Comprehensive Review of Travel Behavior and Mobility Pattern Studies That Used Mobile Phone Data

Mario B. Rojas IV, Eazaz Sadeghvaziri, and Xia Jin

Traditional data acquisition methods, such as surveys and diaries, used in transportation studies have become burdensome and inefficient in comparison to the emerging sources of passively collected data. These newer data sources have the ability to improve data quality and accuracy and the potential to complement conventional data. This paper presents a comprehensive review of studies that have utilized passively collected data, such as data from personal or vehicle GPS devices, mobile phone network data, and—more recently—smartphone GPS data. This review focuses on the data-processing algorithms that have been used to derive travel information from trajectory traces, as well as the variety of applications that have been conducted on the basis of these data. Some applications of these data have included origin–destination estimation, real-time traffic monitoring, and human mobility pattern analysis. Although passively collected data have great potential, issues with possible sample bias and a lack of demographic data require further research. This study may help people interested in employing these data to understand better the current practices, as well as the potential and the challenges associated with the data.

As technologies have advanced, emerging data sources from passive collection methods have shown promise in helping transportation professionals better understand people's movements through space and time. Traditional travel surveys are plagued by low response rates, high respondent burden, and significant implementation costs (*1*). Passively collected data, such as GPS data, mobile network data, and cell phone GPS data, may have the ability to supplement or complement traditional household travel surveys and overcome existing issues.

These data present different opportunities to reflect aspects of people's travel patterns; the data also present challenges in collection and processing. This paper draws a typology of the available data sources and covers different types of data, processing methods, potential applications, and limitations. This review may help people who are interested in employing these data to better understand current practices and the potential and challenges associated with the data.

M. B. Rojas IV and E. Sadeghvaziri, EC3725, and X. Jin, EC3603, Department of Civil and Environmental Engineering, Florida International University, 10555 West Flagler Street, Miami, FL 33174. Corresponding author: X. Jin, xjin1@fiu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2563, Transportation Research Board, Washington, D.C., 2016, pp. 71–79.
DOI: 10.3141/2563-11

PASSIVELY COLLECTED DATA

GPS Data

The first advancement in travel surveys saw the inclusion of GPS applications. GPS devices can provide accurate location information anywhere in the world through the use of satellites in medium-Earth orbit.

A device's position is calculated every 1 to 4 s. The position is calculated on the basis of the distance between the device and the satellites. Connecting to two satellites provides latitude and longitude. Incorporating a third satellite enables the calculation of altitude; the use of additional satellites increases accuracy. GPS has a horizontal accuracy of ~3 m and a vertical accuracy of ~5 m, 95% of the time (*2*). GPS location accuracy suffers when the signal is obstructed, such as in a tunnel or an urban canyon (*3*).

A typical raw GPS data set includes the time stamp, the latitude, the longitude, the altitude, and the speed of each record. The data sets may also include a heading and a measure of accuracy.

A significant body of research has examined the use of GPS data and their capability to supplement or complement household travel surveys (*4*). In most studies, a GPS device was fixed to a participant's vehicle or a participant carried the device daily (*5*). Although these studies indicated that GPS provided detailed travel trajectory data with a sufficient level of accuracy, GPS also had some limitations. The cost of purchasing the GPS units and administering the survey (mailing the units to the participants and retrieving the units) can severely limit the scale and duration of this type of survey. There is also a certain level of respondent burden. For example, a participant could forget to charge the device or leave it at home, either of which would render the device useless.

Mobile Network Data

Other attempts to improve travel surveys saw the incorporation of mobile network data, most commonly the call detail records (CDRs). Similarly to GPS, the location of the cell phone is calculated on the basis of its distance from the surrounding towers. The spacing, the number of towers, and the signal strength directly affect the accuracy of the data. Simply, data are only recorded when the phone is active, such as during a phone call or when sending a message. Through this method it is possible to locate a phone within 50 to 300 m (*6*).

As a cell phone moves, the signal switches to the nearest and strongest of the towers' signals. However, a phone does not need to move to switch towers. A phone can switch between towers, or "oscillate," as a result of network policies on performance optimization or the proximity to a competing cellular tower with equal strength. In travel studies, oscillation can cause the data to indicate false movements; a real movement could also be misinterpreted as an oscillation on the basis of the repetitive nature of the movement.

A typical CDR data set contains the caller ID, the time stamp, the duration of the call or other activity, the longitude, and the latitude. Other data, such as the call receiver's ID, may also be available. As a result of privacy concerns, these IDs are always anonymized, and the formatting varies across carriers.

Because of the proliferation of cellular phones, a large sample of data can be obtained at a comparatively low cost. Recent data indicated a penetration rate of 120% in developed and 92% in developing countries (*7*). These results indicated that a person, particularly in a developed country, may possess more than one cell phone. Caution should be taken, as the sampling of mobile phone networks could introduce bias or overrepresent participants with multiple phones. Not only is the scale of the data sets enormous, this form of data collection also eliminates the respondent's burden; most respondents are not even aware the data is being recorded. However, CDR data tend to be less accurate than GPS data. Because data are only recorded when the phones are in use, CDR data are less frequent and irregular and could therefore leave significant gaps in trajectory traces and complicate the application of the data. However, many studies have used these data successfully.

Cell Phone GPS Data

Recent efforts to improve travel surveys became possible with the advent of cell phone GPS or "assisted GPS" data. This technology merges mobile phone network and traditional GPS data. Similarly to mobile network data, cell phone GPS data have the potential for large-scale applications and a somewhat reduced respondent burden. For example, the recording of the data can reduce battery life, and the retrieval of such data may be burdensome. This source could also cause sample bias. Cell phone GPS data also have an accuracy similar to GPS data of 9 m (*3*).

A cell phone's position is calculated by triangulation. Data points can be recorded through wireless fidelity, GPS satellite, or mobile networks; certain phones give users control over how the data are recorded to help conserve power. Unlike with mobile network data, it is possible to track the phone when it is not being used, and it can be tracked without a cell phone signal if the phone is in view of the satellites.

