The Journal of Transport and Land Use jtlu.org
Vol. 17 No. 1 [2024] pp. 41–65

The use of crowdsourced mobile data in estimating pedestrian and bicycle traffic: A systematic review

Tao Tao (corresponding author)
Carnegie Mellon University
taot@andrew.cmu.edu

Greg Lindsey
University of Minnesota
linds301@umn.edu

Raphael Stern
University of Minnesota
rstern@umn.edu

Michael Levin
University of Minnesota
mlevin@umn.edu

Abstract: To address the need for better non-motorized traffic data, policymakers and researchers are collaborating to develop new approaches and methods for estimating pedestrian and bicyclist traffic volumes. Crowdsourced mobile data, which has higher spatial and temporal coverage and lower collection costs than data collected through traditional approaches, may help improve pedestrian and bicyclist traffic estimation despite their limitations or biases. This systemic literature review documents how researchers have used crowdsourced mobile data to estimate pedestrian and bicyclist traffic volumes. We find that one source of commercial fitness application data (i.e., Strava) has been used much more frequently than other crowdsourced mobile data, and that most studies have used crowdsourced mobile data to estimate bicyclist volumes. Comparatively few studies have estimated pedestrian volumes. The most common approach to the use of crowdsourced counts is as independent variables in direct demand models. Variables constructed from crowdsourced mobile data not only have significant correlations with observed counts in statistical models but also have larger relative importance than other factors in machine learning models. Studies also show that including crowdsourced mobile data can significantly improve estimation performance. Future research directions include application of crowdsourced mobile data in more pedestrian traffic estimations, comparison of the performance of different crowdsourced mobile data, incorporation of multiple data sources, and expansion of the methods using crowdsourced mobile data for non-motorized traffic estimation.

Keywords: Strava, non-motorized traffic, direct demand model, traffic estimation

Article history:
Received: November 30, 2022
Received in revised form: September 28, 2023
Accepted: November 5, 2023
Available online: February 2, 2024

https://doi.org/10.5198/jtlu.2024.2315

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

1 Introduction

Active travel modes, such as walking and biking, are important to address challenging issues in modern society such as auto dependence, air pollution, climate change, obesity, and physical and mental health. However, the share of these active travel modes has been decreasing in recent decades (FHWA, 2018). A better understanding of the spatial and temporal distribution of pedestrian and bicyclist traffic is necessary to promote active travel modes. In contrast to vehicular monitoring programs, historic investments in pedestrian and bicyclist traffic monitoring programs have been very limited (FHWA, 2016). More than half of the states in the US do not have well-established monitoring programs (Ohlms et al., 2019).

To address the lack of comprehensive pedestrian and bicyclist traffic volume data for segments and intersections in networks, engineers have used several approaches to estimation, including expansion factors, direct demand models, four-step models, and agent-based simulation models (Bhowmick et al., 2022; Turner et al., 2017). While these methods utilize different mechanisms, they all share a common requirement for high-quality data as input. For example, direct demand models, which establish the relationships between observed counts and multiple factors such as built environment characteristics, socio-demographics, and roadway geometry, need traffic counts related datasets to carry out the analysis. Four-step models need household travel survey data to estimate the origin and destination (OD) of pedestrian and bicyclist trips among different zones in the studied area.

In recent years, researchers have begun to use crowdsourced mobile data when estimating pedestrian and bicyclist volumes. Crowdsourced mobile data are defined as data collected from a diverse group of individuals with varying levels of expertise through an open call for voluntary participation, using mobile devices and related technology (Estellés-Arolas & González-Ladrón-De-Guevara, 2012). Mobile devices have a large number of users across both time and space. Compared with traditional count datasets, which are collected by sensors or people at a limited number of locations, crowdsourced mobile data have a higher spatial and temporal coverage. The cost of data collection is also lower for crowdsourced mobile data as they are mostly by-products of the services provided by mobile devices. Although new approaches for using crowdsourced mobile data have been proposed, and vendors with proprietary models for estimating bicyclist and pedestrian traffic volumes exist (e.g., StreetLight), routine procedures for using crowdsourced mobile data have yet to be developed and standardized. Given the rapid evolution in the field and the objectives of this project, it is useful to review related studies and provide a synthesis of crowdsourced mobile data and related methods for estimating bicyclist and pedestrian traffic. Our literature review aims to address the following questions:

- What types of crowdsourced mobile data have been utilized to estimate pedestrian and bicyclist traffic volumes?
- Which methods have been employed in the estimation process?
- How does crowdsourced mobile data perform in the estimation?

We use a systematic review approach to address these questions. First, we identify and summarize 10 related literature reviews that have focused on the broader topic of pedestrian and bicyclist traffic estimation, with some touching upon the application of crowdsourced mobile data. These reviews provide the context for our review of empirical papers by summarizing the various applications of crowdsourced data in pedestrian and bicyclist studies. This summary also illustrates the need for more focused reviews on methodological considerations in the use of crowdsourced data in volume estimation, specifically, their performance in estimating measures such as average annual daily pedestrians or bicyclists.

Second, we review 22 empirical papers. Among other elements of the papers, we summarize modes, data sources, modeling approach, how crowdsourced mobile data is used, and the performance of crowdsourced mobile data related variables. Our results provide methodological guidance for researchers interested in using crowdsourced mobile data for pedestrian and bicyclist traffic estimation research and illustrate where additional research will benefit the field.

Following this introduction section, we present the protocols for our review, including criteria for inclusion of papers (Section 2). We then summarize the previous literature reviews (Section 3). We introduce different types of crowdsourced mobile data in Section 4. Sections 5 through 7 summarize key methodological aspects of the empirical papers identified with our protocols. In Section 8, we conclude our findings for this study and discuss implications for future studies.

2 Method

We applied a systematic review approach in this study. A systematic literature attempts to identify all papers written on a topic using explicit inclusion and exclusion criteria and to characterize the state of knowledge based on a comprehensive assessment of evidence (Grant & Booth, 2009; Xiao & Watson, 2019). Traditional literature reviews are mostly narrative as they are more likely to be purposeful and selective, focusing on the progress or development of the topic (Snyder, 2019).

Systematic reviews have several key steps (Xiao & Watson, 2019). First, we defined the three research questions listed in the introduction section. Second, we produced a protocol for conducting the review. The protocol specified the search criteria, selection criteria, and coding categories. Third, we searched several databases of literature with keywords related to our research topic. Fourth, we screened the search results and included or removed studies based on the selection criteria. This step included identification of other review papers and the empirical papers to be summarized in detail. Fifth, we extracted the information defined by the coding categories after thoroughly reviewing the selected studies. Finally, we analyzed the results and presented our findings.

We defined three sets of keywords related to our research questions. The first set includes *pedestrian*, *bicyclist*, and *bike*. These keywords identify the travel modes we want to study. The second set includes *demand*, *traffic*, *volume*, *AADB* (Annual Average Daily Bicyclists), and *AADP* (Annual Average Daily Pedestrians). These keywords are different terms or measures used to describe pedestrian or bicyclist traffic. The last set of keywords includes *forecasting*, *estimation*, and *prediction*. They are closely related to traffic estimation. We did not define keywords related to crowdsourced mobile data as we wanted to include more studies in the search process. We selected those using crowdsourced mobile data in the step of screening. Some less-used keywords were not included, for example, *cyclists* or *non-motorized travel*. We did this for two reasons. First, these keywords are closely related to the keywords we have defined in the protocol. Therefore, studies using these keywords often showed up in the search results. Second, we also included additional studies when we found they are related to our study during the review process. We searched four databases, including *Web of Science*, *Google Scholar*, *Academic Search Premier*, and *PubMed*. We searched for the keywords in the title, abstracts, and keywords of the studies. All the results were searched before March 31, 2023.

Screening search results included two rounds (Figure 1). In the first round, we skimmed the titles, abstracts, and keywords of the studies and only included (1) studies published in English; (2) journal articles, conference proceedings, book sections, theses, and research reports; (3) studies that propose a method or apply a method to predict pedestrian and/or bicyclist traffic with one or more datasets; and (4) literature review studies or empirical studies with comprehensive literature reviews. After this round, we found 172 studies from the search results. In the second round of screening, we reviewed the title, abstract, and introduction of the searched items and removed replicated studies, giving priority to journal articles over reports when they included the same findings. We also removed articles that are not closely related to the topic. In addition, we removed studies that did not incorporate or use crowdsourced mobile data. After two rounds of screening, 19 empirical studies remained for further review and analysis. At the same time, we identified 10 literature review studies and empirical papers with comprehensive reviews for the purpose of establishing the context for this paper.

