Emerging data for pedestrian and bicycle monitoring: Sources and applications

Kyuhyun Lee [a], Ipek N. Sener [b,*]

[a] *Texas A&M Transportation Institute, 2935 Research Parkway, College Station, TX 77845, USA*

[b] *Texas A&M Transportation Institute, 505 E Huntland Drive, Suite 455, Austin, TX 78752, USA*

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

* Corresponding author.

E-mail address: i-sener@tti.tamu.edu. (I.N. Sener).

ABSTRACT

Growing attention on the benefits of non-motorized travel has increased the demand for accurate and timely pedestrian and bicycle travel data. Advancements in technologies and the proliferation of smartphones have created new data sources that can help eliminate limitations related to small sample size and infrequent updates due to limited resources. This study reviews the emerging data sources and their current use, focusing on non-motorized travel monitoring. In this study, the emerging data are categorized into mode-unspecified and mode-specified data based on whether the mode used can be detected with no or little effort. While mode-unspecified data are collected without sorting out non-motorized travelers, mode-specified data at least know who (which mode) is being monitored. So far, commercial vendors provide a vast volume of mode-unspecified data, but their products have been mainly used for motorized trips or are in initial stages of development. Meanwhile, readily available data sources and their applications are more concentrated on mode-specified data, which have enabled varying non-motorized travel studies—including travel pattern identification, route-choice modeling, crash/air pollution exposure estimation, and new facility provision evaluation—but are mostly focused on bicycling. Despite the potential of emerging data, their use also has several challenges, such as limited mode inference, sample bias, and lack of detailed trip/traveler information due to privacy issues. More efforts are needed, such as improving data accuracy and developing robust data fusion techniques, to be able to fully utilize the emerging data sources.

1. Introduction

1.1. Background

Because of growing attention on the benefits of non-motorized travel (i.e., walking and bicycling), the need for accurate, timely pedestrian and bicyclist travel data has increased over the past decades. Non-motorized travel modes have unique characteristics; their trips are more sensitive to the environment (e.g., weather conditions, topography, land use patterns), more variant (i.e., do not always follow dedicated roadways or often make their own paths), and shorter than motorized trips. Collection of walking and bicycling data has traditionally relied on limited location counting and travel surveys for multimodal transportation (Ryus et al., 2014). Because most of the traditional monitoring methods require extensive effort that can be human-resource intensive or time consuming, non-motorized travel data have often been limited by small sample size, time and budget constraints, and infrequent updates.

Over the last 10 years, advancements in technologies and the proliferation of smartphones, along with increased demand for detailed information on non-motorized modes, have led to interest in new monitoring methods. Although still widely reliant on traditionally collected data, researchers and practitioners are investigating emerging methods that can take advantage of mobile devices by which billions of human movement records are automatically, passively collected. In mimicking the current trend of the big data revolution, numerous studies have reviewed relevant technologies and the potential, applications, and challenges related to mobile-device-generated data in the transportation domain (Chen et al., 2016; Lee et al., 2016; Milne and Watling, 2019; Rojas IV et al., 2016; Z. Wang et al., 2018), yet no study has solely focused on non-motorized modes.

This paper provides a review of emerging data sources, applications, and challenges in the context of non-motorized transportation modes with a specific focus on new technology-based data using mobile devices (e.g., smartphones, tablets, watches, and wristbands). Though not intended to be comprehensive, the review provides resources for the transportation community interested in emerging methodologies in non-motorized travel monitoring. The review starts with an overview of pedestrian and bicycle data sources. Emerging data are then examined extensively in terms of their data sources, their current applications, and the remaining challenges transportation professionals face concerning their use. The paper ends with a summary and conclusion.

1.2. Overview of pedestrian and bicycle data sources

To provide a more structured, effective, and easy-to-follow evaluation of data characteristics, Fig. 1 gives an overview of how data sources are classified in this study, followed by a discussion of these data sources.

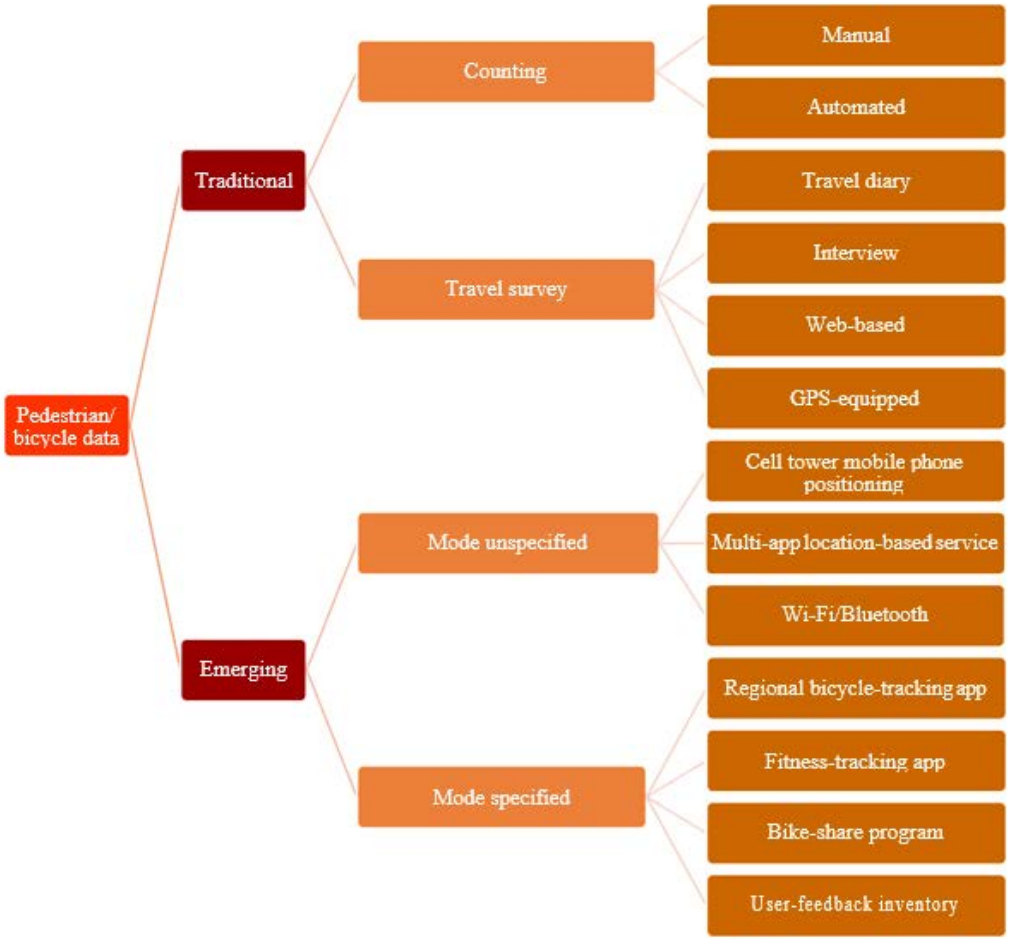

Fig. 1. Classification of pedestrian and bicycle data sources.

1.2.1. Traditional data

When pedestrian and bicycle monitoring programs started, monitoring tools were typically limited to counting and travel surveys. Site counting is the most traditional data-monitoring methodology and is used to directly measure bicycle and pedestrian data. Manual site counting, performed by human data collectors in the field, has been the primary method to collect pedestrian and bicycle traffic volumes (Ryus et al., 2014). Techniques for counting have expanded to automatic methods using video cameras (manual observations of the recorded data may be needed) and diverse sensors, such as pneumatic tubes, inductive loop detectors, passive/active infrared sensors, and radio beams. When traffic volume over a lengthy period of time is needed, such auto-detectors can substitute for human data collectors. Pedestrian signal actuation buttons can be used as a reasonable proxy for roughly determining pedestrian demand (Day et al., 2011). While the main reason for counting is to gather traffic volume data, in some cases counting is capable of collecting additional information beyond simply numbers, such as direct observation of travel behaviors (e.g., helmet wearing, crossing on red) and traveler attributes (e.g., gender, generalized age ranges). Nevertheless, the primary advantage of traditional counting methods over other methods

is that they can provide a near-100% traffic flow count at specific transportation facilities and at specific times. Thus, this method is generally used to measure traffic flows at the micro-scale, and the ground-truth counts can be applied to predict travel demand over areawide network systems.

In traditional travel surveys, subjects are asked to describe their activities and travels in detail via a travel diary, interview, or web-based questionnaire. Travel surveys are preferred over counting when it is critical to understand contextual parameters that can explain travel behaviors (such as trip purpose, demographics, or socioeconomic status). One issue with travel surveys is that asking people to complete questionnaires can trigger survey fatigue and less faithful answers (Lee et al., 2016), and, especially from a non-motorized monitoring perspective, survey participants tend to underreport walking and bicycling. A combination of global positioning system (GPS) loggers and travel surveys can complement such drawbacks to some extent but still suffers from not being able to sample enough to capture the non-motorized activity over the entire network. Although conducting a survey provides comprehensive information pertaining to the participants, it often entails a great deal of preparatory work and post- data processing, high costs, and small sample sizes, which results in infrequent updates. For more information on traditional data collection in the context of non-motorized travel, readers are referred to Turner et al. (2017).