Cell phone GPS data's recording frequencies vary, depending on movement, with fewer data records when the phone is still. For example, while the cell phone is in movement, Google location history data are usually recorded every 30 to 60 s; while the phone is still, the recording rate increases to over 1 min, but rarely exceeds 5 min. As a result, it is common for more than 1,000 points to be recorded in a day.

Google location history data can be accessed as KML or JSON files. Both file types provide information on the time stamp, the longitude, and the latitude; JSON files provide additional information, such as the accuracy level and the activity by mode. Figure 1 shows 1 day of records from the JSON file. Although the activity may not be identified correctly for each point, the map shows the quantity of data points recorded. The relatively large interval between the data points in the horizontal line (located roughly in the middle of the figure) is attributable to high vehicular speeds (50 to 60 mph). This figure shows how the data could be used to detect mobility patterns and other aspects of travel.

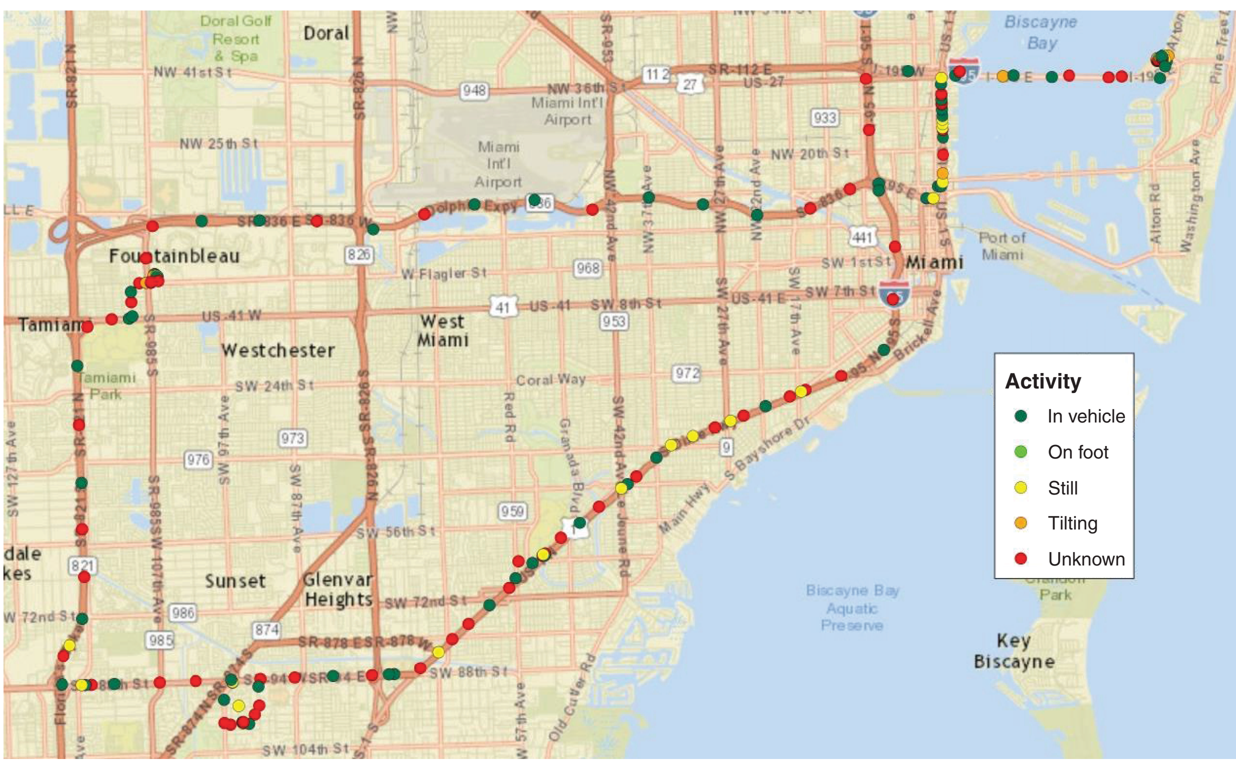

FIGURE 1 JSON file: 1 day of data.

DATA PROCESSING

This section provides an overview of the use of these data in data-processing algorithms from four perspectives: preprocessing, trip end identification, trip mode detection, and trip purpose identification. The intention is to aid understanding of the scope of the data provided and the level of effort involved in the use of these passively collected data.

Preprocessing

Although the content and use may vary, the method for data preprocessing is relatively simple. Generally, the raw data are passed through multiple filters to identify erroneous data points.

Ideally, data should contain a measure of the horizontal dilution of precession as well as the number of satellites, but these measures are usually only found in GPS data sets. When provided with these measures, studies were able to assess the accuracy of all points within the data set directly and filter out data points that had a poor horizontal dilution of precession value or a small number of satellites (*8*). Unfortunately, the horizontal dilution of precession and the number of satellites are not always provided; in these cases, other criteria need to be established to clean the data.

Preprocessing is commonly accomplished through criterion-based data elimination. In the literature, the speed, location, and local relative behavior of the data points were commonly used criteria for data elimination. The most common criterion was to remove data points with unreasonably high speeds; such was the case in Nour et al. (*9*), Huntsinger and Ward (*10*), and Wang et al. (*11*). Although rare, certain freight applications excluded slow speeds (*5*).

The implementation of geographic information system (GIS) tools in preprocessing was also useful. Researchers were able to easily remove data that were outside of the study area or the desired buffer zones [e.g., Sharman and Roorda (*12*)].

At the disaggregate level, the consideration of the distance or time between consecutive data points has proved to be effective (*13*). Sometimes preprocessing also involved the removal of data when a sufficient number of data points could not be obtained (*5*); although labor intensive, manual checking could also be helpful.

