Check for updates

TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–13
© National Academy of Sciences: Transportation Research Board 2020
Article reuse guidelines: sagepub.com/journals-permissions
DOI: 10.1177/0361198120946016
journals.sagepub.com/home/trr
SAGE

Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data

Bahar Dadashova[1], Greg P. Griffin[2], Subasish Das[1], Shawn Turner[1], and Bonnie Sherman[3]

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 or 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 annual average daily bicycle (AADB) counts. This study developed a direct-demand model for estimating AADB in Texas. Data from 100 stations, installed in 12 cities across the state, was used together with the crowdsourced Strava, roadway inventory, and American Community Survey 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 functional class and number of high-income households in a block group, can provide quite an accurate AADB estimate (29% prediction error).

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 or 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 to collect 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 companies aggregate these trips onto a transportation system network, process them for privacy, and then re-sell 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 valuable insights for both the agencies in relation to planning and policy decisions and road users in relation to travel choices. These data sources can be used to reveal quantitative insights into the behavior of non-motorized road users, such as route choice, which can support analysis of safety and mobility outcomes. Although crowdsourced data has a much more extensive coverage compared with non-motorized count stations, nevertheless the data still represent a small percentage of non-motorized users. For instance, researchers found that 3%–9% of bicycle trips counted on trails in Austin

[1]Texas A&M Transportation Institute, College Station, TX
[2]The University of Texas at San Antonio, San Antonio, TX
[3]Texas Department of Transportation, Austin, TX

Corresponding Author:
Bahar Dadashova, B-Dadashova@tti.tamu.edu

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 from the population, therefore affecting the functional form of the underlying process that generates the crowdsourced data. For example, users of activity-based smartphone apps are more consistent; therefore the temporal data produced by these users are stationary, in that the time series data generated from these apps may not exhibit significant seasonal patterns. In contrast, 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, and other issues. This study 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 the second section discusses the previous research on this subject. The third section describes data used in the study and the modeling approach for developing the direct-demand models. The fourth section presents the results of data mining and data analysis. The paper ends with conclusions, acknowledgments, author contributions, and references.

Background and Methodology

Literature Review and Importance of Research

Research on monitoring active transportation modes such as bicycling has supported advancements in practice, including reference-quality counts using permanent traffic recorders (*6*, *7*), and new approaches to crowdsource bicycling activity in addition to passive sensing using smartphones and other digital devices (*8*, *9*). Advancements in different approaches to bicycle counting support performance monitoring, including the critical challenge of comparing collision risk. Permanent counters provide continuous bicycle counts, often in 15-minute bins, but are relatively expensive to install and maintain—and therefore are seldom used to date (*10*). In addition, the permanent counters may record gaps data because of power problems, vandalism, or insect activity (*11*, *12*). However, state departments of transportation are building counting programs with high-quality equipment, improving availability and predictability of reference stations (*13*, *14*). These stations are critical for understanding temporal variation of trips, but do not cover the widespread locations needed for comprehensive safety analysis. Newer sources of big data such as smartphone records can complement these permanent stations by covering large areas, but they represent only a portion of trips at any given location, and introduce bias related to consumer use of the devices being tracked (*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.

Multiple companies aggregate bicycle trips recorded by individuals. However, Strava Metro is the only service that provides a dataset in multiple temporal aggregations for practical analysis in GIS-ready data formats. Evidence from earlier research and practice show predominant representation in Strava by a fitness and recreation-oriented market, nonetheless it supports a range of practical uses, including understanding where bicyclists ride for health (*19*), relative collision exposures (*5*, *20*), and temporal variation (*21*). The Oregon Department of Transportation explored practical use of Strava Metro soon after the service became available, finding it useful to identify routes with high bicycling ridership, but also a need to "expand this information up to total bike riders" (*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 (*24*). Though some studies have combined crowdsourced data with traffic counts and environmental data to understand bicycling contexts better, we suggest that both practitioners and researchers could benefit from a clear approach to expand crowdsourced data to estimate meaningfully the total volume of bicycling trips.

Research on improving bicycle traffic volume data contributes to the challenge of analyzing bicyclist safety by quantifying a denominator for a collision ratio. Bicycle volumes help planners know whether infrastructure changes affect the safety risk of bicycling, in addition to route preferences, equity, and other measures. This study builds on recent work to combine the advantages of emerging big data sources with high-accuracy reference stations. The following method section details the authors' approach in the State of Texas.

Methodology

Models support planners' and researchers' ability to understand transportation trends and scenarios based on limited data and to analyze policy and infrastructure changes for immediate and future contexts. Bicycle transportation models support analysis of the likelihood of cycling in a variety of conditions (*25*), including built environment factors (*26*), seasonal and weather factors (*27*), and temporal variation (*28*). Researchers' continuing model improvements may nonetheless be difficult to replicate or integrate into planning practice.