1.2.2. Emerging data

Technological advancements have created a variety of data acquisition methods that require less effort and fewer resources yet produce a larger volume of data than traditional counting and surveying, as is the case with new technology-based data available through mobile devices. Existing literature has not clearly agreed on the definition of these types of data; instead, it broadly uses the following terminology:

- *Big data*, framed as the 3Vs (volume, velocity, and variety; Romanillos et al., 2016).
- *Crowdsourced data*, which are "user-generated data provided through web and mobile applications, often to address a specific issue or solve a problem" (Smith, 2015).
- *Passive data,* collected passively and automatically, reducing the necessity of direct traveler interaction (Bonnel et al., 2015; Lee et al., 2016).

Embracing all these terminologies, in this study, the term *emerging data* is used as a counterpart to *traditional data.* Emerging data discussed in this study are divided into two categories (*mode unspecified* and *mode specified*) according to the nature of the target population and data collection methodologies.

In regard to the first category, *mode-unspecified* emerging data are not generated from a method targeted at only pedestrians and bicyclists but rather are generated from the general population carrying mobile devices. For instance, telecom companies routinely collect mobile phone positioning (MPP) data for operational purposes. Apps featuring location-based services (LBSs) access location information when people search for restaurants, find routes, tag, and check in. Because the primary purpose for which telecom companies and LBS apps collect data is not to monitor non-motorized travelers, the subject pool is not confined to walking and bicycle trips. Thus, non-

motorized modes must be extracted from raw data sets to be applicable as a meaningful resource. Meanwhile, for *mode-specified* emerging data, as the term implies, the targeted population is at least known as being non-motorized travelers. For instance, fitness-tracking apps and bike-share programs do not monitor auto vehicles. The following sections review emerging data sources and their current use and then discuss the remaining challenges affiliated with the collection and use of these data in detail.

2. Emerging data sources

2.1. Mode-unspecified emerging data sources

This study broadly divides mode-unspecified sources into three categories by different raw data collection mechanisms that generate different data attributes, as discussed below.

2.1.1. Cell tower MPP

When mobile phones connect to cellular networks for communication and internet usage or move from one cell tower boundary to another, signaling between the phones and cell towers takes place. Cellular carriers collect billions of location data points by time for operation and billing reference. From the positioning data, movements of users can be extracted. Commonly recognized spatial precision of MPP data ranges from 200 m to 1000 m (Bowman, n.d.) and is typically better in urban areas than in rural areas due to higher cell tower density.

MPP-generated data consist of information regarding time, duration, and location of an action (e.g., calling, sending a message, using cellular data, moving to another cell tower zone). Secondary data vendors such as Airsage purchase the raw MPP records from telecom carriers and resell specific data after the data extraction. Released data types are origin-destination (OD) pairs (zone- and link-based), traffic speed and volume, imputed trip purpose, home/work location, and demographics, but they are not yet customized for non-motorized modes. In addition, a series of time-space points of raw MPP records can be directly supplied by mobile carriers, such as Orange. A high penetration of mobile phones is beneficial in securing a high level of sampling rates, but some challenges related to passive sampling, signaling techniques, and location precision have thus far restricted applications of MPP data for pedestrian and bicycle monitoring.

2.1.2. Multi-app LBS

A number of apps featuring LBSs, such as location-aware search tools and social networking platforms, are playing a leading role in the big data revolution in many fields, including business, marketing, and transportation, and are greatly facilitating the collection of location information from a large number of populations. LBS apps collect real-time location and displacement of the app users based on diverse wireless techniques such as wireless fidelity (Wi-Fi), Bluetooth, GPS, and device-embedded sensors. For example, when people tag the location they are visiting or find restaurants, their geolocation is positioned and sent to the app servers (e.g., Foursquare, Yelp, TripAdvisor). Even when LBS apps are not in use (but location services are on), some of the apps access geolocation and send a notification of an event or a sale (e.g., Facebook, Groupon). LBS techniques have a high level of locational precision (5–50 m), allowing for locating where people go and stay on a small scale (e.g., roads, bike lanes,

sidewalks, and parks; Bowman, n.d.). Despite the location precision merit, it is hard to obtain sophisticated movement details. For example, even if a pedestrian walking along a street is distinguished, the app may not be allowed to determine whether a pedestrian keeps going in the same direction or crosses the street and then continues walking due to heading accuracy errors (e.g., location errors of smartphone in dynamic movement; Wu et al., 2018).

StreetLight Data and Cuebiq are representative companies that provide aggregated multi-app LBS data sets. Cuebiq's database comes from hundreds of LBS apps, and StreetLight Data partnered with Cuebiq to integrate multi-app LBS data with other data sources (e.g., active mode app data, counts, survey results). By developing algorithms for mode recognition and data fusion, StreetLight Data recently released an easy-to-use-online platform, Bike Ped Essentials (StreetLight Data, 2019). The on-demand analytic service provides OD travel demand (trip volume between OD and within OD), traffic attributes (e.g., volume, distance, time, and speed) for selected time frame (e.g., day of week and time of day) and geometry (e.g., zone, link, or city), and inferred context information (sociodemographics and trip purpose). These data sets can support solving a wide array of mobility questions that have not been actively explored over the last decades, such as identifying/predicting non-motorized traffic flows over the entire city, around transit stations, and at specific facilities and identifying underlying mechanisms explaining variations in traffic flows. However, questions requiring more detailed information (e.g., statistical analysis exploring why each trip route is chosen over all the other options and how individual sociodemographic characteristics affect the decision) cannot be solved because such level of details are not allowed due to privacy invasion issues.

2.1.3. Wi-Fi and Bluetooth

Wi-Fi and Bluetooth are wireless technologies that communicate based on media access control (MAC) addresses. MAC is a universally unique identifier of Wi-Fi- and Bluetooth-enabled devices, allowing it to be anonymously, passively detected by nearby sensors (a MAC address is not associated with any personal information such as a user's identity and phone number). While being in detectable mode within the range longer than the discovery time, surrounding sensors detect the device and record the MAC address, timestamps, and location. Nominally, detectable range is 35 m (indoors) and up to 100 m (outdoors) for Wi-Fi, and the Bluetooth radius is around 100 m (Abedi et al., 2013). Signal discovery time of Wi-Fi is approximately one second, whereas Bluetooth needs a much longer time, around 10 s (Abedi et al., 2013). Although not everyone carries detectable devices and not all devices are discoverable (which affects data quality and richness), using Wi-Fi and Bluetooth data can enhance the capability of monitoring human movement patterns in some specific scenarios.

The primary information that can be obtained is the count of devices within a detection zone. In this sense, this wireless detecting system is very similar to a traditional automated counting system. However, a big difference is in the unique code of each detected device. Thus, by matching the overlapping unique MAC address (and timestamp) captured in multiple scanners along the movements, the travel time and OD matrix can be derived. However, this process is less practical for pedestrians and bicycles mingled with all other modal users in a large network; instead, it is more favorable for a single mode at a fixed link or small zone where entrances and exits are perfectly controlled.

Therefore, for non-motorized traveler monitoring, the optimal practices are measuring activity density, traffic flows, and average dwelling time at building-wide spaces or pedestrian-only corridors or bike paths.

2.2. Mode-specified emerging data sources

In contrast to the emerging mode-uncertainty data sources, mode- specified data at least know that the monitored trips are non-motorized modes. This section describes what types of data sources are available, sorts them into four categories, and introduces example sources that can adequately exemplify the characteristics of the sources based on the authors' research.

2.2.1. Regional bicycle-tracking app

Regional bicycle-tracking apps are dedicated smartphone applications designed by government agencies for travel behavior research purposes and to promote bicycling. These apps basically collect GPS traces and selectively gather demographic information, trip purposes, rider types, bicycle experiences, trip frequency, and so forth. Because most government agencies collect data under the agreement on data usage for research purposes and maintain the servers, it is advantageous to have raw GPS trajectory points that enable transformation of the raw data sets into the desired format for nearly any type of analysis. Meanwhile, in order to utilize the raw GPS points, it is necessary to process, clean, and assign a series of ordered GPS points to wanted geometries (e.g., streets or intersections), which demands heavy computational work.

The first bicyclist monitoring app in this category (in the United States) was CycleTracks, developed in 2009. CycleTracks is a good example that shows how to develop and promote the app and how to utilize the collected data for an advanced level of application. San Francisco County Transportation Authority (SFCTA) developed this app to estimate bicycle trip demand over the region and ultimately assign the bicycle volume to specific streets. To achieve the goal, CycleTracks was designed to collect trip time, space trajectories, and personal information (optional) such as age, gender, zip code, and trip purpose to understand user bias (Hood et al., 2013). SFCTA made a set of open sources freely available online and inspired other regions and countries, such as Cycle Atlanta (deployed in Atlanta, Georgia, in 2012), ORcycle (deployed in Portland, Oregon, in 2014), and Mon RésoVélo (deployed in Montreal, Canada, in 2014). While basically collecting a series of GPS traces and some user profiles, these types of apps have unique functions and thus gather unique information depending on a specific project scheme. For instance, Mon RésoVélo additionally collected reduced greenhouse gas emission and consumed calorie data through bicycle trips taken, whereas ORcycle required its users to provide trip purpose, route comfort, and trip frequency information.

