Transport Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttrv20

Strava Metro data for bicycle monitoring: a literature review

Kyuhyun Lee & Ipek Nese Sener

TRANSPORT REVIEWS
https://doi.org/10.1080/01441647.2020.1798558

Routledge
Taylor & Francis Group

Check for updates

Strava Metro data for bicycle monitoring: a literature review

Kyuhyun Lee and Ipek Nese Sener

Texas A&M Transportation Institute, College Station, TX, USA

ABSTRACT
Monitoring bicycle trips is no longer limited to traditional sources, such as travel surveys and counts. Strava, a popular fitness tracker, continuously collects human movement trajectories, and its commercial data service, Strava Metro, has enriched bicycle research opportunities over the last five years. Accrued knowledge from colleagues who have already utilised Strava Metro data can be valuable for those seeking expanded monitoring options. To convey such knowledge, this paper synthesises a data overview, extensive literature review on how the data have been applied to deal with drivers' bicycle-related issues, and implications for future work. The review results indicate that Strava Metro data have the potential—although finite—to be used to identify various travel patterns, estimate travel demand, analyse route choice, control for exposure in crash models, and assess air pollution exposure. However, several challenges, such as the under-representativeness of the general population, bias towards and away from certain groups, and lack of demographic and trip details at the individual level, prevent researchers from depending entirely on the new data source. Cross-use with other sources and validation of reliability with official data could enhance the potentiality.

ARTICLE HISTORY
Received 1 July 2019
Accepted 14 July 2020

KEYWORDS
Strava; bicycle; crowdsourced data; fitness tracking application; emerging travel data

Introduction

Monitoring active modes of travel (i.e. bicycling and walking) has largely relied on traditional sources such as travel surveys and traffic counts. With the penetration of mobile devices (e.g. smartphones, wearable watches, tablets) and the big data revolution, acquiring active travel information is no longer limited to traditional methods (Lee, Sener, & Mullins, 2016; Lee & Sener, 2020). Among various fitness tracking apps running on global positioning system (GPS)–enabled devices, Strava has continuously collected human movement records since 2009 and launched its commercial data service, Strava Metro, in 2014 (Strava Metro, 2019). Since then, the data have enriched bicycle research opportunities.

Accrued knowledge from colleagues who have already utilised Strava Metro data can provide valuable guidance, especially for cities and agencies seeking expanded data options. This paper aims to be a resource document by reviewing how Strava Metro data have been used so far in support of cycling transport planners and practitioners, how usefulness varies by data characteristics and study questions, and which

CONTACT Ipek Nese Sener i-sener@tti.tamu.edu Texas A&M Transportation Institute, Austin, TX 78752, USA

challenges should be considered to achieve better planning and decision-making outcomes.

To identify relevant literature for this review, four scholarly search engines were selected: Google Scholar, Web of Science, Transport Research International Documentation, and Research Gate. Every post-2009 source that included the term "Strava" was searched. If the resulting record had an abstract that seemed relevant to active travel, researchers checked whether the study used Strava Metro data and the level of significance. When a study had several versions of papers (e.g. a transportation agency's technical reports), published papers or papers having greater significance were selected representatively to avoid overlap. Second, solely health-oriented studies (e.g. examination of heart rates) were beyond the scope of the current review. Third, the current review included only peer-reviewed journal articles, conference proceedings, and technical reports written in English (except for a study by Proulx and Pozdnukhov [2017] that has seminal implications). The searching process resulted in 42 papers. Due to a paucity of literature on pedestrians, the current review focused on bicycling.

This paper starts with an overview of crowdsourced data and Strava Metro data. Next, the current applications of Strava Metro data are extensively discussed, followed by implications for future work. The paper ends with concluding remarks.

Data overview

Crowdsourced data

While still being widely reliant on traditional monitoring methods, the active transportation field began to take advantage of human movement records that smart mobile devices automatically and passively collect. According to the most recent study that extensively reviewed the emerging data sources via mobile devices (crowdsourced data) and their current status of applications in the active travel context (Lee & Sener, 2020), the emerging data sources can be broken down into seven broad categories: mobile phone positioning (MPP); location-based service (LBS); WiFi/Bluetooth; regional bicycle tracking system; fitness tracking app; bike-share programme; and user-feedback inventory. Since their fundamental data collection mechanisms differ, these data sources have different data attributes, potentials, challenges, and adoption rates. Among the seven categories, from a perspective of bicycle monitoring, regional bicycle tracking apps guarantee a generally higher level of data quality, accuracy, and trip/traveller details than the others, but only if time and resources allow for developing the app, recruiting enough users, and handling heavy computational work for massive GPS points. Otherwise, the trip information collected from fitness tracking apps is perhaps an optimal option (but not all the apps offer user-friendly and ready-made databases). Although bike-share programmes and user-feedback inventory are not considered a part of mainstream monitoring methods, they provide valuable insights into, for example, shared bike-use patterns and community needs detection, such as fixing potholes. MPP, LBS, and WiFi/Bluetooth are comparably less feasible for extensive application, mostly due to difficulties in detecting bicyclists from other modes and coarse spatial/temporal sampling resolution (for more detailed information, refer to Lee and Sener [2020]).

As noted by Lee and Sener (2020), there are various vendors providing commercial data service, but "their products have been mainly used for motorised trips or are in initial stages of development" for pedestrian and bicycle monitoring. So far, fitness tracking apps have had the most widespread practical applications for active travel. Among these apps, Strava—through Strava Metro—has vigorously promoted more research endeavours over the last five years, and accordingly, the current review provides a timely analysis.[1] The plausible reason for the popularity of Strava Metro-offering data is its readiness and preparedness in a format that is useable for spatial and statistical analysis, with a rich coverage in time and space yet relatively reasonable price. Likewise, globally dispersed users have resulted in growing interest in use of the data around the world.

Strava metro data

Strava is a fitness tracking platform that supports physical activities by recording GPS trajectories (X and Y coordinates of points passed and the time). Strava Metro cleans the collected GPS trajectories and aggregates to three geometric units—street segments (edges), intersections (nodes), and polygons of trip origin and destination (OD)—to sell them again for public use (i.e. research purposes and transportation planning). Data products consist of spatial files and attribute tables containing information on traveller/trip count and trip/waiting time at the geometries (waiting times are only provided for nodes) for various time frames (e.g. hour, year, weekday, and weekend) by generalised trip purposes (commuting and non-commuting).

Probably the main feature of Strava that has inspired diverse transportation schemes is that it can be—simply and conceptually—defined as a continuous counting system covering the whole region of interest. While the typical counting system is likely to be set at a small number of sites or implemented during a short period, Strava monitoring is almost gapless over both time and space as long as Strava users exist. However, the uncertainty of the representativeness of the general population and innate sampling bias diminish data reliability. Moreover, in the interest of users' privacy, individual demographics and details about discrete trips (i.e. each route from origin to destination) are not provided to third-party data users. Instead, summarised demographic characteristics (e.g. traveller counts by age group/gender) and trip statistics (e.g. average trip time/distance) are offered at the contracted area level.

In general, a complex data-mining process is necessary to systematically convert collected GPS points to a user-friendly format. Customers of Strava Metro can bypass such efforts, but at the same time, already processed datasets bring limitations to the capability of controlling potential errors and omissions. The processed data can be viewed in interactive maps (DataView) without special techniques, but to fully exploit the datasets for advanced and in-depth analyses, geographic information system (GIS) software and skills are necessary. Even with such skills, handling the big data may entail tedious and time-consuming work because file structures and formats may not be very compatible with desired methodologies and existing datasets.

