Transportation Research Part C 68 (2016) 285–299

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Review

The promises of big data and small data for travel behavior (aka human mobility) analysis

Cynthia Chen [a,*], Jingtao Ma [b], Yusak Susilo [c], Yu Liu [d], Menglin Wang [e]

[a] Department of Civil & Environmental Engineering, University of Washington, Seattle, USA
[b] Traffic Technology Services, Portland, USA
[c] Royal Institute of Technology, KTH, Sweden
[d] Institute of Remote Sensing and Geographical Information Systems, Peking University, China
[e] Cambridge Systematics, Chicago, USA

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form 31 March 2016
Accepted 5 April 2016
Available online 23 April 2016

Keywords:
Big data
Small data
Human mobility
Travel behavior
Transportation planning

ABSTRACT

The last decade has witnessed very active development in two broad, but separate fields, both involving understanding and modeling of how individuals move in time and space (hereafter called "travel behavior analysis" or "human mobility analysis"). One field comprises transportation researchers who have been working in the field for decades and the other involves new comers from a wide range of disciplines, but primarily computer scientists and physicists. Researchers in these two fields work with different datasets, apply different methodologies, and answer different but overlapping questions. It is our view that there is much, hidden synergy between the two fields that needs to be brought out. It is thus the purpose of this paper to introduce datasets, concepts, knowledge and methods used in these two fields, and most importantly raise cross-discipline ideas for conversations and collaborations between the two. It is our hope that this paper will stimulate many future cross-cutting studies that involve researchers from both fields.

Contents

* Corresponding author.
E-mail address: qzchen@uw.edu (C. Chen).

http://dx.doi.org/10.1016/j.trc.2016.04.005
0968-090X/

1. Introduction

Research in human movement in time and space has been around for at least over five decades (Weiner, 1999). Motivated by the need to forecast future travel demand to better guide the investment of often mega-scale transportation projects, transportation researchers have long sought to develop models to predict how people travel in time and space and seek to understand the factors that affect travel-related choices. Recently, grand challenges such as global warming and air pollution can all be traced to the over-reliance on automobiles, further motivating transportation researchers and practitioners to develop effective strategies to move toward more sustainable modes of transportation (e.g., public transit and walking and biking). For decades, transportation researchers have largely used data of active solicitation, including, for example, travel surveys where subjects are asked to self-report their activities and travels via paper, web, or phone interviews; travel surveys coupled with GPS loggers during which subjects are asked to both complete questionnaires and carry GPS loggers; and pure GPS-based surveys during which subjects are only asked to carry GPS loggers (Wolf et al., 2001; Axhausen et al., 2003; Hato et al., 2006; Stopher et al., 2008a, 2008b; Bohte and Maat, 2009; Chen et al., 2010; Gong et al., 2011). In the last type, information about subjects' activities and travels still need to be inferred from the collected GPS traces. All these surveys share a common characteristics and that is: active solicitation—subjects and information on their travels are actively recruited. Probably because of this attribute, these surveys are limited by a relatively small sample size (TMIP, 2013). In this paper, we refer to data of active solicitation as *small data*.

Parallel to the continued use of small data in transportation research, the rapid rise and prevalence of mobile technologies have enabled the collection of a massive amount of passive data (*big data*), which have resulted in a surge of studies on human movement (e.g., Gonzalez et al., 2008; Kang et al., 2012a, 2012b; Calabrese et al., 2013). Passive data refers to those data not collected through active solicitation; rather it is generated for purposes that are not intended but can be potentially used for research. Examples include mobile phone sightings generated by phone operators for operation purposes (Calabrese et al., 2011), social media data generated voluntarily by users' online activities (Chen and Schintler, forthcoming), and smart-card data collected at many transit systems worldwide (Pelletier et al., 2011; Ma et al., 2013). Passively collected, such data is very different from data of active solicitation (*small data*) that are familiar to most transportation researchers and thus requires different methods and techniques for processing and modeling. The first purpose of this paper is to introduce passively collected big data to transportation researchers, provide a state of the art review of the methods used, and identify areas of gap that are particularly important for transportation planners.

More importantly, this paper seeks to identify cross-disciplinary concepts and opportunities for both transportation researchers who have traditionally used small data and big data researchers. Our discussion will be on three important subareas of travel behavior research: (1) *behavioral factors* where the interest is identifying factors that explain travel behaviors and uncover the underlying causal mechanisms; (2) *modeling travel behavior* where models are developed to predict human movement behaviors; and (3) *human mobility patterns* where pattern recognition is an important goal. It is our view that the recent advances made with the use of the big data have the potential to drive fundamental advances in research in human mobility and at the same time, knowledge accumulated in transportation research in the past many decades can guide big data studies to answer questions that matter to the society, in particular, those relating to transportation investment decisions and policies in urban environments.

The rest of the paper is organized as follows. In Section 2, an introduction of the big data as well as a review of the current methodologies is provided. Our focus is on passively generated mobile phone dataset. In analyzing human mobility patterns, passively generated mobile phone data has emerged as the most frequently used (and possibly the most reliable) data source (Gonzalez et al., 2008). Other data sources cannot capture the full spectrum of an individual's mobility pattern over multiple days, involving the use of multiple modes of transportation. Examples include the use of taxi data that is mostly suitable for studying drivers' patterns in searching for passengers (Jiang et al., 2009; Liu et al., 2012), use of transit smart card data that only captures the use of transit modes (Long and Thill, 2015), or the use of social media data whose spatial and temporal resolutions are much lower than those of mobile phone data and biased toward certain locations (Cheng et al., 2011; Noulas et al., 2012). The target audience of Section 2 is transportation researchers who are familiar with the small, survey data but not the big data that has been recently utilized. In Sections 3–5, we discuss cross-disciplinary concepts and ideas in the three subareas noted earlier. Concluding discussions are provided in Section 6.

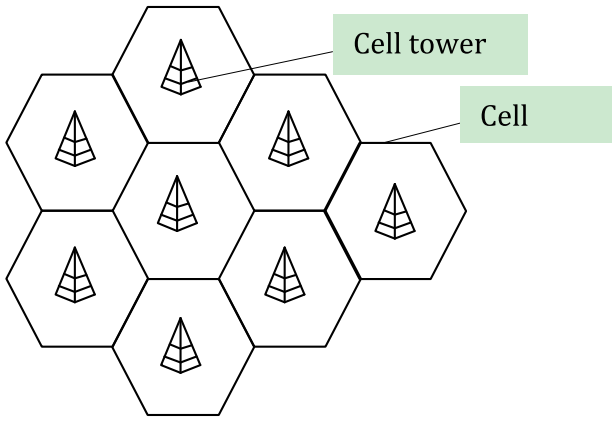

Fig. 1. An example cellular network.

Table 1
Sample records in CDR data.[a]

X	Y	ID	Time	Duration (sec)
195925	32464	J000001	82141	81
195925	32464	J000001	82456	75
195018	31555	J000002	82100	140

[a] XY coordinates are transferred from geographical coordinate system. A conversion can be made to convert them into the absolute latitude and longitude coordinates.

Table 2
An example of the sightings data.

ID	Time[a]	Location[b]
3X35E90	1319242582	34.044162|−112.454400
3X35E90	1319242583	34.044059|−112.455550
3X35E90	1319301785	34.044392|−112.453519

[a] Time is Unix timestamp—defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, Thursday, 1 January 1970.
[b] Location is the longitude and latitude coordinates of mobile phones.

