary List of Important Variables.

After conducting the data mining analysis for the second time, by keeping only the most important ACS variables, we identified the following list of most influential variables (FIGURE 3):

- Strava sample (Strava)
- Male 35-49 (ACS)

- Household income of 200K (ACS)
- OSM Functional Class (Strava)
- Number of Lanes (RHiNO)
- Facility Type (Manual)

As can be observed, we have identified two sets of important variables. The first set of variables include only OSM and ACS factors that are readily available through the Strava database. The second set of variables are from TxDOT roadway inventory (RHiNO) and ACS.



a) OSM Functional Class b) Roadway Functional System

FIGURE 3 Final List of Important Variables.

4.2.2 AADB Prediction Models

After identifying the most important variables, we developed direct demand models for average annual daily bicycles (AADB) using the negative binomial model. The two models have the following functional form:

$$AADB_i = \exp(\beta_0 + \beta_1 \times AADB\ Strava_i + \beta_2 \times \text{Household} > 200\text{K}_i + \\ + \beta_3 \times OSM\ Class_i) \quad (4.1)$$

$$AADB_i = \exp(\beta_0 + \beta_1 \times AADB\ Strava_i + \beta_2 \times \text{Household} > 200\text{K}_i + \\ + \beta_3 \times Func.System_i + \beta_4 \times Num.of\ Lanes_i) \quad (4.2)$$

where, $AADB_i$ – represents the estimated Annual Average Daily Bicycles at segments/edge i; $AADB\ Strava_i$ – represents the Strava sample counts at location i for the given time period; $\text{Household} > 200\text{K}_i$ – represents the number of households with 200K income; $OSM\ Class_i$ – represents the OSM functional class according to Strava; $Func.System_i$ –

represents the roadway functional system according to RHiNO; $Num.of\ Lanes_i -$ represents the number of lanes on the roadway segment; and $\beta_k -$ are the coefficient estimates. We also conducted a prediction analysis to cross-validate the modeling results. FIGURE 4 indicates the prediction intervals of the two models while TABLE 6 reports the error measures for the two models. As can be observed the prediction error of OSM-based model is relatively better than the RHiNO-based model; i.e., 29 vs 38 percent.

TABLE 5 shows the estimation results for both models, as well as the goodness of fit measures. Both models have a relatively lower overdisperion parameter (~1) and higher R^2 values (< 0.7) indicating that both models are good fit for the data.

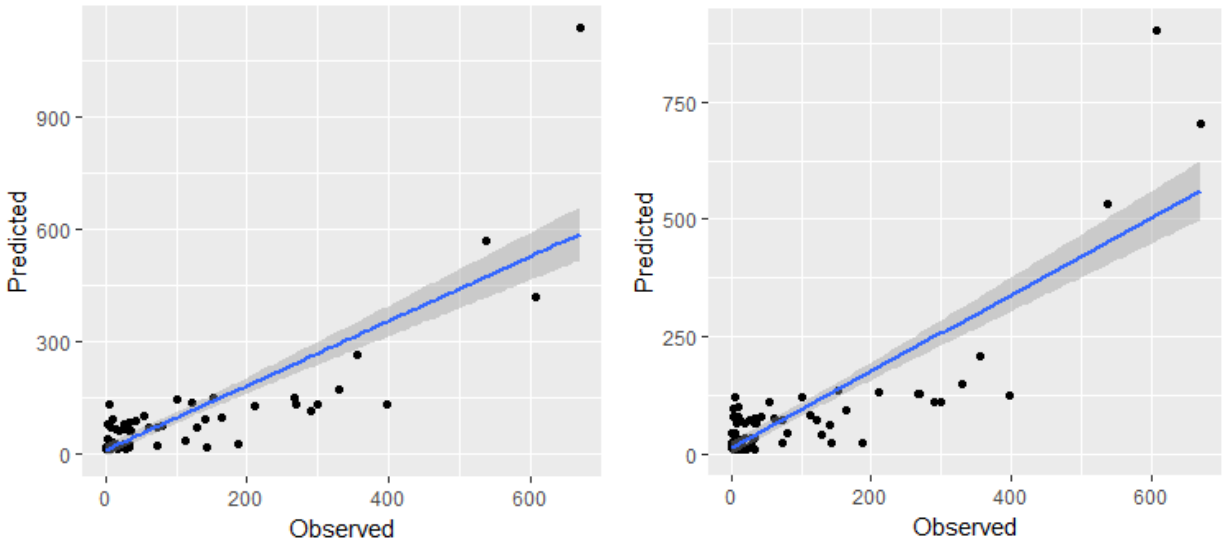
We also conducted a prediction analysis to cross-validate the modeling results. FIGURE 4 indicates the prediction intervals of the two models while TABLE 6 reports the error measures for the two models. As can be observed the prediction error of OSM-based model is relatively better than the RHiNO-based model; i.e., 29 vs 38 percent.