Licence fees are subject to change based on the size of the requested area, the time span of the data needed, and the level of granularity and features in the dataset (Strava Metro, 2019). According to Ohlms, Dougald, and MacKnight (2018), the estimated cost for Virginia was $300,000 for one year (2.5 million activities in 2016 by 110,000 users).

The Oregon Department of Transportation paid $20,000 for one year of bicycle data (June 2013 through May 2014; Chen, Wang, Roll, Nordback, & Wang, 2020). Since budgets might not support this level of expenditure, the cost of obtaining and utilising Strava Metro data should be compared to other monitoring methods based on which research schemes are of interest and how the new data source can support realising the desired scenarios, as discussed in the following section.

Applications

This section synthesises how Strava Metro data have been applied in the existing bicycle literature and proposes current suitability for the corresponding application and open challenges to be considered.

Travel pattern identification

The most fundamental application promoted by the availability of Strava records is to identify and characterise cycling patterns that are unlikely to have been observed through traditional data sources. The insufficient temporal and spatial coverage of traditional monitoring has often hampered capturing a fuller picture of cycling patterns across the network. Once shapefiles (segments, intersections, OD polygons) are joined with attribute files, they can be used to visualise and illustrate bicycle trip patterns. With fine granularity, it is possible to examine from an areawide overview to the directional bicycle flows at a small scale of facilities (e.g. street). For instance, Selala and Musakwa (2016) depicted a year of cycling activity density for the Johannesburg region of South Africa for the first time. Total trip volumes on segments for a certain period (e.g. a year or a week) can show heavily used links overall across networks (Griffin & Jiao, 2015b). When trip identification (i.e. commuting or non-commuting) is applied, researchers can observe how recreational riding shows different patterns from commuting (Lee & Sener, 2019). Seasonal and daytime variations can be monitored as well (Fan & Lin, 2019; Jestico, Nelson, & Winters, 2016). Beyond purely analysing original data, when integrated with other layers of information, Strava-generated data can contribute to the understanding of cycling behaviours. For example, by integrating bicycle trip volumes with official geographical units (i.e. census block groups), researchers determined which sociodemographic/built environment factors were associated with the different levels of bicycle activity density (Griffin & Jiao, 2015b; Hochmair, Bardin, & Ahmouda, 2019). In terms of going beyond identifying Strava bicycle characteristics or examining Strava-oriented bicycle patterns, more targeted planning strategies can be achieved, as discussed next.

Travel demand estimation

Bicycle volumes are rudimentary measure for understanding existing travel conditions and predicting ridership in a range of planning scenarios. Since observing every rider is practically impossible, network-wide ridership often must be determined from modelled volumes. Making accurate predictions is a shared challenge, which leads to continuous investigation of more powerful variables to include in travel demand estimation. In this

Table 1. Application overview in travel demand estimation.

Reference	Summary description
Dadashova and Griffin (2020)	Estimated mixed-effects models with autocorrelated errors of daily bicycle volume at 34 count stations.
Jestico et al. (2016)	Predicted categorised daily cycling volumes (low, medium, high) across a network from Poisson regression models that estimated daily cycling volumes at 18 manual count locations.
Proulx and Pozdnukhov (2017)	Predicted link-level bicycle flows across a network from geographically weighted data fusion models of average hourly cycle volumes collected during the peak period (4–7 *pm*) at 536 directional links.
Roll (2018)	Estimated negative binomial models of annual average bicycle travel demand at 52 count locations.
Roy et al. (2019)	Predicted bias-corrected AADB counts for all streets from Poisson regression models that estimated AADB volumes at 60 manual count locations.

respect, Strava data may become a valuable input in modelling bicycle travel demand and help improve the model accuracy, as concluded by the reviewed studies (Table 1).

Roll (2018) developed negative binomial models for average annual daily bicycle (AADB) traffic in the Central Lane Metropolitan Planning Organization in Oregon. Among four different combinations of variables (Strava rider counts, street and bicycle facility classification, accessibility to amenities, network density), model specifications including the Strava variable produced the highest pseudo R-squared (0.75 and 0.77) compared to those without Strava (from 0.59–0.68). In a study by Dadashova and Griffin (2020) that accounted for daily bicycle flows at count stations in Texas, Strava user counts had the greatest impacts on daily bicycle counts. Jestico et al. (2016) predicted cycling volumes across the entire network in Victoria, Canada, utilising Poisson regression outcomes, which estimated daily cycling volumes by incorporating peak-period Strava bicycle volume (across 7–9 am and 3–6 pm) and four other variables (slope, speed, on-street parking, time of year). The overall average model error (with 100 times 90–10 cross-validation split) was 38%, and the authors concluded that Strava could bring added value to modelling bicycle traffic flows.

Alongside efforts to estimate travel demand, some researchers have proposed a new framework to mitigate sample bias in the fitness-app-generated data. Roy, Nelson, Fotheringham, and Winters (2019) examined geographical covariates significant for correcting bias by applying a variable selection technique, including household income, distance to residential areas/green spaces, and traffic speed. Then they used the Poisson regression model to predict all bicyclists across street segments in Tempe, Arizona. When the prediction accuracy was verified by observed counts at 60 locations, for 86% of segments, predicted AADB counts were correct to within a margin of ±100. Proulx and Pozdnukhov (2017) laid out a novel approach fusing various datasets (manual/automated counts, Strava Metro, bike-share programme usage, two regional travel demand model outcomes) to predict network-wide bicycle flows in San Francisco, California. The data fusion approach began with applying a criterion value with a weight of 1 to average hourly cycle counts at 536 directional links. Then the weighting matrix for the other four datasets against the weight of 1 was determined based upon distance between observations and link similarity (e.g. bicycle facility similarity). Resting on the determined weighting matrix, the authors estimated link-level bicycle flows and assessed model performance for all possible combinations of four datasets. Predictive accuracy only improved significantly when the Strava data were included.

Although the results of the reviewed studies demonstrate promising applications, it is arguably unclear whether Strava data are feasible for regional travel forecasting models because the current data structure does not offer complete tour information including trip start/end and route taken at an individual level—a limitation not specific to travel demand forecasting practice but likely to disrupt other applications.

Route choice analysis

Until GPS trackers are prevalent, understanding why people take certain routes over a set of route choices has exclusively relied on stated preference surveys. Given that real-world decisions may differ from stated choices, ground-truth pathways of Strava cyclists are appealing to transportation system modellers. Certainly, real footprints of Strava can help understand bicycle trip behaviours, but they are unlikely to be one size fits all to modellers. Several studies investigated route choices of Strava cyclists and variables related to the cycle behaviours (see Table 2).

Sun, Du, Wang, and Zhuang (2017) explored how environmental characteristics influenced hourly aggregated recreational bicycling on weekdays in 2015 for 119 streets in Glasgow, United Kingdom. Strava cyclists tended to favour streets with a short length, low vehicle traffic flow, better connectivity, or surrounded by residential land use. Orellana and Guerrero (2019) examined the effect of street network structure on variations in total cyclists, weekday cyclists, and weekday commuting activities between September 2014 and September 2015 in the city of Cuenca, Ecuador. Their results indicated influential effects of roadway hierarchy, household density, living conditions index, land use mixture, segregated bicycle paths, intersections, and slope on cyclist volumes across the three dependent variables.