Trip End Identification

Various attributes have been utilized in past studies to indicate that a trip has ended or that an intermediate point has been reached. When the data set was complete (no signal loss), many studies used time-constrained rules. Trip ends have been assumed when the speed was zero or very low (*8*). Other studies used the "dwell duration," or the amount of time spent at a certain location, which can be calculated as the difference in the arrival and departure times (*14, 15*). Many of these studies used conditional statements, in which both conditions must be satisfied to identify a trip end. Although these rules have been the most commonly applied, care should be taken to choose the thresholds wisely.

Many studies also used the capabilities of GIS to help identify trip ends; these capabilities included data clustering (*5*), georeferencing (*12*), and the retrieval of the first and last recorded points in a trajectory (*16*).

Some of the less often used methods included the identification of a change in the latitude, the longitude, or the constant heading threshold (*8*); the application of filters to CDR data to smooth the trajectories (*17*); probabilistic methods in which activity locations were identified on the basis of the frequency, the duration, and the time of day (*18*); and point density or other clustering methods in which points were assigned to a cluster on the basis of relative distance (*19*). Visual map checking was also noted (*8*). One study showed great potential for the application of a model-based clustering algorithm to identify clusters that were then divided into trip end clusters and travel clusters (*20*).

During instances when there was no signal, many studies used the time between consecutive points as a proxy dwell time for the detection of trip ends (*21*). For example, it is possible to assume that if the location of Point A and Point B does not change significantly over a time interval (when no signal is available), then the location is a trip end. However, if the distance between the two points changes significantly during this time interval, it is likely that the missing data are best represented by a trip rather than by a trip end.

Trip Mode Detection

The literature provided a plethora of methods that had been successfully employed to detect travel modes. The speed-based method has been applied successfully and is applicable to most situations. Commonly used criteria included the average or maximum speed (*4*), a range of average or maximum speeds (*22*), the statistical mode of speed (*8*), and the average acceleration for each mode (*23*).

GIS tools have also been used to consider built environment characteristics. For example, pedestrians must walk on links that are accessible; the duration must not exceed 60 s (*21*), and the speed should not exceed 10 km/h (*4*). Other studies created buffer zones around bus stops or rail stations and considered proximity to the stops to identify these modes (*5*).

Several other studies used probabilistic methods to determine which mode of transportation was used. Some studies developed a probability matrix (*9, 24, 25*); other studies employed the fuzzy logic method (*21*). These methods have been proved to be effective in identifying cycling and walking modes but have struggled to differentiate between motorized modes (*26*).

Machine learning is the emerging approach in this area because of the approach's high level of accuracy (*27*). Some common methods have included decision tree (*28*), Bayesian network (*29*), support vector machine (*23, 30, 31*), conditional random field (*32*), random forest (*33*), and multilayer perceptron neural network (*34, 35*). The machine learning method was the most effective, and it has been suggested that the method be used when calibration data are available (*26*).

Trip Purpose Identification

Similarly to mode detection, purpose identification has been accomplished by three general methods: criterion based, probabilistic, and machine learning. In most studies, trip purpose identification was the most difficult step. Many studies employed GIS to facilitate this process.

In the application of the criterion-based method, some studies used land use coding (*4*), and others focused on land use–purpose matching rules (*36*). The inclusion of data matching, which entailed the verification of locations, such as home, for use throughout the study, was also noted (*23*). General trip end rules, such as assigning the purpose as "shopping" if a trip end was located near a known mall, were also used but often relied on information provided by the participant (*10, 14, 15, 18, 19, 37, 38*). The time of day and the duration were also used in some studies after proximity rules were applied (*4, 39*).

Several studies employed the probabilistic approach. One study employed a multinomial logit model in high-density areas and a single deterministic matching method in low-density areas (*4*). Another explored the use of nested multinomial logit models, which were calibrated on the basis of existing, revealed participant information (*26*). Trip purpose was also determined on the basis of probabilistic calculations related to trip distance (*40*).

Machine learning, specifically the category model and decision trees, were also employed (*16, 41*). Because of their complexities, probabilistic and machine-learning methods were used less frequently. Of all the procedures, the identification of trip purpose was the area with the most room for improvement.

Table 1 presents a brief summary of the studies based on cell phone data and includes the data type (CDR and GPS), the sample size, the data-processing methods, and the major findings related to the effectiveness and accuracy of the methods used.

MODELING APPLICATIONS

The section discusses some of the applications based on passively collected data. Only studies that used CDRs or cell phone GPS appear in Table 2. In most cases, some form of validation—usually *N*-fold cross validation, participant verification or correction, or other existing data—was used to assess the accuracy of the applications.

Origin–Destination Matrix Applications

Several studies were able to apply the available data to produce different facets of origin–destination (O-D) tables. The use of different data, GPS and CDR, proved that both types of data could be used to reproduce trips accurately (*10, 42*). One study was able to estimate Florida's statewide O-D truck flows (*13*). Similar studies proposed a method that used CDR data to infer trips and then estimate large-scale O-D matrices (*11, 18, 43, 44*); these studies found that the data were most effective at the aggregate level.

In an attempt to capture more-detailed information, one study used similar CDR data in combination with Foursquare check-in data to reproduce O-D matrices (*45*). Through the incorporation of time of day data, another study found that it was possible to create O-D sample characteristics, mobile O-D flow distributions, directional patterns, and spatial analysis; flow analysis for each O-D pair was also conducted (*46*).

Traffic Monitoring Applications

Another common application encountered in the literature was traffic monitoring. Through the use of a limited GPS data sample and embedded road traffic sensors, one study explored the possibility of estimating the fuel consumption and the emissions of different modes (*47*). A similar study used GPS data of commercial and private vehicles to better understand the emissions and fuel consumption by link and time of day (*48*). The generation of a traffic performance measure was also a frequented topic (*49–53*).