1

TABLE 5. Estimation Results

Variables		**Model 1**			**Model 2**		
		Estimate	**St. D.**	**p-value**	**Estimate**	**St. D.**	**p-value**
OSM Highway Functional Class	Primary	4.138	0.053	< 0.001			
	Secondary	2.590	0.060	< 0.001			
	Tertiary	3.078	0.062	< 0.001			
	Residential	2.862	0.037	< 0.001			
	Path	4.271	0.031	< 0.001			
	Cycleway	4.144	0.027	< 0.001			
	Footway	3.323	0.062	< 0.001			
Functional System	Collector (Minor)				3.211	0.078	< 0.001
	Local Road				2.506	0.083	< 0.001
	Minor Arterial				2.987	0.118	< 0.001
	Principal Arterial				3.929	0.116	< 0.001
	Shared Path or Trail				4.270	0.035	< 0.001
AADB Strava		0.038	0.000	< 0.001	0.031	0.000	< 0.001
Number of Households with >200K income		0.002	0.000	< 0.001	0.002	0.000	< 0.001
Number of Lanes					-0.066	0.027	< 0.05
R^2 (Model Accuracy)			75%			70%	
Overdispersion			0.967			1.172	

2



a) Model 1 b) Model 2

FIGURE 4. Predicted vs Observed AADB.

TABLE 6. Relative Accuracy per Strava Percentage Categories.

Prediction Error Measure	Model 1	Model 2
Mean Absolute Percentage Error	29%	38%
Mean Squared Error	5,855	4,836
Mean Absolute Error	41	42

4.3 Implications and Discussion

The direct demand models indicate that crowdsourced Strava data together with roadway functional class (or system) and the number of high-income households can provide a relatively accurate estimate of annual average daily bicycle counts. This traffic estimation technique is designed to work even with zero Strava activities, by using minimal values observed with manual counts throughout the state. TABLE 7 can be used to review against estimates with low Strava sample counts.

TABLE 7. Estimated Number of Bicycle Counts Given Strava Sample and Roadway Class.

Strava Sample Counts	OSM Functional Class						
	Primary	**Secondary**	**Tertiary**	**Residential**	**Path**	**Cycleway**	**Footway**
0	63	13	22	17	72	63	28
5	76	16	26	21	87	76	34
10	92	19	32	26	105	92	41
20	134	29	46	37	153	135	59

There are several reasons why this model might over-or-under predict bicycle traffic. Strava use itself may be particularly high or low in a certain area. It might over-estimate such if a major event was routed through the area during the Strava sampling period, or under-estimate if Strava use is particularly low. Researchers expect higher fluctuations in rural areas with lower overall Strava use, as compared with urban areas.

Changes in segment classification over time, such as upgrading a street from a tertiary to secondary segment, could significantly impact bicycle traffic estimation values. Similarly, any errors in the classification will expand the error of the traffic estimate. High-income households have a relatively minor, yet statistically significant, role in scaling Strava activities to estimate totals. However, there may be areas that do not respond to residential income in an average manner, such as bicycling loops in large parks. Use of the route in the park may be rather homogenous, but nearby residential income could skew traffic estimates when they do not, in practice, impact bicycling rates.