Bicycle count data forms the basis for model calibration, and the model accuracy requires balance with available resources (*29*). 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 (*30*) and examining bicycle traffic over larger areas through crowdsourced data collected by smartphone users (*31*). Crowdsourcing was first popularized in transportation planning as a public participation method to collect ideas from a broad range of people, and the approach is becoming more prevalent to monitor traffic (*32*). Regardless of input traffic data, bicycle traffic models can be assessed and improved through rigorous evaluation (*33*).

Random Forests. Because the list of potential factors for including in the regression is very comprehensive, this study used a 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 and is considered to be one of the most efficient classification methods (*34*). Instead of using support vector machine or other machine learning tools, RF was used in this study because of its variable importance measure, one of the most significant byproducts of RF. The classification accuracy and Gini impurity measure variable importance ranking. 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 important.

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

$$\mathrm{MDA}(x_i) = \frac{\sum_{\mathrm{tree}=1}^{B} \mathrm{MDA}^{\mathrm{tree}}(x_i)}{B} \qquad (1)$$

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

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):

$$\mathrm{MDG}^{n}(x_i) = 1 - \sum_{k=1}^{K} p(k|n) \qquad (2)$$

where $\mathrm{MDG}^{n}(x_i)$ 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. This study used the RF method to select the most important factors affecting the AADB demand.

Bicyclist Direct-Demand Model. Traditionally, bicycle demand has been estimated using several approaches, such as adjustment factors (*35*), ordinary least squares 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 follow a Poisson distribution, which is a discrete probability distribution; for example, the 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 relatively big data, however, most of the data sets used in non-motorized data studies are relatively small. Therefore most researchers use a negative binomial model, which is a standard choice for basic count data. The negative binomial regression model has the following functional form:

$$\mathrm{AADB}_i = \exp\left(\sum_{k=1}^{K} \beta_k \times X_{k,i} + \varepsilon_i\right) \qquad (3)$$

where $\mathrm{AADB}_i -$ is the AADB number at segment i; $\beta_k -$ is the coefficient estimate, $X_{k,i} -$ 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 the negative binomial model is then assumed to follow a gamma distribution with mean variance α^2, $\exp(\varepsilon_i) \sim G(1, \alpha^2)$. α is also referred to as an overdispersion parameter; lower overdispersion indicates a better model fit.

Data Overview

Count Data Collection and Quality Assurance

The bicycle count data used in this study was collected as part of Texas Department of Transportation (TxDOT)

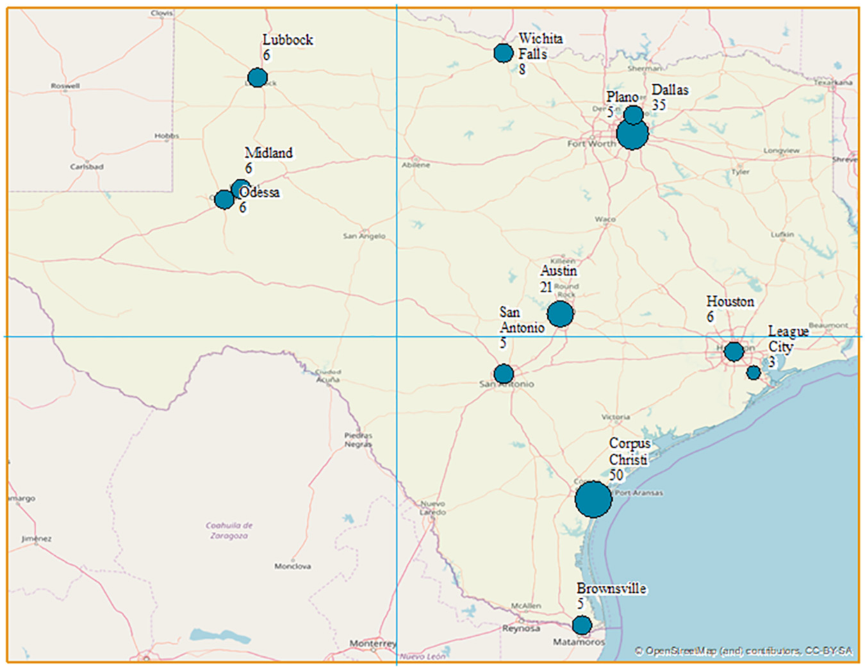

Figure 1. Map of Texas showing number of permanent and temporary counters per city.

project 0-6927 (*37*) and is readily available through the Texas Bicycle and Pedestrian Count Exchange Program website (*38*). This section briefly describes the data collection and quality control process. A more comprehensive data collection and quality check process can be found in Turner et al. (*39*).