Similar studies targeted flow and density models. A study that used GPS data tested six microscopic traffic flow models (*54*). This study used the genetic algorithm–based approach to estimate model parameters for two cases: speed and headway data. Generally, all models performed better with speed data than with headway data.

Other studies were interested in real-time traffic monitoring; one attempted to use CDR data and traffic counts (*55*). However, inconsistencies between the cell tower handover rate and the traffic volume counts prevented accurate volume estimation. In a similar study, researchers employed cell phone GPS data to study traffic conditions in real time (*56*). To overcome privacy concerns, this study employed virtual trip lines, which is a technology that only transmits data at certain locations. The study suggested that a 2% to 3% penetration rate of GPS-enabled cell phones was sufficient to duplicate the results.

Choice Model Applications

Route choice was the focus of several studies. One study sought to explore the application of route choice portfolios, which had the potential to solve the traffic assignment problem (*57*). Traditional GPS data were used, and the results indicated that the participants did not have a single dominant route. Moreover, the study suggested that route choice portfolios better suited a traveler who sought to optimize route decisions under uncertain conditions.

Another study considered the application of general route choice models on the basis of real-world GPS data (*58*). This study had three main findings. First, the observed route choice percentages varied from those derived through the use of stochastic user equilibrium expectations but approached specific values. Second, four types of heterogeneous driver learning and choice evolution pattern were identified. Third, driver and choice situation variables could predict the identified learning patterns.

Another study combined GPS and GIS data (*59*). The study considered three models, which contained different choice set sizes of five, 10, and 15. Estimations of the effects of free-flow travel time, left turns, right turns, intersections, and circuity on the attractiveness of different route alternatives proved to be statistically significant and reasonable. Also, the factors' sensitivity varied on the basis of trip and traveler characteristics.

Another study used GPS data to estimate a utility function that reflected cyclists' evaluations of paths (*60*). The study used logit models to determine the relative importance of four statistically significant path parameters: length, auto speed, grade, and the presence of bike lanes. The results indicated the possibility of generating a relatively robust path and mode model that could be included in multimodal travel forecasting models.

One study successfully converted GPS data into routes to characterize route choice variability and compare the least-cost route to the actual route (*61*). Generally, discretionary trips displayed greater intraindividual variability; work and study trips displayed greater interindividual variability and deviation from the least-cost routes.

Multiday Applications

Multiday GPS Travel Behavior Data for Travel Analysis contained four case studies in which multiday GPS data were analyzed (*62*). The first case study explored how drivers' choices were affected by auto network reliability. O-D pairs were estimated on the basis of the GPS data, and the day-to-day travel time variation was examined. The

TABLE 1 Summary of Data-Processing Methods for Cell Phone Data

Study	Type	Data Set	Data-Processing Algorithm	Findings
37	CDR	500,000 participants; 12 months	Home and work anchor points detected by time and frequency	Comparison to existing data showed method's accuracy.
39	CDR	61 billion location data points; 2 months	Home and work identified by time, duration, and frequency; kernel density estimation used for space–time activity density	Data were viable for analysis of human activity patterns in space–time.
11	CDR	5 weeks	Identified trip ends by duration time; calculated O-D and demand	CDR gave viable dynamic O-D traffic flow estimate (intercity trips).
38	CDR	829 million location estimations	Home location identified by frequency and time	Mobile phone data represented reasonable proxy for individual mobility.
20	CDR	7,989 mobile devices and survey data combined (simulated phone)	Model-based clustering method; distinguished activity–travel cluster with logit model; location type detected with use of set of behavior-based algorithms	Home was identified within 100 m (70%) and 1,000 m (97%); work identified within 100 m (65%) and 1,000 m (86%).
18	CDR	18 million participants; 1 month	Home and work identified by time, duration, and frequency; trips detected with fuzzy classification method	Successfully generated large scale O-D information that matched output of traditional methods
17	CDR	8 participants; 1 day	Outliers removed with use of filters; RNF, RLAF, and KF	RNF and RLAF improved speed and position significantly; KF only improved speed estimation significantly.
14	CDR	18,000 participants; 2 weeks	Stops identified by time and distance; home and work identified by frequency and time	Network travel times reduced 10%; waiting time increased 2%.
10	CDR	24 h/day; 1 month	Classified participants (resident, visitor) on basis of inferred home location; identified purpose (HBW, HBO, NHB)	Data were robust enough to develop and estimate external trip models.
15	CDR	3,600 participants; 2 months	Points clustered by distance; corrected oscillations; activity locations detected by duration	Low entropy: 20 locations for 60% accuracy, 50 for 70%, and 100 for 80%; high entropy: 20 locations for 40% accuracy and 100 for 50%
28	GPS	6 participants	Mode classified by *k*–nearest neighbor, naive Bayes, decision trees, and support vector machines, hidden Markov model, and decision trees with hidden Markov	Decision trees with hidden Markov model was most accurate (98%–99%).
35	GPS	114 trips (38 car, 38 bus, 38 walk)	Mode classified by neural network; considered all GPS points rather than only critical points	Acceleration and speed might be best indicator for mode detection; analysis with critical points produced accurate estimations of mode.
16	GPS	16 participants	Mode detection with accelerometer (multiple machine learning methods); trip detection with Markov decision process; accuracy increased with Gaussian mixture model	Support vector machine made the best mode predictions (93%–95%).
27	GPS	6 participants; 3 weeks	Used GIS data for mode detection; tested multiple machine learning methods	Random forest method was able to achieve 93.70% accuracy with GIS and 76% accuracy without GIS.
24, *25*	GPS and other sensors	14 participants; 266 h 15 participants; 355 h	Mode classified by probabilistic means on basis of speed	Nonmotorized modes showed greater accuracy (bike, 98% and walking, 92% versus railway, 80%).
19	GPS and CDR	111 participants; 3 months	Preprocessed data; trip ends identified by clustering by space and time; home and work identified by frequency and time	Social contract influenced number of trips; communication usage influenced travel intensity but not distance.
30	GPS	3 participants; 7 h	Mode detection through support vector machine learning	98.86% accurate with sensor and GPS data, 97.89% without GPS; difficult to differentiate bus from car and bike modes
34	GPS	Microsoft's GeoLife data set	Simulated near-real time multilayer perceptron neural network for mode detection	Incorporation of spatial information helped achieve higher accuracy in mode detection to 93%–95%.
33	GPS	35 participants; 2 weeks	Mode identification through random forest	Accuracy: bus (87.93%), car (97.68%), and walking (90.33%); instantaneous speed and GPS accuracy most influential
9	GPS	658 trips	Mode detected with probabilistic methods that used speed, acceleration, and acceleration changes	Accuracy: walking (98%), bike (55%), transit (9%), and auto (72%)