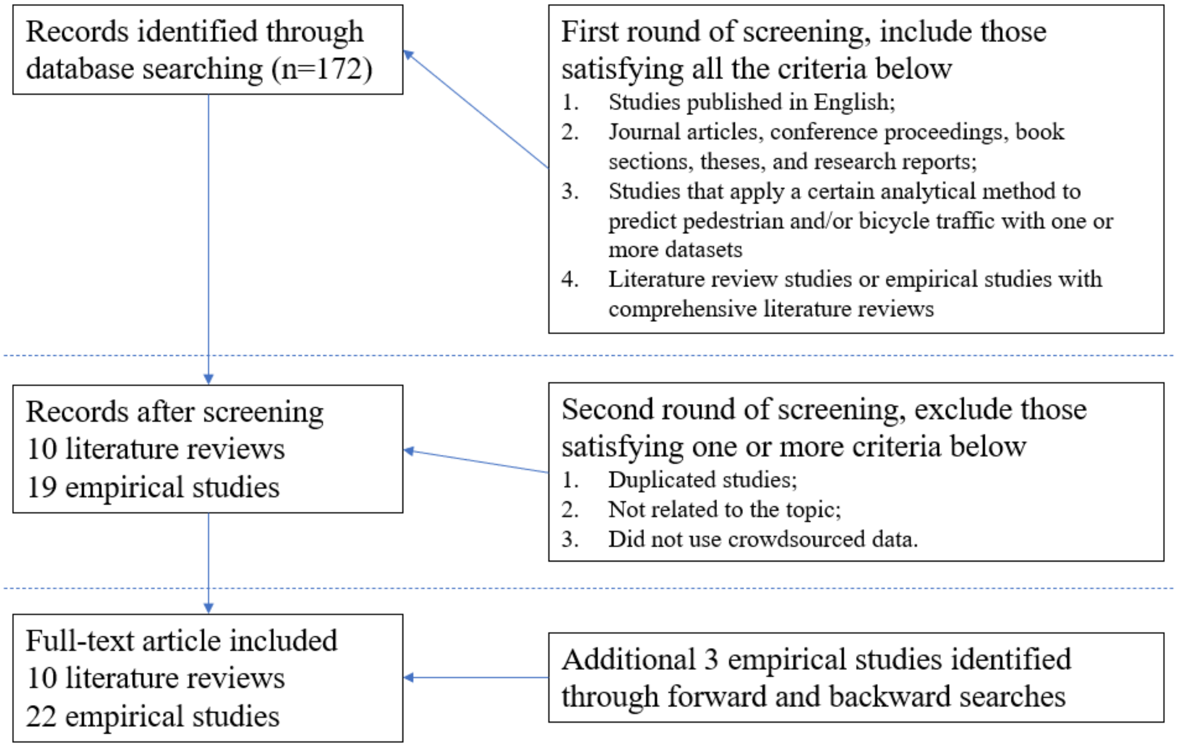
Records identified through database searching (n=172)
First round of screening, include those satisfying all the criteria below
1. Studies published in English;
2. Journal articles, conference proceedings, book sections, theses, and research reports;
3. Studies that apply a certain analytical method to predict pedestrian and/or bicycle traffic with one or more datasets
4. Literature review studies or empirical studies with comprehensive literature reviews
Records after screening
10 literature reviews
19 empirical studies
Second round of screening, exclude those satisfying one or more criteria below
1. Duplicated studies;
2. Not related to the topic;
3. Did not use crowdsourced data.
Full-text article included
10 literature reviews
22 empirical studies
Additional 3 empirical studies identified through forward and backward searches

Figure 1. Searching and screening results

Last we extracted key information from each of the selected literature reviews and empirical studies. For the 10 literature reviews, the information includes general information about the studies (e.g., authors, year, and title), travel modes, type of review, number of papers reviewed, focus of review, and selected findings. As to the 19 empirical studies, the information includes general information, study locations, travel modes, application of their results, types of crowdsourced mobile data, information about pedestrian or bicyclist count data, methods, and performance of the crowdsourced mobile data related variables. During this extraction process, we found three additional empirical studies in the reference lists, bringing the total number of empirical studies in this review to 22 (Figure 1).

3 Previous literature reviews related to pedestrian and bicyclist volume estimation

Table 1 presents key information from the 10 literature reviews identified in the search process. These data include travel mode, type of review, number of papers reviewed, focus, and selected findings. Although the exact number of empirical papers included in each review was difficult to determine, the number ranged from as few as 14 to more than 40. Of these reviews, five concentrate on both pedestrian and bicycle modes, while three focus solely on bicyclists, and two specifically on examining pedestrians. Most of the studies (7) were traditional narrative reviews that selectively examine the progress or development in the field or across research domains (Snyder, 2019). Three of them used a systematic review approach (Grant & Booth, 2009; Xiao & Watson, 2019). In addition, six studies were published as independent review articles. For the other four papers, literature review comprises only one section and serves for the empirical analysis.

These literature review papers cover a wide range of topics related to pedestrian and bicyclist traffic estimation (Table 1). Three studies reviewed various methods used to model pedestrian and bicyclist volumes (Bhowmick et al., 2022; Turner et al., 2017; Yasmin et al., 2021). In general, these methods include direct demand models, four-step regional models, and agent-based models, with direct demand models described most often (Bhowmick et al., 2022; Yasmin et al., 2021). Three studies reviewed models used in direct demand models (e.g., linear regression, generalized linear regression, and machine learning), including the independent variables associated with traffic volumes (Munira & Sener, 2017; Schneider et al., 2021; Singleton et al., 2021). Two early studies reviewed the practices on four-step regional models (Grant & Booth, 2009; Turner et al., 1997). Finally, two studies reviewed the general applications of crowdsourced data (Lee & Sener, 2020b; Nelson, Ferster, et al., 2021). These applications include estimating travel demand, assessing safety, and others.

Although some of previous literature review studies note the application of crowdsourced data in estimating pedestrian and bicyclist traffic estimation, most of the reviews are narrative, concentrate more on the various applications of crowdsourced data, and do not present or assess more detailed methodological considerations or comparisons of model performance. For example, Lee and Sener (2020b) reviewed seven elements of applications of Strava Metro data, including identifying travel pattern, estimating travel demand, analyzing route choices, etc. (Table 1). For estimating travel demand, they provided a useful overview of five studies but did not include specifications or detailed descriptions of models. Furthermore, the previous literature reviews did not offer systemic review of the performance of crowdsourced data in estimating pedestrian and bicyclist volumes. For example, when reviewing the application of crowdsourced data in estimating travel demand, Nelson, Ferster, et al. (2021) focused on the types of crowdsourced data that have been applied but did not extend to discussion on how crowdsourced data improved the estimation of volumes.

Table 1. Summary of previous literature reviews on pedestrian and/or bicyclist volume estimation

Author	Mode	Type of review	# Of papers reviewed	Focus of review	Selected findings of review
Turner et al. (1997)	Pedestrian and bicycle	Narrative review (Independent review paper)	Not clear in citations	Four-step regional models for estimating volumes	• The study summarized four types of modified or improved regional models
Liu et al. (2012)	Pedestrian and bicycle	Narrative review (Independent review paper)	Not clear in citations	Four-step regional models for estimating volumes	• Refined zones and improved variable measurement could enhance the four-step regional models • More calibrated and validated pedestrian and bicyclist count data are needed
Munira and Sener (2017)	Pedestrian and bicycle	Systemic review (Independent review paper)	22	Direct demand models and independent variables for estimating volumes	• Pedestrian and bicyclist volumes should be modeled and interpreted separately • Negative binomial model is suitable for modeling pedestrian and bicyclist volumes
Turner et al. (2017)	Pedestrian and bicycle	Narrative review (Part of a paper)	16	Methods for estimating volumes	• Geographic scale is important for deciding the pedestrian and bicyclist volume estimation methods and there exist four types of geographic scales: 1) regional; 2) network; 3) road segment; 4) point
Lee and Sener (2020b)	Bicycle	Systemic review (Independent review paper)	27	Applications of Strava data in bicycling research and practice	• Strava data have been applied in seven types of applications: 1) identifying travel pattern; 2) estimating travel demand; 3) analyzing route choice; 4) evaluating the impact of new infrastructure; 5) controlling for exposure in crash analysis; 6) assessing air pollution
Nelson, Ferster, et al. (2021)	Bicycle	Narrative review (Independent review paper)	21	Applications of crowdsourced data in bicycling research and practice [a]	• Crowdsourced data have been applied in three applications: 1) mapping ridership; 2) assessing safety; 3) tracking attitudes
Schneider et al. (2021)	Pedestrian	Narrative review (Part of a paper)	14	Direct demand models and independent variables for estimating volumes	• The study summarized the methods and independent variables that have been applied in direct demand models to estimate pedestrian volume
Singleton et al. (2021)	Pedestrian	Narrative review (Part of a paper)	32	Direct demand models and independent variables for estimating volumes	• Many direct demand models for pedestrian volume rely on manually collected, short-duration counts
Yasmin et al. (2021)	Pedestrian and bicycle	Narrative review (Part of a paper)	30	Method and independent variables for estimating volumes	• Segments and intersections are most used analysis units • Majority of the temporal units used are daily and hourly
Bhowmick et al. (2022)	Bicycle	Systemic review (Independent review paper)	41	Methods for estimating link-level volumes	• Direct demand modeling is the mostly applied approach • Studies vary significantly in reporting of the results

Note:
[a] Crowdsourced data in general, including data such as fitness application data, Open Street Map data, and social media data.