Bicycle count data were collected 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). Permanent bicycle counts were provided by city and metropolitan planning organizations, while temporary counts (at least seven days) were collected by the project research team members using equipment owned by TxDOT. Of these stations, 118 are permanent, automatic bicycle (inductive loop) and pedestrian (infrared) counters operated by cities, metropolitan planning organizations, and special districts such as the San Antonio River Authority. The remaining count locations used mobile bicycle (pneumatic tube) and pedestrian (infrared) equipment. Bicycle counts were collected from a wide variety of facility types, including shared-use paths, bicycle lanes, shoulders, sidewalks, other paths, unpaved facilities, and shared roadways. In urban areas, agencies generally attempt to count recurring locations on an annual basis; however, many have noted that because of resource constraints, counts are either sporadic or occur every other year.

The bicycle count database was developed using the Federal Highway Administration's Traffic Monitoring Guide (*40*). 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 station's ID per travel direction, and latitude and longitude, among other variables.

The data quality and consistency were checked by examining the location information as well as the trend and seasonal patterns observed in daily counts. Count data obtained from agencies were already quality checked. Therefore the quality check was primarily carried out for the temporary counts. The daily bicycle counts were assigned to three groups and were either removed or kept in the database:

- Valid counts. Bicycle counts are assumed to be valid when data only appear during the days the counter is installed at the facility, and there are no sudden spikes or zero values. Other properties of valid data points are associated with strong weekend and consistent nightly uses. These data points were kept in the database.
- Abnormal but valid (ABV) counts. ABV refers to bicycle counts observed during special events (e.g., festivals and races) and abnormal weather. These data were adjusted and added to the database.

Table 1. Example of Count Location Information

Traffic Monitoring Guide 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	Northbound
Travel direction 2	Southbound

- Invalid counts. Invalid counts occur when the count data also appear during the days when the equipment is not being used, and there are sudden significant increase or decrease in bicycle counts that are not associated with a special event or abnormal weather. These errors can happen because of several reasons such as the installation of the counter, miscoding of the metadata, poor or careless maintenance of the metadata, and actual counting errors. The invalid data were removed from the final database.

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." 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. Note that in Strava, the roadway segments are labeled as "edges" while the intersections and segment endpoints (e.g., cul-de-sac) are labeled as "nodes." The number of athletes and activities shows the number of bicyclists and pedestrians on a given segment/intersection at the given year, day, hour, and minute. The number of activities indicates the total activity on a given edge or node, while the number of athletes indicates the number of unique user IDs on that edge or node. The difference between the two indicators is that the athlete number is adjusted such that if the same user appears on the edge or node more than once, then it is recorded only once. In contrast, the number of activities reports all the activities, regardless of the user ID.

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 bicyclists traveling the default direction of travel
Reverse athlete	Number of bicyclists traveling the opposite direction of travel
Activity	Number of bicyclists/pedestrians traveling the default direction of travel
Reverse activity	Number of bicyclists/pedestrians traveling the opposite direction of travel
Total activity	Number of total bicyclists/pedestrians on a given Strava segment/intersection

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, the following equation was used:

$$A = 180 + \arctan\left(\frac{Y2 - Y1}{X2 - X1}\right) \times \frac{180}{\pi} \tag{4}$$

$$\text{Cardinal Direction} = \begin{cases} WB & \text{if } 1 \leq A < 90 \\ SB & \text{if } 90 \leq A < 180 \\ EB & \text{if } 180 \leq A < 270 \\ NB & \text{if } 270 \leq A \leq 360 \end{cases} \tag{5}$$

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

Table 3 presents the descriptive statistics of the percentage of bicyclists using the Strava app per OSM functional class. As can be observed, the mean percentage of Strava users varies from 6% to 16%, according to OSM functional class.

Socioeconomic Factors and Roadway Data

A list of potentially important variables that can help to explain the relationships between the observed bicycle counts and Strava activity was compiled. For this

Table 3. Proportion of Strava to Bicycle Counts per Open Street Map (OSM) Functional Class (Annual Average Daily Counts)

OSM functional system	Sample size (n)	Strava user percentage			
		Min.	Max.	Mean	SD
Primary	5	1%	35%	8%	0.04
Secondary	20	0%	19%	6%	0.02
Tertiary	11	0%	70%	16%	0.13
Residential	29	0%	100%	7%	0.19
Path	9	0%	75%	8%	0.06
Cycleway	19	0%	100%	7%	0.09
Footway	7	0%	100%	6%	0.12

Note: Min. = minimum; Max. = maximum; SD = standard deviation.

purpose, the American Community Survey and the TxDOT roadway inventory database were used.