5. CONCLUSIONS AND RECOMMENDATIONS

We explored several different approaches to leverage crowdsourced data from Strava Metro to estimate bicycle volumes across the state of Texas, focusing on data that practitioners can regularly obtain and implement their own estimates following this guide. Therefore, we limited the data used to Strava Metro's standard data product, the Texas Department of Transportation's (TxDOT's) Roadway Inventory, and American Community Survey data. Following the recommended practice, we used negative binomial regression to develop the direct demand model for estimating AADB (*38*, *39*)

We found that functional classification, or the type of roadway or trail segment, is a key factor for estimating total use with crowdsourced data. This makes sense because Strava is marketed toward a recreation/fitness-oriented user base, and researchers expected these users to more often choose off-street paths based on previous research (*19*) Therefore, we expected Strava data to represent a relatively smaller proportion of users on urban arterial streets, where bicyclists may ride more often for work or shopping, rather than recreational trips logged using

Strava. We included functional classification to characterize the type of infrastructure on a given segment in the models. We found the model using the OSM classification had a lower prediction error compared to the roadway classification offered through by TxDOT roadway inventory data. This result indicates that the methodology can be easily adopted or calibrated by other states.

Preliminary model testing showed the number of households with income more than $200,000 a year was positively associated to the number of bicycle trips recorded on Strava. This finding reinforces expectations of a high-income bias to trip counts crowdsourced with this platform (*40*). Hence, transportation professionals should consider the role of an income bias in trip estimates, and that factors from this study may have different interactions in other contexts.

To develop the AADB models, we have used the ground counts collected from 100 count stations. The ground counts were mainly collected from urban areas and shared-use paths. Moreover, as indicated earlier, Strava uses OSM as the basemap. OSM classifies the roadways into 22 categories, while the sites used in this study only represent 7 of them. Although the model goodness of fit measures are within an acceptable range (i.e. 29% error margin, and 70% accuracy level), we suggest that the practitioners take caution when implementing these models to estimate the bicycle counts for rural segments and OSM functional classes that are not included in this study.

ACKNOWLEDGMENTS

The research summarized in this paper was funded by TxDOT contract DTFH61-12-D-00046 and is described in more detail in Report 0-6961. The authors acknowledge and appreciate the assistance and feedback of Chris Glancy, Phil Lasley, Joan Hudson, Haynes Bunn and Robert Benz.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: Bahar Dadashova, Shawn Turner, and Greg Griffin; data collection: Bahar Dadashova and Subasish Das. Author; analysis and interpretation of results: Bahar Dadashova, Greg Griffin, Shawn Turner, Subasish Das and Bonnie Sherman. Author; draft manuscript preparation: Bahar Dadashova, Greg Griffin, Shawn Turner, Subasish Das and Bonnie Sherman. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Griffin, G. P., and J. Jiao. Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods. *URISA Journal*, Vol. 27, No. 1, 2015, pp. 57–66.

2. 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.

3. Conrow, L., E. Wentz, T. Nelson, and C. Pettit. Comparing Spatial Patterns of Crowdsourced and Conventional Bicycling Datasets. *Applied Geography*, Vol. 92, No. December 2017, 2018, pp. 21–30. https://doi.org/10.1016/j.apgeog.2018.01.009.

4. Proulx, F. R., and A. Pozdnukhov. Bicycle Traffic Volume Estimation Using Geographically Weighted Data Fusion. 2017, pp. 1–14.

5. Sanders, R. L., A. Frackelton, S. Gardner, R. Schneider, and M. Hintze. Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2605, 2017, pp. 32–44. https://doi.org/10.3141/2605-03.