4 Crowdsourced mobile data sources

Research question 1 in this study explores the various types of crowdsourced mobile data employed for estimating pedestrian and bicycle volume. General crowdsourced data, as defined by Estellés-Arolas and González-Ladrón-De-Guevara (2012), encompasses contributions from the general public and can include a diverse array of information, including pedestrian and bicyclist volume from Strava, network infrastructure from Open Street Map, and sentiment and opinion from Twitter (Nelson, Ferster, et al., 2021). This study narrows its focus to crowdsourced mobile data, which refers to data collected from a heterogeneous population with varying expertise levels through an open call for voluntary participation, using mobile devices and associated technology. Crowdsourced mobile data discussed in this study are mainly used to generate insights into pedestrian and bicyclist travel patterns, such as counts, origins and destinations, and associated demographics.

Four principal types of crowdsourced mobile data are found in the existing literature: fitness tracking application data, bicycle sharing data, application location-based service data, and cellular signal data (Table 2). These categories are distinguished based on two key criteria: the availability of travel mode information and the mechanism for raw data collection (Lee & Sener, 2020a). Bicycle sharing and fitness tracking application data are generated along with the information of travel mode. Bicycle sharing data is collected through the synergy between smartphone application, station docking system, and potentially GPS (Geographical Positioning System) devices, while fitness tracking data are collected mostly through smartphone applications. Application location-based service and cellular signal data are produced without the travel mode information. Application location-based service data are generated when users engage with location-based services on their smartphones, while cellular signal data is collected through their communication with mobile phone devices.

Table 2. Comparison among different types of crowdsourced mobile data and traditional data

	Advantages	**Disadvantages**
Traditional observed count data	• Representation of the whole population • Accurate count data • Travel modes are known	• Small coverage of geographical area and temporal range • High cost of data collection • Malfunction of facilities
Bicycle sharing data	• Large coverage in temporal range • Low cost of data collection • Travel modes are known	• Only cover urban core area • Biased toward bicycle sharing users • Additional workload of data pre-processing
Fitness tracking application data	• Large coverage in geographical area and temporal range • Low cost of data collection • Travel modes are known	• Biased toward runners and bicyclists • Some data sources need additional workload of data pre-processing
Application location-based service data	• Large coverage in geographical area and temporal range • Low cost of data collection	• Biased toward smartphone users • Additional workload of data pre-processing • Travel modes are unknown
Cellular signal data	• Large coverage in geographical area and temporal range • Low cost of data collection	• Biased toward mobile phone users • Additional workload of data pre-processing • Travel modes are unknown • Low spatial precision

Bicycle sharing data is generated from bicycle sharing programs, including both station-based and dockless ones. Station-based bicycle sharing programs mainly document the starting/ending time and origin/destination information (i.e., rental stations) of the bicycle trips. Dockless bicycle sharing programs and some station-based programs equip their bicycles with GPS sensors and can record the trip trajectories. Compared with traditional observed count data, bicycle sharing data has a wider temporal coverage and a lower cost to collect. Most bicycle sharing programs, however, only cover the urban core areas (Nelson, Ferster, et al., 2021), and not all metropolitan areas or cities operate bicycle sharing programs. Because most bicycle sharing programs only provide the origin/destination data (and not trip trajectories), modelers can only provide measures of demand for the analysis zones where the stations are located, which manifests in high volumes that are not reflected in connecting zones. Although modelers could use shortest-path algorithms or other route-choice models to assign bicycle sharing trips to the network (Kothuri et al., 2022; Proulx, 2016), these approaches introduce error, reducing the validity of the volume estimates. Bicycle sharing trajectories, like trajectories of fitness tracking applications, do not introduce this error. As with measures of bicyclist demand provided by fitness tracking applications, bicycle sharing data only contains trips of bicycle sharing users, which is only a portion of all bicyclists.

Fitness tracking application data refers to the running, walking, and cycling trajectory information gathered from fitness applications on smartphones. The travel modes of these data are known, which sets it apart from other location-based service data. One popular example of these smartphone fitness applications is Strava. While the main function of these fitness applications is to help users record their GPS trajectories during physical activities, the information of these trajectories can be utilized to support pedestrian and bicyclist traffic volume estimation. For example, Strava provides licenses upon which the trajectory data will be aggregated to road segments. The aggregated data include trip numbers by different directions, age groups, and genders. Since 2022, Strava has started to provide the origin and destination data of the trips among standard H3 hexagonal zones (Gamez, 2022). Some fitness applications do not provide aggregated data but GPS trajectories directly, such as Mon ResoVelo (Strauss et al., 2015). In this case, data users need to aggregate the GPS trajectories by themselves which involves heavy work of data processing. Many studies have compared the Strava data with observed count data directly. Lee and Sener (2020b) summarized the correlations between Strava and observed bicyclist counts in the literature and found that the correlations range from 0.36 to 0.83, most of which are larger than 0.6. Compared with traditional count data, fitness application data have a larger geographical and temporal coverage and lower cost of data collection. However, as their users are mostly runners and cyclists, fitness application data bias toward these populations. In addition, among these subpopulations, there is a disproportionate representation of people wealthy enough to purchase the devices on which they can be used.

Application location-based service data are the geographical location data collected from smartphone applications. When people use these applications, their GPS locations, user ID, and the corresponding time will be uploaded to the cloud servers (Nishi et al., 2014). For example, when using Yelp to search for a nearby restaurant, the location and time of the user will be recorded. The data aggregated from multiple smartphone applications can provide valuable support for pedestrian and bicyclist volume estimation. Unlike fitness tracking application data, raw location-based service data from other apps typically does not include information about the mode of transportation used during activities. When dealing with these data, data users need to develop algorithms to identify walking or biking trips from all activities recorded by the applications, which is a challenging task. Compared with traditional observed count data, application location-based service data cover a larger geographical area and a longer temporal range and have a lower cost when collecting data. However, these data are biased toward smartphone users.

Cellular signal data are geographical location data collected from mobile devices when they connect to cellular networks or move across the cell tower boundaries. Cellular service providers record and maintain these data for operation and billing purposes. Similar to application location-based services, analytical algorithms such as statistical regression or machine learning are necessary to detect walking or biking trips from all activities as the travel modes are not available with the raw data. Compared with traditional count data, cellular signal data have a more thorough coverage in terms of geographical area and temporal range. The cost of the corresponding data collection is lower. However, cellular signal data have a lower spatial precision, ranging from 200 to 1000 meters. This limitation makes it difficult to recognize walking or cycling trips from these data, as these trips are usually very short. In addition, cellular signal data is biased toward mobile phone users.

Some companies, such as StreetLight and Cuebiq, purchase data from multiple sources and generate pedestrian and bicyclist trip information for their customers. Although these datasets themselves do not fit the definition of crowdsourced mobile data because they are not provided voluntarily in response to a call, we include them in our discussion for two reasons. First, their generation process utilizes crowdsourced mobile data. For example, StreetLight uses application location-based service data (StreetLight Data Inc, 2023a, 2023b). Second, some of these data are meant to be approximations of or surrogates for segment-level counts of pedestrians and bicyclists in networks (i.e., the data are indicators that potentially may be used in ways analogous to segment-specific estimates that are predicted with direct demand models). For example, StreetLight used to provide StreetLight pedestrian or bicyclist index travelling through traffic analysis zones by a specific time unit (e.g., hour or day). The StreetLight index is a pedestrian or bicyclist trip "count" estimated by its algorithm. It is strictly not an actual count but is strongly correlated with observed counts. Additional trip information includes travel time, length, speed, and circuity. Simple demographics of travelers, such as education and ethnicity, also can be offered. The models used to generate these estimates tend to be proprietary, which means that users do not know exactly how the estimates were produced. This characteristic potentially can limit the ability of analysts to develop custom applications and can create dependencies on particular vendors. One study compared the StreetLight bicyclist index with observed bicyclist counts of 32 locations from six cities in Texas and found that the correlations between these two data are 0.62 and 0.69 on weekdays and weekends respectively (Turner et al., 2020). In the same study, the scholars also compared the StreetLight bicyclist index with bicyclist miles traveled calculated with Strava trip counts. They established a linear regression, in which the StreetLight bicyclist index is the dependent variable and the Strava based bicyclist miles traveled is the independent variable. They found that the correlation between these two variables is 0.94 (Turner et al., 2020). Since May 2022, StreetLight has provided the pedestrian or bicyclist volume instead of the old StreetLight index. StreetLight volume is pedestrian or bicyclist trip counts estimated from multiple data sources and calibrated by observed historical counts (StreetLight Data Inc, 2023a, 2023b). Compared with StreetLight index, StreetLight volume is more comparable to observed counts. A challenge faced by public agencies in using data from sources like StreetLight is that algorithms are prone to change, which can complicate long-term trend monitoring and other comparative analyses.