NOTE: O-D = origin–destination; RNF = recursive naive filter; RLAF = recursive look-ahead filter; KF = Kalman filter; HBW = home-based work; HBO = home-based other; NHB = non–home based.

TABLE 2 Summary of Applications Based on Cell Phone Data

Study	Type	Sample	Applications	Findings
11	CDR	5 weeks	Dynamic O-D estimation using probabilistic rules	Good estimate of traffic volumes and traffic flow of special events; best for long-distance or intercity trips
55	CDR	Multiple samples	Traffic monitoring with MNL and ANN	Accuracy of MNL was 76.4% and ANN was 78.1%.
63	CDR	4 months	Criterion-based method determined interaction characteristics and developed mobility profiles and forecasted weekly mobility patterns; Monte Carlo simulation estimated evolution of daily states.	Social media was viable as a stand alone or supplement to estimate the mobility behavior of individuals.
38	CDR	829 million location estimations	Daily mobility pattern calculated with a multivariate regression model	Job accessibility and distance to nonwork destination influenced variations in individual and vehicular total trip lengths.
14	CDR	18,000 participants; 2 weeks	Mobility patterns; home and work and routes fed into transit optimization model	Decreased systemwide travel time by 10%
18	CDR	18 million participants; 1 month	O-D estimation; rules-based home and work detection; mode classified by fuzzy classification method	Methodology generated O-D tables for large cities with large populations and had acceptable level of accuracy.
67	CDR	1,310 participants; 12 months	Spatial travel behavior; *k*-means clustering to divide sample; intrapersonal CV applied as dependent variable in some univariate general linear model; GLM models used to assess seasonal effects	Activity space varied more than number of activity locations; individual factors controlled monthly spatial behavior variation; significant intrapersonal monthly variability.
64	CDR	744 participants; 2 months	Location variability by time of day; entropy (measure of location variability); temporal profile of location variability; model-based clustering; linear regression on panel data	Time-of-day effect accounted for 36% of variations in location variability; smallest location variability was early in morning.
43	CDR	2.87 million participants; 1 month	O-D estimation; determine scaling factors (MITSIMLab); route choices of drivers based on discrete choice-based probabilistic model	Relatively effective and economical; provided viable method
44	CDR	2 cases: 2.8 million and 2.0 million participants	Demand estimation; traditional 4-step model	Results were similar to existing data and O-D matrices.
10	CDR	24 h/day; 1 month	External travel demand; average weekday trip tables were disaggregated into three trip purposes primarily on basis of home and work locations	Passively collected mobile phone data can be good source for development of external trip models.
46	CDR	2 months (232 GB)	O-D estimation; trip identification algorithm used	Predicted demand patterns for different day types; showed time-varying features of demand by time of day; revealed directional patterns by time of day
51	CDR	17.5 million participants; 1 month	Rail transit use patterns; algorithm used to match data to transit line	Usage patterns changed on basis of transit dependency; spatiotemporal patterns were very different; demand of residents outside city center could not be satisfied.
45	CDR and Foursquare data	CDR (515,557 participants) Foursquare (13 days)	O-D estimation; location data input into characterization model; cellphone sample to TAZ population ratio was expansion factor; sample O-D multiplied by TAZ's expansion factor to estimate final O-D table; Foursquare provided trip production or attraction at TAZs	Trip attraction and production were reasonable; CDR O-D matrices varied because of disproportionate variation and incomplete covariation; Foursquare O-D matrices varied because of incomplete covariation.
56	GPS	100 vehicles; 8 h	Traffic monitoring	Data viable for traffic monitoring; 2%–3% penetration rate was sufficient to provide accurate velocity measurements
52	GPS	Not specified	Average speed and average travel time	Android outperformed iOS.

NOTE: MNL = multinomial logit; ANN = artificial neural network; CV = coefficient of variation; GLM = general linear model; GB = gigabyte; TAZ = traffic analysis zone.

results indicated that reliability was significantly impacted by trip and household characteristics. Also, there was a lack of definitive proof of a direct correlation between reliability and travel frequency at the household level.

The second case study used a 3-day, person-based GPS data set to detect day-to-day variations in the number, type, and level of dispersion (distance) of the destinations visited. Patterns of variability were discovered through the use of latent class cluster analysis. Distance and location were identified as influential over the variation type. The types of variability were also a function of spatial attributes.

In the third case study, the authors used the same data set as the previous case study to analyze day-to-day variations in mode choice. The authors successfully classified individuals into groups on the basis of mode changes and frequency. These groups were connected to specific personal and household characteristics. Relationships between participant characteristics and modality were also considered. Correlations between modality style and other characteristics (household income, number of workers, individual education level, employment status, and gender) were confirmed, but to different extents. The use of multiple modes was noted for those with greater transit access (greater proximity). Age and the presence of children in a household indicated that there was a preference for only using one mode, specifically auto. However, statistical significance was not achieved because of the sample size.

The last study used a multiday, person-based GPS data set to study the deviation in travel time between the shortest path and the actual path, as well as the frequency of use of the shortest path for home-to-work trips (*62*). The study revealed that participants did not make the same home-to-work auto trip frequently over multiple days and suggested that multiday studies required a large sample size. It was also noted that most participants did not use the shortest path, which could be a result of trip circuity, the number of turns, or the age of the driver.