2.2.2. Fitness-tracking app

Fitness-tracking apps collect opt-in users' information on physical activities such as walking, running, and bicycling using diverse sensors (e.g., GPS sensors, gyroscopes, accelerometers) built into the devices. Because the primary initiative of these apps is to support athletes who want feedback and motivation with respect to their fitness activities, use of the collected data for transportation research is contingent on the fundamental function—

which is closely related to sample bias—that data contributors are more likely to be recreational-oriented populations.

There are numerous fitness-tracking apps that support workouts, and these apps can be divided into two groups based on whether preprocessed data are commercially available or not. At the time of this writing, only one tracker, Strava, sold a license to allow access to walking and bicycling datasets for public use—that is, for research purposes and transportation planning. Strava's data service, Strava Metro, provides three licenses that can be purchased based upon data aggregation units: node (point), street (segment), and OD (polygon). The product format includes shapefiles and database files designed for use by a geographic information system (GIS) that enables various analyses in the GIS environment. While trip purpose filtering is possible (commute and noncommute) at the aggregate level, trip, and demographic information is not available at the discrete level due to privacy issues.

The second group of apps do not allow commercial access to their data sets despite the wealth of physical activity data. For instance, Fitbit, Garmin, and Endomondo measure personal health-oriented activity metrics such as step count, heart rate, sleep quality, and GPS traces (on devices having GPS tracking functionality) but do not offer aggregated and anonymized data sets to third parties for commercial purposes like Strava Metro. Instead, some of the apps provide web application programming interfaces (APIs) in their cloud data repositories so that people of interest can download the records logged on by the app users with consent. While the public API-based data acquisition requires intensive data development processes (including data downloading, filtering, extracting, and map matching), it is free to use and may supply some attributes not available in the ready-made data products of Strava Metro, such as individual levels of OD pairs, gender, and age.

2.2.3. Bike-share programs

Because of the expansion of bike-share programs, incidental usage records provide another source for monitoring bicycle trips in different schemes. Most bike-share services are operated based on smartphone apps that make it convenient for users to find, rent, and return bikes and also help transportation planners identify where residents and tourists go with a public bicycle. Bikes equipped with GPS can provide details on the route taken between every pair of stations (Wergin and Buehler, 2017); even for bikes with no GPS sensor, check-in and check-out records at stations can be used for trip OD data (Faghih-Imani et al., 2014). Data sets collected by these bike-share programs include information on trips (trip OD and trip start/end time) and station characteristics (capacity and coordinates; Faghih-Imani and Eluru, 2016). Depending on operational policies (e.g., compulsory memberships), bike-share data may provide bicyclist information such as age, gender, and membership type. As a new generation of bike-sharing programs—free-floating bike-share systems—have rapidly expanded in major cities in the world (e.g., Dropbike and Mobike), it has become possible to examine more dynamic bike trip patterns (Du et al., 2019).

2.2.4. User-feedback inventory

User-feedback inventory (web and app-based) is a typical crowdsourced platform used to engage citizens in the planning process by gathering their localized knowledge and experiences (often called a volunteered geographic

information [VGI] platform). Community members volunteer to report, for example, needed improvements for infrastructure, desired change proposals, and hot spots where collisions occur. This type of platform plays the role of a digital channel for direct interactions between citizens and government employees (Le Dantec et al., 2015), thereby delivering useful implications to transportation planning fields (LaMondia and Watkins, 2017). More diverse inventories have been emerging, but some of the main examples are presented below.

OpenStreetMap (OSM), an editable and free-to-use world map that is built and maintained by global volunteers, is one of the most popular examples of VGI. Within OSM, 1 million individuals have contributed to a set of geographic data that include roads, cycle paths, and trails used (OpenStreetMap, 2018). OSM road network features may have particular tags for filtering (e.g., bicycle-only and car-accessible ways) and can be extracted for a wide range of applications (Hochmair et al., 2015).

BikeMaps.org is a safety data collection website and mobile app tool that citizens can use to report crash locations and information such as crash time and injury severity. This geo-crowdsourcing, wherein citizens make a bicycle incident map by adding the collision location information to the map, is able to collect minor injury and crash occurrences that formal bicycle incidents (collected by police officers and insurance companies) are more likely to miss.

Knowledge Based Systems Inc. (2018), in cooperation with the City of College Station, Texas, and the Texas A&M Transportation Institute, developed a community-driven app for sidewalk inventory and condition assessment data originally called MySidewalk (renamed WalkOn™). Basically, app users track their walking, and when faced with missing or damaged sidewalks, they can report the locations of the defects and submit both descriptions and photos of damage. The app also can detect informal pedestrian paths frequently used by the public. The uploaded geospatial data can be downloaded using GIS software.

3. Emerging data applications

3.1. Mode-unspecified emerging data applications

While mode-unspecified sources have great potential to enhance research opportunities for non-motorized modes because of vast volumes, high sampling rates, and little or no need for direct interaction with data donors, applications have been mostly concentrated on vehicle travel research. Thus far, cell tower MPP data have been limited to estimating vehicle traffic parameters, such as travel time/speed and traffic volume/ flow, and modeling trip OD pairs and regional travel demand (Calabrese et al., 2013; Çolak et al., 2015; Huntsinger and Donnelly, 2014).

Nonetheless, multi-app LBS data have a few cases of being applied for pilot projects that examine the potential to support data-driven decisions. State Smart Transportation Initiative programs in the United States analyzed walking trip patterns around light-rail stations in Sacramento, California, using non-motorized mode data for OD trips to/from the stations provided by StreetLight Data. Despite failing to include all trips generated irrespective of trip distances (i.e., trips less than 500 m were not included), the pilot project results were able to suggest implications for policymakers and planners—for instance, the need for improving connectivity at crossing points that pedestrians and

bicyclists pass (McCahill, 2017; State Smart Transportation Initiative, 2017). In addition, the California Department of Transportation initiated a pilot project in partnership with StreetLight Data in 2018. Utilizing multi-app LBS data, the project aims to produce statewide information on people's active transportation behavior (StreetLight Data, 2018).

Wi-Fi and Bluetooth sensing data have also been widely used in vehicle monitoring studies, such as for estimating travel volume/time along roadways (Barcelö et al., 2010; Bhaskar et al., 2014; Li et al., 2018). In addition, an increasing volume of studies has delved further into monitoring crowds—that is, foot traffic in a small area and network. For instance, Kurkcu and Ozbay (2017) estimated pedestrian densities, flows, and average wait times using Wi-Fi and Bluetooth traces collected from six sensors in a public transit terminal. Application cases are also found for monitoring pedestrians in other public spaces, such as airports, shopping malls, and campuses (Meneses and Moreira, 2012; Oosterlinck et al., 2017; Schauer et al., 2014), or for approximating population counts at specific locations in a city (Traunmueller et al., 2018), but they are not widespread for large and complex networks or for bicycle traffic.

Applying mode-unspecified emerging data to active transportation planning is still in the initial phase and is limited, but as third-party data vendors begin to add walking and bicycling into multimode mobility analytics, research opportunities will be enriched in such a way that the challenges of traditional monitoring methods, such as small sample size and limited monitoring coverage, will be overcome.

3.2. Mode-specified emerging data applications

Varying mode-specified emerging data sources have promoted an increasing amount of research on bicycle travel behaviors. Due to the numerous applications, it is not feasible to list all use cases; thus, representative example applications are presented in this section, providing a broad picture of applications by sub-category. One caveat that is noted is that although the original initiative of the current review work was not to exclude applications for walking, few applications for pedestrians were found in the literature. Due to the small number of studies on pedestrian monitoring, data applications provided in the literature review are more concentrated on bicycling than walking.

3.2.1. Travel pattern identification

The first step for designing in-depth analysis and making relevant decisions is to determine when and where people go. Identifying travel patterns is the most representative, basic level of the application of mode-specified emerging data. The intrinsic deficiency in temporal and spatial coverage of traditional monitoring methods has often hindered capturing a fuller picture of non-motorized travel patterns. However, by taking advantage of the high spatial and temporal resolution of GPS traces, researchers are now able to visualize and identify travel patterns in a versatile spatiotemporal window for various purposes. For instance, Apasnore et al. (2017) located roads with high bicycling frequency to determine adequate study sites. Selala and Musakwa (2016) depicted a year of bicycle flow areawide for Johannesburg, South Africa, for the first time in the region. Using ridership data per segment provided

by Strava Metro, Griffin and Jiao (2015) and Hochmair et al. (2019) aggregated bicycle kilometers traveled (BKT) at the census block scale and determined the effects of socioeconomic and built-environment features on BKT. High-resolution data collected through a GPS-enabled smartphone can also give planners and engineers deeper insight into targeted bicycle travel patterns. For instance, bicycle trip data from CyclePhilly, a regional bicycle-tracking app in Philadelphia, Pennsylvania, was applied to detect wrong-way bicycle riders (riding the wrong direction on one-way streets, which may put riders in danger) and identify attributes associated with a high chance of such risky behaviors (Dhakal et al., 2018).