5 Overview of selected empirical papers

Authors, study locations, travel mode, application, and method are summarized in Table 3. The earliest study was published in 2014. Sixteen studies focused on the context of North America. Twenty studies focused on the travel mode of bicycling. Although the same protocols were used to search for pedestrian studies, only two were identified. The imbalance between bicycle and pedestrian volumes is noteworthy, especially given that more people walk and are exposed to risk compared to those who bicycle. As to the applications of the studies, 18 studies were for traffic volume estimation. The other four studies applied the results of traffic volume estimation for crash analysis and/or health analysis. As to the method, the majority of the studies (19) used direct demand model. One used the data fusion approach to combine data from multiple sources. The remaining three used aggregation analysis and Strava user rate expansion.

Table 3. General information of the selected studies

Author	Study location	Travel mode	Application/Purpose	Method
Nishi et al. (2014)	Japan	Pedestrian	Traffic volume estimation	Aggregation analysis
Strauss et al. (2015)	Montreal, Canada	Bicyclist	Crash analysis	Direct demand model
Jestico et al. (2016)	Victoria, BC, Canada	Bicyclist	Traffic volume estimation	Direct demand model
Proulx (2016)	San Francisco, US	Bicyclist	Traffic volume estimation	Direct demand model
Haworth (2016)	London, UK	Bicyclist	Traffic volume estimation	Direct demand model
Sanders et al. (2017)	Seattle, US	Bicyclist	Traffic volume estimation	Direct demand model
Lißner et al. (2018)	Dresden, Germany	Bicyclist	Traffic volume estimation	Direct demand model
Roll (2018)	Eugene, Springfield, Coburg, OR, US	Bicyclist	Crash analysis and health analysis	Direct demand model
Roy et al. (2019)	Maricopa County, AZ, US	Bicyclist	Traffic volume estimation	Direct demand model
Kwigizile et al. (2019)	Ann Arbor and Grand Rapids, MI, US	Bicyclist	Traffic volume estimation	Direct demand model
Saad et al. (2019)	Orange County, US	Bicyclist	Crash analysis	Direct demand model
Dadashova and Griffin (2020)	TX, US	Bicyclist	Traffic volume estimation	Direct demand model
Dadashova et al. (2020)	TX, US	Bicyclist	Traffic volume estimation	Direct demand model
Lin and Fan (2020)	Charlotte, US	Bicyclist	Traffic volume estimation	Direct demand model
Pogodzinska et al. (2020)	Krakow, Poland	Bicyclist	Traffic volume estimation	Direct demand model
Camacho-Torregrosa et al. (2021)	Spain	Bicyclist	Crash analysis	Strava user rate expansion
Nelson, Roy, et al. (2021)	Boulder, Ottawa, Phoenix, San Francisco, Greater Victoria, North America	Bicyclist	Traffic volume estimation	Direct demand model
Munira (2021)	Austin, US	Bicyclist	Traffic volume estimation	Direct demand model and data fusion approach
Huo et al. (2022)	Nanjing, China	Pedestrian	Traffic volume estimation	Aggregation analysis
Miah, Hyun, Mattingly, Broach, et al. (2022)	Portland, US	Bicyclist	Traffic volume estimation	Direct demand model
Miah, Hyun, Mattingly and Khan (2022)	Portland, US	Bicyclist	Traffic volume estimation	Direct demand model
Kothuri et al. (2022)	Portland, Bend, Eugene, Charlotte, boulder, Dallas, US	Bicyclist	Traffic volume estimation	Direct demand model

6 Methods for estimation

Research question 2 concerns the types of methods used in estimation of pedestrian and bicycle traffic volumes with crowdsourced mobile data. The direct demand model has a wide application in selected studies as shown in Table 3. Direct demand models assume that pedestrian or bicyclist traffic is correlated with several types of variables and that these correlational relationships can be estimated through statistical modeling or machine learning techniques. Analysts then can use the estimated models to predict pedestrian or bicyclist traffic. Direct demand models offer two advantages. First, they are easy to interpret. The relationship between non-motorized traffic and independent variables can be inferred from the coefficients of statistical models or relative importance of machine learning models. Secondly, they are straightforward to implement. Statistical modeling and machine learning can be performed using a wide range of software and packages, making it a flexible and accessible method.

The independent variables used in the selected studies include crowdsourced mobile data (C), built environment (BE), socio-demographics (SD), traffic facility (TF), time (T), weather (W), and other variables (O). Equation (1) below presents the conceptual direct demand model.

$$Traffic = f(C, BE, SD, TF, T, W, O), \tag{1}$$

where, f specifies a modeling function, which could be a statistical model (e.g., ordinary least squared model or generalized linear model) or machine learning model (e.g., random forest).

We present the modeling information of the selected studies with direct demand model in Table 4. All these studies estimated bicyclist volume. The 19 studies vary in their use of crowdsourced mobile data in modeling traffic in terms of analytic units, dependent variables, counts and sources of crowdsourced estimates, sample sizes, specific modeling approach or technique, and measures of evaluation. Specifically, 13 studies used road segment and eight studies used intersection or count location as the analytic unit. Nine studies, which are about half of the selected studies, estimated AADB. The other ten studies estimated daily, hourly, or short-time bicyclist volume. Besides observed bicyclist count, nearly all studies used data from smartphone fitness applications (Strava and Mon ResoVelo). Five studies used bicycle sharing data. Two study also applied the StreetLight bicyclist count (Kothuri et al., 2022), which is similar to the StreetLight bicyclist volume. It is evident that Strava data are dominant in research. There are two reasons for its wide application. Firstly, Strava data offer counts at the segment level. Since many scholars use road segments as their analysis units, they can directly incorporate segment-level Strava data into their research. Secondly, Strava provides free access to scholars and practitioners working in active travel planning, such as pedestrians and bicyclists (Strava, 2020). However, other data sources such as StreetLight are often not available to academic scholars and come at a cost.

Table 4. Information of selected studies with direct demand model

Author (Year)	**Analysis unit**	**Dependent Variable**	**Main data**	**Sample size**	**Estimation model**	**Evaluation** [a]
Strauss et al. (2015) [b]	Intersection Road segment	AADB	Observed bicyclist count Mon ResoVelo GPS trip data	Signalized intersection model (638)	Linear regression	R^2: 0.7
				Non-signalized intersection model (438)		R^2: 0.58
				Cycle track model (70)		R^2: 0.52
				Bicycle lane model (14)		R^2: 0.76
				No facility model (36)		R^2: 0.48
Jestico et al. (2016)	Road segment	Daily bicyclist volume (7-9 am and 3-6 pm combined)	Observed bicyclist count Strava bicycle trip count	612	Poisson regression	NA [c]
Proulx (2016)	Road segment	Peak hour bicycle volume (4-7 pm weekday)	Observed bicyclist count Strava bicycle trip count Bicycle sharing trip count	77	Geographically weighted regression	NA

Author (Year)	Analysis unit	Dependent Variable	Main data	Sample size	Estimation model	Evaluation [a]
Haworth (2016)	Road segment	Hourly bicyclist volume	Observed bicyclist count Strava bicycle trip count	4,172	Linear regression	R^2: 0.68
Sanders et al. (2017)	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	46	Poisson regression	R^2: 0.62
Lißner et al. (2018)	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	NA	Linear regression	MAPE: 36% R^2:0.75
Roll (2018)	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	52	Negative binomial regression	R^2: 0.75
Roy et al. (2019)	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	44	Poisson regression	R^2: 0.64
Kwigizile et al. (2019) [d]	Road segment	Hourly bicyclist volume	Observed bicyclist count Strava bicycle trip count	1,520	Negative binomial regression	NA
					Random forest	NA
Saad et al. (2019)	Intersection	Daily bicyclist volume	Observed bicyclist count Strava bicycle trip count	171	Linear regression	R^2: 0.80
Dadashova and Griffin (2020)	Road segment	Daily bicyclist volume	Observed bicyclist count Strava bicycle trip count	8,813	Mixed effect regression	MAPE: 29%
Dadashova et al. (2020)	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	100	Generalized linear regression with log link	R^2: 0.75; MAPE: 29%
Lin and Fan (2020)	Road segment	Short-time period bicyclist volume	Observed bicyclist count Strava bicycle trip count	NA	Linear regression	R^2: 0.61
Pogodzinska et al. (2020)	Count location	Daily bicyclist volume	Observed bicyclist count Bicycle sharing trip count	99	Linear regression	R^2: 0.92
Nelson, Roy, et al. (2021) [e]	Road segment	AADB	Observed bicyclist count Strava bicycle trip count	Boulder model (15)	Poisson regression	NA
	Intersection			Ottawa model (1058)		
	Road segment			Phoenix model (35)		
	Road segment			San Francisco model (53)		
	Road segment			Greater Victoria model (54)		
Munira (2021)	Intersection	AADB	Observed bicyclist count Strava bicycle trip count	NA	Negative binomial regression and weighted voting approach	NA