2. The big data

2.1. Nature of passively-generated anonymous mobile phone data

2.1.1. Data types

The location information in a passive mobile phone dataset is generated as the result of a phone's communicating with the cellular network maintained and operated by cellular network operators (e.g., phone providers); this process is called positioning. A cellular network is one that enables mobile phones to communicate with each other; it comprises multiple base stations, each serving an area, which is called a cell (Fig. 1). Each cell has a unique cell ID.

Mobile phone positioning is required when a user communicates with the network (Ficek and Kencl, 2012). When a user initiates a network connection event (e.g. a voice-call), the cellular network operator needs to know his/her location in order to determine the cell tower used to channel this event. Thus, the positioning data (containing information on users' locations) is generated when an event occurs. Such data is automatically and passively generated for cellular network operators' own purposes, including billing information collection and network management.

Positioning can result in two types of mobile phone data. The CDR data (CDR) is probably the most widely used today (e.g., Gonzalez et al., 2008; Kang et al., 2012a, 2012b). Every record in a CDR data represents a phone call, with information on the caller, callee, the starting time of the call, the duration of the call, and the XY coordinates of the tower that first channeled the call when the call was first initiated. A sample of the CDR data is shown in Table 1, where ID refers to a user who initiates or receives a call, X and Y are coordinates of the tower that channels the user's call, and TIME and DURATION

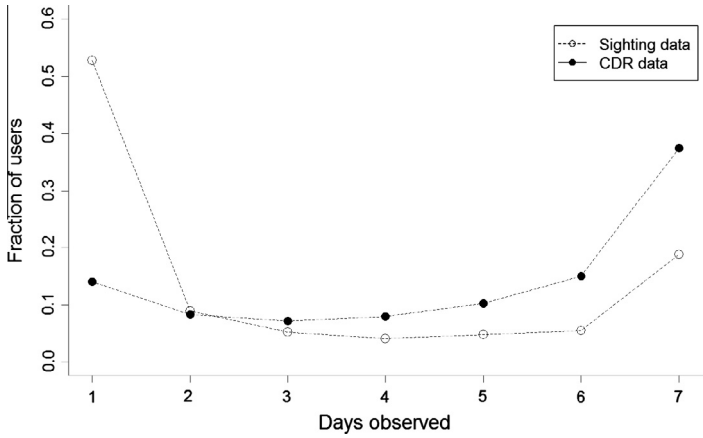

Fig. 2. Number of days on which at least one record is observed on a single day.

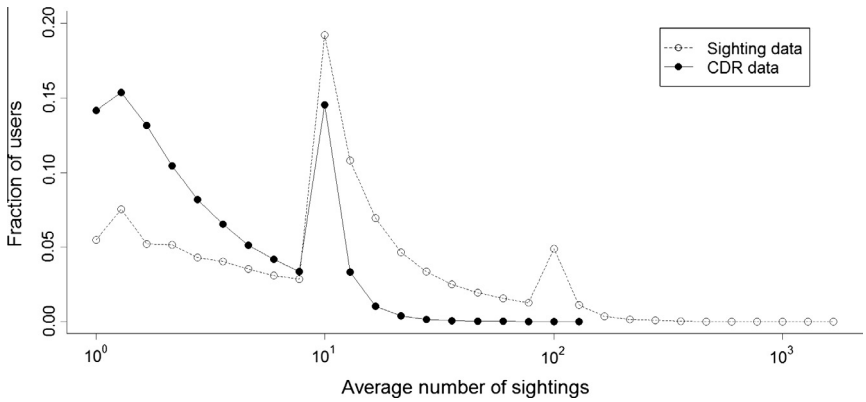

Fig. 3. Average number of records per day per user.

note that starting time and the duration of the call. For example, the first record represents that user "J000001" had a call at 8:21:41 at the location (195925, 32464) lasting for 81 sec.

A less frequently used mobile phone dataset is called "sightings data" (e.g., Ma et al., 2012; Calabrese et al., 2013; Chen et al., 2014) (see Table 2). A sighting is generated each time a phone is positioned. Thus, each row in the dataset represents a sighting. One may think of sightings data as somewhat a processed version of the CDR data. It differs from the CDR data in three aspects: (1) *Temporal resolution*: sightings data likely has a higher level of temporal resolution than CDR data—while a single phone call will generate one record in the CDR data, the same call may generate multiple sightings. (2) *Spatial resolution*: sightings data likely has a higher level of spatial resolution than CDR data. The locations reported in CDR are cell phone tower locations and thus depend on the density of the cellular network, which varies from as little as a few hundred meters in metropolitan areas to a few kilometers in rural regions (e.g., Calabrese et al., 2013; Chen et al., 2014). On the other hand, the locations reported in the sightings data are the results of triangulation of multiple towers (AIRSAGE, 2013) and these locations are widely viewed as device locations as opposed to tower locations in the CDR data; (3) *User interactions*: information on user interactions can be directly observed from the CDR data, since both the caller and the callee are recorded for a single phone call. This is not the case for the sightings data.

2.1.2. Number of records

The number of days during which a phone is observed at least once on a single day varies greatly between different users and this is the case for both types of data. Fig. 2 provides some insights in this regard, using two independent samples: a CDR sample and a sightings sample.[1] Both distributions show a U shape.

[1] The CDR sample is of 7 days (5 weekdays and 2 weekend days) and the sightings data sample is of 61 days. We randomly removed two weekend days from the CDR sample, so that the weekend–weekday ratio in both samples is approximately 0.4. Number of days observed (see Fig. 2) in the sightings data was contracted by a factor of $\frac{61}{7}$ to make it comparable with the number of days in the CDR sample.

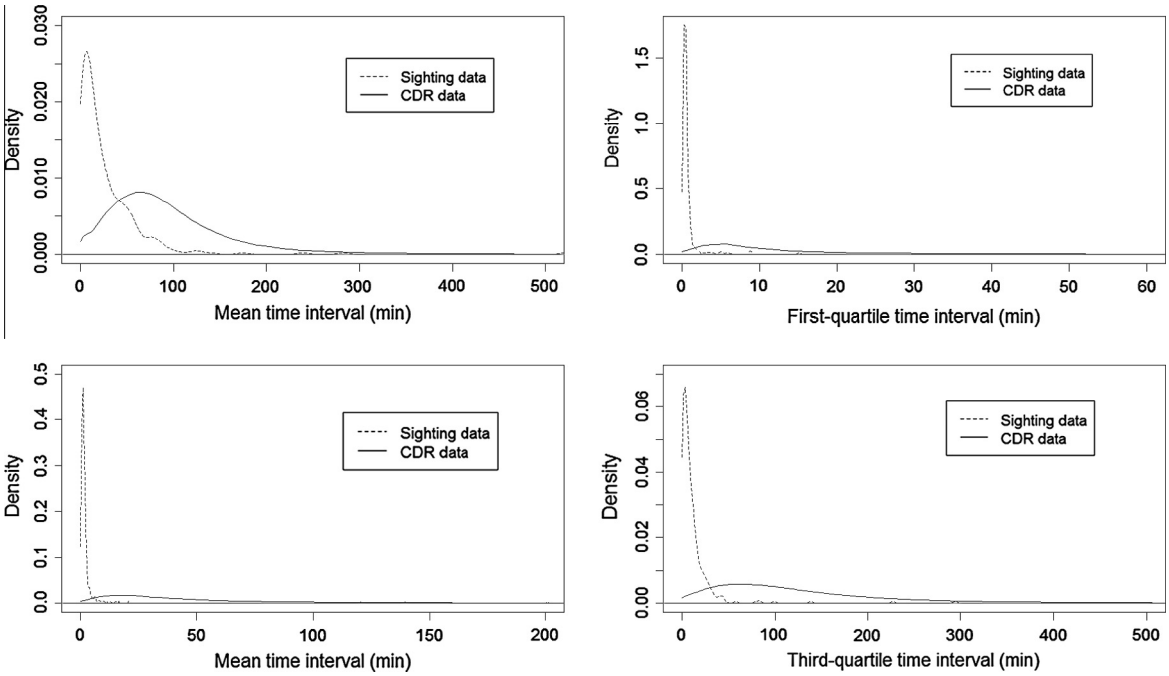

Fig. 4. Time intervals between consecutive records.