Several CDR studies focused on human mobility patterns. One presented a technique to use social sensing to gauge human mobility (*63*). Another study compared individual mobility and vehicular mobility to understand daily mobility; the major influences on total trip lengths were observed to be job accessibility and distance to nonwork destinations (*38*). An optimized network design model for public transit that decreased systemwide travel times was also proposed (*14*). Entropy was used as a measure of location variability to explore the effect of time of day on travel behavior (*64*). From that study, the authors determined that the time of day affected location variability and that location variability was smallest during the morning hours.

Other Applications

In addition to the groups previously discussed, other applications were noted. One study used anonymous GPS data to construct activity–travel pattern characteristics, which were combined with land use data to estimate various models of demographic characteristics (*65*). The proposed method successfully identified several characteristics (work status, education level, age, possession of license, presence of children) but struggled with others (gender, household size, number of vehicles). The results were generally positive and demonstrated an ability to reconstruct some socioeconomic demographic data.

CDR and GPS data were also used to explore human mobility patterns (*38, 66*). CDR data were used to demonstrate that variations in individual and vehicular mobility were mainly attributable to accessibility rather than to population density and land use (*38*). A different study employed GPS taxi trajectories to serve as a proxy for individuals (*66*). The study discovered that, unlike most models, the travel distance and the elapsed time of these data were best fit by an exponential distribution, and human mobility tended to be sporadic.

One year of large-scale CDR data were used in an attempt to study human activity–travel behavior with respect to temporal trends (*67*). The results indicated that monthly variation in unique activity locations displayed seasonal trends, and spatial distribution varied greatly. The study also revealed that inter- and intrapersonal factors were more influential than seasonal impacts. Also, the daily variation of activity locations remained relatively constant throughout the study, and the participants' activity more than doubled during the summer months.

DISCUSSION OF FINDINGS

Many studies have explored the potential of passively collected data to supplement traditional surveys; this exploration may be the beginning of a paradigm shift. Moving forward, it will be possible to tap into the full potential of these data sources and supplement them with minor surveys. This approach has the potential to decrease respondent burden and cost while improving data quality and prediction accuracy.

Irrespective of accuracy, traditional GPS data have proved to be less useful on a large scale because of their cost. Mobile network data are a cheaper alternative to traditional GPS data. Mobile network data can provide anonymous information for millions of users. Scale alone can be misleading, though, because data are only recorded when the phone is used. When this factor is considered, along with laws in parts of the United States concerning phone usage while driving, the potential of this data source may also be limited. Although cell phone GPS data have no direct cost associated with them, there may be some respondent burden for participants who lack computer literacy. However, this respondent burden could easily be overcome by providing instructions, as has been noted by the authors. Cell phone GPS data provide nearly the same accuracy as traditional GPS data and have the large-scale, low-cost, high penetration rate, and low respondent burden benefits of mobile network data.

Given the passive nature of these data, a considerable amount of effort is needed to derive useful trip information, such as the trip ends, the mode, and the location type. Studies have employed various methods to process and transform the data points into meaningful representations of human movement. On the basis of this information, many applications have become feasible, including O-D estimation, traffic monitoring, and the understanding of spatiotemporal human mobility patterns. These data sources have a level of detail that could afford researchers opportunities to create real-time representations of congestion, and therefore emissions, throughout the transportation network. These data hold the promise of helping produce more accurate transportation measures and representative models of human behavior and usher in a new era for activity-based modeling.

Although very detailed, these data have a few limitations. The most impactful of these limitations is the lack of demographic information, which is critical in travel studies and demand analysis. However, some studies have shown the potential for the derivation of demographic information with the aid of supplementary data (such as land use or census data). Another limitation is the potential sample bias of these data, as not all people carry smartphones and some people may possess multiple phones. This issue may diminish as technology advances and smartphones get more common; however, studies that use these data still need to be aware of the sampling issue. Besides the incorporation of demographic information through supplemental data sources and spatial analysis techniques, other areas for improvement and research may include validation methods to verify the trip information and the transferability analysis of the data-processing algorithms and derived findings.