Author (Year)	**Analysis unit**	**Dependent Variable**	**Main data**	**Sample size**	**Estimation model**	**Evaluation [a]**
			StreetLight bicycle trip count Bicycle sharing trip count			
Miah, Hyun, Mattingly, Broach, et al. (2022)	Count location	Daily bicyclist volume	Observed bicyclist count Strava bicycle trip count	Utilitarian use locations (1463)	Classification and regression tree	NA
				Mixed use locations (957)		NA
Miah, Hyun, Mattingly and Khan (2022)	Count location	Daily bicyclist volume	Observed bicyclist count Strava bicycle trip count Bicycle sharing trip count	6,746	Deep neural network	R^2: 0.82 MAPE: 86%
Kothuri et al. (2022) [f]	Count location	AADB	Observed bicyclist count Strava bicycle trip count StreetLight bicycle trip count Bicycle sharing trip count	Permanent count locations in all cities (60)	Poisson regression	R^2: 0.82 MAPE: 77%
				Permanent count locations in Dallas (23)		R^2: 0.97 MAPE: 38%
				All count locations in all cities (311)		R^2: 0.78 MAPE: 124%
				All count locations in three Oregon cities (227)		R^2: 0.82 MAPE: 107%
				All count locations in Portland (88)		R^2: 0.85 MAPE: 77%
				All count locations in Eugene (76)		R^2: 0.81 MAPE: 72%
				All count locations in Bend (63)		R^2: 0.47 MAPE: 131%
				All count locations in Boulder (39)		R^2: 0.90 MAPE: 144%
				All count locations in Charlotte (14)		R^2: 0.81 MAPE: 180%
				All count locations in Dallas (31)		R^2: 0.96 MAPE: 40%
				All count locations in all cities (311)	Random forest	NA
				All count locations in three Oregon cities (227)		MAPE: 150%
				All count locations in Portland (88)		MAPE: 104%
				All count locations in Eugene (76)		MAPE: 85%

Notes:

[a] Evaluation indices only include R^2 and MAPE (Mean Absolute Percentage Error), which can be compared across models.

[b] Strauss et al. (2015) estimated models for signalized intersections, non-signalized intersections, cycle track, bicycle lane, and road segments with no facility, respectively.

[c] NA indicates the information is not available in the study.
[d] Besides negative binomial regression, Kwigizile et al. (2019) applied several machine learning models, among which random forest has the best fitness to the data.
[e] Nelson, Roy, et al. (2021) estimated models for Boulder, Ottawa, Phoenix, San Francisco, and Greater Victoria, respectively.
[f] Kothuri et al. (2022) estimated Poisson regression models for permanent count locations in all six cities and in Dallas, respectively. They also estimated Poisson regression models for all count locations (including permanent and short-term ones) in all six cities, in three Oregon cities, in Portland, in Eugene, in Boulder, in Charlotte, and in Dallas, respectively. Besides Poisson regression models, they also estimated random forest models for all count locations in all six cities, in three Oregon cities, in Portland, and in Eugene, respectively. The estimation models applied by the selected studies in Table 4 could be categorized into two types: traditional statistical models and advanced machine learning models. Statistical models use specific probability distributions, such as normal distribution, Poisson distribution, and negative binomial distribution, to fit the observed data. Linear regression, Poisson regression, and negative binomial regression are examples of statistical models. Machine learning models use advanced approaches other than probability distributions, such as decision tree, support vector machine, and neural network, to fit the observed data. Random forest, support vector machine, and neural network are all within the family of machine learning models.

Statistical models were more widely applied in the studies. Among the 19 selected studies in Table 4, 17 used various types of statistical models, including linear regression (6), Poisson regression (5), negative binomial regression (3), generalized linear regression with log link (1), geographically weighted regression (1), and mixed effect regression (1). For these studies, the sample sizes, which indicate the number of locations or number of locations by time units, range from 14 to 8,813. Generally, studies that estimated short-time volumes have larger sample sizes. The R^2 of these models ranges from 0.48 to 0.97, meaning that the models can explain about half to more than 90% of the variation in actual, observed counts. The mean absolute percentage error (MAPE) in estimation ranges from 29% to 180%.

Machine learning models were applied in four selected studies and three of them were published in 2022. Note that two studies (Kwigizile et al., 2019; Miah, Hyun, Mattingly, & Khan, 2022) applied multiple machine learning models and selected the one with the best performance and

Table 4 only listed the best model. Kwigizile et al. (2019) compared random forest, K nearest neighbors, regression tree, neural network, and support vector machine, and found that random forest performed best in terms of RMSE (root mean squared error). Miah, Hyun, Mattingly and Khan (2022) estimated models with shallow neural network, deep neural network, random forest, and extreme gradient boosting (XGBoost), and found that deep neural network had the best performance in terms of RMSE, MAPE, and MAE (mean absolute error). The other two studies applied only one machine learning model each, specifically random forest (Kothuri et al., 2022) and classification and regression tree (Miah, Hyun, Mattingly, Broach, et al., 2022). It is worth noting that decision tree based approaches, such as random forest, regression tree, and XGBoost, were used more than other approaches. This might be because tree-based methods usually have a better performance in fitting structured tabular data while other approaches, such as neural networks, are better at fitting unstructured data, including pictures, voices, and texts. The sample sizes of these four studies range from 76 to 6,746. The MAPE ranges from 85% to 150%.

Besides the direct demand model, three other methods were applied in the literature: data fusion, aggregation analysis and Strava user rate expansion. Because they were only applied in three studies, we briefly describe the methods in this review. Munira (2021) applied a data fusion approach called weighted voting to aggregate the bicyclist volumes estimated from multiple data sources. Aggregation analysis was used by Nishi et al. (2014) and Huo et al. (2022) to estimate pedestrian traffic with application location-based service data and cellular signal data, respectively. The general idea is to extract pedestrian activities from these data and aggregate them into defined analysis zones. Strava user rate expansion was proposed by Camacho-Torregrosa et al. (2021) to compute bicyclist traffic. Strava user rate indicates the ratio between Strava bicyclist count and observed bicyclist count. The authors assumed that Strava user rate is consistent for locations sharing similar characteristics. Therefore, they can calculate the bicyclist traffic based on Strava user rate and Strava trip count.

7 The role of crowdsourced variables

Research question 3 concerned the performance of crowdsourced mobile data related variables in producing valid estimates of pedestrian and bicycle volumes. One of the advantages of statistical models is that they provide significance levels of the independent variables through statistical testing. The information of significance levels is very helpful in determining what factors are important to estimating the pedestrian and bicyclist volumes. Correlation indicates the direction of the linear relationship between the dependent and independent variables. Elasticity measures the percentage change in the dependent variable corresponding to a 1% change in the independent variable. After reviewing the 17 studies using statistical models, we summarized the significance levels, correlations and elasticities of crowdsourced mobile data (Table 5).

Table 5. Correlation and elasticity of crowdsourced mobile data in statistical models [a]

Author (Year)	**Crowdsourced mobile variable**	**Elasticity [b]**	**Ranking**
Strauss et al. (2015)			
• Signalized intersection model	Mon ResoVelo bicycle trip count (+)	NA [c]	NA
• Non-signalized intersection model	Mon ResoVelo bicycle trip count (+)		
• Cycle track model	Mon ResoVelo bicycle trip count (+)		
• Bicycle lane model	Mon ResoVelo bicycle trip count (+)		
• No bicycle facility model	Mon ResoVelo bicycle trip count (+)		
Jestico et al. (2016)	Strava bicycle trip count (+)	NA	NA
Proulx (2016)	Strava bicycle trip count (NP)	NA	NA
	Bicycle sharing trip count (NP)		
Haworth (2016)	Strava bicycle trip count (+)	NA	NA
Sanders et al. (2017)	Strava bicycle trip count (+)	NA	NA
Lißner et al. (2018)	Strava bicycle trip count (+)	NA	NA
Roll (2018)	Strava bicycle trip count (NP)	NA	NA
Roy et al. (2019)	Strava bicycle trip count (+)	NA	NA
Kwigizile et al. (2019)	Strava bicycle trip count (+)	0.02%	13/14
Saad et al. (2019)	Strava bicycle trip count (+)	NA	NA
Dadashova and Griffin (2020)	Strava bicycle trip count (+)	NA	NA
Dadashova et al. (2020)	Strava bicycle trip count (+)	0.18%	8/9
Lin and Fan (2020)	Strava bicycle trip count (+)	NA	NA
Pogodzinska et al. (2020)	Bicycle sharing trip count (+)	NA	NA
Nelson, Roy, et al. (2021)			
• Boulder model	Strava bicycle trip count (+)	NA	NA
	Percentage of Strava commute bicycle trips (-)		
• Ottawa model	Strava bicycle trip count (+)		
	Percentage of Strava commute bicycle trips (+)		
• Phoenix model	Strava bicycle trip count (+)		
	Percentage of Strava commute bicycle trips (+)		
• San Francisco model	Strava bicycle trip count (+)		
	Percentage of Strava commute bicycle trips (-)		
• Greater Victoria model	Strava bicycle trip count (+)		
	Percentage of Strava commute bicycle trips (-)		
Munira (2021)	Strava bicycle trip count (NP)	NA	NA
	StreetLight bicycle trip count (NP)		
	Bicycle sharing trip count (NP)		
Kothuri et al. (2022) [d]			
• Permanent count locations in all cities	Strava and StreetLight bicycle trip count (+)	0.26%	NA
• Permanent count locations in Dallas	Strava and StreetLight bicycle trip count (+)	0.36%	NA
• All count locations in all cities	Strava bicycle trip count (+)	0.54%	2/7
	StreetLight bicycle trip count (+)	0.18%	5/7
• All count locations in three Oregon cities	Strava bicycle commute trip count (+)	0.62%	2/10
• All count locations in Portland	Strava bicycle trip count (+)	0.63%	4/9
	StreetLight bicycle trip count (+)	0.21%	7/9
• All count locations in Eugene	Strava bicycle trip count (+)	0.65%	2/5
	StreetLight bicycle trip count (+)	0.18%	4/5
• All count locations in Bend	No significant variable	NA	NA
• All count locations in Boulder	Strava bicycle trip count (+)	0.46%	2/3
• All count locations in Charlotte	No significant variable	NA	NA
• All count locations in Dallas	Strava and StreetLight bicycle trip count (+)	0.39%	2/6