Two other independent studies estimated discrete choice models to measure the likelihood of each street segment being chosen by Strava users (Strava user counts on links were split into five categories from low to high; LaMondia & Watkins, 2017; Lin & Fan, 2020). Strava users' route choices were associated with roadway characteristics/household income (in both studies); accessibility to shopping areas and restaurants (LaMondia & Watkins, 2017); and slope, time of day, and bicycle facilities (Lin & Fan, 2020).

While Strava tracking records have been utilised to examine route preferences, technically they are not suitable to build path-based route choice models due to the way that Strava Metro provides data products (other than sampling-related issues). Because

Table 2. Application overview in route choice analysis.

Reference	Summary description
Huber and Lißner (2019)	Proposed a method to derive a single route from Strava data.
LaMondia and Watkins (2017)	Estimated ordinal logistic model to measure the likelihood of each street segment being chosen by Strava users (458 cyclists).
Lin and Fan (2020)	Estimated ordered probit model to measure the likelihood of each street segment (N = 237,673) being chosen by Strava users.
McArthur and Hong (2019)	Proposed a method to compute the shortest commuting routes between trip beginnings and trip ends from Strava data.
Orellana and Guerrero (2019)	Estimated a negative binomial model to examine the effect of street network structure on variations in cycling activity counts on street segments (N = 19,103).
Sun et al. (2017)	Estimated mixed-effects model to examine associations between environmental characteristics and the rate of recreational bicycle trips on streets (N = 119).

aggregated forms do not allow identification of a single trip route, a set of alternative choices for each trip generation-destination pair cannot be generated.

Some scholars have attempted to address this limitation using OD data provided by Strava Metro. McArthur and Hong (2019) computed the shortest commuting routes between trip beginnings and trip ends that cyclists would take to minimise trip distance. While the study found the counterfactual shorted link flows, the individual trip route actually taken was still missing, which means failure to form a choice set. In another attempt, Huber and Lißner (2019) suggested a new method to derive a single route and a route alternative from Strava Metro data by assigning and matching bicycle volumes in trip origin zone, destination zone, and passing-through zone to street segments that intersect with the zones. This method proposed how to disaggregate the original Strava data (in aggregate format) so that the derived route could be used for predicting a route choice model, but far more intensive validation is needed because the size of the zones and route search algorithms affect the robustness of the estimation results. In addition, even if a chosen single route can be derived from the original dataset, studies that necessitate personal information about a trip maker (e.g. Hood et al. [2011] and Dhakar and Srinivasan [2014]) are unlikely to be feasible due to privacy protection.

Infrastructure evaluation

Cities continue to improve the cycling environment, but the scarcity of suitable data sources hinders measuring the effectiveness of all efforts. Moreover, given that a change in one location might generate ripple effects in broader areas, a rich dataset is needed for thorough evaluation. Such issues might be managed with Strava to some extent. Table 3 provides an overview of applications related to infrastructure evaluation.

Heesch and Langdon (2016) examined the usefulness of Strava-collected bicycle trips in assessing the impact of a bikeway expansion in Queensland, Australia. Based on spatial and numerical comparisons, they detected monthly changes in Strava cyclists on the expanded bikeway and surrounding area over three months. They also, however, argued that cross-reference with other sources (e.g. on-site counts) was needed due to spatially differential Strava usage. Another study (Heesch, James, Washington, Zuniga, & Burke, 2016) in the same country conducted a practical evaluation of a newly opened bikeway in Brisbane by comparing monthly Strava counts over a year (sixth months before/after the opening). The comparison provided evidence that some cyclists shifted from the pre-existing unsafe route to the new safe bikeway. The study results also

Table 3. Application overview in infrastructure evaluation.

Reference	Summary description
Heesch and Langdon (2016)	Examined the usefulness of Strava-collected data in evaluating the impact of a bikeway expansion based on visual and numerical comparisons.
Heesch et al. (2016)	Evaluated a newly opened bikeway by comparing monthly Strava counts before and after the opening over a year.
Boss et al. (2018)	Analysed changes in network-wide ridership through local indicators of spatial autocorrelation (local Moran's I) from monthly aggregated Strava cycling counts.
Hong et al. (2019)	Evaluated the effects of big infrastructure investments by developing a fixed-effects Poisson panel regression model for four years of monthly Strava counts.
Hong et al. (2020)	Investigated interactions between cycling infrastructure and adverse weather conditions by developing a fixed-effects regression model for hourly Strava bicycle trips.

showed that the new bikeway motivated people to switch transport modes to cycling, but it was not appealing for women. These contextual findings, however, were obtained from concurrently carried out field observations and intercept surveys, not from Strava.

The above level of infrastructure evaluation targeting a few locations has application cases with traditional counting systems (e.g. automated sensors), which could be even more accurate (Deenihan, Caulfield, & O'Dwyer, 2013; Skov-Petersen, Jacobsen, Vedel, Thomas Alexander, & Rask, 2017). However, the greater the study dimension, the more difficult empirical analysis may become. In some studies, Strava has shown potential for city-wide explorations that are unlikely to be plausible with the traditional method. For instance, Boss, Nelson, Winters, and Ferster (2018) analysed changes in network-wide ridership in Ottawa-Gatineau from monthly aggregated Strava cycling counts per segment for a year. In this study, Boss et al. discerned ridership changes with respect to the newly installed infrastructure itself and around it, as well as the temporary closing of a route caused by the construction project. Hong, McArthur, and Livingston (2019) appraised big infrastructure investments partly prepared for the 2014 Commonwealth Game in Glasgow. For the analysis, a fixed-effects Poisson panel regression model was developed from four years of monthly Strava counts. It was found that among four new cycling routes, three had positive effects on increasing the monthly total volume of cycling.

A fine temporal scale of Strava-generated data also enabled investigating interactions between the benefits of cycling infrastructure and bad weather conditions in a city with a high level of precipitation. Hong, McArthur, and Stewart (2020) estimated a fixed-effects regression model using hourly bicycle trips between 6 am and 11 pm taken from Strava in 2016 and found that safe cycling infrastructure (i.e. segregated lanes and shared off-road lanes) was not effective in neutralising the adverse impact of rain in Glasgow, Scotland.

The aforementioned studies exemplify how to utilise the fitness tracking records being updated every day to evaluate the impacts of infrastructure involving longitudinal analysis, not only at several sites but also in wide areas. However, in this research stream, the analytical approach should be considered in terms of ways to control extraneous factors such as seasonal variations, spatial autocorrelation, natural increments in Strava users, and atypical events (e.g. cycle races).

Crash exposure control

Preventing crashes is a primary task for most transportation authorities, and systemising a tool for crash analysis is a vital part of safety interventions. When identifying risk factors for cycling-involved crashes, researchers must determine directly measured exposure to attain proper safety implications, but doing so is a common challenge (Turner et al., 2017). To date, substantial safety modelling efforts have used surrogate measures such as population, employment, or bicycle mode share (e.g. Amoh-Gyimah, Saberi, & Sarvi, 2016; Cai, Lee, Eluru, & Abdel-Aty, 2016; Nashad, Yasmin, Eluru, Lee, & Abdel-Aty, 2016), or counts taken from automatic sensors or human collectors have been directly used or expanded for analysis scales (e.g. Guo, Osama, & Sayed, 2018; Osama & Sayed, 2017; Prato, Kaplan, Rasmussen, & Hels, 2016). However, a growing number of studies are demonstrating that Strava Metro data can be integrated with the approach to control bicycle crash exposure, as shown in the Table 4.

Table 4. Application overview in crash exposure control.