One can observe that there is a great amount of heterogeneity in average number of times that a phone is being sighted (sightings data) or a phone call is made (CDR data), as shown in Fig. 3. For both datasets, there is a peak around 10 records; overall (without this peak), the distribution shows an exponential pattern, with the majority of the users having fewer than 10 records and a small number of users having a large number of records.[2]

2.1.3. Temporal resolution

In general, CDR data is expected to be coarser than sightings data in its temporal resolution. In one paper, CDR data was found to have an average inter-event time of 8.2 h for 100,000 individuals over a course of six months (Gonzalez et al., 2008). Using a sightings dataset, Calabrese et al. found an average inter-event time of 260 min, which is much lower than that in González et al. (Calabrese et al., 2011). They further characterized time interval between consecutive sightings by its first, second and third quartiles. The authors reported the arithmetic average of the medians as 84 min and found the temporal resolution of their data was fine enough to detect changes of location where the user stops for as little as 1.5 h.[3]

Fig. 4 shows the distributions of time intervals between consecutive records, expressed in mean, median, first- and third-quartiles for the two independent samples. The peaking patterns of the two datasets differ from each other, which reflects some unique patterns associated with the two kinds of data—for the sightings data, the sightings generated tend to be clustered together since a single phone call will trigger the generation of multiple sightings; for the CDR data, the intervals accurately reflect the intervals between consecutive calls.

2.2. Current methodologies in big data research

2.2.1. Pre-processing

At any given location in a cellular network, there may be several cell towers whose radio signals reach a device. If these multiple cell towers have similar signal strengths, the connection of a device may hop between multiple towers even when the device is stationary. In this case, it may appear that the user travels for several kilometers in just a few seconds. This phenomenon is known as *oscillation* in a cellular network. Reflected in the data is a sequence of records that hops between several towers. A few methods have been proposed to address the oscillation problem. The speed-based correction method (Iovan et al., 2013) is the most frequently used and works as follows: if tower A is recorded between multiple tower B records and the switch speed between A and B is larger than a predetermined threshold, oscillation is then detected. This method is

[2] We do not attempt to make a comparison between two samples, as they are from different localities. The point here is to provide some illustrations of the two datasets.

[3] This suggests that stops of less than 1.5 h may be missed. Many household travel surveys define a stop of more than 5 min as an activity that needs to be recorded.

based on the observation that oscillation results in a location change characterized with an abnormally high speed. The challenge with this method is the choice of a speed threshold that distinguishes between normal and abnormal speeds. Alternatively, a pattern-based method has been applied in some studies. This method recognizes the unique pattern in location updates associated with oscillation—frequent switches between pairs of locations. Lee and Hou (2006) defined the occurrence of oscillation as when three consecutive switches between a pair of locations are observed. Once oscillation is identified, all locations involved in these switches are replaced with the location in the pair with which the user has been associated most of the time. A similar method was used by Bayir et al. (2010). This pattern-based method has the risk of mistaking the actual movements of a user who frequently travels between two locations for oscillation. Wang (2014) proposed a hybrid method that works as follows: first, sub-sequences seemingly resulting from oscillation are detected based on pattern-based approaches; then, switching speeds between pairs of locations are determined for each sub-sequence; at last, sub-sequences are only updated if the switching speed is beyond a speed threshold as determined in speed-based approaches.

2.2.2. From traces to activity locations

The spatial uncertainty associated with the big data (either CDR or sightings data) mean that on their own, the location traces in the data do not represent locations visited or passed through. In other words, these traces need to be processed, aggregated and derived for different types of activity locations individuals visit.

Generally speaking, there are two main approaches to aggregate the location traces. The first one is to label activity locations by frequency. This approach is only applicable to CDR data where the number of cell phone towers in an area is finite. For example, the home location is identified if a phone is being located during night time (e.g., from 9 pm to 5 am) at a frequency exceeding a pre-determined threshold (e.g., 70%) (Wang, 2014).

The frequency method cannot apply to sightings data, as a location record is the result of triangulation and consequently, each record is unique. Thus, clustering is often applied. A number of studies use distance-based clustering. For instance, Ye et al. clustered the location records within a distance of 200 m (2009) and Calabrese et al. used 1 km as the distance threshold in clustering (2010). Clustering methods such as k-means and hierarchical clustering require the number of clusters as the input. In many applications, the required inputs described above are usually unavailable and the arbitrarily selected ones bring uncertainty to the clustering results (Ester et al., 1996). Ester et al. proposed a density based clustering method which uses local density instead of distance to determine the clusters (1996). This method does not require the number of clusters as an input and is able to detect the outliers that do not belong to any cluster. However, there is uncertainty on the border points that could belong to multiple clusters.

Another flexible algorithm that does not require the number of clusters as an input is model-based clustering. In this method, the location records are assumed to be generated from a statistical model. The Gaussian mixture model (GMM) is one of the most applied models for the clustering analysis (Fraley and Raftery, 2002). A GMM is a weighted sum of multiple individual Gaussian distributions with unknown parameters and each individual distribution mathematically captures a cluster. To run this method, only the maximum number of clusters M is required as an input and for each number m from 1 to M, the M clusters are estimated by expectation–maximization algorithm (EM) with the Bayesian Information Criteria (BIC). At last, the clustering result with the highest BIC is chosen from the M results as the best one. Another advantage of this method is that its result can be evaluated by the probability that a given observation does not belong to its currently assigned cluster. Chen et al. utilized this method to aggregate the traces and obtained promising results (2014).

2.2.3. Inferring activity locations (trip purposes)

Most big data encounter a "trajectory data rich but activity information poor" problem (Gong et al., 2015). Clustering results in inferred activity locations, but not their types. To derive the types of locations, researchers have used simplistic methods, such as the frequency-based approach (Phithakkitnukoon et al., 2010; Alexander et al., 2015), which assigns the most frequently visited locations during day time and night time as home and work locations, respectively, or the model-based approach (Chen et al., 2014), which predicts location types with a statistical model. Additionally, the land use information surrounding an inferred activity location can also be used to determine the likelihood of corresponding activity. The basic idea is to link a location with the nearby POIs (point of interests), by considering several predefined empirical rules (e.g. attractiveness of POIs, service hours of POIs, and stop durations) (Xie et al., 2009; Huang et al., 2010; Spinsanti et al., 2010; Gong et al., 2015). Similar methods have also been widely applied by transportation researchers with the GPS data (e.g., Chen et al., 2010).