Notes:
[a] + denotes a significant and positive relationship; - denotes a significant and negative relationship; NP denotes no p-value, indicating that the variable was included in the final model, but the study did not provide the corresponding p-value.
[b] Calculation of elasticity is based on the method introduced in Ewing and Cervero (2010). For a log-linear form regression (i.e., the dependent variable is transformed with logarithm function and the independent variable is in its original form), the elasticity is $\beta \times \bar{x}$. For a log-log form regression, the elasticity is β. For a linear-linear form regression, the elasticity is $\beta \times \frac{\bar{x}}{\bar{y}}$. β is the coefficient and $\bar{x}$ is the average value of the corresponding independent variable. $\bar{y}$ is the average value of the dependent variable. Some studies have no elasticities listed because they did not provide the information of $\bar{x}$ and $\bar{y}$ or did not use log-log form in the statistical models.
[c] NA indicates the related information is not available in the study.
[d] All models considered five crowdsourced mobile variables: Strava bicycle trip count, StreetLight bicycle trip count, Strava and StreetLight bicycle trip count, Strava commute bicycle trip count, and Strava non-commute bicycle trip count. All variables of trip counts were transformed with the logarithm function before including in the modeling process. The table did not include the logarithm form to simplify the results and improve the readiness of the table.

In Table 5, all 17 studies applied crowdsourced mobile data as independent variables to help estimate or predict bicyclist traffic volumes. A commonly used form of crowdsourced mobile data is an aggregated bicycle trip count on road segment. For intersections, Strauss et al. (2015) used the bicycle trip count aggregated from linked road segments equipped with different bicycle facilities, and Saad et al. (2019) used total bicyclists entering the intersections from all linked road segments. Kothuri et al. (2022) not only considered independent Strava or StreetLight trip count but also tested the sum of these two data sources in their models. They also transformed these trip count variables with the logarithm function to better incorporate the potential nonlinear relationships. In most cases, at least one crowdsourced bicycle trip count variable is significant and has a positive relationship with observed bicyclist traffic except for two cases. Kothuri et al. (2022) considered three types of crowdsourced trip counts, including Strava, StreetLight, and sum of Strava and StreetLight, in their models with all count locations in three Oregon cities, in Bend, or in Charlotte. However, none of these variables are significant in the models for unknown reasons. One possible explanation is the relatively small sample size but large variance of the observed bicyclist counts. In addition, Proulx (2016), Roll (2018), and Munira (2021) did not provide information of significance level.

Besides the trip count, two studies considered commuting bicycle trip count in their models as the crowdsourced mobile data provide the information about whether the trip is commuting or not. Nelson, Roy, et al. (2021) included percentage of Strava bicycle trips that are commuting trips in their city models and three models showed that this variable is positively correlated with bicyclist traffic. Kothuri et al. (2022) considered Strava commute bicycle trip count and non-commute bicycle trip count in their models. However, only the model with all count locations in three Oregon cities shows there exists a significant and positive relationship between Strava commute bicycle trips and bicyclist traffic volume. The various results of the commuting bicycle trip count variable might be due to its potential high correlation with total crowdsourced bicycle trip count.

Among the 17 papers that reported statistical models, only three reported data from which elasticities could be calculated. Across the models reported in these papers, the elasticities associated with the crowdsourced mobile variables range from 0.02% to 0.65%, with the majority falling within the narrower interval of 0.18% to 0.65%. In comparison to other variables within the same statistical models, the elasticities of crowdsourced mobile variables occupy varying ranks. In two studies (Dadashova et al., 2020; Kwigizile et al., 2019), the elasticities of Strava bicycle trip counts were notably low, registering at 0.02% and 0.18%, respectively. In both studies, these elasticities were ranked second lowest among all variables examined. However, in Kothuri et al. (2022), the elasticities of Strava bicycle trip counts were among the highest-ranked variables.

In Table 6, we present the crowdsourced mobile variables applied in machine learning models and their corresponding relative importance. Although machine learning models cannot provide significance levels, they can generate relative importance of the independent variables to measure their contributions to estimating the dependent variable. Relative importance is the percentage of variance reduction by one independent variable among the total variance reduction by all the independent variables (Molnar, 2020). With that said, its value is between 0 and 100%. A larger value of relative importance indicates a larger contribution of the independent variable. Two studies reported the relative importance of the crowdsourced variables considered in their machine learning models. Miah, Hyun, Mattingly and Khan (2022) reported the relative importance of Strava bicycle trip count in their two machine learning models. In the random forest model, the relative importance is 16%, ranking first among all 45 variables considered in the model. In the XGBoost model, the relative importance is 24% with a ranking of second place. Kothuri et al. (2022) found that most of the crowdsourced mobile variables considered in their random forest are ranked top among all 99 variables, with the relative importance ranging from 2% to 11%.

Table 6. Relative importance of crowdsourced mobile variables in machine learning models

Author (Year)	Crowdsourced mobile variable	Relative importance	Ranking
Kwigizile et al. (2019)	Strava bicycle trip count	NA [a]	NA
Miah, Hyun, Mattingly, Broach, et al. (2022)	Strava bicycle trip count	NA	NA
Miah, Hyun, Mattingly and Khan (2022) [b]	Strava bicycle trip count	Random forest: 16%	1/45
		XGBoost: 24%	2/45
Kothuri et al. (2022) [c]			
• All count locations in all cities	Sum of log of Strava bicycle trip count and log of StreetLight bicycle trip count	10%	1/99
	Log of Strava commute bicycle trip count	9%	2/99
	Strava commute bicycle trip count	8%	3/99
• All count locations in three Oregon cities	Sum of log of Strava bicycle trip count and log of StreetLight bicycle trip count	10%	1/99
	Log of Strava commute bicycle trip count	8.2%	2/99
	Strava commute bicycle trip count	7.5%	3/99
• All count locations in Portland	Sum of log of Strava bicycle trip count and log of StreetLight bicycle trip count	11%	1/99
	Log of Strava commute bicycle trip count	10%	2/99
	Log of Strava bicycle trip count	9%	3/99
• All count locations in Eugene	Log of Strava commute bicycle trip count	7%	3/99
	Strava commute bicycle trip count	6%	4/99
	Sum of log of Strava bicycle trip count and log of StreetLight bicycle trip count	4%	7/99

Notes:
[a] NA indicates the related information is not available in the study.
[b] The authors only provided the relative importance of random forest and XGBoost models while they applied multiple machine learning approaches.
[c] All models considered five crowdsourced mobile variables: Strava bicycle trip count, StreetLight bicycle trip count, Strava and StreetLight bicycle trip count, Strava commute bicycle trip count, and Strava non-commute bicycle trip count. Both the original and logarithm forms of these crowdsourced mobile variables were included in the modeling process. To save space, we only listed the three most important crowdsourced mobile variables.

In addition to the measures of correlation, elasticity, and relative importance used to assess independent variables, several studies found that models with crowdsourced mobile data perform better than those without (i.e., they mostly have better fitness and their predicted volumes are more accurate, see Table 7). Proulx (2016) found that inclusion of Strava data reduces RMSE by 30.5 to 39.4 and inclusion of bicycle sharing data reduces RMSE by 1.2 to 8.5. Sanders et al. (2017) found their Poisson regression model with Strava bicycle trip count variable not only has a better fitness in terms of R^2 but also included fewer independent variables (i.e., a more parsimonious form). Kothuri et al. (2022) had similar findings. Roll (2018) found that their models with Strava bicycle trip count perform better in terms of R^2. Kwigizile et al. (2019) applied both statistical and machine learning models to construct models with and without Strava bicycle trip counts. They found that all models with Strava data have lower RMSE and higher R^2 than those without. Miah, Hyun, Mattingly, Broach, et al. (2022) had similar comparison results with their machine learning models. In addition, Proulx (2016) showed that inclusion of both Strava and bicycle sharing data could further improve the model performance than inclusion of single type of data. However, Kothuri et al. (2022) found that inclusion of bicycle sharing data can hardly improve model performance when the model has already included Strava and StreetLight data.