Reference	Summary description
Aldred et al. (2018)	Developed case-control crash occurrence models (multilevel binary logistic) by comparing injury sites ($N = 567$) with control sites ($N = 6,046$); exposure was derived from the cycling flow estimation model where Strava counts were an explanatory variable.
Chen et al. (2020)	Developed crash frequency models (negative binomial) at intersections ($N = 209$); categorised Strava bicycle counts as an exposure.
Raihan et al. (2019)	Developed crash frequency models (zero-inflated negative binomial) at segments and intersections ($N = 397$); categorised Strava bicycle volume as an exposure.
Saad et al. (2019)	Developed crash frequency models (negative binomial) at intersections ($N = 171$); categorised adjustment factor-applied Strava bicycle counts as an exposure.
Saha et al. (2018)	Developed crash frequency models (conditional autoregressive) at the block group level ($N = 11,355$); categorised Strava bicycle miles travelled and bicycle trip intensity as an exposure.
Sanders et al. (2017)	Proposed simplified exposure estimating models (Poisson) at intersections ($N = 46$); used Strava's annual activity counts on segments as an explanatory variable.
Sener et al. (2019)	Developed crash frequency and severity models (negative binomial and binary logit) at the block group level ($N = 1,053$); categorised Strava bicycle trip intensity as an exposure.
Y. Wang et al., (2018)	Developed crash occurrence models (binary logit) at segments ($N = 188$) and intersections ($N = 184$); categorised Strava bicycle counts as an exposure.

The methods to integrate Strava data (as bicycle exposure) into crash models are split into two types: two-stage inclusion involving an additional estimation process or immediate inclusion. Based on the example of the two-stage inclusion method, Aldred, Goodman, Gulliver, and Woodcock (2018) used systematically estimated cycling flow from Cynemon (a cycling network model that estimates traffic flows across London, with Strava cycle data used as an explanatory variable) to control for bicycle exposure when investigating injury risk factors of cycling injuries in London in 2013–2014. Sanders, Frackelton, Gardner, Schneider, and Hintze (2017) proposed abbreviated exposure estimating models that can reduce data burden and be ready for practice by integrating Strava bicycle data. Adding Strava data (annual activity cycle count) to bicycle exposure estimation models of Poisson regression could improve explanatory power (pseudo R-squared from 0.57–0.62) while diminishing the need to collect other datasets. The authors noted that the simplified modelling process can support local practitioners having limited access to full-scale exposure estimations.

The approach involving an additional estimation process is likely to be much more reliable. However, probably due to resource constraints, most studies have chosen the immediate inclusion method. Immediate inclusion studies can be further subdivided contingent upon whether manipulation to moderate Strava's shortcomings was included. First, in the studies that did not employ manipulation, total bicycle volume for a year or converted daily average was applied as the facility-specific exposure at the intersection or street (Chen et al., 2020; Y. Wang et al., 2018). While handy in practice, this method must assume Strava users are evenly distributed, keeping the same proportion to total bicyclists across the facilities. However, the proportion is very likely to vary spatially, which can affect the robustness of the direct facility-specific exposure measurement.

To mitigate the effects of spatially uneven distribution and problematic arbitrary numbers, several studies aggregated and classified the original bicycle counts into three categories: low, medium, or high. Raihan, Alluri, Wu, and Gan (2019) counted the total segment bicycle trips within census block groups and stratified the values. Subsequently, the intersections and segments (unit of analysis) were assigned to one of the three classes. Similarly, Saha, Alluri, Gan, and Wanyang (2018) calculated bicycle miles travelled and bicycle trip intensity at the census block group level and then incorporated the measures (as low, medium, or high) to bicycle crash models. Sener, Lee, Hudson, Martin, and Dai (2019) also applied the categorised trip intensity measures to control for exposure when estimating crash frequency/severity models. Still, this form of manipulation is likely to violate another assumption—that is, Strava records are socio-demographically randomly sampled and equivalently represent whole population groups.

Saad, Abdel-Aty, Lee, and Cai (2019) suggested two adjustment factors using official data to overcome the sampling-related shortcomings. First, they created a population adjustment factor that calibrates percentages of cyclists by age and gender based on the cycle data of the National Household Travel Survey (NHTS). They also determined a field data adjustment factor from actual observations that was then applied to estimating more accurate bicycle trip volume. In crash frequency model specifications, when the two adjustment factors were adopted, the model performance turned out to have better goodness of fit than without any adjustment factors.

Overall, Strava-generated exposure variables have shown statistical significance in conventional crash models (e.g. Poisson, negative binomial, zero-inflated negative binomial, binary logit models), thereby improving model performance. Although it cannot simply be concluded that Strava can serve as accurate exposure, it obviously has the potential to act as a surrogate or proxy solution in the safety task. To maximise that potential, it is recommended to adjust the original Strava data with official data rather than use it directly.

Air pollution exposure assessment

Adverse health impacts of cycling in a polluted atmospheric environment is another parameter that a growing body of literature suggests considering when planning and operating a transport system. Without adequate data sources, analysing exposure to air pollution is another difficult task. However, the following studies show that expanded data options through Strava may facilitate consideration of air pollution exposure (Table 5).

Sun and Mobasheri (2017) measured average instantaneous exposures to particulate matter with an aerodynamic diameter below 2.5 μm ($PM_{2.5}$) and 10 μm (PM_{10}) for riders at nodes and compared how the level of momentary exposure differs by discrete trip purposes (commuting and recreation) in Glasgow, United Kingdom. The estimation results suggested that recreational bicyclists are less likely to be exposed to air pollutants because they tend to prefer the city outskirts. A similar study by Sun, Moshfeghi, and Liu (2017) approximated the inhaled doses of air pollutant for all the cycling and

Table 5. Application overview in air pollution exposure assessment.

Reference	Summary description
Lee and Sener (2019)	Investigated potential exposure of bicyclists on roadways ($N = 3{,}501$) to traffic-related air pollution: Strava bicycle counts were the dependent variable of spatial autocorrelation regression models.
Sun and Mobasheri (2017)	Estimated instantaneous air pollution exposure at nodes while cycling for commuting and non-commuting (mathematic calculation and local Moran's I).
Sun et al. (2017)	Estimated average inhaled dose of air pollutant during a single cycling and walking trip (mathematic calculation and local Moran's I).

walking trips recorded in 2015. The amount of inhaled air pollutant was estimated as a product of $PM_{2.5}$, exposure duration (combined trip time at nodes and moving in segments), and general ventilation rate while riding and walking. On average, the total inhaled dose while moving in segments was greater than two times that at nodes for both, whereas a single cyclist was likely to inhale four times the amount of $PM_{2.5}$ that a pedestrian breathed in. Another study by Lee and Sener (2019) estimated the potential exposure of bicyclists on roadways to traffic-related air pollution across El Paso, Texas, by developing spatial autocorrelation regression models. The authors indicated significant associations between greater bicycle volume and higher levels of $PM_{2.5}$ emissions, implying the need for appropriate guidance for healthy riding.

With the aid of Strava, researchers have been able to attain new insights into exposure to air pollution of cyclists, but the evidence from the study results is less obvious than that from empirical analyses that recruit participants and let them carry individual instruments (e.g. Good et al., 2016; Matt et al., 2016). The Strava app may be used to recruit and collect movement trajectories for empirical endeavours, such as to measure an individual's direct exposure, but observations from already collected/aggregated Strava Metro data are likely to be limited to "probable exposure" to air pollution at the population level.

Implications for future work

The literature discussed in the previous section shows various application cases of Strava Metro data and implies what should be considered when planning use of the new source or initially framing a study based on the current state of practice. This section summarises such implications according to six topics that can be instructive for transportation communities.