In travel pattern analysis studies, one of the most important aspects is how well emerging data represent the general population because this aspect affects the validity and acceptability of the analysis results. Many scholars analyzed the representativeness of emerging data sources by making a comparison with manual/automated counts and concluded that the information has the potential to be used to understand travel patterns. For instance, several studies assessed correlations between Strava bicycle counts and ground-truth data, which resulted in acceptable to almost perfect synchronization, from 0.40 to 0.96 of the R-squared values (Boss et al., 2018; Conrow et al., 2018; Hong et al., 2019; Jestico et al., 2016). At the same time, the comparisons suggest that the level of representativeness can vary by temporal and spatial frame. For instance, Conrow et al. (2018) compared Strava bicycle counts to manual counts at 122 locations in Sydney and found that Strava better represented the general bicyclists in certain areas (e.g., the highest spatial match in the central business district). Jestico et al. (2016) indicated that bicycle volumes aggregated for the peak period showed the highest R-squared value (0.58) among a different set of hourly aggregations, and thus they used the AM and PM peak totals for further analysis. Therefore, it is important to decide the best time and spatial frames depending on the study purpose to identify more valid travel patterns. Also, heavily relying on emerging data sources with no validation process may lead to less valid analysis results depending on the required accuracy of data. Nevertheless, if the research interest lies in examining variations in travel patterns over time or by season or location, rather than estimating absolute numeric values, then emerging data can provide valuable information.

3.2.2. Route-choice modeling

When predicting a rider's decision-making from a set of route options, revealed preference can have stronger prediction power than stated preference. Because GPS trajectories can provide information on ground-truth footprints of actual behaviors, GPS data collected from tracking apps are appealing to route-choice modelers. The first application of GPS records collected through a regional bicycle-tracking app took place in San Francisco, California (Hood et al., 2013). In Hood et al.'s (2013) study, actual bicycle routes recorded in CycleTracks were modeled and added to a regional travel demand model, enabling assignment of bicycle trips to specific streets. LaMondia and Watkins (2017) estimated which street segments and facilities are preferred by bicyclists. For their study in two regions (Auburn in Alabama and Atlanta in Georgia), the researchers adopted diverse data collection methodologies, including asking volunteers to offer their Strava records and developing regional bicycle-tracking apps (CycleDixie in Alabama and Cycle Atlanta in Georgia). Using a regional bicycle-tracking app—ORcycle

in Portland, Oregon—Blanc, and Figliozzi (2016) incorporated cyclists' comfort level with route-choice modeling. They surveyed bicyclists via the ORcycle app about their trip purpose, riding frequency, and concerns while riding. Malleson et al. (2018) presented a novel approach to understanding pedestrians' route-choice variations between outbound and return trips using individual-level walking traces from an anonymous smartphone app in the area of greater Boston, Massachusetts.

How and why people take different routes is essential information when designing and managing transportation systems, but the above applications using emerging data sources cannot be generalizable unless individual- level trip information (e.g., raw GPS points for each trip) is collected. Since the above apps were developed by the researchers (or they were given access to the server), they were able to get the raw data sets, but heavy computational tasks such as cleaning (e.g., removing noise data), filtering (e.g., ruling out extremely short or long trips), and map matching (assigning a series of GPS points to routes with an associated timestamp) were also the responsibility of the researchers.

3.2.3. Travel demand prediction

When estimating travel demand across networks, mode-specified emerging data sources used as a continuous counting system covering the entire network can play an important role in improving model performance and prediction power. Jestico et al. (2016) estimated cycling volumes using manual counts and Strava Metro data. In this analysis, the researchers categorized Strava bicycle volumes into low, medium, and high rather than directly using the actual value, and they indicated that Strava counts were a good proxy for predicting the relative level of cycling volumes. To improve traffic volume prediction capabilities, Proulx and Pozdnukhov (2017) fused data from different sources: Strava Metro, bike-share program usage, manual and automated counts, and two regional full-population travel demand model estimations. The results highlighted that discrete data sources cover different types of travels, trip purposes, and spatial variations, and combining the given data improved model predictive accuracy. The researchers, however, also indicated that the main challenge of the combination may lie in matching various data sets that have heterogeneous attributes. The Oregon Department of Transportation predicted daily bicycle traffic volumes across the network in the Central Lane Metropolitan Planning Organization (Roll, 2018). Strava bicycle trips were applied to the modeling approach as an explanatory variable with other variables, including infrastructure, amenity accessibility, and network density. In the diagnostic tests that measure the effects of bicycle paths, shortest routes, and Strava data, an inclusion of Strava ridership data most significantly improved the model performance.

3.2.4. Crash exposure estimation

Improving the active travel environment is one of the major concerns of transportation agencies and authorities. Efforts to unveil contributing factors on pedestrian- and bicycle-involved crashes must involve controlling for exposure, which is key to attaining proper safety implications. Nevertheless, safety modeling efforts have often relied on “surrogate measures of bicycle exposure, such as population, employment, and vehicular traffic volume”

(Saha et al., 2018). As Y. Wang et al. (2018) indicated, "The primary challenge to quantifying the risk for pedestrians and cyclists is the missing measures of exposure." However, a growing number of studies have demonstrated that emerging data can enhance the capabilities of controlling for bicycle crash exposure by making efforts to mitigate the data limitations.

Strauss et al. (2015) predicted annual average daily bicycle (AADB) volumes from both field counts and GPS records of a regional bicycle-tracking app (Mon RésoVélo) in Montreal. They then validated injury occurrence models using the two different AADB volumes. The validation results convinced the authors that the bicycle-tracking app data can be used as a reliable source of bicycle flow data for safety analyses. In addition, Strava Metro data have been widely used to control for exposure in macro-level crash models. Saha et al. (2018) estimated models of four years (2011–2014) of bicycle crashes over 11,355 census block groups in Florida. In this modeling approach, volumes at street segments were aggregated as bicycle trip miles (BMT) and bicycle intensity at the census block groups. Instead of the direct BMT and bicycle intensity values, the researchers added categorical breakdowns (low, medium, and high) into the models. Another study by Sener et al. (2019) adopted the same method of categorizing bicycle volume as an exposure variable in crash models. Even though the categorized value indirectly and partially controlled for bicycle volume, Strava- based exposures had significant associations with more crashes. The Oregon Department of Transportation used daily average bicycle counts on roadway segments, and the inclusion of the Strava data significantly improved the prediction power of crash models (Y. Wang et al., 2018). Saad et al. (2019) adopted two adjustment factors (population representation and field observation) to overcome sample bias. In the above studies, despite the inherent sampling problems, Strava data explicitly showed potential as a reasonable exposure metric in developing crash models. Even with a relatively simple level of manipulation, the capability can be enhanced.

3.2.5. Air pollution exposure estimation

The adverse health impacts of polluted air while engaged in physical activity have been receiving growing interest. Despite the importance of this field of study, limited data on active travelers may constrain population-level analysis over a broad region. Several studies have demonstrated the potential of Strava Metro data to assess the exposure to air pollution of active travelers.

Sun and Mobasheri (2017) estimated air pollution exposure at the moment when riders are at nodes (intersections) by discrete trip purposes (commuting and recreation) in Glasgow, United Kingdom. Another study in the same area by Sun et al. (2017) calculated the amount of inhaled doses of air pollutants while riding and walking (waiting at nodes plus moving in edges). On average, a single cyclist was found to inhale four times the amount of air pollution that a pedestrian inhaled. When edges and nodes were separately considered, the total inhaled dose in edges was more than two times that in nodes for both cyclists and pedestrians. These studies could not measure immediate air pollution exposure of a discrete bicyclist (because Strava Metro does not offer information on how long each trip takes and how many times each cyclist makes a trip) but rather showed aspatial association between air pollution and active travel over a broad region. Lee and Sener (2019) also explored the potential exposure of bicyclists

to traffic-related air pollution across the city of El Paso by developing spatial autocorrelation regression models of link-level Strava bicycle volumes. Lee and Sener pointed out that for the entire network-level analysis, additional data processing was needed, even though Strava Metro delivers preprocessed data sets. For example, they needed to merge adjacent segments to avoid overrepresentation of very short segments (e.g., 1 m) and remove presumably noisy segments (e.g., highways and expressways where bicyclists are not allowed).

3.2.6. Infrastructure evaluation

Other forms of mode-specified emerging data applications have come from evaluating the impacts of new bicycle facilities. Continuously collected records are advantageous for monitoring variations in travel patterns over time that are probably caused by infrastructure interventions. Several researchers have compared bicycle volumes before and after the provision of the bike infrastructure. For instance, Heesch et al. (2016) compared Strava bike volumes pre- and post-opening of a new bikeway in Brisbane and found that the new bikeway was effective in attracting more bicyclists. While that study focused on a few specific routes, the following studies exemplify how the effect of infrastructure can be measured at the citywide level. Boss et al. (2018) identified citywide ridership changes caused by an installation or temporary closure of cycling infrastructure in Ottawa- Gatineau, Canada. The research team was able to locate the streets with an unexpected increase/decrease in ridership. Hong et al. (2019) were able to appraise how people reacted to big infrastructure investments using monthly aggregated Strava bicycle counts in Glasgow. For the analysis, a fixed-effects panel regression model was developed to consider overall time trends (over four years) as well as areawide effects. It was found that among the four new cycling routes, three had positive effects on increasing the monthly total volume of cycling trips. These studies exemplify how to utilize emerging data being updated every day to longitudinal analysis that is hardly available without continuous monitoring, not only at several points but also in the broad urban area.