Table 7. Model performance improved by crowdsourced mobile data

Author (Year)	Model performance improvement by crowdsourced mobile data
Proulx (2016)	• Inclusion of Strava data reduces RMSE by 30.5 to 39.4 • Inclusion of bicycle sharing data reduces RMSE by 1.2 to 8.5 • Inclusion of both Strava and bicycle sharing data improves RMSE by 32.2 to 43.3
Sanders et al. (2017)	• The inclusion of Strava data improves R^2 from 0.57 to 0.62. • At the same time, the model with Strava data variable include fewer independent variables.
Roll (2018)	• Inclusion of Strava data improves R^2 from 0.68 to 0.77
Kwigizile et al. (2019)	Negative binomial regression: • Inclusion of Strava data reduces RMSE by 0.97 and increases R^2 by 0.08. Machine learning: • Inclusion of Strava data reduces RMSE by 0.66 to 0.87 and increases R^2 by 0.05 to 0.08.
Miah, Hyun, Mattingly, Broach, et al. (2022)	Inclusion of Strava data reduces median absolute percentage error • by 9 to 80 percentage points for utilitarian use locations • by 1 to 12 percentage points for mixed usage locations
Kothuri et al. (2022)	Poisson regression: • Inclusion of Strava data improves R^2 by 0.003 to 0.2 • Inclusion of StreetLight improves R^2 by 0.01 to 0.18 • Inclusion of both Strava and StreetLight improves R^2 by 0.04 to 0.24 Random forest: • Inclusion of bicycle sharing data did not improve performance when there exist Strava and StreetLight data

8 Discussion and conclusion

In this study, we applied a systematic literature review approach to explore the types of crowdsourced mobile data and methods that have been used to estimate pedestrian and bicyclist traffic volume. After searching the scholarly databases with the keywords related to our topics and screening the search results with several criteria in two rounds, we summarized 10 related literature reviews and included 22 empirical studies in our review. We then extracted key information related to our topics from the selected studies.

With respect to our first research question (i.e., types of crowdsourced mobile data used in traffic volume estimation), we identified four distinct types of crowdsourced mobile data: bicycle sharing data, fitness tracking application data, application location-based service data, and cellular signal data.

Compared with traditional observed count data, crowdsourced mobile data have larger coverage of geospatial and temporal range and lower cost data collection. However, they are biased toward a certain population based on their mobile data sources. For example, fitness tacking application data are biased toward runners and cyclists as they are collected from smartphone fitness applications such as Strava and Mon ResoVelo. In addition, cellular signal data have relatively lower spatial precision. Strava data have been used more than other data sources in the studies analyzed in this review. An important finding is that the vast majority of studies using crowdsourced data have focused on bicycle volumes and not pedestrian volumes.

Researchers have applied four different methods to use crowdsourced mobile data to estimate pedestrian and bicyclist volumes, including direct demand modeling, data fusion, aggregation analysis, and Strava user ratio expansion. The direct demand modeling approach dominates among studies identified for this review. All studies with direct demand modeling estimated bicyclist traffic volume. Statistical models are more frequently used to construct the relationships between various factors and bicyclist traffic volumes while machine learning models have been applied in recent years.

Our review results showed that crowdsourced mobile data related variables not only have significant correlations with observed bicyclist count in statistical models but also have relatively larger relative importance than other independent variables in machine learning models. Furthermore, the inclusion of crowdsourced mobile data improves the model prediction performance in both statistical models and machine learning models. These results confirm the important role of crowdsourced mobile data in the estimation of bicyclist volumes. However, we also found that a majority of papers have not reported or included in their papers the data needed for estimation of key metrics such as elasticities. Overall, researchers and practitioners alike will be better served with better information about the independent variables in models that have the greatest effects on observed volumes.

Based on the review results, we identified four research topics related to the application of crowdsourced mobile data in the estimation of pedestrian and bicycle traffic that deserve more attention in future studies.

8.1 Increase crowdsourced mobile data use in estimating pedestrian traffic

One finding from this review is that most of the selected studies (20) used crowdsourced mobile data to estimate bicyclist traffic. For studies with direct demand modeling, all of them estimated bicyclist traffic. A research gap is that very few studies have incorporated crowdsourced mobile data in pedestrian traffic volume estimation. Many crowdsourced mobile data provide pedestrian-related information. For example, Strava provides trip counts related to walking, running, and hiking, and StreetLight provides estimated pedestrian volumes. Similar to bicycle trip counts generated from crowdsourced mobile data, these pedestrian trip counts share the advantages in terms of large spatial and temporal coverage and low collection cost, and the disadvantages of biased population characteristics. However, due to the different patterns between pedestrian and bicyclist trips, pedestrian trip counts generated by crowdsourced mobile data may perform differently in traffic estimation models with crowdsourced bicyclist trip counts. For example, how would crowdsourced mobile data related variables perform in the estimation of pedestrian traffic? How much improvement can crowdsourced mobile data bring to pedestrian estimation models? Addressing these questions and similar ones could help improve the current works in estimating pedestrian traffic and provides better support for works such as pedestrian facility planning and pedestrian crash analysis.

8.2 Compare crowdsourced data performance in non-motorized traffic estimation

Crowdsourced mobile data vary in spatial coverage, temporal patterns, and accuracy, potentially affecting their performance in estimating pedestrian and bicyclist traffic across different modes, contexts, and regions. Most of the studies, however, only apply the Strava data in estimating pedestrian and bicyclist estimation. While Strava has advantage such as providing free access to scholars and offering segment-level counts, other crowdsourced mobile data can also provide valuable contributions as they have different user demographics, geographic coverage, and data granularity. For instance, StreetLight and Cuebiq data combine multiple data sources. Therefore, they can capture a more diverse user base and offer a more representative sample for pedestrian and bicyclist estimation than Strava. Their performance could be different in terms of travel mode, local context, regional text, and more. How do these crowdsourced mobile data perform differently in estimating pedestrian and bicyclist traffic? How do they perform differently in locations with different land-use patterns (e.g., commercial area, residential area, and open space)? How do they perform differently in different cities or regions? Among the reviewed studies, very few of them compare the performance of different types of crowdsourced mobile data in the estimation of non-motorized traffic. Exceptions include Proulx (2016) and Kothuri et al. (2022), both of whom compared multiple sources (e.g., bicycle sharing, Strava, and StreetLight) of data in estimating bicyclist traffic. More studies are still needed in more areas with different contexts to test the generalizability of their findings. Local transportation agencies may have access to multiple crowdsourced mobile data sources or are in the process to purchase/request additional data sources. Understanding the heterogeneity of the crowdsourced mobile data in estimating pedestrian and bicyclist traffic could help local transportation agencies in deciding which data sources they need most or how they can achieve better performance with less investment in purchasing the licenses.

8.3 Incorporate multiple types of crowdsourced data to further improve model performance

Leveraging multiple types of crowdsourced mobile data potentially could yield more nuanced and comprehensive results compared with using just one data type in estimating pedestrian and bicyclist traffic (Proulx, 2016). When different data categories are combined, the resulting data offers a richer, multi-dimensional view of user behavior and travel patterns. However, this approach cannot guarantee model performance improvement. Some crowdsourced mobile data have a larger coverage in terms of time, space, and population than others. For example, both Strava and StreetLight data have a larger area coverage than bicycle sharing data. Inclusion of bicycle sharing data to models with Strava and StreetLight data may result in limited improvement of model performance (Kothuri et al., 2022). In addition, the marginal improvement in model performance needs to be assessed against the marginal costs of model refinement and improvement. We hypothesize that, as more types of data are integrated into a demand model, the marginal improvements from each new data source may diminish. There clearly remains a need to explore how different combinations of crowdsourced mobile data improve model performance and whether this improvement could generate benefits that exceed the associated costs. Such analyses would provide valuable guidance for researchers, policymakers, and industry stakeholders in making informed decisions about the optimal use of crowdsourced mobile data for estimating pedestrian and bicyclist traffic volumes.

8.4 Expand methods using crowdsourced data for non-motorized traffic estimation

Crowdsourced mobile data could be applied with other popular methods to estimate pedestrian and bicycle traffic, including four-step models and simulation-based models. In four-step models, the first two steps are estimating the number of trips generated and determining where these trips are likely to go. Traditionally, household travel survey data are used to carry out these two steps. However, household surveys are expensive to collect and can only provide reliable estimates at the census block group or tract levels. Crowdsourced mobile data may outperform household travel surveys in certain respects. For example, the related collection cost for crowdsourced mobile data is low and the data provides granular estimates in terms of trip counts and origin and destination (OD) information in refined zones. Agent-based models use agents with specific behaviors, preferences, and decision-making rules to simulate walking and cycling trips in the real world. Crowdsourced mobile can provide frequency and duration of walking and cycling trips, the speed and distance traveled, and the types of routes preferred by cyclists and pedestrians for calibrating the behavior of individual agents in the model, such as their travel patterns and route selections.