Representativeness

Bicyclists do not always ride on dedicated bikeways, instead often pioneering their own paths, which makes for complex movement traces across networks. In this sense, Strava seems likely to be a mirror that reflects real footprints. However, a reasonable suspicion that may arise is how well the data represent the general population. According to the reviewed literature in this paper, Strava Metro data typically represent 1–5% of total bicycle volume (Cesme, Dock, Westrom, Lee, & Barrios, 2017; Chen et al., 2020; Griffin & Jiao, 2015a; Turner et al., 2019; Z. Wang et al., 2018). Several studies have also qualified correlations between Strava samples and actual count data and found overall strong correlations greater than 0.75 (see Table 6). For example, Roy et al. (2019) developed a linear regression model between AADB counts at 44 locations and corresponding Strava bicycle flow and found an R-squared value of 0.76 in Arizona. Conrow, Wentza, Nelson, and Pettit (2018) obtained a higher R-squared value (0.79) by linearly correlating manual counts at 122 sites and Strava ridership volumes in the greater Sydney area. However, the magnitude of correlation may vary by local context (possibly including the number of observation sites). In a study by Fan and Lin (2019), correlation with manual counts at seven locations was much lower, at 0.36.

While the overall high correlation implies that Strava riders could be a good proximation of cyclists, the level of correlation is likely to change according to which temporal and

Table 6. Correlations between official counts and Strava counts.

Reference	Official count	Strava sample	Correlation	Study area
Conrow et al. (2018)	Manual count collected at 122 locations between 7 and 9 am on March 1, 2016	Bicycling rider volume in month of March 2016	$R = 0.79$	The Greater Sydney area
Fan and Lin (2019)	Manual counts at 7 locations	Cycling trips from December 2016 to November 2017	Adjusted $R^2 = 0.356$	Charlotte, North Carolina
Haworth (2016)	Mix of manual and video counts at 164 sites collected over a 4-week period in April and May 2013	Strava data in 2013	Adjusted $R^2 = 0.62$	London, United Kingdom
Hochmair et al. (2019)	Average number of cyclists per weekday from video imaging at 32 sites based on 2-day periods between October and December 2016	Average number of Strava activity counts per day between January and June 2016	$R = 0.55$	Miami-Dade County, Florida
Hong et al. (2019)	Manual counts at 35 locations for 2 days in September 2014	Bicycle trips for the corresponding times	$R^2 = 0.67$ (daily correlation), $R^2 = 0.82$ (hourly correlation)	Glasgow, United Kingdom
Huber, Lißner, and Francke (2019)	Hourly traffic volume automatically collected from January 2015 to June 2016 and temporal manual counts from May to June 2016	Bicycle trips between June 2015 and June 2016	$R^2 = 0.754$ (hourly correlation)	Dresden, Germany
Jestico et al. (2016)	Manual counts of cyclists at 18 locations for 6 days in January, 8 days in May, 6 days in July, and 14 days in October 2016	Cyclist count in 2016	$R^2 = 0.40$ (7–9 am), $R^2 = 0.56$ (3–6 pm), $R^2 = 0.58$ (a.m. and *p*.m. combined)	Victoria, Canada
Roy et al. (2019)	AADB extrapolated from automated 2-week period counts at 44 locations in April, May, October, and November in 2016	Street-level AADB for 2016	$R^2 = 0.76$	Maricopa County, Arizona
Sun et al. (2017)	AADB volume at 119 streets offered by the UK Department for Transport	AADB in 2015	$R = 0.83$	Glasgow, United Kingdom

spatial window is adopted, even within the same sample set. For example, Hong et al. (2019) linearly correlated the number of cyclists who passed counting sites ($N = 25$) for two days in 2014 and the corresponding Strava bicycle counts in the same year. As the level of aggregation interval increased from hourly to daily, the R-squared significantly improved, from 0.67 to 0.82. Roy et al. (2019) compared Strava counts to official counts at daily, monthly, and annual levels and obtained the highest R-squared (0.76) from the annual comparison. Beyond a temporal correlation, Conrow et al. (2018) investigated how Strava data spatially correspond to actual bicycle trips using local Moran's I. Although Strava data showed overall good spatial correspondence, they did not match for all the locations, and certain areas showed a stronger spatial match, such as in the central business district.

As scholars suggested, despite the uncertainty in representativeness, Strava bicycle behaviour can be generalisable to the entire population (but depending on local context). However, which temporal aggregation is applied and where the analysis site is may affect the validity and acceptability of the analysis results, so adequate time and spatial frame should be taken.

Sample bias

Strava's sampling is inherently not random, either for individuals or their activities. The characteristics expected from the sampled population are that they own a smartphone (or equivalent device), can operate it, and are motivated to track their activities, which leads to lopsided demographic distribution. In terms of demographic features, as almost every reviewed piece of literature indicated, data contributors are more likely to be male (in general, more than three-fourths) and a certain age group between 25 and 44 (roughly half). Also, Strava policy that restricts app users to people aged 16 years or over completely excludes younger members of society (Strava, 2018a). As for activity, the app is intrinsically predisposed to being recreational rather than utilitarian, primarily because Strava is a fitness tracking app. Overall, studies show that 20–40% of Strava cycling is for commuting (though the Sun, Du, Wang, and Zhuang [2017] study reported 60%).

However, oversampling certain groups (e.g. economically active, tech savvy, and younger) also appears in most crowdsourced data being collected through mobile devices, not just Strava Metro data, a fact that has been broadly acknowledged by other literature reviews of big data in transportation (Lee et al., 2016; Milne & Watling, 2019; Z. Wang et al., 2018). Also, some of the skewness may be in accordance with the nature of general bicyclists. For instance, Sanders et al. (2017) found general bicyclists in the study region were skewed towards male and 25- to 44-year-old riders, suggesting "[bias] concern was not sufficient not to consider use of Strava."

Nevertheless, this sample bias must be anticipated, and analysts must deliberate on how this expected problem will influence the study results. Further, when decision-making is involved, under-observed groups with no access to mobile devices or Strava, such as minority communities and the elderly, must be deliberately considered.

Data fusion

While representativeness and sampling bias are undeniable handicaps of Strava app data, they might be addressed to some extent by fusing multiple datasets. The most straightforward and feasible way is cross-referencing with traditional sources. For instance, Saad et al. (2019) adjusted Strava bicycle flow with automatic counts, but it should be noted that robustness of this combination is also subject to the number of counter stations. As in the study by Heesch et al. (2016), intercept surveys can make a connection between Strava opt-in users and general cyclists by asking purpose-oriented questions—for example, whether the individual is a regular Strava cyclist. As discussed previously, nationwide official data such as the NHTS or American Community Survey provide another source that can complement the lack of sociodemographic information—not directly or at an individual scale but indirectly and at an aggregated level, such as census block groups.

Beyond this simple level of cross-use of different datasets, the more sources that are fused, the more comprehensive, reliable insights that can be achieved because a discrete data source offers unique advantages by covering different types of travellers, trip purposes, and spatial variations. For instance, bike-share systems are likely to be more favourably used by visitors around tourist attraction sites, and fitness tracking app users are more likely to be a recreation-oriented population (Proulx & Pozdnukhov, 2017). In addition, site-based datasets give information on trips typically passing network links or intersections in numbers, whereas GPS records reveal a series of movement traces. As data fusion methods, such as machine learning and data analytics, get more elaborate and advanced, the limitations of Strava data (and other mobile-generated crowdsourced data) in certainty, accuracy, representativeness, and reliability may be gradually overcome. While all-inclusive data fusion is appealing, the main challenges may lie in how to fit the discrete datasets with their heterogeneous attributes. As Proulx and Pozdnukhov (2017) suggested, for a successful data fusion, different datasets should be compatible and complementary to each other in the following six dimensions:

- Population scope (full population or a subset).
- Trip aggregation (individual or aggregated).
- Temporal scale (temporal extent, time series, or event).
- Temporal resolution (fineness of slices of time).
- Spatial scale (site-based, traces, OD points, and OD zones).
- Demographics (descriptions of the trip and trip maker).