2.2.3.1. Deriving origin–destination matrices. OD matrix estimation, one of the most important steps in travel demand forecasting, has been traditionally performed by directly observing people's travel patterns with hardware setups to sample all or a portion of the drivers entering and leaving the study cordon boundary. These technologies belong to a class of active probing. Observing travel patterns at a large scale has only become possible with the emergence of the passively collected big data, such as smartphones and smart cards like transit passes and toll tags.

After deriving activity locations, OD estimation involves the calculation of travel times between pairs of origins and destinations. This involves mapping origins and destinations inferred from the cell phone data (see above) to the transportation network. Depending on the types of mobile phone data sources, this mapping can be done using different approaches. If the location data is presented in cell tower voronoi polygons (e.g., CDR data) (Iqbal et al., 2014; Larijani

et al., 2015), the nearest road network node or metro station from the voronoi polygon is often selected to connect the derived locations to the network. If the location data are in triangulated and processed latitude and longitude positions (e.g., sightings data), aggregation to a pre-determined zone structure is required (e.g., grids, census tracts, or traffic analysis zones) (Calabrese et al., 2011). Aggregation to finer geographer areas could lead to noisy and unbalanced OD representations (Chung and Kuwahara, 2007; Zandlebergen, 2009). This problem is alleviated when zone sizes become larger (Alexander et al., 2015).

OD estimation also involves the calculation of scaling factors that convert mobile phone derived OD trips (which is still sample-based) to population-level counts. Several methods have been proposed, for example, using a ratio of census-based population counts and observed cell phone users (Calabrese et al., 2011; Wang et al., 2012, 2014; Alexander et al., 2015), although this method can be biased when the ratio is high (e.g., over 20 for certain zones with low penetration rates of cell phones (Calabrese et al., 2011). This method is analogous to the singly-constrained method that scales to either origins or destinations, as opposed to the doubly-constrained method that seeks to scale to both origins and destinations counts at the same time (Ortuzar and Willumsen, 2011). A more advanced scaling method involves the calculation of the scaling factors together with the traffic assignment step, either iteratively (Ma et al., 2012) or as an optimization problem (Iqbal et al., 2014; Toole et al., 2015).

2.2.4. Inferring mode and route choices

Studies that infer mode and route choices from the big data are much more limited, compared to those to obtain activity locations and infer their types (see above). Inferring route choices typically relies on two kinds of points available in the mobile phone data: (1) points that indicate activity locations, or origins and destinations; and (2) intermediate points between identified origins and destinations. These points are overlaid with the transportation network data (roadway or transit) to determine the most likely route choices. Some studies use intermediate points to infer route choices (Schlaich et al., 2010; Ma et al., 2012; Tettamanti et al., 2012). These studies first generated a choice set comprising multiple paths between an origin and a destination. The definitions used to define what accounts for an origin/destination vary greatly, including census block groups (Ma et al., 2012), Voronoi polygons (Tettamanti et al., 2012), or a large location area (e.g., LA) (Schlaich et al., 2010). These studies also used different metrics to measure the proximity to road or transit networks. In other cases, those intermediate data points were not used; rather an incremental traffic assignment technique was applied (Jiang et al., 2013).

Studies inferring mode choices from the mobile phone data are even sparser than those for route choices. Prior to inferring mode choices, points obtained from the mobile phone data are first overlaid with the network data (e.g., roadway or transit networks) as is the case for route choices. Afterward, one can identify, for example, air travels, by geo-referencing trip ends around airport locations within a certain range (e.g., 3 miles) and with certain trip duration (Ma et al., 2012). An alternative, more direct method is to apply the unsupervised *k*-means clustering algorithm to partition the records (in the mobile phone data) into car or transit travels by their travel speeds, if these points fall within the same origin and destination areas (Wang et al., 2010).

2.3. Unresolved issues (for transportation planning applications)

2.3.1. Validation

Unlike travel surveys during which subjects self-report locations visited and modes of transportation used, passively generated datasets lack ground truth to be validated against. Probably due to this reason, only a few studies using passive data have addressed this to some extent. The limited amount of studies conducted in this regard suffer the ecological fallacy—"a conclusion about individual behavior drawn from data about aggregate behavior" (Freedman, 2002). These studies compare inferred results from the big data (aggregated over all users) to aggregated results from an independent sample. In some studies, the two samples compared do not match (Calabrese et al., 2013). Though representing a very important first step toward the right direction, it is worthy to note that the inferred results at the individual level can have a great amount of errors even though a high level of accuracy is observed at the aggregate level.

Future studies shall develop ways to validate results at the individual level. Potentially, there can be a number of ways. In 2014, Chen et al. (2014) demonstrated a method through the use of simulations. Using two independent sources: a regional household travel survey and a real-world mobile phone dataset, they generated a simulated mobile phone dataset that has ground truth information and exhibits similar spatial and temporal patterns as the real-world mobile phone dataset. This allowed them to assess the accuracy of the inferred results at the individual level. A model-based approach was also developed to identify activity locations and infer their types. The results are quite promising—the mean number of unique locations visited with the model-based approach is 2.73 compared to 2.82 as ground truth and 4.62 using the prevailing approach in the literature (Calabrese et al., 2013). In addition, 70% and 65% of identified home and work places are within 100 m from the true ones. Other methods shall also be explored. For example, transportation researchers have long collected both GPS traces and survey data, using the latter as the ground truth for validation purposes (Chung and Shalaby, 2005; Bohte and Maat, 2008, 2009; Chen et al., 2010; Gong et al., 2011). Similar methods can be adopted for the purpose of validating inferred results from the big data, facilitated by the prevalence of mobile devices in many populations.

2.3.2. Representativeness

A key issue that has largely been neglected in the current big data research is the representativeness of the data (Liu et al., 2015). It is widely recognized that mobile phone data is not representative and multiple reasons contribute to this, including proximity of mobile phones (Patel et al., 2006), multiple mobile phones (Ahonen and Moore, 2006), penetration rates that vary greatly by cell phone carriers (Experian Simmons, 2011), and sample selection (Wang, 2014). Because the frequency of phone use varies greatly across users (see Fig. 4), more active users are likely represented in the sample and those users have been found more mobile than others (Couronné et al., 2013; Ranjan et al., 2012; Iovan et al., 2013). Therefore, sample selection based on phone usage would potentially result in an overestimation of mobility levels. On the other hand, some studies (Iovan et al., 2013) also suggested that some mobility measures seem to be immune to this sampling bias. A question then becomes what mobility measures tend to be immune to this bias, under what circumstances and what information can be inferred from the mobility measures.

For the big data to be utilized in transportation planning for investment and policy related decisions, the issue of representativeness must be addressed. Linking mobility patterns to socio-demographics is important, not just at the area level (e.g., Shi et al., 2015), but more importantly at the individual level. A new study on predicting individual-level wealth and poverty status using the CDR data of the same individual shows promising results (Blumenstock et al., 2015), suggesting that population bias associated with mobile phone data can be potentially corrected. As we transition from Section 2 on the nature of the data to Sections 3–5 on three important subareas in travel behavior (human mobility) analyses, it is important to keep those data-related issues in mind, as they contribute to both questions and answers.