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 unique features of using ACS is its ability to produce estimates on a wide range of geographies, including low geographic levels such as block groups. Block group level ACS data for Texas was collected. As ACS contains an extensive list of variables, the variable selection was conducted by using RF (discussed above).

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.

The acquired databases were conflated on the Strava network using ArcMap 10.5.1. It is important to note that observed bicycle count data is a point data, Strava and RHiNO are polynomial, and ACS is polyline data. Table 4 presents the descriptive statistics of all the variables and data sources considered for the analysis.

The following steps were followed 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. If a Strava segment passes through two or more block groups, assign mean values of the block group level information to the Strava segment.
3. Conflate RHiNO roadway level data to the Strava segments.

Results

After removing the sites with missing data and with short-term counts (i.e., one week), 100 out of 155 stations were used to develop the direct-demand models for estimating AADB. The counts from short-term stations were used to cross-validate the estimation results; and the leave-one-out approach was used to cross-validate the AADB models. Finally prediction analysis was conducted using the estimation results, and the observed and predicted AADB were compared.

Selection of the Most Influential Factors

RF methodology was used 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 according to two important measures discussed above: mean decrease accuracy (Figure 2*a*) and Gini impurity (Figure 2*b*).

The initial analysis results indicate that household income and demographic variables are very influential for explaining bicyclist counts (Figure 2). Because most of these variables belong to the same category, the most important variables from each category were selected and RF analysis was conducted again. Figure 3 shows the results of the second RF test.

Finally, the following variables were found to be the most important for explaining the AADB counts: Strava sample (Strava), OSM functional class (Strava), the male population in the age group 35–49 (ACS), number of households with income of more than $200,000 per annum (ACS), number of lanes (RHiNO), and roadway facility type (RHiNO).

AADB Direct-Demand Models

As can be observed, two sets of important variables have been identified. The first set of variables includes only OSM functional class and ACS factors. The second set of variables is from the TxDOT roadway inventory (RHiNO). Therefore two direct-demand models were developed based on the need and availability of data. The first model, which is also more parsimonious, included only OSM and ACS variables. The second model included OSM, ACS, and RHiNO variables. The estimation results of the two models indicated that the male population in the age group 35–49 was not statistically significant in either. In the second model, the OSM functional class was not found to be statistically

Table 4. Descriptive Statistics of Variables

Variable name	Source	Unit of analysis	Min.	Max.	Mean	SD
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
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	Manual	Point	1	669	66.93	127.68
Non-motorized facility width	Manual	Polyline	4	25	8.47	4.13
Non-motorized facility buffer width	Manual	Polyline	0	5	2.91	0.84
Median width	RHiNO	Polyline	0	16	4.6	2.66
Number of lanes	RHiNO	Polyline	0	6	2.75	1.09
Posted speed limit	RHiNO	Polyline	0	55	17.35	16.75
Inside shoulder width	RHiNO	Polyline	0	10	0.3	1.71
Outside shoulder width	RHiNO	Polyline	0	20	0.6	2.95
Surface width	RHiNO	Polyline	0	76	33.85	16.33
Average activity (AvgActivity)	Strava	Polyline	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, Wichita Falls
Non-motorized facility type	Manual	Polyline	Shared-use path; on-street bike lane
Parking	Manual	Polyline	No on-street parking; parallel parking
Pavement condition	Manual	Polyline	Poor; fair; good; excellent
Pavement type	Manual	Polyline	Asphalt; concrete; crushed granite/gravel
Place of interest (POI) within 50 miles	Manual	Polyline	High school; university
Shade	Manual	Polyline	None; partial; full
Street lighting	Manual	Polyline	None; one side; both sides; partial
Transit	Manual	Polygon	No; yes
Functional classification	RHiNO	Polyline	Principal arterial; minor arterial; collector; local; shared path or trail
OSM functional system (CLAZZ)[a]	Strava	Polyline	15 = Primary; 21 = Secondary; 31 = Tertiary; 32 = Residential; 72 = Path; 81 = Cycleway; 91 = Footway

Note: ACS = American Community Survey; RHiNO = Roadway Highway Inventory Network Offload; Min. = minimum; Max. = maximum; SD = standard deviation.
[a]Definition of Open Street Map (OSM) functional class or highway link can be found in this link: https://wiki.openstreetmap.org/wiki/Key:highway.