Potential errors

GPS point data collected via mobile devices have intrinsic errors that may not be removed or corrected while being preprocessed. Strava Metro cleans propositioning uncertainty and noise, but the likelihood of some flaws (e.g. spatial mismatching) remains. Although end data users do not have the full ability to check and/or control such potential errors, additional data cleaning efforts might be helpful. For instance, Conrow et al. (2018) indicated that "some bi-directional streets were represented as two separated line feature segments," requiring exclusion of duplicated segments. Strava Metro announced they have improved trip alignment algorithms to reduce the double counting problems (Strava, 2018b), but the need to check for errors may remain. As indicated by LaMondia and Watkins (2017), there may be abnormal numbers to be verified. For instance, a segment might have only 10 riders, but an immediately neighbouring segment may have 100 riders without any trip generators (e.g. restaurants).

Lee and Sener (2019) revealed another possible noise that can stem from Strava users. In their study, trips were reported for highways where bicyclists seldom ride, generating speculation that people may have "forgotten to turn off the app even after reaching their destination, and the app recorded trips in vehicles." Turner et al. (2019) also noted that "because Strava users must manually deactivate the app at the end of a ride, it is possible that a small number of automobiles (i.e. those that were higher speed) are recorded as cycling or pedestrian activities, potentially increasing overall speed and distance averages on recorded routes." In this sense, unrealistically long trip times and distances might need to be filtered in advance.

Spatial and temporal autocorrelation

Strava provides location-based data that are susceptible to spatial correlation. Ignoring spatial dependence can lead to misleading analysis results in modelling approaches to estimate parameters related to travel behaviours. The reviewed studies improved model estimation power by controlling for spatial effects and indicated that spatial interactions between nearby events in Strava activities should be assessed (Boss et al., 2018; Hochmair et al., 2019; Lee & Sener, 2019; Saha et al., 2018). In a similar vein, the common temporal pattern of Strava cyclists across reviewed literature is there are two peaks a day (i.e. commuting time), more trips on weekdays, and reduced trips in adverse weather conditions (extremely low or high temperatures and rainy days). If the explicit aim of a Strava data-involved study is time-series examination (e.g. pre- and post-evaluation of bicycle infrastructure), then temporal variations, seasonal effects (if the study region has months of extreme weather), and natural increments or decrements in users should be considered because interactions with extraneous factors may mask the real effects of variables of interest.

Privacy protection and trade-offs

Strava data are obtained from passive and continuous monitoring and then handed over to third parties, which poses a risk to privacy. To address this problem, Strava (2018a) implements several privacy control management strategies. First, individuals are given the option to not share their information with others and not include their records on Strava Metro's server. Second, consented records are released in an aggregated form in which individual identities are disguised. Even with this privacy protection setting, a potential risk remains for individuals and movement patterns to be identified in areas that have few Strava users.

Due to such risk, as of July 2018, Strava no longer provides minute-by-minute data but has instead started to provide hourly counts and aggregated counts in five-count buckets (instead of absolute numbers); it also no longer shows routes with counts of fewer than three users (Strava, 2018b). Although these settings do not require review by an institutional review board (Lee & Sener, 2019), they generate trade-offs between sample size and data quality. Filtering out the information of those users who do not want to donate their records, excluding routes that do not exceed three users, and binning counts into five buckets all reduce sample size, which may seriously impact robust analyses at a fine scale. For instance, if 100 cyclists pass a street but only 2% of them use the Strava app and agree to donate data, this street is shown as an absolute zero route even though 100 actual cycling activities took place on the route. This omission will likely result in many zero activities, which is more serious in the fine analytical unit (e.g. links and hourly aggregation). Given that the level of temporal aggregation (e.g. yearly, peak time) and the type of spatial aggregation (e.g. census block group or by infrastructure type) may help avoid excessive zeros, it is important to look into the given dataset first and then choose appropriate aggregation methods.

More critically, aggregating traveller/trip counts makes it unavoidable that important details on individual travel activity will be lost, and anonymising data donors excludes their demographic information and consequently blocks the opportunity to scale sample bias. Accordingly, it is hardly possible to retain essential information from the Strava Metro product such as discrete trip routes, trip purposes, and sociodemographic specifications at the individual level, all of which are essential to understanding the parameters behind travel behaviours and are conventionally collected through traditional travel surveys. Since the limited information may not fit existing methods and could potentially reduce insights, it is necessary for researchers to carefully check which information of interest is available when framing research goals.

Concluding remarks

Strava is leading the big data revolution in the active transportation sphere, where a lack of ample datasets has often constricted a vigorous study environment. The new source of Strava Metro data, with its unique attributes and advantages, has spurred growing research endeavours, thereby cultivating a favourable environment for active transportation professionals. As the volume of relevant literature has increased, the available scenarios and analytical approaches have diversified. At the same time, methods to understand the data properties and the challenges faced when using the data have gradually accumulated. The current study sought to synthesise and convey the accrued knowledge over the last five years, with a focus on how Strava Metro data have been applied in the existing literature and their implications for future work.

While there are still doubts about the use of Strava data, many studies demonstrated acceptable or strong correlations with field-observed data. Beyond the correlation investigation, the comparative advantages expected from Strava are to visualise and illustrate revealed spatiotemporal travel flows across the wide study region. In predicting travel demand, Strava counts may play an important role to enhance model performance as an explanatory variable. The impacts of infrastructure improvement could be assessed by comparing before and after travel volumes, not only for a single location but also on a citywide scale. Until now, the most vigorously applied subdomain was safety-related work. When developing crash models, inclusion of accurate exposure is one of the key qualifications to acquire robust results. As eight of the reviewed studies noted, crowdsourced bicycle data taken from Strava have the potential to estimate and control for crash exposure. The fitness tracking records also provided new insight about the cross-section between active transportation and air pollution at the population level.

While Strava Metro data have supplemented studies of methodologies established with traditional data or further developed new approaches, there are limitations on the use of the data as a full substitution to traditional monitoring systems. First, the Strava counting system does not collect all counts. Namely, there is a discrepancy between Strava samples and general populations, raising concerns about the under- and over-representativeness of populations. In addition, unlike conventional travel surveys, this monitoring process was not created to offer quality data to transportation planners and practitioners, which means the data do not contain all the desired details, such as person-level user profiles and individual full-route traces. Thus, research that requires such information has not been attempted heretofore and may not even be feasible—even with great advances in data-mining skills—due to the nature of Strava and privacy protection. However, considering that "current trends suggest that it is inevitable that automatically recorded, digital data will come into mainstream use both for academic study and for the practical planning of transport systems" (Milne & Watling, 2019), and that "no known dataset has full resolution into the spatial and temporal dimensions of the entire population's travel patterns" (Proulx & Pozdnukhov, 2017), it might be a more rational choice to compensate for the shortcomings rather than abandon the imperfect data.