3. Behavioral factors

One challenge in understanding and predicting human mobility patterns is to explain the factors that give rise to the observed behavioral patterns. Decades of travel behavior research can shed great light in this. For nearly all types of travel behaviors, three categories of factors have been consistently found important: socio-demographics, the built environment (e.g., density, diversity, design, and the distance decay effect) (e.g., Liu et al., 2014 on the distance decay effect), and trip and alternative related factors (e.g., trip purpose, travel time and costs, reliability). The various mobility patterns identified from the big data potentially suggest different search strategies being employed by people—for example, the identified Truncated Levy Flight (TLF)[4] suggests a search pattern affected by the built environment, offering corroborations to the vast amount of studies that examines the role of the built environment on travel behaviors (e.g., Chen et al., 2008). More importantly, new studies on how the mobility patterns may vary under different built environments will be beneficial in helping us understand how people's search strategies may change under different circumstances. In fact, the finding of different mobility patterns (e.g., TLF, exponential, etc.) from recent big data research (Kang et al., 2012a, 2012b) suggests that human movement is *not* simply the result of a mechanical process, but to a great extent influenced by a number of factors that vary from place to place and from population to population.

The excitement in this field is not just on what factors affect travel decisions but more on how these factors have shed light in the underlying complexities involved in travel decisions. In statistical language, it is about uncovering the underlying causality mechanisms. In transportation research, alternative theories (from the theory of random utility maximization) have been tested, including, for example, prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1986), bounded rationality as a part of decision making processes (Rasouli and Timmermans, 2015), and regret theory (Chorus et al., 2008). In broader and longer terms of analysis, theory of travel planned behavior (Ajzen, 1991; Bamberg et al., 2003), the norm-activation theory (Schwartz, 1977; Stern, 2000) and a combination of both (Bamberg et al., 2007) have been used as an alternative way to investigate and explain individual choice behaviors. At the same time, more recent studies using big data have started to inquire the underlying mechanisms that result in observed patterns (Jiang et al., 2009; Schneider et al., 2013). However, these studies still largely attempt to identify (physical) signature patterns that together make up the observed mobility patterns. These results (e.g., human mobility motifs, preferential return patterns) are insightful, but they do not explain the underlying mechanisms for individual travel choices. Yet on the other hand, based on these identified signature patterns, we can develop theories and hypotheses about how people make travel related choices and apply simulation models to validate those hypotheses. For example, one can ask the question "whether the identified signature patterns are the results of any of the behavioral theories noted above (e.g., bounded rationality, prospect theory), or perhaps a combination of them?" Answers to this question will not only advance our understanding in people's travel behaviors, but also our predictive capabilities in travel demand forecasting. And such research requires collaboration from both small and big data researchers.

Because a single trip is most likely made in connection to other trips made before or after, trip chaining has long been a focus of many studies, whether viewed from a daily perspective (e.g., daily activity and travel patterns), or a tour level (e.g., a sequence of trips that start and end at the same location). Hence, there has been active development in activity-based models, attempting to account for connections between trips within a day (Pas, 1985; Kitamura et al., 1997; Miller et al., 2004). Because trips are made in the context of where one lives and works, there has been very active research in

[4] A Levy Flight pattern describes a pattern characterized by many small movement steps connected by a few long-distance relocations. A Levy Flight Pattern is truncated (thus, Truncated Levy Flight, or TLF) as movements can only take place in finite space.

understanding and modeling the linkage between long-term choices (e.g., residential location choices) and trip making behaviors (Chen et al., 2008). One subject that has received substantial attention in the last decade is residential self selection—people self-select into neighborhoods with characteristics that match their lifestyles (Cao et al., 2006). Because of the existence of residential self-selection, studies evaluating the effect of the built environment on travel behavior must control residential self-selection. A number of studies have also brought in the life-cycle perspective, seeking to understand the root of residential self-selection (Chen et al., 2009; Chen and Lin, 2011). Because of its longitudinal nature,[5] big data can shed insights in this particular area by identifying individuals who relocated (home or work locations) and analyze how the consequent mobility patterns may have changed. To disentangle residential self-selection (from travel behavior to the built environment) from the built environment effect (from built environment to travel behavior), advanced statistical models that travel behavior researchers are familiar with (e.g., structural equations models) can be employed (Bagley and Mokhtarian, 2002; Chen et al., 2008; Mokhtarian and Cao, 2008).

Research in big data mining can help identify additional factors that explain our travel decisions and uncover the underlying decision mechanisms. For example, a number of studies show the coupling between one's social networks and travel behavior in both time and space (Eagle et al., 2009; Phithakkitnukoon et al., 2012). Given this, we may ponder whether one's social relations lead to consequent travel behaviors or vice versa? Perhaps the relationships exist in both ways. Answers to these questions have several important implications, including shedding new light in additional factors that give rise to observed behaviors, offering potential intervention opportunities to modify one's travel behavior through his or her social network, and motivating or altering social interactions through travel behavior changes.[6]

Advances in travel behavior research can also at the same time propel the big data research in analyzing human mobility patterns. Travel behavior research, for example, has long shown that a person's attitudes and preferences can play an important role in trip-making choices (Gärling et al., 1998; Cao et al., 2008). And contrary to the long-held belief that attitudes remain stable over time, recent research showed that one's attitudes can change after a behavior change (Wang and Chen, 2012). Perhaps these results can motivate big data research to move beyond the current state: primarily analyzing the observed movement patterns, and neglecting the experiences generated and attitudes formed through those movements. In travel behavior research, attitudes are often solicited from a set of statements written by researchers, while the natural forms available through big data (e.g., social media twitter data) may reveal new insights about attitudes and values and stimulating methodological development (e.g., natural language processing). Additionally, a step beyond mostly understanding correlations, but into uncovering causality mechanisms will be welcome. In this regard, research designs (e.g., quasi-experimental designs) (Shadish et al., 2002) and methods (e.g., structural equations systems) widely used in travel behavior research for unpacking causality may be deployed (Bollen, 1989; Golob, 2003).

4. Model development

In transportation research, discrete choice models (Ben-Akiva and Lerman, 1994) are the cornerstone model developed to predict travel behaviors, more specifically choices on modes of transportation, time of day, destinations, and even routes. Discrete choice models are developed based on the theory of random utility maximization (RUM). RUM assumes that when choosing from a set of discrete alternatives (e.g., mode choices and destination choices), the decision maker is aware of all feasible alternatives and their associated attributes, willing to make tradeoffs across attributes, and given these conditions, he/she then chooses the choice that will maximize his or her satisfaction (Ben-Akiva and Lerman, 1994). Both anecdotal and scientific evidence suggest that these assumptions only apply to a certain set of behaviors under certain circumstances (Chorus et al., 2006; Ramos et al., 2011, 2013, 2014). The standard discrete choice models, e.g. Multinomial Logit Model, usually hold some common assumptions, i.e.: (1) the random components of the utilities of the different alternatives are independent and identically distributed (IID) with a type I extreme-value (or Gumbel) distribution; (2) it maintains homogeneity in responsiveness to attributes of alternatives across individuals (i.e., an assumption of response homogeneity); and (3) the error variance–covariance structure of the alternatives is identical across individuals (i.e., an assumption of error variance–covariance homogeneity) (Bhat, 2000). However, these assumptions would be violated when if a travelers perceive a higher utility to all transit modes (bus, train, etc.) because of the opportunity to socialize on board with other travelers (Andrews et al., 2012) or if the public transport system offers different levels of comfort (an unobserved variable) on different routes. To address this problem, three classes of discrete choice models that relax one or more of the assumptions discussed above have now been commonly used. The first class of models (labeled as "heteroscedastic models") is relatively restrictive: they relax the identically distributed (across alternatives) error term assumption, but do not relax the independence assumption (part of the first assumption above) or the assumption of response homogeneity (second assumption above). The second class of models (labeled as "mixed multinomial logit (MMNL) models") and the third class of models (labeled as "mixed generalized extreme value (MGEV) models") are very general; models in this class are flexible enough to relax the independence and identically distributed (across alternatives) error structure of the MNL as well as to relax the assumption of response homogeneity. For further discussion on this, please see Bhat (2000).