REFERENCES

1. Wolf, J., M. Oliveira, and M. Thompson. Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System–Enhanced Household Travel Survey. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1854,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 189–198.
2. U.S. Department of Defense. *Global Positioning System Standard Positioning Service Performance Standard.* Sept. 2008. http://www.gps.gov/technical/ps/2008-SPS-performance-standard.pdf. Accessed July 2015.
3. Lindsey, G., S. Hankey, X. Wang, J. Chen, and A. Gorjestani. *Feasibility of Using GPS to Track Bicycle Lane Positioning.* Report CTS 13-16. Intelligent Transportation Systems Institute, Center for Transportation Studies, University of Minnesota, Minneapolis, 2013.
4. Chen, C., H. Gong, C. Lawson, and E. Bialostozky. Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned From New York City Case Study. *Transportation Research Part A: Policy and Practice,* 2010, pp. 830–840.
5. Bohte, W., and K. Maat. Deriving and Validating Trip Destinations and Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in the Netherlands. *Transportation Research Part C: Emerging Technologies,* 2009, pp. 285–297.
6. Federal Communication Commission. 911 Wireless Services. 2014. https://www.fcc.gov/guides/wireless-911-services. Accessed July 2015.
7. International Telecommunication Union. ITU Statistics. 2015. http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx. Accessed July 2015.
8. Stopher, P., C. FitzGerald, and J. Zhang. Search for a Global Positioning System Device to Measure Person Travel. *Transportation Research Part C: Emerging Technologies,* 2008, pp. 350–369.
9. Nour, A., J. Casello, and B. Hellinga. Developing and Optimizing a Transportation Mode Inference Model Utilizing Data from GPS Embedded Smartphones. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
10. Huntsinger, L.F., and K. Ward. Using Mobile Phone Location Data to Develop External Trip Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2499,* Transportation Research Board, Washington, D.C., 2015, pp. 25–32.
11. Wang, M.-H., S.D. Schrock, N.V. Broek, and T. Mulinazzi. Feasibility of Using Cellular Phone Location Data in Traffic Survey on Intercity Trips. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.
12. Sharman, B., and M. Roorda. Analysis Tool to Process Passively Collected GPS Data for Commercial Vehicle Demand Modeling Applications. Presented at Innovations in Freight Demand Modeling and Data Symposium, Washington, D.C., 2010.
13. Zanjani, A.B., A.R. Pinjari, M. Kamali, A. Thakur, J. Short, V. Mysore, and S.F. Tabatabaee. Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data: Application for Florida Statewide Model. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2494,* Transportation Research Board, Washington, D.C., 2015, pp. 87–95.
14. Berlingerio, M., F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M.L. Sbodio. AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
15. Wang, M., C. Chen, and J. Ma. Time-of-Day Dependence of Location Variability: Application of Passively-Generated Mobile Phone Dataset. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
16. Lu, H., J. Yang, Z. Liu, N.D. Lane, T. Choudhury, and A.T. Campbell. The Jigsaw Continuous Sensing Engine for Mobile Phone Applications. 2010. http://www.cs.dartmouth.edu/~campbell/papers/jigsaw.pdf. Accessed July 2015.
17. Horn, C., S. Klampfl, M. Cik, and T. Reiter. Detecting Outliers in Cell Phone Data: Correcting Trajectories to Improve Traffic Modeling. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2405,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 49–56.
18. Fang, J., M. Xue, and T.Z. Qiu. Anonymous Cellphone-Based Large-Scale Origin–Destination Data Collection: Case Studies in China. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
19. Ythier, J., J.L. Walker, and M. Bierlaire. The Influence of Social Contacts and Communication Use on Travel Behavior: A Smartphone-Based Study. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.
20. Chen, C., L. Bian, and J. Ma. From Traces to Trajectories: How Well Can We Guess Activity Locations from Mobile Phone Traces? *Transportation Research Part C: Emerging Technologies,* 2014, pp. 326–337.
21. Schuessler, N., and K.W. Axhausen. Processing Raw Data from Global Positioning Systems Without Additional Information. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2105,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 28–36.
22. Bohte, W., K. Maat, W. Quak, and T. Deleft. A Method for Deriving Trip Destination and Modes for GPS-Based Travel Surveys. In *Urbanism On Track: Application of Tracking Technologies In Urbanism,* Delft University Press, Netherlands, 2008, pp. 127–143.
23. Pereira, F., C. Carrion, F. Zhao, C.D. Cottrill, C. Zegras, and M. Ben-Akiva. The Future Mobility Survey: Overview and Preliminary Evaluation. Presented at 10th International Conference of Eastern Asia Society for Transportation Studies, Taipei, Taiwan, 2013.
24. Nitsche, P., P. Widhalm, S. Breuss, and P. Maurer. A Strategy on How to Utilize Smartphones for Automatically Reconstructing Trips in Travel Surveys. *Procedia Social and Behavioral Sciences,* 2012, pp. 1033–1046.
25. Nitsche, P., P. Widhalm, S. Breuss, N. Brändle, and P. Maurer. Supporting Large-Scale Travel Surveys with Smartphones: A Practical Approach. *Transportation Research Part C: Emerging Technologies,* 2014, pp. 212–221.
26. Wolf, J., W. Bachman, M.S. Oliveira, J.A. Auld, A. Mohammadian, and P. Vovsha. *NCHRP Report 775: Applying GPS Data to Understand Travel Behavior: Background, Methods, and Tests.* Transportation Research Board of the National Academies, Washington, D.C., 2014.
27. Stenneth, L., O. Wolfson, P.S. Yu, and B. Xu. Transportation Mode Detection Using Mobile Phones and GIS Information. In *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems,* Association for Computing Machinery, New York, 2012, pp. 54–63.
28. Reddy, S., J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Determining Transportation Mode on Mobile Phones. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers,* IEEE, New York, 2008, pp. 25–28.
29. Moiseeva, A., J. Jessurun, and H. Timmermans. Semiautomatic Imputation of Activity Travel Diaries: Use of Global Positioning System Traces, Prompted Recall, and Context-Sensitive Learning Algorithms. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2183,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 60–68.
30. Jahangiri, A., and H. Rakha. Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
31. Zhang, L., S. Dalyot, D. Eggert, and M. Sester. Multi-Stage Approach to Travel-Mode Segmentation and Classification of GPS Traces. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* Volume XXXVIII-4/W25, 2011, pp. 87–93.
32. Zheng, Y., Q. Li, Y. Chen, X. Xie, and W.Y. Ma. Understanding Mobility Based on GPS Data. In *Proceedings of the 10th International Conference on Ubiquitous Computing,* Association for Computing Machinery, New York, 2008, pp. 312–321.
33. Ansari Lari, Z., and A. Golroo. Automated Transportation Mode Detection Using Smart Phone Applications via Machine Learning: Case Study Mega City of Tehran. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.