As such, cross-use of Strava data with multiple sources may have great potential for expanding applications. Validation with official data is also helpful to justify data reliability. While it is surely advantageous to have an official counting system in fusing and validating data, various opportunities remain for cities and municipalities in the absence of such supplementary validation datasets. For instance, simply locating study sites (Apasnore, Ismail, & Kassim, 2017), selecting counting sites (Brum-Bastos, Ferster, Nelson, & Winters, 2019; Duncan, 2017), or identifying overall bicycle patterns throughout a city (Selala & Musakwa, 2016) does not necessarily require rigorous validation.

Although the current review originally intended to encompass pedestrian trips, few applications were found in the literature (only Sun, Moshfeghi, and Liu [2017] analysed walking trips), possibly because of high uncertainty and unsure reliability. This phenomenon seems to be created by the characteristics of walking, rather than exclusively by Strava. Walking is much more divergent than other transport modes in terms of trip purpose, time, distance, and real route taken, thus requiring a large enough sample size to draw significant implications, but the proportion of walking trips reported to the fitness tracking app is likely to be very small. The challenges of pedestrian monitoring through Strava may not be different from what traditional methodologies encounter. However, another big data provider, StreetLight Data, recently started to offer bicycle and pedestrian analytics with advanced metrics, which may demonstrate welcome progress to active travel planners. Given that the big data revolution is still ongoing with technological improvement, the knowledge gap related to non-motorised travel may gradually be filled.

Note

1. Another commercial vendor providing crowdsourced data products for active modes of travel is Street Light Data (2019). It combines multiple sources collected from multi-app LBSs, fitness tracking apps, counters, and traditional surveys, which is likely to improve data quality, but its applications have lagged behind Strava Metro due to its recent launch in 2019.

Acknowledgment

The authors would like to acknowledge Dawn Herring for her editorial review as well as two anonymous reviewers and the editor for their insightful feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially funded by a grant from the Safety through Disruption (Safe-D) National University Transportation Center (UTC) and supported by the U.S. Department of Transportation through the UTC programme.

ORCID

Kyuhyun Lee http://orcid.org/0000-0003-2314-8326
Ipek Nese Sener http://orcid.org/0000-0001-5493-8756

References

Aldred, R., Goodman, A., Gulliver, J., & Woodcock, J. (2018). Cycling injury risk in London: A case-control study exploring the impact of cycle volumes, motor vehicle volumes, and road characteristics including speed limits. *Accident Analysis and Prevention*, *117*, 75–84. doi:10.1016/j.aap.2018.03.003

Amoh-Gyimah, R., Saberi, M., & Sarvi, M. (2016). Macroscopic modeling of pedestrian and bicycle crashes: Across-comparison of estimation methods. *Accident Analysis & Prevention*, *93*, 147–159. doi:10.1016/j.aap.2016.05.0010001-4575

Apasnore, P., Ismail, K., & Kassim, A. (2017). Bicycle-vehicle interactions at mid-sections of mixed traffic streets: Examining passing distance and bicycle comfort perception. *Accident Analysis and Prevention*, *106*, 141–148. doi:10.1016/j.aap.2017.05.003

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

Brum-Bastos, V., Ferster, C. J., Nelson, T., & Winters, M. (2019, November). Where to put bike counters? Stratifying bicycling patterns in the city using crowdsourced data. *Transport Findings*, doi:10.32866/10828

Cai, Q., Lee, J., Eluru, N., & Abdel-Aty, M. (2016). Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. *Accident Analysis & Prevention*, *93*, 14–22. doi:10.1016/j.aap.2016.04.018 C

Cesme, B., Dock, S., Westrom, R., Lee, K., & Barrios, J. A. (2017). Data-driven urban performance measures: Case study application in the district of Columbia. *Transportation Research Record: Journal of the Transportation Research Board*, *2605*, 45–53. doi:10.3141/2605-04

Chen, C., Wang, H., Roll, J., Nordback, K., & Wang, Y. (2020). Using bicycle app data to develop safety performance functions (SPFs) for bicyclists at intersections: A generic framework. *Transportation Research Part A: Policy and Practice*, *132*, 1034–1052. doi:10.1016/j.tra.2019.12.034

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

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

Deenihan, G., Caulfield, B., & O'Dwyer, D. (2013). Measuring the success of the great Western Greenway in Ireland. *Tourism Management Perspectives*, *7*, 73–82. doi:10.1016/j.tmp.2013.03.004

Dhakar, N. S., & Srinivasan, S. (2014). Route choice modeling using GPS-based travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, *2413*, 65–73. doi:10.3141/2413-07

Duncan, A. (2017). A comparative analysis of cyclists' paths through shared space and non-shared intersections in Coventry, England. *Journal of Urban Design*, *22*(6), 833–844. doi:10.1080/13574809.2017.1336062

Fan, W., & Lin, Z. (2019). *Evaluating the potential use of crowdsources bicycle data in North Carolina*. Center for Advanced Multimodal Mobility Solutions and Education, The University of North Carolina at Charlotte. Retrieved from https://cammse.uncc.edu/sites/cammse.uncc.edu/files/media/CAMMSE-UNCC-2018-UTC-Project-Report-03-Fan-Final.pdf

Good, N., Mölter, A., Ackerson, C., Bachand, A., Carpenter, T., Clark, M. L., … Volckens, J. (2016). The Fort Collins commuter study: Impact of route type and transport mode on personal exposure to multiple air pollutants. *Journal of Exposure Science and Environmental Epidemiology*, *26*, 397–404. doi:10.1038/jes.2015.68

Griffin, G. P., & Jiao, J. (2015a). Crowdsourcing bicycle volumes: Exploring the role of volunteered geographic information and established monitoring methods. *URISA Journal*, *27*(1), 57–66.

Griffin, G. P., & Jiao, J. (2015b). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport and Health*, *2*(2), 238–247. doi:10.1016/j.jth.2014.12.001

Guo, Y., Osama, Y., & Sayed, T. (2018). A cross-comparison of different techniques for modeling macro-level cyclist crashes. *Accident Analysis & Prevention*, *113*, 38–46. doi:10.1016/j.aap.2018.01.015

Haworth, J. (2016). Investigating the potential of activity tracking app data to estimate cycle flows in urban areas. In *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (pp. 515–519). International Society for Photogrammetry and Remote Sensing.

Heesch, K. C., James, B., Washington, T. L., Zuniga, K., & Burke, M. (2016). Evaluation of the Veloway 1: A natural experiment of new bicycle infrastructure in Brisbane. *Australia. Journal of Transport and Health*, *3*(3), 366–376. doi:10.1016/j.jth.2016.06.006

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

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

Hong, J., McArthur, D. P., & Livingston, M. (2019). The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. *Transportation*, doi:10.1007/s11116-019-09988-4

Hong, J., McArthur, D. P., & Stewart, J. L. (2020). Can providing safe cycling infrastructure encourage people to cycle more when it rains? The use of crowdsourced cycling data (Strava). *Transportation Research Part A*, *133*, 109–121. doi:10.1016/j.tra.2020.01.008

Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco. *California. Transportation Letters*, *3*(1), 63–75. doi:10.3328/TL.2011.03.01.63-75

Huber, S., & Lißner, S. (2019). Disaggregation of aggregate GPS-based cycling data—How to enrich commercial cycling data sets for detailed cycling behavior analysis. *Transportation Research Interdisciplinary Perspectives*, *2*, 100041. doi:10.1016/j.trip.2019.100041