6. El Esawey, M., A. I. Mosa, and K. Nasr. Estimation of Daily Bicycle Traffic Volumes Using Sparse Data. *Computers, Environment and Urban Systems*, Vol. 54, 2015, pp. 195–203. https://doi.org/10.1016/j.compenvurbsys.2015.09.002.

7. Johnstone, D., K. Nordback, and S. Kothuri. Annual Average Nonmotorized Traffic Estimates from Manual Counts: Quantifying Error. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2672, No. 43, 2018, pp. 134–144. https://doi.org/10.1177/0361198118792338.

8. Cao, C., Z. Liu, M. Li, W. Wang, and Z. Qin. Walkway Discovery from Large Scale Crowdsensing. *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2018, pp. 13–24. https://doi.org/10.1109/IPSN.2018.00009.

9. Griffin, G. P., and J. Jiao. The Geography and Equity of Crowdsourced Public Participation for Active Transportation Planning. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. in press, 2019. https://doi.org/10.1177/0361198118823498.

10. Griffin, G. P., K. Nordback, T. Götschi, E. Stolz, and S. Kothuri. *Monitoring Bicyclist and Pedestrian Travel and Behavior*. Transportation Research Board, Washington, D.C., 2014.

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

12. 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, NCHRP Report 797*. Transportation Research Board, Washington, D.C., 2014.

13. Lindsey, G., K. Nordback, and M. A. Figliozzi. Institutionalizing Bicycle and Pedestrian Monitoring Programs in Three States: Progress and Challenges. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, pp. 1–22.

14. Turner, S., R. Benz, J. Hudson, G. P. Griffin, P. Lasley, B. Dadashova, and S. Das. *Improving the Amount and Availability of Pedestrian and Bicyclist Count Data in Texas*. Austin, TX, 2018.

15. Norman, P., C. M. Pickering, and G. Castley. What Can Volunteered Geographic Information Tell Us about the Different Ways Mountain Bikers, Runners and Walkers Use Urban Reserves? *Landscape and Urban Planning*, Vol. 185, No. February, 2019, pp. 180–190. https://doi.org/10.1016/j.landurbplan.2019.02.015.

16. Shearmur, R. Dazzled by Data: Big Data, the Census and Urban Geography. *Urban Geography*, No. August 2015, 2015, pp. 1–4. https://doi.org/10.1080/02723638.2015.1050922.

17. Hankey, S., G. Lindsey, X. Wang, J. Borah, K. Hoff, B. Utecht, and Z. Xu. Estimating Use of Non-Motorized Infrastructure: Models of Bicycle and Pedestrian Traffic in Minneapolis, MN. *Landscape and Urban Planning*, Vol. 107, No. 3, 2012, pp. 307–316. https://doi.org/10.1016/j.landurbplan.2012.06.005.

18. Misra, A., and K. Watkins. Modeling Cyclist Route Choice Using Revealed Preference Data: An Age and Gender Perspective. *Transportation Research Record: Journal of the Transportation Research Board*, 2018, p. 036119811879896. https://doi.org/10.1177/0361198118798968.

19. Griffin, G. P., and J. Jiao. Where Does Bicycling for Health Happen? Analysing Volunteered Geographic Information through Place and Plexus. *Journal of Transport & Health*, Vol. 2, No. 2, 2015, pp. 238–247. https://doi.org/10.1016/j.jth.2014.12.001.

20. Saad, M., M. Abdel-Aty, J. Lee, and Q. Cai. Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2019, p. 036119811983676. https://doi.org/10.1177/0361198119836764.

21. Boss, D., T. Nelson, M. Winters, and C. J. Ferster. Using Crowdsourced Data to Monitor Change in Spatial Patterns of Bicycle Ridership. *Journal of Transport & Health*, 2018. https://doi.org/10.1016/j.jth.2018.02.008.

22. Figliozzi, M., and B. Blanc. *Evaluating the Use of Crowdsourcing as a Data Collection Method for Bicycle Performance Measures and Identification of Facility Improvement Needs*. Salem, OR, 2015.