3.2.7. Bike-sharing pattern analysis

As bike-share programs have become more popular, operational management records have enriched layers of information on the bicycle community. Motivations, interactions with built-environment conditions, trip time, and distance of bike-share users may deviate from conventional travel behavior theory or just show analogies. Understanding multifaceted characteristics, including differences and similarities, is helpful for agencies and cities looking to adopt a new bike-share system, extend/facilitate an existing system, or propose a solution to operational issues.

Morency et al. (2017) collected the first six years of data from the bike- share program BIXI in Montreal, Canada, which started in 2009, and conducted a longitudinal analysis of how the system had evolved and matured. They indicated important parameters of the levels of BIXI usage by building a demand model of bicycle trip volumes. Wergin and Buehler (2017) analyzed the bike-sharing trip records of CaBi, which is a docking-based bike-share system in Washington, DC. The analysis results indicated variations in trip attributes according to different membership types (e.g., 24-hour and 3-day members were more likely to make longer trips in distance than monthly and annual

membership users) and showed that greater CaBi usage was related to greater transit ridership. Nolan et al. (2018) examined how bike-sharing trips were associated with land use, subways, and bicycle lanes in New York City by analyzing shared-bike usage data downloaded from the Citi Bike website. Nolan et al. suggested that big data, despite lacking socioeconomic information common in travel analysis, could provide useful insights to planners.

Because of the boom of free-floating bike-sharing in recent years, a growing interest in the dynamic mode of trip patterns has emerged. A crucial issue of the free-floating bike-share systems is how to ensure system balance (supply and demand) and adjust the distribution of bikes. Though still in the early stage, several case studies have demonstrated how to develop a model framework to discover usage patterns from the usage records. One example case study of interest that can help transportation professionals was done by Du et al. (2019) and demonstrates the creation of a model framework if raw usage data (geolocation, timestamp of the trips) are accessible. Du et al. developed a wide range of data-mining techniques (including probability modeling, supervised machine-learning algorithms, and cluster-based time-domain analysis); however, results implied that the unique characteristics (free check-ins and check-outs) of free- floating bike travel patterns may require more complex data-mining procedures to thoroughly understand them.

3.2.8. Community needs detection

One of the promising aspects of emerging data sources is the capability to figure out problems that people are faced with while walking or bicycling. User-feedback inventory collects community needs from people who actually walk and cycle, involving them in collaborative decision- making. For instance, Qin et al. (2018) located barriers and obstacles in pedestrian networks that need to be repaired through an accessibility mapping system, GMU-GcT. They further prioritized repair locations to maximize benefits under the fixed budgets by developing optimization models. The feedback mechanisms can also be applied to gather niche information. For instance, Nelson et al. (2015) developed BikeMaps.org, a global web-mapping tool by which citizens map cycling collisions and municipalities identify the location of hazards. Through the web-mapping tool, Nelson et al. were able to collect near-miss collisions that have been typically less accessible in order to perform an exhaustive safety study. How- ever, these applications should be cautiously addressed due to possible inequality in civic participation. Pak et al. (2017) identified how geographic distribution of street fixing requests (on FixMyStreet) varied across districts in Brussels, Belgium. Considerable differences in the level of participation were suggested between the districts, implying that low-income and ethnically diverse communities are less likely to report.

4. Emerging data challenges

4.1. Mode detection

Mode-unspecified emerging data are not collected primarily for pedestrians and bicyclists, which admittedly contributes to a high level of passivity in monitoring human movement patterns and at the same time begets challenges to transportation researchers and practitioners. To make use of these data, it is necessary to extract walking and bicycling trips from messy and muddled raw data sets. One of the challenges of the emerging data

wherein all trips and all modes are intermingled is to pick walking and bicycling out from the raw data sets by relying not on the aid of traveler input but on machine-learning algorithms. A number of trials have attempted to infer transportation modes from raw data sets collected through MPP. Overall, non-motorized modes (walking and bicycling) can be successfully differentiated from motorized modes due to low speed (Anderson and Muller, 2006; Lin et al., 2013; Nikolic and Bierlaire, 2017), but challenges still remain. For instance, similarities between slow walking and the stationary mode and between bicycling and slow cars present obstacles to accurate mode detection. Moreover, because there is no standard defining the success of mode detection, proposed solutions and disciplines are specific to each study (Prelipcean et al., 2017).

Even if such issues are successfully addressed, as Wang et al. (2010) indicated, the MPP data sets "cannot be used to infer transportation modes for very short trips," primarily because locational precision and sampling granularity of the MPP data are too coarse to be usable for tracking walking and bicycling trips. Collected geolocation information is from the nearest cell towers, and MPP data do not necessarily represent the exact X and Y coordinates of places visited or passed (Chen et al., 2016). Accordingly, positional accuracy of MPP data (approximately 200 m to 1000 m) is insufficient to cover shorter distances than the coverage distances. Moreover, since sampling frequency depends on the frequency of the actions (e.g., calling, texting), a mobile phone in idle mode generates little information. Wide acceptance of MPP data in active transportation schemes may require greater efforts, such as data science or joint use with other sources (e.g., joint use of MPP and Wi-Fi signals by Mun et al., 2008).

4.2. Data validity

The widely recognized challenges of emerging data are associated with data reliability and validity in terms of representativeness of the general population and sampling bias skewed toward a certain population group. A possible question that may arise about emerging data concerns how well the data represent the general active travel population. For instance, although over 10% of the adults in the United States can be reflected in the LBS data sets provided by StreetLight Data (10% is much higher than the approximately 1% reflected in the National Travel Household Survey in the United States), the pure rate of sampling for only pedestrians and bicyclists may be much lower. Further, fitness-tracking apps do not necessarily sample all of the general population. For example, Strava Metro typically represents 1–5% of total bike volumes of a city, county, or state (Y. Wang et al., 2018). However, if information of interest is relative levels of activity density (e.g., where the most heavily used streets are, when the peak periods and busiest hours are), then concerns about representativeness can be alleviated to some extent.

In addition, sampling bias is an obvious concern across both emerging data sets. First, studies using mobile phone data have already acknowledged variations in levels of phone ownership and use across the population (Rojas IV et al., 2016). Mobile phone (and app) users tend to be economically active, tech savvy, and younger members of society (Milne and Watling, 2019). Second, sampled populations for fitness-oriented apps are likely to be certain population groups, such as males, younger individuals, and opt-in app users (Blanc and Figliozzi, 2016; Hochmair et al., 2019; Hood et al., 2013; Jestico et al., 2016). When using mobile-based emerging data, it is necessary to consider how to deal with a specific population that is less likely to use a smartphone

(e.g., older users, minority community members with no or little economic/technological access to mobile phones). In terms of bike-share programs, casual users might be different from regular membership users; casual trips are more likely to be for tourism (Buck et al., 2013; Wergin and Buehler, 2017). These sampling issues, therefore, must be expected in advance, and their probable influence on study results and policy implications must be deliberately considered.

While sampling limitations may restrict wide applications of emerging data, fusing multiple data sets may be a feasible way to overcome the challenges. The most straightforward data fusion is cross-use of emerging data with traditional sources (e.g., observed counts and travel survey results). For instance, bicycle flows collected via fitness-tracking apps can be validated by field counts (Saad et al., 2019) or complemented by intercept surveys (Heesch et al., 2016). Beyond the simple level of data fusion, when multiple data sets are combined, more comprehensive and reliable insights can be achieved given that different sources cover different types of travel activities, journey purposes, and spatial variations (e.g., bike-share systems are more favorably used by visitors near attraction sites, and fitness- tracking app users are more likely to be recreation-oriented). Proulx and Pozdnukhov (2017) exemplified direct travel demand estimation over the entire bicycle network in San Francisco, California, fusing various data sets: Strava Metro, bike-share program usage, manual and automated counts, and two regional travel demand model outcomes. This research area likely offers great potential to improve the capabilities of emerging mobile-generated data. As data fusion methods become more advanced and solid, the limitations of emerging data related to under- and overrepresentation can be expected to be overcome to some extent.

4.3. Privacy protection

Protecting individual privacy is an inherent issue in the collection and use of passive data. Activity records collected through passive monitoring and handed over to third parties have limited access to personal information due to the risk of a privacy leak. To forestall this problem, data providers implement several privacy control managements. For instance, individuals are given the option to not share their information with others and not include their records on data storage servers. Second, only consented records are stored and released. Third, traveler/trip information is only provided in an aggregated form in which individual identities are disguised. In these privacy protection settings, using the deidentified data does not require a research ethics review, such as by an Institutional Review Board, which is a group that reviews research details to ensure protection of the rights of human subjects. Despite the settings that secure individual privacy, there may still be a possibility for individuals and movement patterns to be identified in areas that do not have many non-motorized travelers or trips. Due to such risk, one commercial data vendor, Strava Metro, announced it would no longer provide routes with counts of fewer than three users. Preserving privacy may be a serious issue when raw GPS trajectories need to be postprocessed by researchers. The most commonly applied approach is to remove identifiers of individuals, and beyond that basic method, more complicated approaches have been suggested by scholars, including the obfuscation method (Wightman et al., 2011) and the anonymity method (Calabrese et al., 2015).