Huber, S., Lißner, S., & Francke, A. (2019). Utility of GPS data for urban bicycle traffic planning in Germany: Potentiality, limitations and prospects. *International Journal of Transport Development and Integration*, *3*(1), 1–14. doi:10.2495/TDI-V3-N1-1-14

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

LaMondia, J. J., & Watkins, K. (2017). *Using crowdsourcing to prioritize bicycle route network improvements* (Project 2013–083). Southeastern Transportation Research, Innovation, Development and Education Center. Retrieved from https://rosap.ntl.bts.gov/view/dot/36713/dot_36713_DS1.pdf

Lee, K., & Sener, I. N. (2019). Understanding potential exposure of bicyclists on roadways to traffic-related air pollution: Findings from El Paso, Texas, using Strava Metro data. *International Journal of Environmental Research and Public Health*, *16*(371), 1–20. doi:10.3390/ijerph16030371

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

Lee, R. J., Sener, I. N., & Mullins IIIJ. A. (2016). An evaluation of emerging data collection technologies for travel demand modeling: From research to practice. *Transportation Letters*, *8*(4), 181–193. doi:10.1080/19427867.2015.1106787

Lin, Z., & Fan, W. (2020). Bicycle ridership using crowdsourced data: Ordered probit model approach. *Journal of Transportation Engineering, Part A: Systems*, *146*(8), 04020076. doi:10.1061/JTEPBS.0000399

Matt, F., Cole-Hunter, T., Donaire-Gonzalez, D., Kubesch, N., Martínez, D., Carrasco-Turigas, G., & Nieuwenhuijsen, M. (2016). Acute respiratory response to traffic-related air pollution during physical activity performance. *Environment International*, *97*, 45–55. doi:10.1016/j.envint.2016.10.011

McArthur, D. P., & Hong, J. (2019). Visualising where commuting cyclists travel using crowdsourced data. *Journal of Transport Geography*, *74*, 233–241. doi:10.1016/j.jtrangeo.2018.11.018

Milne, D., & Watling, D. (2019). Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*, *76*, 235–244. doi:10.1016/j.jtrangeo.2017.11.004

Nashad, T., Yasmin, S., Eluru, N., Lee, J., & Abdel-Aty, M. A. (2016). Joint modeling of pedestrian and bicycle crashes: Copula-based approach. *Transportation Research Record: Journal of the Transportation Research Board*, *2601*(1), 119–127. doi:10.3141/2601-14

Ohlms, P. B., Dougald, L. E., & MacKnight, H. E. (2018). *Assessing the feasibility of a pedestrian and bicycle count program in Virginia* (Report No. VTRC 19-R4). Virginia Transportation Research Council.

Orellana, D., & Guerrero, M. L. (2019). Exploring the influence of road network structure on the spatial behaviour of cyclists using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, *46*(7), 1314–1330. doi:10.1177/2399808319863810

Osama, A., & Sayed, T. (2017). Investigating the effect of spatial and mode correlations on active transportation safety modelling. *Analytic Methods in Accident Research*, *16*, 60–74. doi:10.1016/j.amar.2017.08.0032213-6657

Prato, C. G., Kaplan, S., Rasmussen, T. K., & Hels, T. (2016). Infrastructure and spatial effects on the frequency of cyclist-motorist collisions in the Copenhagen Region. *Journal of Transportation Safety & Security*, *8*(4), 346–360. doi:10.1080/19439962.2015.1055414

Proulx, F., & Pozdnukhov, A. (2017). *Bicycle traffic volume estimation using geographically weighted data fusion*. Retrieved from http://faculty.ce.berkeley.edu/pozdnukhov/papers/Direct_Demand_Fusion_Cycling.pdf

Raihan, M. A., Alluri, P., Wu, W., & Gan, A. (2019). Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis and Prevention*, *123*, 303–313. doi:10.1016/j.aap.2018.12.009

Roll, J. (2018). *Bicycle count data: What is it good for? A study of bicycle travel activity in Central Lane Metropolitan Planning Organization* (Report No. FHWA-OR-RD-18-16). Oregon Department of Transportation. Retrieved from https://www.oregon.gov/ODOT/Programs/ResearchDocuments/304-76120Bicycle20Counts20Travel20Safety%20Health.pdf

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*(62), 1–20. doi: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. doi:10.1177/0361198119836764

Saha, D., Alluri, A., Gan, A., & Wanyang, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis and Prevention*, *118*, 166–177. doi:10.1016/j.aap.2018.02.014

Sanders, R. P., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington potential option for resource-constrained cities in an age of big data. *Transportation Research Record: Journal of the Transportation Research Board*, *2605*(1), 32–44. doi:10.3141/2605-03

Selala, M. K., & Musakwa, W. (2016). The potential of Strava data to contribute in non-motorised transport (NMT) planning in Johannesburg. In *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (pp. 587–594). International Society for Photogrammetry and Remote Sensing.

Sener, I. N., Lee, K., Hudson, J. G., Martin, M., & Dai, B. (2019). The challenge of safe and active transportation: Macro-level examination of pedestrian and bicycle crashes in the Austin District. *Journal of Transportation Safety & Security*, doi:10.1080/19439962.2019.1645778

Skov-Petersen, H., Jacobsen, J. B., Vedel, S. E., Thomas Alexander, S. N., & Rask, S. (2017). Effects of upgrading to cycle highways—An analysis of demand induction, use patterns and satisfaction before and after. *Journal of Transport Geography*, *64*, 204–210. doi:10.1016/j.jtrangeo.2017.09.011

Strava. (2018a). *Strava terms of service*. Retrieved from https://www.strava.com/legal/terms

Strava. (2018b). *Strava data changes: Notification from Strava Metro to Strava licensees* (communication received through Texas Department of Transportation on July 9, 2018).

Strava Metro. (2019). *Strava FAQs*. Retrieved from https://metro.strava.com/faq/

StreetLight Data. (2019). *StreetLight Data offers complete, stand-alone dashboard of bicycle and pedestrian transportation metrics*. Retrieved from https://www.streetlightdata.com/bicycle-and-pedestrian-metrics-alternative-to-iot-bike-counters/

Sun, Y., Du, Y., Wang, Y., & Zhuang, L. (2017). Examining associations of environmental characteristics with recreational cycling behaviour by street-level Strava data. *International Journal of Environmental Research and Public Health*, *14*(644), 1–12. doi:10.3390/ijerph14060644

Sun, Y., & Mobasheri, Y. (2017). Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava data. *International Journal of Environmental Research and Public Health*, *14*(274), 1–19. doi:10.3390/ijerph14030274

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

Turner, S., Benz, R., Hudson, J., Griffin, G., Lasley, P., Dadashova, B., & Das, S. (2019). *Improving the amount and availability of pedestrian and bicyclist count data in Texas* (Report No. FHWA/TX-19/0-6927-R1). Texas A&M Transportation Institute. Retrieved from https://static.tti.tamu.edu/tti.tamu.edu/documents/0-6927-R1.pdf

Turner, S., Sener, I. N., Martin, M., Das, S., Shipp, E., Hampshire, R., … Robinson, S. (2017). *Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities* (Report No. FHWA SA-17-041). U.S. Department of Transportation. Retrieved from https://safety.fhwa.dot.gov/ped_bike/tools_solve/fhwasa17041/index.cfm

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, *11*, 141–155. doi:10.1016/j.tbs.2017.02.005

Wang, Y., Monsere, C. M., Chen, C., & Wang, H. (2018). Development of a crash risk-scoring tool for pedestrian and bicycle projects in Oregon. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*, 30–39. doi:10.1177/0361198118794285