References

Bhowmick, D., Saberi, M., Stevenson, M., Thompson, J., Winters, M., Nelson, T., … & Beck, B. (2022). A systematic scoping review of methods for estimating link-level bicycling volumes. *Transport Reviews*, 43(4), 622–651. https://doi.org/10.1080/01441647.2022.2147240

Camacho-Torregrosa, F. J., Llopis-Castelló, D., López-Maldonado, G., & García, A. (2021). An examination of the Strava usage rate—A parameter to estimate average annual daily bicycle volumes on rural roadways. *Safety*, *7*(1), 8. https://doi.org/10.3390/safety7010008

Dadashova, B., & Griffin, G. (2020). Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. *Transportation Research Part D: Transport and Environment*, *84*, 102368. https://doi.org/10.1016/j.trd.2020.102368

Dadashova, B., Griffin, G., Das, S., Turner, S., & Sherman, B. (2020). Estimation of average annual daily bicycle counts using crowdsourced Strava data. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*(11), 390–402. https://doi.org/10.1177/0361198120946016

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. https://doi.org/10.1177/0165551512437638

Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

FHWA. (2016). *The traffic monitoring guide*. Retrieved from https://www.fhwa.dot.gov/policyinformation/tmguide/tmg_fhwa_pl_17_003.pdf

FHWA. (2018). *Summary of Travel Trends: 2017 National Household Travel Survey*. Washington, DC: FHWA.

Gamez, C. (2022). *Origins and destinations data export and download*. Retrieved from https://stravametro.zendesk.com/hc/en-us/articles/8187514314263

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Haworth, J. (2016). Investigating the potential of activity tracking app data to estimate cycle flows in urban areas. *ISPRS — International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B2*, 515–519. https://doi.org/10.5194/isprsarchives-XLI-B2-515-2016

Huo, J., Fu, X., Liu, Z., & Zhang, Q. (2022). Short-term estimation and prediction of pedestrian density in urban hot spots based on mobile phone data. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10827–10838. https://doi.org/10.1109/tits.2021.3096274

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

Kothuri, S., Broach, J., McNeil, N., Hyun, K., Mattingly, S., Miah, M. M., … & Proulx, F. (2022). *Exploring data fusion techniques to estimate network-wide bicycle volumes*. Retrieved from https://pdxscholar.library.pdx.edu/trec_reports/234/

Kwigizile, V., Oh, J.-S., & Kwayu, K. (2019). *Integrating crowdsourced data with traditionally collected data to enhance estimation of bicycle exposure measure*. Retrieved from https://rosap.ntl.bts.gov/view/dot/44138

Lee, K., & Sener, I. (2020a). Emerging data for pedestrian and bicycle monitoring: Sources and applications. *Transportation Research Interdisciplinary Perspectives*, *4*, 100095. https://doi.org/10.1016/j.trip.2020.100095

Lee, K., & Sener, I. (2020b). Strava metro data for bicycle monitoring: A literature review. *Transport Reviews*, *41*(1), 27–47. https://doi.org/10.1080/01441647.2020.1798558

Lin, Z., & Fan, W. D. (2020). Modeling bicycle volume using crowdsourced data from Strava smartphone application. *International Journal of Transportation Science and Technology*, *9*(4), 334–343. https://doi.org/10.1016/j.ijtst.2020.03.003

Lißner, S., Francke, A., & Becker, T. (2018). Modeling cyclists traffic volume — Can bicycle planning benefit from smartphone-based data. Paper presented at the 7th Transport Research Arena TRA, April 16–19, Vienna, Austria.

Liu, F., Evans, J. E. J., & Rossi, T. (2012). Recent practices in regional modeling of nonmotorized travel. *Transportation Research Record: Journal of the Transportation Research Board*, *2303*(1), 1–8. https://doi.org/10.3141/2303-01

Miah, M. M., Hyun, K. K., Mattingly, S. P., Broach, J., McNeil, N., & Kothuri, S. (2022). Challenges and opportunities of emerging data sources to estimate network-wide bike counts. *Journal of Transportation Engineering, Part A: Systems*. https://doi.org/10.1061/jtepbs.0000634

Miah, M. M., Hyun, K. K., Mattingly, S. P., & Khan, H. (2022). Estimation of daily bicycle traffic using machine and deep learning techniques. *Transportation*, *50*, 1631–1684. https://doi.org/10.1007/s11116-022-10290-z

Molnar, C. (2020). *Interpretable machine learning — A guide for making black box models explainable*. lulu.com. Retrieved from https://christophm.github.io/interpretable-ml-book/

Munira, S. (2021). *Fusing nonmotorized traffic data: A decision fusion framework.* College Station, TX: Texas A&M University.

Munira, S., & Sener, I. N. (2017). *Use of direct-demand modeling in estimating nonmotorized activity: A meta-analysis*. College Station, TX: Texas A&M University.

Nelson, T., Ferster, C., Laberee, K., Fuller, D., & Winters, M. (2021). Crowdsourced data for bicycling research and practice. *Transport Reviews*, *41*(1), 97–114. https://doi.org/10.1080/01441647.2020.1806943

Nelson, T., Roy, A., Ferster, C., Fischer, J., Brum-Bastos, V., Laberee, K., … & Winters, M. (2021). Generalized model for mapping bicycle ridership with crowdsourced data. *Transportation Research Part C: Emerging Technologies*, *125*, 102981. https://doi.org/10.1016/j.trc.2021.102981

Nishi, K., Tsubouchi, K., & Shimosaka, M. (2014). Hourly pedestrian population trends estimation using location data from smartphones dealing with temporal and spatial sparsity. Paper presented at the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Nov. 1–4, Seattle, WA.

Ohlms, P. B., Dougald, L. E., & Macknight, H. E. (2019). Bicycle and pedestrian count programs: Scan of current U.S. practice. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(3), 74–85. https://doi.org/10.1177/0361198119834924

Pogodzinska, S., Kiec, M., & D'Agostino, C. (2020). Bicycle traffic volume estimation based on GPS data. *Transportation Research Procedia*, *45*, 874–881. https://doi.org/10.1016/j.trpro.2020.02.081

Proulx, F. (2016). *Bicyclist exposure estimation using heterogeneous demand data sources.* Berekely, CA: University of California.

Roll, J. (2018). *Bicycle count data: What is it good for? A study of bicycle travel activity in central lane metropolitan planning organization*. Retrieved from https://rosap.ntl.bts.gov/view/dot/36255

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Science*, *3*(2), 62. https://doi.org/10.3390/urbansci3020062

Saad, M., Abdel-Aty, M., Lee, J., & Cai, Q. (2019). Bicycle safety analysis at intersections from crowdsourced data. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(4), 1–14. https://doi.org/10.1177/0361198119836764

Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington. *Transportation Research Record: Journal of the Transportation Research Board*, *2605*(1), 32–44. https://doi.org/10.3141/2605-03

Schneider, R. J., Schmitz, A., & Qin, X. (2021). Development and validation of a seven-county regional pedestrian volume model. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(6), 352–368. https://doi.org/10.1177/0361198121992360

Singleton, P. A., Park, K., & Lee, D. H. (2021). Varying influences of the built environment on daily and hourly pedestrian crossing volumes at signalized intersections estimated from traffic signal controller event data. *Journal of Transport Geography*, *93*, 103067. https://doi.org/10.1016/j.jtrangeo.2021.103067

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

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

Strava. (2020). *Strava announces Strava Metro, the largest active travel dataset on the planet, is now free and available to cities everywhere*. Retrieved from https://blog.strava.com/press/metro/

StreetLight Data Inc. (2023a). *StreetLight bicycle volume methodology and validation white paper*. https://www.streetlightdata.com/whitepapers/

StreetLight Data Inc. (2023b). *StreetLight pedestrian volume methodology and validation white paper*. https://www.streetlightdata.com/whitepapers/

Turner, S., Hottenstein, A., & Shunk, G. (1997). *Bicycle and pedestrian travel demand forecasting: Literature review*. Austin, TX: Texas Department of Transportation. https://static.tti.tamu.edu/tti.tamu.edu/documents/1723-1.pdf

Turner, S., Martin, M., Griffin, G., Le, M., Das, S., Wang, R., … & Li, X. (2020). *Exploring crowdsourced monitoring data for safety*. Retrieved from https://safed.vtti.vt.edu/projects/exploring-crowdsourced-monitoring-data-for-safety/

Turner, S., Sener, I. N., Martin, M. E., Das, S., Hampshire, R. C., Fitzpatrick, K., … & Wijesundera, R. K. (2017). *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities*. Washington, DC: Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/36098

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research*, *39*(1), 93–112. https://doi.org/10.1177/0739456x17723971

Yasmin, S., Bhowmik, T., Rahman, M., & Eluru, N. (2021). Enhancing non-motorist safety by simulating trip exposure using a transportation planning approach. *Accident Analysis & Prevention, 156*, 106128. https://doi.org/10.1016/j.aap.2021.106128