4.4. Contextual information

The lack of detailed contextual information is another challenge for most emerging data sets (except cases that collect and release such information under an agreement of the data donors) in light of the importance of understanding parameters behind travel behavior. Non-purpose-oriented emerging data are unlikely to contain all the desired information. As mentioned previously, activity records collected through passive monitoring and handed over to third parties have limited access to personal information due to the risk of a privacy leak; to preclude that risk, data providers operate privacy protection policies, such as prior consent for the collection and use of personal data and aggregating data with anonymization. These privacy protection approaches beget trade-offs in sample size and data quality. In other words, information on those individuals who do not grant third-party access to their records is filtered out, leading to a reduction in sample size. Aggregating athlete/trip counts makes it inevitable that details on travel activities at the individual level will be lost, and anonymization (lack of demographic information) precludes the opportunity to scale sample bias. Accordingly, most mobile-generated emerging data are unlikely to contain information pertaining to trip purposes, sociodemographic specifications, or modes used (for mode-unspecified emerging data) at the individual level, all of which is usually collected in traditional travel surveys. While a certain type of emerging data set (e.g., Strava Metro data) offers demographic statistics (e.g., the total number of trips made by age group and gender in the contracted study area), trip summaries (e.g., average trip time and distance), and simplified reason for trips (e.g., commuting or noncommuting), these data are not available at the individual level due to privacy issues (Romanillos et al., 2016). Because the limited information provided typically reduces potential insights, it is recommended that researchers carefully check which information of interest is available or unavailable at the stage of framing the research goals.

4.5. Non-free data

In general, raw GPS points or LBS records collected via mobile devices require complex data-mining processes to clean and transform them into a user-friendly format (e.g., positioning GPS points to street segments). Compared to traditional monitoring methods, some commercially available emerging data sets have a higher level of readiness and preparedness. The data sets provided by data vendors (e.g., StreetLight Data and Strava Metro) do not require completion of initial data-mining processes; under contracted data licenses, access to data sets already cleaned, smoothed, and matched to network geometry by analytic teams is enabled. The cost of licenses is subject to change according to various parameters, including the scale of study area of interest, time range of the data needed, and level of granularity and features in the data set. For example, the estimated cost for Virginia of Strava Metro data was $300,000 for one year (2.5 million activities in 2016 by 110,000 users; Ohlms et al., 2018). This level of expenditure may not be feasible under limited budgets, thus restricting access to the ready-made data sets. By contrast, when access to the raw GPS trajectories are available, then researchers do not need to pay for purchasing the data, but they must complete a series of data-mining processes to systematically extract information from the GPS points. Usually, these processes require analytic expertise, computational skills, and hands-on experience, which may also require human or financial resources to some extent. Therefore, the cost of obtaining and utilizing emerging data

should be compared with other monitoring methods.

4.6. Walking uncertainty

Compared to bicycle data, emerging pedestrian data have been notably less used by transportation researchers and practitioners. A high level of variability and uncertainty in walking trip data may be the reason for diminished data reliability and applicability. However, limited applications of emerging pedestrian data in part stem from the unique features of walking, not solely from the drawbacks of emerging data. Since walking trips are the most divergent among all transportation modes (including automobiles, bikes, and trains) in terms of trip purpose, time, distance, and route, it is imperative to have a large enough sample size to draw significant implications on the characteristics. Moreover, it is challenging to monitor every walking trip through a mobile device (though this is a challenge for not only emerging data but also for traditional data) and to estimate how many walking trips are omitted from the collected data sets. A potential feasible approach is to merge multiple databases (e.g., site counts and fitness-tracking app records; Turner et al., 2019) until more progress can be made to overcome such shortcomings. Regardless, putting more effort into using emerging data for pedestrian monitoring and filling the gap between pedestrian and bicycle data research is essential.

5. Summary and discussion

This summary section recapitulates the information discussed in Sections 2–4, putting more focus on synthesizing the current state of emerging data within the context of non-motorized modes in terms of data availability and accessibility, data types available, key challenges to be considered, and current suitability for application.

Because different emerging data sources have different fundamental mechanisms, variations in data attributes, advantages, challenges, and degree of power to spur applications in the non-motorized transportation field exist. In brief, a higher passivity in the data collection process is adversely associated with the overall capability of distinguishing pedestrians and bicyclists. Consequently, mode-specified sources currently have more viable options for non-motorized monitoring.

Among the four suggested mode-specified data categories, fitness- tracking apps—more precisely, Strava Metro data—have been most widely, vigorously welcomed by transportation agencies and researchers. Several reasons possibly explain the popularity. First, it is a ready-made product that saves time and resources when preprocessing raw GPS trajectories. Second, Strava is one of the most popular pioneer fitness trackers globally, which means a number of app users contribute to an accumulated database and there are many customers around the globe. Third, it is delivered in flexible formats for data management—useful geometries (point, segment, OD polygons), shapefiles for customizing analyses in a GIS software, and finer spatial/temporal resolution—all of which are broadly applicable from a small scale to large scale and compatible with other data sources. Fourth, it offers extensive data coverage in time and space at a relatively reasonable cost. These advantages have greatly promoted research endeavors and analytic methodologies and thus have diversified available scenarios over the past

five years. However, despite the widespread use of the data, studies that need each trip's route information and detailed individual user profiles have not been attempted and may not even be feasible because of privacy protection policies. In addition, underrepresented general populations and overrepresented certain populations are the most frequently mentioned sampling drawbacks that must be considered when framing research questions and interpreting results.

Although applications have lagged behind due to its recent service launch, StreetLight Data provides access to data sets that have significantly comparable attributes to Strava Metro data overall. Multi-app LBS data by StreetLight Data supply OD-based travel demand aggregated or averaged traffic parameters (e.g., volume, distance, time, and speed) for selected settings (time and geometry), and deduced contextual information (e.g., trip type and income levels). The multi-app LBS data product is superior to the other data sources in terms of sampling bias since it is integrated and validated with various sources of data sets (e.g., active mode app, in-road sensor, video reader, and traditional travel survey). This progress in mode-unspecified emerging data is encouraging to many agencies and scholars, but the app still fails to meet the data demand at the individual person/trip level like Strava Metro.

However, if researchers have access to raw GPS tracepoints, it is possible to obtain individual trip information that the third-party data vendors do not offer. For example, since regional bicycle-tracking apps developed for public purposes collect and store the GPS trajectories (and some user profiles, provided agreement has been given), researchers can handle the raw data and directly manipulate it into a tailor-made database, such as OD trip routes by each user. Thus, regional bike-tracking app data in general guarantee a higher dimension of data quality in terms of information details and more advanced applications over the other categories of emerging data. One of the great utilizations of such discrete trip-level information is developing route-choice models that can be applied to travel demand precision, assignment, and calibration for the entire network, which is very limited unless access to a series of GPS trajectories is ensured. However, efforts to develop the apps, limited scalability to a larger area beyond the region, sampling problems (low sampling rates and bias), and technical burdens posed to the data end users remain as challenges.

Although not belonging to mainstream methods to monitor general pedestrians and cyclists, bike-share programs and user-feedback inventory provide important insights that serve monitoring efforts in a more exhaustive manner. Available data types depend heavily on the operation system, but possible databases from bike-sharing programs include information on membership type (and other user profiles), check-in/turn-in location and time, and trip routes (if GPS traces are accessible). By analyzing the data sets, transportation agencies and municipalities can identify public bike- use patterns for better operation and further integration with public transportation and tourism. User-feedback inventory also enhances authorities' ability to monitor and address problems that community members encounter while walking and bicycling, such as unsafe intersections, potholes, and damaged streetlamps.

While Wi-Fi and Bluetooth signaling also collect human movement trajectories (time, geolocation, and MAC address), optimal application is limited in the aspect of spatial scale. Because the discoverable radius is roughly between

35 m and 100 m, detecting movements beyond that range requires a deployment of a series of sensors. In other words, the density of sensors decides spatial scalability, and thus applications for non-motorized modes have mostly been constrained to counting people and estimating dwell time (and other time categories) in small spaces, short corridors, and buildings. The least promising emerging data source currently is MPP data be- cause this source suffers from uncertainty in mode detection. Although available data types are similar to those of Strava Metro and StreetLight Data, usage has concentrated on motorized vehicles, and data are still not possible for walking and bicycling monitoring due to difficulties in differentiating from other types of trips and to coarse spatial/temporal sampling resolution.

Growing availability of various emerging data sources over time, space, and volume have reinforced studies conducted using traditional data and methodologies and further broadened the opportunities for research, which have been limited due to inadequate data availability and quality. However, challenges must still be dealt with to more extensively utilize the emerging data. The challenges may prevent immediate adoption of emerging data without hesitation, but cross-use of multiple data sets or validation with traditional data sources has great potential to address the challenges (in part or completely) and expand available applications. Even if these shortcomings remain unsolved, cities and municipalities that have no reliable monitoring systems can take advantage of the new source of data to improve their capabilities in understanding pedestrians and bicyclists.