34. Das, R. D., N. Ronald, and S. Winter. A Simulation Study on Automated Transport Mode Detection in Near-Real Time Using a Neural Network. Presented at Research at Locate 2015, Brisbane, Australia, 2015.
35. Gonzalez, P.A., J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi, and R. Perez. Automating Mode Detection for Travel Behavior Analysis by Using Global Positioning Systems Enabled Mobile Phones and Neural Networks. *IET Intelligent Transport Systems,* 2010, pp. 37–49.
36. Schönfelder, S., and U. Samaga. Where Do You Want to Go Today? More Observations on Daily Mobility. Presented at 3rd Swiss Transport Research Conference, Monte Verita; Ascona, Italy, 2003.
37. Cartwright, W., G. Gartner, L. Meng, M.P. Peterson, and K. Rehrl. *Location Based Services and TeleCartography II from Sensor Fusion to Context Models: Lecture Notes in Geoinformation and Cartography.* Springer-Verlag, Berlin, 2008, pp. 301–315.
38. Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies,* 2013, pp. 301–313.
39. Cheng, X., W. Li, F. Jia, D. Yang, and Z. Duan. Analyzing Human Activity Patterns Using Cellular Phone Data: Case Study of Jinhe New Town in Shanghai, China. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.
40. Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, and K.W. Axhausen. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1870,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 46–54.
41. McGowen, P., and M.G. McNally. Evaluating the Potential to Predict Activity Types from GPS and GIS Data. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
42. Thakur, A., A.R. Pinjari, A.B. Zanjani, J. Short, V. Mysore, and S.F. Tabatabaee. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2529,* Transportation Research Board, Washington, D.C., 2015, pp. 66–73.
43. Iqbal, M.S., C.F. Choudhury, P. Wang, and M.C. González. Development of Origin–Destination Trip Matrices Using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies,* 2014, pp. 63–74.
44. Çolak, S., L.P. Alexander, B.G. Alvim, S.R. Mehndiratta, and M.C. González. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations, and Opportunities. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2526,* Transportation Research Board, Washington, D.C., 2015, pp. 126–135.
45. Rokib, S.A., M.A. Karim, T.Z. Qiu, and A. Kim. Origin–Destination Trip Estimation from Anonymous Cell Phone and Foursquare Data. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
46. Li, H. Using Mobile Phone Data to Analyze Origin–Destination Travel Flow Dynamics for City of Pasadena, CA and Surrounding Area. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
47. Barth, M.J., E. Johnston, and R.R. Tadi. Using GPS Technology to Relate Macroscopic and Microscopic Traffic Parameters. In *Transportation Research Record 1520,* TRB, National Research Council, Washington, D.C., 1996, pp. 89–96.
48. Venter, C.J., and J.W. Joubert. Use of Multisource Global Positioning System Data to Characterize Multiday Driving Patterns and Fuel Usage in a Large Urban Region. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2338,* Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 1–10.
49. Bachman, W., M. Oliveira, J. Xu, and E. Sabina. Household-Level Global Positioning System Travel Data to Measure Regional Traffic Congestion. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2308,* Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 10–16.
50. Liao, C.-F. Generating Reliable Freight Performance Measures with Truck GPS Data: Case Study in Twin Cities Metropolitan Area, Minnesota. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2410,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 21–30.
51. Li, W., X. Cheng, D. Yang, Z. Duan, and G. Guo. Mining Rail-Transit Use Patterns from Mobile Phone Data: Case Study of Three Communities in Shanghai. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
52. Yin, E., and T.Z. Qui. Using GPS-Enabled Smartphones and Geofence to Capture Traffic Data on Urban Freeways and Arterials. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
53. Ma, Y., J. van Dalen, C. de Blois, and L. Kroon. Estimation of Dynamic Traffic Densities for Official Statistics: Combined Use of Data from Global Positioning System and Loop Detectors. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2256,* Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 104–111.
54. Ranjitkar, P., T. Nakatsuji, and M. Asano. Performance Evaluation of Microscopic Traffic Flow Models with Test Track Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1876,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 90–100.
55. Demissie, M.G., G.H. de Almeida Correia, and C. Bento. Intelligent Road Traffic Status Detection System Through Cellular Networks Handover Information: An Exploratory Study. *Transportation Research Part C: Emerging Technologies,* 2013, pp. 76–88.
56. Herrera, J.C., D.B. Work, R. Herring, X. Ban, Q. Jacobson, and A.M. Bayen. Evaluation of Traffic Data Obtained Via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment. *Transportation Research Part C: Emerging Technologies,* 2010, pp. 568–583.
57. Levinson, D., and Z. Shanjiang. A Portfolio Theory of Route Choice. *Transportation Research Part C: Emerging Technologies,* 2013, pp. 232–243.
58. Tawfik, A.M., and H.A. Rakha. Network Route-Choice Evolution in a Real-World Experiment: Necessary Shift from Network- to Driver-Oriented Modeling. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2322,* Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 70–81.
59. Dhakar, N.S., and S. Srinivasan. Route Choice Modeling Using GPS-Based Travel Surveys. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2413,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 65–73.
60. Casello, J.M., and V. Usyukov. Modeling Cyclists' Route Choice Based on GPS Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2430,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 155–161.
61. Spissu, E., I. Meloni, and B. Sanjust. Behavioral Analysis of Choice of Daily Route with Data from Global Positioning System. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2230,* Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 96–103.
62. *Multiday GPS Travel Behavior Data for Travel Analysis.* Report FHWA-HEP-015-026. FHWA, U.S. Department of Transportation, 2015.
63. Gkiotsalitis, K., F. Alesiani, and R. Baldessari. Educated Rules for the Prediction of Human Mobility Patterns Based on Sparse Social Media and Mobile Phone Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
64. Wang, M., C. Chen, and J. Ma. On Making More Efficient Location Prediction. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
65. Auld, J., A. Mohammadian, M.S. Oliveira, J. Wolf, and W. Bachman. Demographic Characterization of Anonymous Trace Travel Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2526,* Transportation Research Board, Washington, D.C., 2015, pp. 19–28.
66. Liang, X., X. Zheng, W. Lv, T. Zhu, and K. Xu. The Scaling of Human Mobility by Taxis Is Exponential. *Physica A: Statistical Mechanics and Its Applications,* 2012, pp. 2135–2144.
67. Järv, O., R. Ahas, and F. Witlox. Understanding Monthly Variability in Human Activity Spaces: A Twelve-Month Study Using Mobile Phone Call Detail Records. *Transportation Research Part C: Emerging Technologies,* 2014, pp. 122–135.

The Standing Committee on Transportation Planning Applications peer-reviewed this paper.