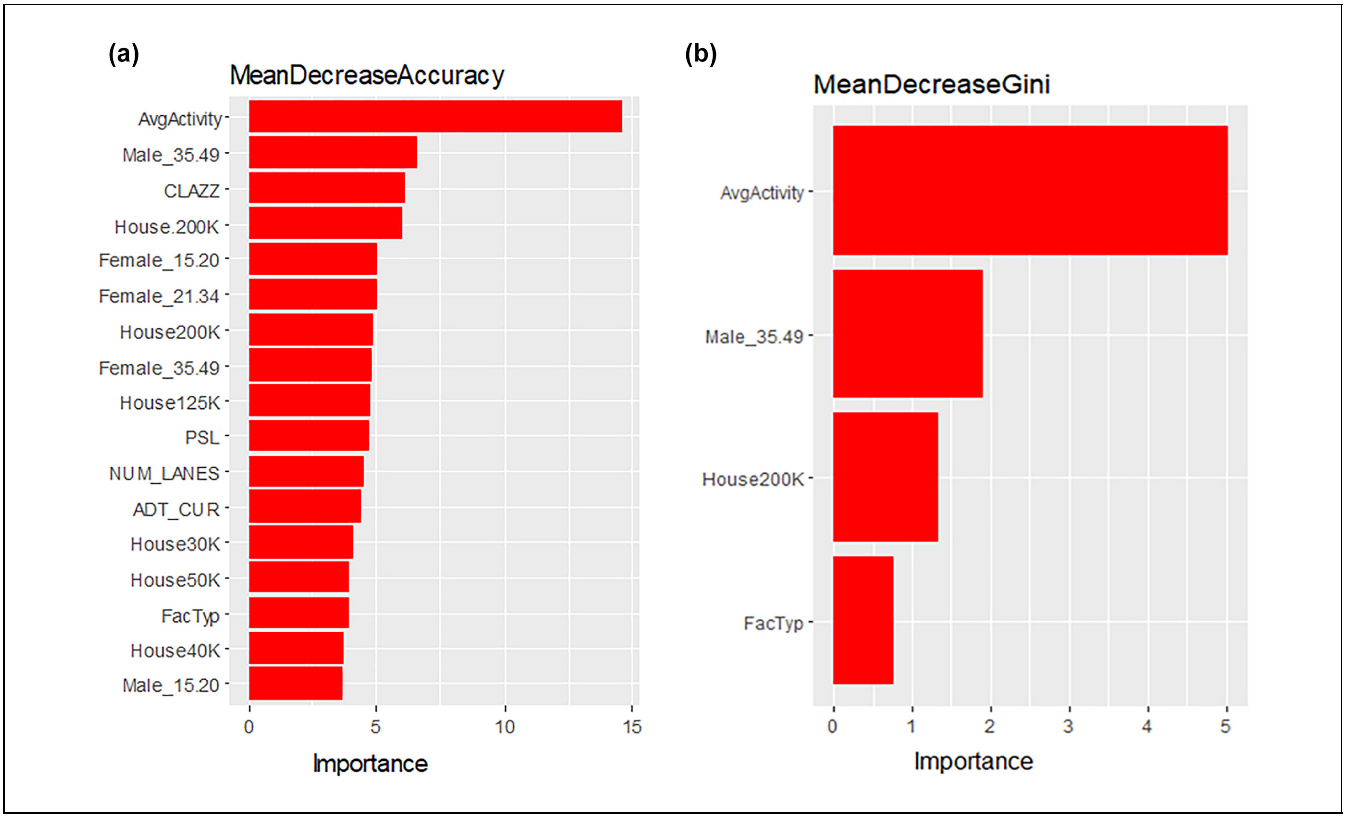

Figure 2. Preliminary list of important variables: (*a*) mean decrease accuracy and (*b*) mean Gini impurity.