6. Conclusion

This paper reviewed current emerging data collected through mobile devices for pedestrian and bicyclist monitoring. The review included an examination of both mode-unspecified and mode-specified emerging data and recognized their potential to enrich pedestrian and bicycle studies. While mode-unspecified data sources do not target only pedestrians and bicyclists during data collection (but rather all populations carrying mobile devices), mode-specified data sets at least know which mode is being monitored. The data collection mechanism differentiates the pace at which transportation planners and agencies adopt the emerging data. So far, emerging monitoring tools for non-motorized travelers are more concentrated on the mode-specified sources, especially for bicycling. How- ever, while bicycle studies have vigorously applied the mode-specified emerging data, relatively fewer applications can be found for pedestrians in the existing literature.

Although this paper provided a broad overview of emerging data, the aim of the study did not allow the authors to explain each data source in detail and discuss the differences between sources in depth, even though these discussions are valuable since the use of mobile devices to collect movement traces is becoming more elaborate and diverse. Future research may expand on the independent discussion of specific data sources and applications or compare how diverse data sets are applied to different non- motorized studies according to their characteristics (e.g., the level of accuracy to detect non-motorized trips). Also, an intensified discussion of how to overcome the suggested challenges related to emerging data with a focus on non-motorized travelers will be of value.

Acknowledgments

The authors would like to acknowledge the valuable comments of two anonymous reviewers on an earlier version of this paper. The authors would also like to thank Dawn Herring of TTI for her editorial review.

Funding

The project was partially funded by the Safety through Disruption (Safe-D) National UTC, a grant from the U.S. Department of Transportation's University Transportation Centers Program (Federal Grant Number: 69A3551747115).

References

Abedi, N., Bhaskar, A., Chung, E., 2013. Bluetooth and Wi-Fi MAC Address Based Crowd Data Collection and Monitoring: Benefits, Challenges and Enhancement. Australasian Transport Research Forum, Brisbane.

Anderson, I., Muller, H., 2006. Practical activity recognition using GSM data. Technical Report CSTR-06-016. Department of Computer Science, University of Bristol.

Apasnore, P., Ismail, K., Kassim, A., 2017. Bicycle-vehicle interactions at mid-sections of mixed traffic streets: examining passing distance and bicycle comfort perception. Accid. Anal. Prev. 106, 141–148. https://doi.org/10.1016/j.aap.2017.05.003.

Barcelö, J., Montero, L., Marqués, L., Carmona, C., 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on Bluetooth traffic monitoring. Transp. Res. Rec. J. Transp. Res. Board 2175, 19–27. https://doi.org/10.3141/2175-03.

Bhaskar, A., Qu, M., Chung, E., 2014. Bluetooth vehicle trajectories by fusing Bluetooth and loops: motorway travel time statistics. IEEE T Intell Transp 16, 113–122. https://doi.org/10.1109/TITS.2014.2328373.

Blanc, B., Figliozzi, M., 2016. Modeling the impacts of facility type, trip characteristics, and trip stressors on cyclists' comfort levels utilizing crowdsourced data. Transp. Res. Rec. J. Transp. Res. Board 2587, 100–108. https://doi.org/10.3141/2587-12.

Bonnel, P., Hombourger, E., Olteanu-Raimond, A., Smoreda, Z., 2015. Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations. 10th International Conference on Transport Survey Methods, Leura, Australia, pp. 381–398. https://doi.org/10.1016/j.trpro.2015.12.032.

Boss, D., Nelson, T., Winters, M., Ferster, C.J., 2018. Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. J. Transp. Health 9, 226–233. https:// doi.org/10.1016/j.jth.2018.02.008.

Bowman, N., n.d. Big Data for Transportation Models. https://connect.ncdot.gov/projects/planning/TPB%20Model%20User%20Groups/NCMUG_2016-11-16_Present_W2_Street- Light_11-18_NealBowman.pdf. Accessed August 11, 2018.

Buck, D., Buehler, R., Happ, P., Rawls, B., Chung, P., Borecki, N., 2013. Are bikeshare users different from regular cyclists? Transp. Res. Rec. J. Transp. Res. Board 2387, 112–119. https://doi.org/10.3141/2387-13.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. C Emerg. Technol. 26, 301–313. https://doi.org/10.1016/j.trc.2012.09.009.

Calabrese, F., Ferrari, L., Blondel, V.D., 2015. Urban sensing using mobile phone network data: a survey of research. ACM Computing Surveys 47 (2), 25. https://doi.org/ 10.1145/2655691.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. C Emerg. Technol. 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing cell phone location data

for urban travel: current methods, limitations, and opportunities. Transp. Res. Rec. J. Transp. Res. Board 2526, 126–135. https://doi.org/10.3141/2526-14.

Conrow, L., Wentz, E., Nelson, T., Pettit, C., 2018. Comparing spatial patterns of crowdsourced and conventional bicycling datasets. Appl. Geogr. 92, 21–30. https://doi.org/10.1016/j.apgeog.2018.01.009.

Day, C.M., Premachandra, H., Bullock, D.M., 2011. Rate of pedestrian signal phase actuation as a proxy measurement of pedestrian demand. Transportation Research Board 90th Annual Meeting. DC, Washington https://doi.org/10.5703/1288284316056.

Dhakal, N., Cherry, C.R., Ling, Z., Azad, M., 2018. Using CyclePhilly data to assess wrong-way riding of cyclists in Philadelphia. J. Saf. Res. 67, 145–153. https://doi.org/10.1016/j. jsr.2018.10.004.

Du, Y., Deng, F., Liao, F., 2019. A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system. Transp. Res. C Emerg. Technol. 103, 39–55. https://doi.org/10.1016/j.trc.2019.04.006.

Faghih-Imani, A., Eluru, N., 2016. Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: a case study of New York CitiBike System. J. Transp. Geogr. 54, 218–227. https://doi.org/10.1016/j.jtrangeo.2016.06.008.

Faghih-Imani, A., Eluru, N., El-Geneidy, A.M., Rabbat, M., Haq, U., 2014. How land-use and urban form impact bicycle flows: evidence from the Bicycle-Sharing System (BIXI) in Montreal. J. Transp. Geogr. 41, 306–314. https://doi.org/10.1016/j.jtrangeo.2014.01.013.

Griffin, G.P., Jiao, J., 2015. Where does bicycling for health happen? Analyzing volunteered geographic information through place and plexus. J. Transp. Health 2, 238–247. https://doi.org/10.31235/osf.io/5gy3u.

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. J. Transp. Health 3, 366–376. https://doi.org/10.1016/j.jth.2016.06.006.

Hochmair, H.H., Zielstra, D., Neis, P., 2015. Assessing the completeness of bicycle trail and lane features in OpenStreetMap for the United States. Trans. GIS 19, 63–81. https:// doi.org/10.1111/tgis.12081.

Hochmair, H.H., Bardin, E., Ahmouda, A., 2019. Estimating bicycle trip volume for Miami- Dade County from Strava tracking data. J. Transp. Geogr. 75, 58–69. https://doi.org/ 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. Transp. Advance online publication. https://doi.org/10.1007/s11116-019-09988-4.

Hood, J., Sall, E., Charlton, B., 2013. A GPS-based bicycle route choice model for San Francisco. California. Transp. Letters 3, 63–75. https://doi.org/10.3328/ tl.2011.03.01.63-75.

Huntsinger, L.F., Donnelly, R., 2014. Reconciliation of regional travel model and passive device tracking data. Transportation Research Board 93rd Annual Meeting. DC, Washington.

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data.

J. Transp. Geogr. 52, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006.

Knowledge Based Systems Inc., 2018. Decentralized, Public, and Mobile-based Sidewalk Inventory Tool. https://www.sbir.gov/sbirsearch/detail/1168625, Accessed date: 13 August 2018.

Kurkcu, A., Ozbay, K., 2017. Estimating pedestrian densities, wait times, and flows with Wi-fi and Bluetooth sensors. Transp. Res. Rec. J. Transp. Res. Board 2644, 72–82. https://doi. org/10.3141/2644-09.

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. https://rosap.ntl.bts.gov/view/dot/36713/dot_36713_ DS1.pdf.

Le Dantec, C.A., Asad, M., Misra, A., Watkins, K.E., 2015. Planning with crowdsourced data: rhetoric and representation in transportation planning. 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1717–1727. https://doi. org/10.1145/2675133.2675212.

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. Intern. J. Environ. Res. Public Health 16 (371), 1–20. https://doi.org/10.3390/ijerph16030371.

Lee, R.J., Sener, I.N., Mullins III, J.A., 2016. An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. Transp. Letters 8, 181–193. https://doi.org/10.1080/19427867.2015.1106787.

Li, P., Mirchandani, P.B., Zhang, L., Zhang, L., 2018. Highway Travel Time Estimation with Captured In-vehicle Wi-Fi Mac Addresses: Mechanism, Challenges, Solutions and Applications. Final Report, SOLARIS Consortium, Tier 1 University Transportation Center for Advanced Transportation Education and Research, Department of Civil and Environmental Engineering, University of Nevada. https://www.unr.edu/Documents/engineering/solaris/Task%20C%20Report%20for%20SOLARIS.pdf.

Lin, M., Hsu, W.J., Lee, Z.Q., 2013. Detecting modes of transport from unlabeled positioning sensor data. J. Locat. Based Serv. 7, 272–290. https://doi.org/10.1080/ 17489725.2013.819128.

Malleson, N., Vanky, A., Hashemian, B., Santi, P., Verma, S.K., Courtney, T.K., Ratti, C., 2018. The characteristics of asymmetric pedestrian behavior: a preliminary study using passive smartphone location data. Trans. GIS 22, 616–634. https://doi.org/10.1111/tgis.12336.