[5] To capture long-term choices such as residential location changes, the length of the data will need to be longer than a few months, which is the most typical length of most mobile phone datasets.

[6] Privacy concerns relating to revealing social patterns through phone data will need to be addressed before these connections can be made.

The discovery of certain mobility patterns from the big data offers us an opportunity to identify the links between microscopic individual choices and emergent macroscopic behaviors and to re-examine the decision rules used to model travel related choices. A very first question one can ask is: if every individual makes travel-related decisions as specified by the widely used random utility theory and consequently discrete choice models, what types of mobility patterns will emerge at the population level? What alternative mathematical frameworks will generate patterns that more closely mimic the one observed at the population level? Agent-based simulations may be employed to answer these questions (Kitamura et al., 2000).

Moreover, there is not a cornerstone model (like the discrete choice models) to predict how individual travel behaviors may have changed over time. In this regard, development in recent big data studies can offer insights. Song et al. (2010) showed that the frequency of an individual's visiting the kth most visited location f_k could be approximated as $f_k \sim k^{-\zeta}$, where $\zeta \approx 1.2 \pm 0.1$. They also calculated the maximum predictability, which is the probability that an appropriate prediction algorithm can predict correctly a user's whereabouts, peaks around 0.93. Similar results are found by other researchers (Lu et al., 2012, 2013).

Another potential area of development relates to the formulation of choice sets for destination choice models. It has been a convention that the choice set for discrete choice models includes all alternative locations within an area (McNally, 2000). This practice has been applied to modeling both long-term residential location choices and short-term choices for shopping and recreational activities. It has been widely recognized that such an all-inclusive approach is unrealistic and likely does not conform to how people search in the real world (Chen and Lin, 2012). Studies in transportation research have attempted to extend the discrete choice framework to also model the choice set generation process to account for some factors (e.g., attitudes, perceptions and elimination criteria) (Ben-Akiva and Boccara, 1995). Here, results from the big data research may provide a specific distribution from which alternative choices can be generated. As an example, some big data studies showed a Truncated Levy Flight (TLF) pattern for human mobility when displacement (between cell towers) is used, suggesting that small-scale movements are more randomly distributed (more a circular pattern) while large-scale movements are more likely a function of spatial distribution of the opportunities in the area (more directional bursts). This implies that the formation of choice sets likely varies, depending on the types of choices interested—perhaps the existing all-inclusive approach is more appropriate for short-term choices such as shopping and social recreational activities and for long-term choices such as residential and work locations, what is available in an area may play a more important role. These questions remain yet unanswered.

5. Human mobility patterns

Primarily because of the lack of longitudinal data, the majority of the previous studies in travel behavior research use cross-sectional data. A limited number of studies have collected small data over an extended period, ranging from 3 to 4 days to over one month or longer. Recent examples include the British Household Panel Survey, which is now merged into the UK Household Longitudinal Survey (Clark et al., 2014) and several other datasets (Chatterjee and Ma, 2009; Yanez et al., 2010) which were collected for special purposes. The metrics used include daily trip rates (Pas, 1987; Pas and Koppelman, 1987; Pas and Sundar, 1995), daily travel time (Pas and Sundar, 1995), daily trip distance (Stopher et al., 2007) or combination of some or all (Axhausen et al., 2002), action space (Susilo and Kitamura, 2005), activity time use (Kang and Scott, 2010), and "unique" trip sequences (Joh et al., 2002; Moiseeva et al., 2014). These studies reveal both a substantial amount of repetition and variability, when compared between consecutive days (Jou and Mahmassani, 1997; Ma and Goulias, 1997; Goulias, 1999; Simma and Axhausen, 2001; Srivastava and Schoenfelder, 2003; Buliung et al., 2008; Roorda and Ruiz, 2008), or examined in the combinations of two attributes (e.g., shopping and car) (Hanson and Huff, 1982, 1986; Schlich et al., 2004). The amount of intra-person variability is found to vary from over 20% to about 80%, depends on the metrics used, the type of the trips and also the days of the analysis (Kang and Scott, 2010; Chikaraishi et al., 2011; Susilo and Axhausen, 2014). Studies that developed similarity indexes incorporating multiple dimensions of a daily activity and travel patterns also reach the same conclusion (Hanson and Huff, 1982, 1986; Pas, 1983, 1988; Schlich and Axhausen, 2003).

Studies using the big data tend to focus on identifying mobility patterns from which predictions can be made. Instead of using meaningful metrics such as trip rates, travel distance or travel time,[7] these studies analyze displacements between inferred stops (e.g., between two cell power locations) and treat them as a series of random events whose spatial and temporal distributions are assumed to possess certain statistical regularities (Kang et al., 2012a, 2012b). These results, while informative, are derived using metrics (e.g., displacements) whose meanings are not clear. In majority of the datasets involving cell phone towers, displacements are straight-line distances between two consecutively observed cell phone towers. An additional note is that measurements of straight-line distances are less meaningful than network distances as both humans and vehicles follow defined ways on earth, under most common circumstances. In datasets involving cell phone sightings, the meaning of a displacement has more clarity, as one must first infer activity locations (or origins and destinations) from the data. However, since most of the studies have tried to mine activity locations instead of trajectories[8] and few studies have attempted validating (Calabrese et al., 2013; Chen et al., 2014), the meaning of a displacement still is unclear. In other words, the direct use of displacements in transportation applications is inappropriate, with consequences unknown.

[7] For transportation planning applications, travel time and travel distances need to be measured between two meaningful locations.

[8] Though mining activity locations constitutes a very first step prior to mining mobility trajectories, the two are not equivalent.

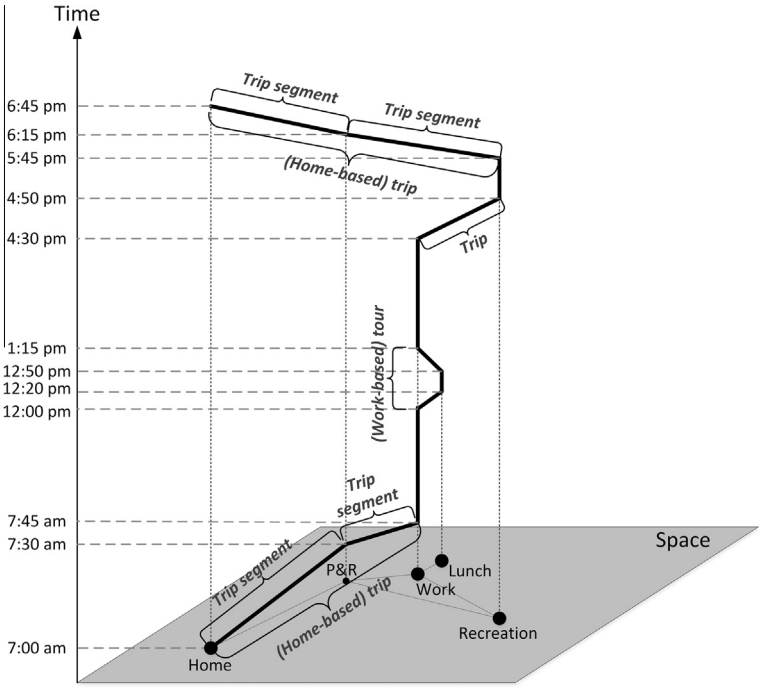

Fig. 5. A 3D-illustration of trips, trip segments, tours and daily activity and travel patterns.