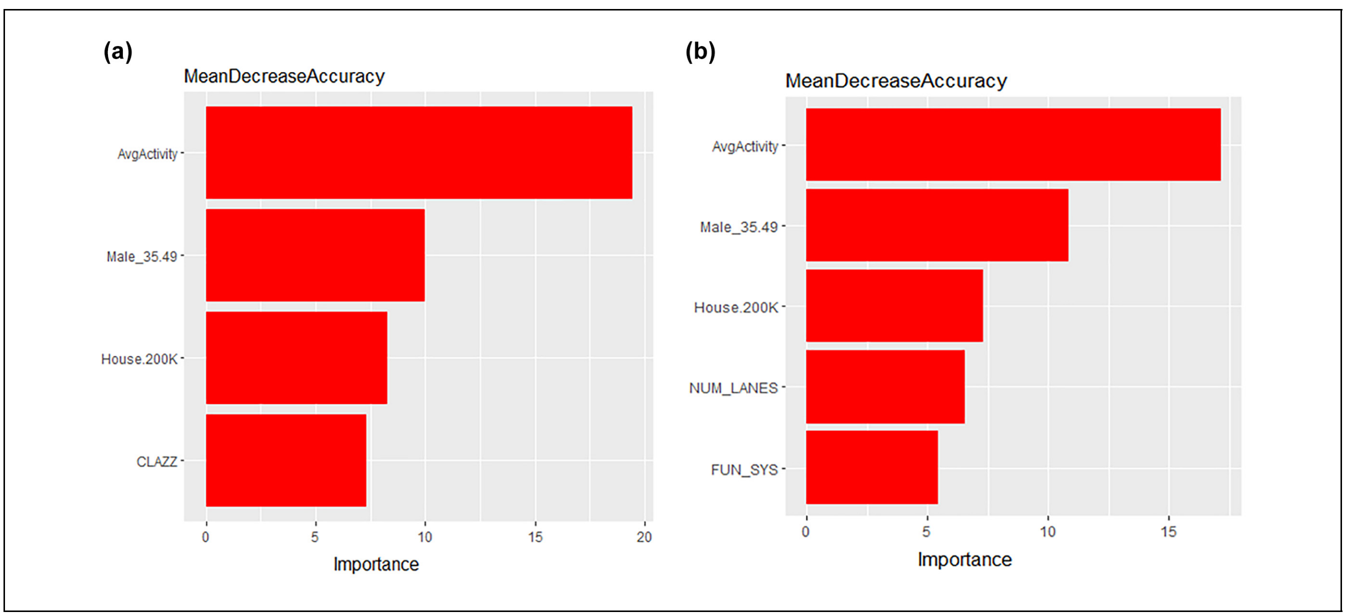

Figure 3. Final list of important variables: (*a*) Open Street Map functional class and (*b*) roadway functional system.

Table 5. Direct-Demand Model Estimation Results

Variables		Model 1			Model 2		
		Estimate	SD	*p*-value	Estimate	SD	*p*-value
Open Street Map highway functional class	Primary	4.138	0.053	<0.001	na	na	na
	Secondary	2.590	0.060	<0.001	na	na	na
	Tertiary	3.078	0.062	<0.001	na	na	na
	Residential	2.862	0.037	<0.001	na	na	na
	Path	4.271	0.031	<0.001	na	na	na
	Cycleway	4.144	0.027	<0.001	na	na	na
	Footway	3.323	0.062	<0.001	na	na	na
Functional system	Collector (Minor)	na	na	na	3.211	0.078	<0.001
	Local road	na	na	na	2.506	0.083	<0.001
	Minor arterial	na	na	na	2.987	0.118	<0.001
	Principal arterial	na	na	na	3.929	0.116	<0.001
	Shared path or trail	na	na	na	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		na	na	na	–0.066	0.027	<0.05
LOOCV error			187			586	
Overdispersion			0.967			1.172	
R^2 (model accuracy)			75%			70%	

Note: SD = standard deviation; AADB = annual average daily bicycle count; LOOCV = leave-one-out cross-validation; na = not applicable.

significant, therefore it was removed from this model. The resulting models have the following functional form:

$$\mathrm{AADB}_i = \exp(\beta_0 + \beta_1 \times \text{AADB Strava}_i + \beta_2 \times \text{Household}{>}200\text{K}_i + \ + \beta_3 \times \text{OSM Class}_i) \tag{6.1}$$

$$\mathrm{AADB}_i = \exp(\beta_0 + \beta_1 \times \text{AADB Strava}_i + \beta_2 \times \text{Household}{>}200\text{K}_i + \ + \beta_3 \times \text{Func. System}_i + \beta_4 \times \text{Num. of Lanes}_i) \tag{6.2}$$

where $\mathrm{AADB}_i -$ represents the estimated AADB at segment/edge i; AADB Strava_i represents the annual average daily Strava users at location i for the given time period; $\text{Household}{>}200\text{K}_i$ represents the number of households with more than $200,000.00 annual income; OSM Class_i represents the OSM functional class of the Strava segment; 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.

The leave-one-out method was used to cross-validate the estimation results. In this approach, the negative binomial models are developed by using all but one observation. Therefore a total of 100 models are developed, and the MSE is calculated by comparing the predicted and observed value of the remaining observation. Finally the leave-one-out cross-validation (LOOCV) error was calculated by averaging the prediction error of all 100 models. Table 5 shows the estimation results for both models, together with the LOOCV error, model overdispersion, and R-squared value. Both models have a relatively lower overdispersion parameter (˜1) and higher R^2 values (<0.7), indicating that both models are a good fit for the data.

Prediction Analysis. Finally, using the model estimation results, the AADB counts were predicted and compared with the observed counts. 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 the OSM-based model is relatively better than the RHiNO-based model (29% versus 38%).