McCahill, C., 2017. Improving Last-mile Connections to Transit: An Exploration of Data and Analysis Tools. https://www.cnu.org/sites/default/files/2017_NewUrbanResearch_ImprovingLastMileConnectionstoTransit_McCahill.pdf.

Meneses, F., Moreira, A., 2012. Large scale movement analysis from WiFi based location data. The 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, Australia. https://doi.org/10.1109/IPIN.2012.6418885.

Milne, D., Watling, D., 2019. Big data and understanding change in the context of planning transport systems. J. Transp. Geogr. 76, 235–244. https://doi.org/10.1016/j. jtrangeo.2017.11.004.

Morency, C., Trépanier, M., Paez, A., Verreault, H., Faucher, J., 2017. Modelling bike sharing usage in Montreal over 6 years. No. CIRRELT-2017-33, Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation.

Mun, M., Estrin, D., Burke, J., Hansen, M., 2008. Parsimonious mobility classification using GSM and WiFi traces. 5th Workshop on Embedded Networked Sensors (HotEmNets), Charlottesville, Virginia.

Nelson, T.A., Denouden, T., Jestico, B., Laberee, K., Winters, M., 2015. BikeMaps.org: a global tool for collision and near miss mapping. Front. Public Health 3 (53), 18. https://doi.org/ 10.3389/fpubh.2015.00053.

Nikolic, M., Bierlaire, M., 2017. Review of transportation mode detection approaches based on smartphone data. 17th Swiss Transport Research Conference, EPFL CONF-229181.

Nolan, R.B., Smart, M.J., Guo, Z., 2018. Bikesharing trip patterns in New York City: associations with land use, subways, and bicycle lanes. Intern. J. Sustain. Transp. https://doi. org/10.1080/15568318.2018.1501520.

Ohlms, P.B., Dougald, L.E., MacKnight, H.E., 2018. Assessing the Feasibility of a Pedestrian and Bicycle Count Program in Virginia (VTRC 19-R4). Virginia Transportation Research Council.

Oosterlinck, D., Benoit, D.F., Baecke, P., de Weghe, N.V., 2017. Bluetooth tracking of humans in an indoor environment: an application to shopping mall visits. Appl. Geogr. 78, 55–65. https://doi.org/10.1016/j.apgeog.2016.11.005.

OpenStreetMap, 2018. 1 Million Map Contributors! https://blog.openstreetmap.org/2018/ 03/18/1-million-map-contributors/, Accessed date: 13 April 2019.

Pak, B., Chua, A., Vande Moere, A., 2017. FixMyStreet Brussels: socio-demographic inequality in crowdsourced civic participation. J. Urban Technol. 24 (2), 65–87. https://doi.org/ 10.1080/10630732.2016.1270047.

Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O., 2017. Transportation mode detection—an in-depth review of applicability and reliability. Transp. Rev. 37, 442–464. https://doi.org/ 10.1080/01441647.2016.1246489.

Proulx, F., Pozdnukhov, A., 2017. Bicycle Traffic Volume Estimation Using Geographically Weighted Data Fusion. http://faculty.ce.berkeley.edu/pozdnukhov/papers/Direct_Demand_Fusion_Cycling.pdf.

Qin, H., Curtin, K.M., Rice, M.T., 2018. Pedestrian network repair with spatial optimization models and geocrowdsourced data. GeoJournal 83, 347–364. https://doi.org/10.1007/ s10708-017-9775-x.

Rojas IV, M.B., Sadeghvaziri, E., Jin, X., 2016. Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data. Transp. Res. Rec. J. Transp. Res. Board 2563, 71–79. https://doi.org/10.3141/2563-11.

Roll, J., 2018. Bicycle count data: What is it good for? A study of bicycle travel activity in Central Lane Metropolitan Planning Organization. FHWA-OR-RD-18-16. Oregon Department of Transportation.

Romanillos, G., Austwick, M.Z., Ettema, D., De Kruijf, J., 2016. Big data and cycling. Transp. Rev. 36, 114–133. https://doi.org/10.1080/01441647.2015.1084067.

Ryus, P., Ferguson, E., Laustsen, K.M., Schneider, R.J., Proulx, F.R., Hull, T., Miranda-Moreno, L., 2014. Guidebook on pedestrian and bicycle volume data collection. NCHRP Report 797. National Cooperative Highway

Research Program. https://doi.org/10.17226/22223.

Saad, M., Abdel-Aty, M., Lee, J., Cai, Q., 2019. Bicycle safety analysis at intersections from crowdsourced data. Transportation Transp. Res. Rec. J. Transp. Res. Board. Advance online publication. https://doi.org/10.1177/0361198119836764.

Saha, D., Alluri, P., Gan, A., Wu, W., 2018. Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. Accid. Anal. Prev. 118, 166–177. https://doi.org/10.1016/j.aap.2018.02.014.

Schauer, L., Werner, M., Marcus, P., 2014. Estimating crowd densities and pedestrian flows using Wi-Fi and Bluetooth. The 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 171–177.

Selala, M.K., Musakwa, W., 2016. The potential of Strava data to contribute in non-motorized transport (NMT) planning in Johannesburg. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLI-B2, pp. 587–594. https://doi. org/10.5194/isprs-archives-xli-b2-587-2016.

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. J. Transp. Saf. Secur. https://doi.org/10.1080/19439962.2019.1645778.

Smith, A., 2015. Crowdsourcing Pedestrian and Cyclist Activity Data. Pedestrian and Bicycle Information Center White Paper Series. http://www.pedbikeinfo.org/-cms/downloads/ PBIC_WhitePaper_Crowdsourcing.pdf.

State Smart Transportation Initiative, 2017. Understanding trip-making with big data. https://www.ssti.us/wp/wp-content/uploads/2017/07/SSTI_Connecting_Sacramento_ Tripmaking.pdf, Accessed date: 13 August 2018.

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. Accid. Anal. Prev. 83, 132–142. https://doi.org/10.1016/j.aap.2015.07.014.

StreetLight Data, 2018. Real-world Big Data for Active Transportation Planning. https://3yemud1nnmpw4b6m8m3py1iy-wpengine.netdna-ssl.com/wp-content/uploads/2019/ 01/SLD-Flyer-CalTrans-v4.pdf.

StreetLight Data, 2019. StreetLight Data Offers Complete, Stand-alone Dashboard of Bicycle and Pedestrian Transportation Metrics. https://www.streetlightdata.com/bicycle-and-pedestrian-metrics-alternative-to-iot-bike-counters/, Accessed date: 3 December 2019.

Sun, Y., Mobasheri, A., 2017. Utilizing crowdsourced data for studies of cycling and air pollution exposure: a case study using Strava data. Intern. J. Environ. Res. Public Health 14, 274. https://doi.org/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. J. Transp. Health 6, 93–104. https://doi.org/10.1016/j.jth.2017.06.004.

Traunmueller, M.W., Johnson, N., Malik, A., Kontokosta, C.E., 2018. Digital footprints: using WiFi probe and locational data to analyze human mobility trajectories in cities. Comput. Environ. Urban. 72, 4–12. https://doi.org/10.1016/j.compenvurbsys.2018.07.006.

Turner, S., Sener, I.N., Martin, M., Das, S., Shipp, E., Hampshire, R., Fitzpatrick, K., Molnar, L., Wijesundera, R., Colety, M., Robinson, S., 2017. Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities. FHWA Report SA-17-041. U.S. Department of Transportation.

Turner, S., Benz, R., Hudson, J., Griffin, G., Lasley, P., 2019. Improving the amount and availability of pedestrian and bicyclist count data in Texas. FHWA Report TX-19/0-6927-R1. Texas Department of Transportation.

Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., 2010. Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records. Intelligent Transportation Systems (ITSC), 13th International IEEE Conference, Funchal, Portugal, pp. 318–323. https://doi.org/10.1109/itsc.2010.5625188.

Wang, Z., He, S.Y., Leung, Y., 2018a. Applying mobile phone data to travel behaviour research: a literature review. Travel Behav. Society 11, 141–155. https://doi.org/ 10.1016/j.tbs.2017.02.005.

Wang, Y., Monsere, C.M., Chen, C., Wang, H., 2018b. Development of a crash risk scoring tool for pedestrian and bicycle projects in Oregon. Transp. Res. Rec. J. Transp. Res. Board 2672, 30–39. https://doi.org/10.1177/0361198118794285.

Wergin, J., Buehler, R., 2017. Where do bikeshare bikes actually go? An analysis of Capital Bikeshare trips using GPS data. Transp. Res. Rec. J. Transp. Res. Board 2661, 12–21. https://doi.org/10.3141/2662-02.

Wightman, P., Coronell, W., Jabba, D., Jimeno, M., Labrador, M., 2011. Evaluation of location obfuscation techniques for privacy in location-based information systems. 2011 IEEE Third Latin-American Conference on Communications, pp. 1–6. https://doi.org/10.1109/latincom.2011.6107399.

Wu, D., Xia, L., Geng, J., 2018. Heading estimation for pedestrian dead reckoning based on robust adaptive Kalman filtering. Sensors 18, 1970. https://doi.org/10.3390/ s18061970.