A very first direction for big data researchers is the clarification of the meaning of a displacement and this requires two essential steps: first, validation of methods used to mine activity locations and trajectories from the big data (Chen et al., 2014) and second, the use of meaningful terms such as segments, trips, tours, or patterns. As noted earlier, the limited amount of work on validation is mostly at the area level, for example, comparing population density calculated from the big data against the census data (Calabrese et al., 2013), or comparing shares of trip purposes against a travel survey (Gong et al., 2015).[9] While these initial attempts are promising, the field needs many more studies, in particular, those that can do the validation at the individual level, explicitly answering questions such as "what percentage of the individuals in a sample that have their activity locations and trajectories correctly or wrongly inferred?"

To understand some of the meaningful travel-related metrics frequently used by transportation researchers, Fig. 5 provides a 3D illustration of what is, to a transportation researcher, a segment, a trip, a tour, and a daily pattern. A trip refers to a movement between two meaningful locations, for example, from home to work in the morning, which is also a home-based trip (defined as trips whose one end is home). In Fig. 5, the trip from home to work has two segments, comprising a drive to a P&R lot and a bus ride (depicted by two slopes for different travel speeds). At noon, the person goes from her work place to a lunch place, eats lunch there and then returns to work. This sequence of two trips (from work to lunch, and then back to work) constitutes a work-based tour (defined as a sequence of trips that start and end at the same location). Both of these two trips are non-home-based trips. After work, this person goes to a recreation place, spend some time there, and then returns home. The trip from work to recreation is again non-home-based, but the trip from recreation to home is home-based. Together, all trips from the first one leaving home to the last one returning home constitutes a daily activity and travel pattern.

In addition, whenever possible, real-world transportation networks shall be brought into the picture, so that meaningful network distances are measured. Once these steps are completed, many important follow-up questions can be asked, including for example:

1. When measured with clear origins and destinations and in meaningful metrics (e.g., travel time, travel/network distances), what particular mobility pattern will emerge from the big data?
2. How does an observed pattern vary when different modes of transportation, or different trip purposes are examined?
3. What is the role of the underlying built environment in regulating the observed patterns?

It is our view that the full potential of the longitudinal nature of the big data (often lasting over 1 month) is still yet to be fully explored. This is particularly important when meaningful metrics are used. We ask some sample questions and by no

[9] This particular study inferred trip purposes from taxi trajectory data and compared them against a travel survey.

means these questions encompass the full set of questions that will significantly advance our knowledge in human mobility patterns:

1. What are the various regularity rhythms (e.g., daily vs weekly) associated with different types of trips and how are these cycles *affected*[10] by the built environment?
2. How do these various rhythms interact with each other in forming one's mobility patterns over time?
3. How does intra-person variability change over time?

6. Discussions

This paper is written with a simple purpose of bringing together two groups of researchers in analyzing human mobility patterns: travel behavior researchers who have long relied on household travel surveys (small data) and big data researchers who have recently used passively-generated data (big data) (Chen et al., 2015). To achieve this, we introduce some key concepts and developments in each field and raise some cross-disciplinary ideas in three sub-areas: (1) behavioral factors (Section 3); (2) model development (Section 4); and (3) movement patterns (Section 5).

The list of ideas proposed is by no means an exhaustive one, but serves the important purpose of stimulating a discussion from both groups of researchers and forming collaborations between the two. Indeed, recent papers have shown a convergence between the two groups. Some of the well-known theories in travel behavior are re-discovered in big data research, for example, the spatial and temporal fixities associated with various activities due to various constraints (e.g., biological) (Hägerstrand, 1970; Golledge and Stimson, 2007; Schwanen et al., 2008), the history and time of day dependence (Kitamura et al., 1997, 1998), and the importance of understanding activities behind the trips (Pas, 1985; Kitamura, 1988). These re-discoveries are welcoming, but the field will benefit much more by having more communications between the two groups. In addition, recent big data studies have started moving toward an analysis scale at the individual level, recognizing that there is likely a great amount of heterogeneity in a population (Liu et al., 2014), corroborating the decades-long history in individual-based, activity-based research by transportation researchers (Pas, 1985; Kitamura, 1988).

The paper has likely, inevitably left out additional areas that collaborations between the two groups would be mutually beneficial. One area is the determination of zones within a study area. While transportation researchers have long used Traffic Analysis Zones (TAZ), big data researchers have often relied upon uniform grids or lattices. Conversations between the two groups can potentially answer some fundamental questions such as what constitutes a zone in the context of analyzing human mobility patterns and what methods can be applied to identify a zone using the data that is available? Great benefits can be derived if both transportation and big data researchers can attempt to answer these questions together. The resulting collaborations between the two can result in high-quality studies whose analyses are based on sound understanding of how individuals move in time and space. Equally important, such studies will be directly relevant to the development of transportation decisions and policies, in terms how we can develop policies (short-term and long-term) to encourage more environmentally-friendly travel behaviors and build toward a sustainable future.

An additional note of caution as we race toward big data research is the awareness of the amount of fallacies we will create along the way. In Michael Jordan's words, the analogy of predicting something with mass amount of data is like "having billions of monkeys typing and one of them will write Shakespeare" (Gomes, 2014). It is our view that conceptualization of frameworks and hypotheses of ideas as long used by transportation researchers with the small data shall *not* be abandoned and continued applications of important research techniques will only strengthen big data research. It is no doubt that big data research will help advance the field of transportation research and at the same time, it is equally important to be aware of their pitfalls. It is indeed those pitfalls that make collaborations with transportation researchers potentially cross-cutting.

Acknowledgements

Funding for this research is provided by a NSF (National Science Foundation) Grant (CMMI 1200275) and a NIH (National Institute of Health) grant (1R01GM108731-01A1) to Cynthia Chen. The authors sincerely thank the help of Xiangyang Guan, a PhD student in the Department of Civil and Environmental Engineering, University of Washington (Seattle), for his help in creating the figures.

References

Ahonen, T.T., Moore, A., 2006. A Mobile Phone for Every Living Person in Western Europe: Penetration Hits 100%. Communities Dominate Brands, 2013.
AIRSAGE, 2013. How Accurate is the Data? <http://wwwairsage.com/Technology/Accuracy/> (retrieved July 29, 2013).
Ajzen, I., 1991. The theory of planned behavior. Organ. Behav. Hum. Decis. Process. (50), 179–221
Alexander, L., Jiang, S., Murga, M., Gonzalez, M., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C 58, 240–250.

[10] By italic, we emphasize that this question can only be answered by longitudinal dataset. Current studies touching upon this issue have largely answered the question "how do human mobility patterns vary by the built environment?