Model Interpretation and Discussions

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 AADB 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 Strava sample counts in Texas for counts taken between 2016 and 2018, or adapted for other contexts using the methods proposed in this paper. Note that all these

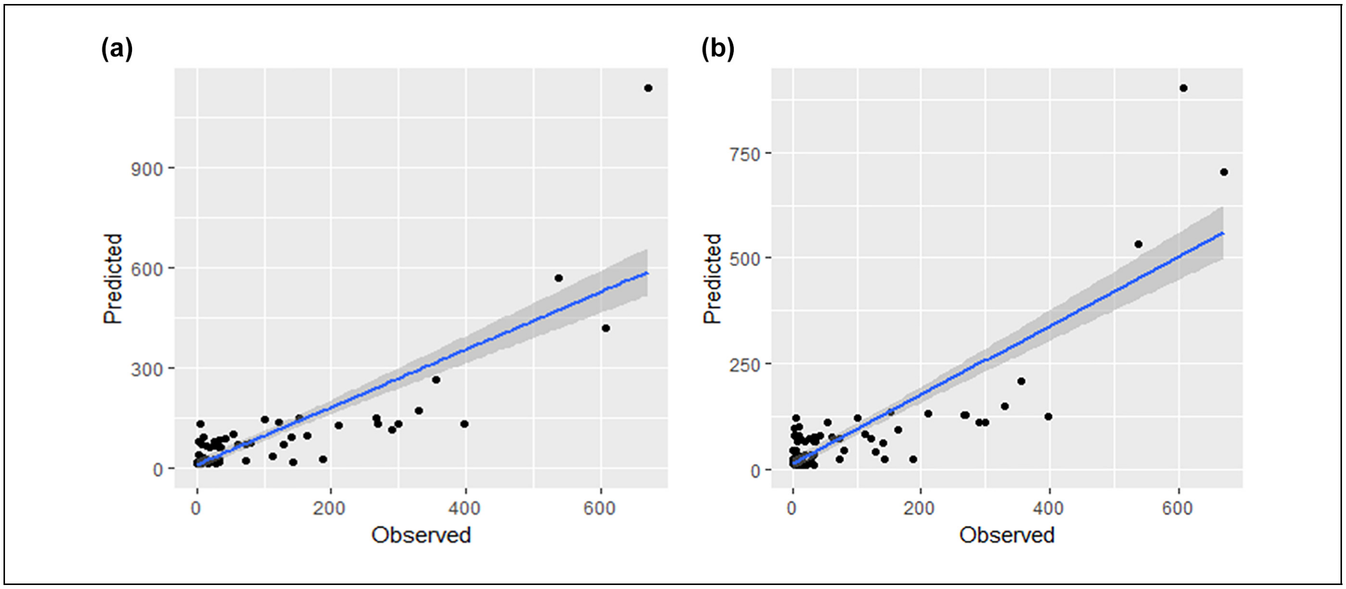

Figure 4. Predicted versus observed annual average daily bicycle count: (*a*) Model 1 and (*b*) Model 2.

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

estimates are associated with a 29% error rate and are not directly transferrable to other contexts.

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 particular 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. Though the model is calibrated to on-ground traffic counts, future research should further evaluate model accuracy through cross-validation using more counting sites as they become available.

Changes in segment classification over time, such as upgrading a street from a tertiary to secondary segment, could significantly affect 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 household income in an average manner, such as bicycling loops in large parks. The 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, affect bicycling rates.

Conclusions and Recommendations

Several different approaches to leverage crowdsourced data from Strava Metro to estimate bicycle volumes across the State of Texas were explored, focusing on data

Table 7. Estimated Number of Bicycle Counts Given Strava Sample and Roadway Class in Texas, 2016–2018

	Open Street Map functional class						
Strava sample counts	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

that practitioners can regularly obtain and implement in their estimates following this guide. Therefore, the data used was limited to Strava Metro's standard data product, TxDOT's roadway inventory, and ACS data. Following the recommended practice, negative binomial regression was used to develop the direct-demand model for estimating AADB (*41*, *42*)

It was 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 the researchers expected these users to choose off-street paths more often, based on previous research (*19*). Therefore, Strava data was expected 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. Functional classification was included to characterize the type of infrastructure on a given segment in the models. It was found that the model using the OSM classification had a lower prediction error than the roadway classification offered by the TxDOT roadway inventory data. This result indicates that the methodology can be readily adopted or calibrated by other states. To reduce the estimation error increasing the sample size of observed counts is recommended. Moreover, using more sites from diverse types of bicycle facilities may help to improve the accuracy for different functional classes.