23. Romanillos, G., M. Zaltz Austwick, D. Ettema, and J. De Kruijf. Big Data and Cycling. *Transport Reviews*, Vol. 36, No. 1, 2016, pp. 114–133. https://doi.org/10.1080/01441647.2015.1084067.

24. 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.

25. Turner, S., P. Lasley, and B. Sherman. Texas Bicycle and Pedestrian Count Exchange. 2019.

26. Turner, S., Benz, R., Hudson, J., Griffin, G., Lasley, P., Dadashova, B. and Das, S., 2019. *Improving the Amount and Availability of Pedestrian and Bicyclist Count Data in Texas*(No. FHWA/TX-19/0-6927-R1).

27. Le, H. T. K., R. Buehler, and S. Hankey. Correlates of the Built Environment and Active Travel: Evidence from 20 US Metropolitan Areas. *Environmental Health Perspectives*, Vol. 126, No. 7, 2018, p. 077011. https://doi.org/10.1289/EHP3389.

28. Lu, T., A. Mondschein, R. Buehler, and S. Hankey. Adding Temporal Information to Direct-Demand Models: Hourly Estimation of Bicycle and Pedestrian Traffic in Blacksburg, VA. *Transportation Research Part D: Transport and Environment*, Vol. 63, No. May, 2018, pp. 244–260. https://doi.org/10.1016/j.trd.2018.05.011.

29. Schmiedeskamp, P., and W. Zhao. Estimating Daily Bicycle Counts in Seattle, Washington, from Seasonal and Weather Factors. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2593, 2016, pp. 94–102. https://doi.org/10.3141/2593-12.

30. Kuzmyak, J. R., J. Walters, M. Bradley, and K. M. Kockelman. *NCHRP Report 770, Estimating Bicycling and Walking for Planning and Project Development: A Guidebook*. Transportation Research Board of the National Academies, Washington, D.C., 2014.

31. Figliozzi, M., P. Johnson, C. M. Monsere, and K. Nordback. Methodology to Characterize Ideal Short-Term Counting Conditions and Improve AADT Estimation Accuracy Using a Regression-Based Correcting Function. *Journal of Transportation Engineering*, Vol. 140, No. 5, 2014.

32. McArthur, D. P., and J. Hong. Visualising Where Commuting Cyclists Travel Using Crowdsourced Data. *Journal of Transport Geography*, Vol. 74, No. November 2018, 2019, pp. 233–241. https://doi.org/10.1016/j.jtrangeo.2018.11.018.

33. Meyer, M. D. *Transportation Planning Handbook*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2016.

34. Ermagun, A., G. Lindsey, and T. Hadden Loh. Bicycle, Pedestrian, and Mixed-Mode Trail Traffic: A Performance Assessment of Demand Models. *Landscape and Urban Planning*, Vol. 177, No. May, 2018, pp. 92–102. https://doi.org/10.1016/j.landurbplan.2018.05.006.

35. Nordback, K., W. E. Marshall, and B. N. Janson. *Development of Estimation Methodology for Bicycle and Pedestrian Volumes Based on Existing Counts*. Colorado Department of Transportation (CDOT), Denver, CO, 2013.

36. Dadashova, B., G. P. Griffin, S. Das, S. Turner, and M. Graham. *Guide for Seasonal Adjustment and Crowdsourced Data Scaling*. College Station, Texas, 2018.

37. Breiman, L. Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32.

38. 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*, No. March, 2018, pp. 1–13. https://doi.org/10.1016/j.compenvurbsys.2018.03.002.

39. Wang, H., C. Chen, Y. Wang, Z. Pu, and M. B. Lowry. *Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions*. Seattle, WA, 2017.

40. Hochmair, H. H., E. Bardin, and A. Ahmouda. Estimating Bicycle Trip Volume for Miami-Dade County from Strava Tracking Data. *Journal of Transport Geography*, Vol. 75, No. January, 2019, pp. 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013.