Preliminary model testing showed the number of households with annual income of more than $200,000 was positively associated with the number of bicycle trips recorded on Strava. This finding reinforces expectations of a high-income bias to trip counts crowdsourced with this platform (*43*). Therefore, 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, the ground counts collected from 100 count stations were used. The ground counts were mainly collected from urban areas and shared-use paths. Moreover, as indicated above, Strava uses OSM as the base map. OSM classifies the roadways into 22 categories, while the sites used in this study represent just seven of them. Although the model goodness of fit measures are within an acceptable range (29% error margin, and 70% accuracy level), the authors suggest that practitioners use 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 authors acknowledge and appreciate the assistance and feedback of Chris Glancy, Phil Lasley, Joan Hudson, Haynes Bunn, and Robert Benz. The authors would like to thank the editor and four anonymous reviewers for the feedback and comments that helped to improve the quality of the paper significantly.

Author Contributions

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

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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.

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, 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. *Journal of Transportation Geography*, 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*, 2017. 2605: 32–44.
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, 2018. 2672: 134–144.
8. Cao, C., Z. Liu, M. Li, W. Wang, and Z. Qin. Walkway Discovery from Large Scale Crowdsensing. *Proc., 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*, Porto, Portugal, IEEE, New York, 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*, 2019. 2673: 460–468.
10. Griffin, G. P., K. Nordback, T. Götschi, E. Stolz, and S. Kothuri. *Transportation Research Circular E-C183: Monitoring Bicyclist and Pedestrian Travel and Behavior*. Transportation Research Board of the National Academies, 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*, 2014. 2468: 64–73.
12. Ryus, P., E. Ferguson, K. M. Laustsen, R. J. Schneider, F. R. Proulx, T. Hull, and L. Miranda-Moreno. *NCHRP Report 797: Guidebook on Pedestrian and Bicycle Volume Data Collection*. Transportation Research Board of the National Academies, 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. 2443: 134–142.
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*. Texas A&M Transportation Institute, 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, 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*, Vol. 36, No. 7, 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. 2672: 145–154.
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. 2673: 1–14.
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*, Vol. 9, 2018, pp. 226–233. 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*. Oregon Department of Transportation, Salem, OR, 2015.
23. Romanillos, G., M. Z. 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. Thakuriah, P. V., P. Metaxatos, J. Lin, and E. Jensen. An Examination of Factors Affecting Propensities to Use Bicycle and Pedestrian Facilities in Suburban Locations. *Transportation Research Part D: Transport and Environment*, Vol. 17, No. 4, 2012, pp. 341–348. https://doi.org/10.1016/j.trd.2012.01.006.
26. 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.
27. 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*, 2016. 2593: 94–102.
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, 2018, pp. 244–260. https://doi.org/10.1016/j.trd.2018.05.011.
29. 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.
30. 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, p. 04014014.
31. McArthur, D. P., and J. Hong. Visualising Where Commuting Cyclists Travel using Crowdsourced Data. *Journal of Transport Geography*, Vol. 74, 2018, 2019, pp. 233–241. https://doi.org/10.1016/j.jtrangeo.2018.11.018.

32. Meyer, M. D. *Transportation Planning Handbook*. John Wiley & Sons, Inc., Hoboken, NJ, 2016.
33. Ermagun, A., G. Lindsey, and T. H. Loh. Bicycle, Pedestrian, and Mixed-Mode Trail Traffic: A Performance Assessment of Demand Models. *Landscape and Urban Planning*, Vol. 177, 2018, pp. 92–102. https://doi.org/10.1016/j.landurbplan.2018.05.006.
34. Breiman, L. Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32.
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, Denver, CO, 2013.
36. Dadashova, B., G. P. Griffin, S. Das, S. Turner, and M. Graham. *Guide for Seasonal Adjustment and Crowdsourced Data Scaling*. Technical Report 0-6927-P6. Texas Department of Transportation, College Station, TX, 2018.
37. 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*. Technical Report 0-6927-R1. Texas Department of Transportation, College Station, TX, 2019. https://static.tti.tamu.edu/tti.tamu.edu/documents/0-6927-R1.pdf.
38. Turner, S., P. Lasley, and B. Sherman. *Texas Bicycle and Pedestrian Count Exchange*. Texas A&M Transportation Institute, College Station, TX, 2019.
39. Turner, S., P. Lasley, J. Hudson, and R. Benz. *Guide for Pedestrian and Bicyclist Count Data Submittal*. Technical Report 0-6927-P7. Texas Department of Transportation, College Station, TX, 2019.
40. Federal Highway Administration. *Traffic Monitoring Guide*. U.S. Department of Transportation, Bureau of Transportation, Washington, D.C., 2016.
41. 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*, Vol. 70, 2018, pp. 125–137. https://doi.org/10.1016/j.compenvurbsys.2018.03.002.
42. 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.
43. 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, 2019, pp. 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013.