Andrews, G., Parkhurst, G., Susilo, Y.O., Shaw, J., 2012. 'The Grey Escape': How and why are older people using their free bus passes? J. Transp. Plann. Technol. 35 (1), 3–15.
Axhausen, K.W., Schonfelder, S., Wolf, J., Oliveira, M., Samaga, U., 2003. 80 weeks of GPS-traces: approaches to enriching the trip information. Arbeitsbericht Verkehrs- und Raumplanung. ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme, 178.
Axhausen, K.W., Zimmerman, A., Schonfelder, S., Rindsfuser, G., Haupt, T., 2002. Observing the rhythms of daily life: a six week travel diary. Transportation 29, 95–124.
Bagley, M.N., Mokhtarian, P.L., 2002. The impact of residential neighborhood type on travel behavior: a structural equations modeling approach. Ann. Reg. Sci. 36, 279–297.
Bamberg, S., Hunecke, M., Blohbaum, 2007. Social context, morality, and the use of public transportation: two field studies. J. Environ. Psychol. 27 (3), 190–203.
Bamberg, S., Rolle, D., Weber, C., 2003. Does habitual car use not lead to more resistance to change of travel mode? Transportation 30, 97–108.
Bayir, M.A., Demirbas, M., Eagle, N., 2010. Mobility profiler: a framework for discovering mobility profiles of cell phone users. Pervasive Mobile Comput. 6, 435–454.
Ben-Akiva, M., Boccara, B., 1995. Discrete choice models with latent choice sets. Int. J. Res. Mark. 12, 9–24.
Ben-Akiva, M., Lerman, S., 1994. Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge, MA.
Bhat, C., 2000. Flexible model structures for discrete choice analysis. In: Hensher, D., Button, K. (Eds.), Handbook of Transport Modeling. Pergamon, Amsterdam, pp. 71–89.
Blumenstock, J., Cadamuro, C., On, R., 2015. Predicting poverty and wealth from mobile phone metadata. Science 350 (6264), 1073–1076.
Bohte, W., Maat, K., 2008. Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. In: 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability. Annecy, France.
Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp. Res. Part C 17, 285–297.
Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.
Buliung, R.N., Roorda, M.J., Remmel, T.K., 2008. Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey (TTAPS). Transp. Res. Part A 35, 697–722.
Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011. Real time urban monitoring using cell phones: a case study in Rome. IEEE Trans. Intell. Transp. Syst. 12 (1), 141–151.
Calabrese, F., Diao, M., Lorenzo, G.D., Ferreira Jr., J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. Part C 26, 301–313.
Calabrese, F., Lorenzo, G.D., Pereira, F.C., Liu, L., Ratti, C., 2010. Analyzing Cell-phone Mobility and Social Events. NetMob-Analysis of Mobile Phone Networks, Cambridge, MA.
Cao, X., Handy, S.L., Mokhtarian, P.L., 2006. The influences of the built environment and residential self-selection on pedestrian behavior: evidence from Austin, TX. Transportation 33, 1–20.
Cao, X., Mokhtarian, P., Handy, S., 2008. Differentiating the influence of accessibility, attitudes, and demographics on stop participation and frequency during the evening commute. Environ. Plann. B 35, 431–442.
Chatterjee, K., Ma, K., 2009. Time taken for residents to adopt a new public transport service: examining heterogeneity through duration modeling. Transportation 36, 1–25.
Chen, C., Batty, M., van Vuren, T., 2015. Special issue: emerging, passively generated datasets for travel behavior and policy analysis. Transportation 42 (4), 537–540.
Chen, C., Bian, L., Ma, J., 2014. From sightings to activity locations: how well can we guess the locations visited from mobile phone sightings. Transp. Res. Part C 46 (10), 326–337.
Chen, C., Chen, J., Timmermans, H., 2009. Historical deposition influence in residential location decisions: a distance-based GEV model for spatial correlation. Environ. Plann. A 41 (11), 2760–2777.
Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. Transp. Res. Part A 44 (10), 830–840.
Chen, C., Gong, H., Paaswell, R., 2008. Role of the built environment on mode choice decisions: additional evidence on the impact of density. Transportation 35 (3), 285–299.
Chen, C., Lin, H., 2011. Decomposing residential self-selection via a life course perspective. Environ. Plann. A 43 (11), 2608–2625.
Chen, C., Lin, H., 2012. How far do people search? Analyzing the roles of housing supply, intra-household dynamics, and the use of information channels. Housing Stud. 27 (7), 898–914.
Chen, Z., Schintler, L.A., 2015. Sensitivity of location-sharing services data: evidence from American travel pattern. Transportation 42 (4), 669–682.
Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z., 2011. Exploring millions of footprints in location sharing services. In: 5th International AAAI Conference on Web and Social Media. Barcelona, Spain, pp. 81–88.
Chikaraishi, M., Fujiwara, A., Zhang, J., Axhausen, K.W., Zumkeller, D., 2011. Changes in variations of travel time expenditure: some methodological considerations and empirical results from German mobility panel. Transp. Res. Rec. 2230, 121–131.
Chorus, C.G., Arentze, T., Timmermans, H., 2008. A random regret-minimization model of travel choice. Transp. Res. Part B 42 (1), 1–18.
Chorus, C.G., Molin, E.J.E., van Wee, G.P., 2006. Response to transit information among car drivers: regret-based modeling and simulations. Transp. Plann. Technol. 29 (4), 249–271.
Chung, E., Kuwahara, M., 2007. Mapping personal trip OD from probe data. Int. J. ITS Res. 5 (1), 1–6.
Chung, E.-H., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys. Transp. Plann. Technol. 28 (5), 381–401.
Clark, B., Chatterjee, K., Melia, S., Knies, G., Laurie, H., 2014. Life events and travel behaviour: exploring the interrelationship using UK household longitudinal study data. J. Transp. Res. Rec. 2413, 54–64.
Couronné, T., Raimond, A.O., Smoreda, Z., 2013. Chatty Mobiles:Individual mobility and communication patterns. CoRR, abs/1301.6553.
Eagle, N., Pentland, A., Lazer, D., 2009. Inferring social network structure using mobile phone data. Proc. Natl. Acad. Sci. (PNAS) 106 (36), 15274–15278.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd.
Experian Simmons, 2011. The 2011 Mobile Consumer Report <http://www.experian.com/assets/simmons-research/white-papers/experian-simmons-2011-mobile-consumer-report.pdf> (retrieved July 13, 2013).
Ficek, M., Kencl, L., 2012. Inter-call mobility model: a spatio-temporal refinement of call data records using a Gaussian mixture model. Proceedings of INFOCOM, IEEE.
Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. 97, 611–632.
Freedman, D., 2002. The Ecological Fallacy <http://www.stat.berkeley.edu/~census/ecofall.txt> (retrieved July 1, 2015).
Gärling, T., Gillholm, R., Gärling, A., 1998. Reintroducing attitude theory in travel behavior research. Transportation 25, 129–146.
Golledge, R., Stimson, R.J., 2007. Spatial Behavior. The Guilford Press, New York.
Golob, T.F., 2003. Structural equation modeling for travel behavior research. Transp. Res. Part B 37, 1–25.
Gomes, L., 2014. Machine-learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts <http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts> (retrieved July 7th, 2015).
Gong, H., Chen, C., Bialostozky, E., Lawson, C., 2011. A GPS/GIS method for travel mode detection in New York City. Comput. Environ. Urban Syst. 36 (2), 131–139.

Gong, L., Liu, X., Wu, L., Liu, Y., 2015. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartogr. Geogr. Inform. Sci. 43 (2), 103–114.
Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (5), 779–782.
Goulias, K.G., 1999. Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. Transp. Res. Part B 33, 535–557.
Hägerstrand, T., 1970. What about people in regional science? Papers Reg. Sci. Assoc. 24 (1), 6–21.
Hanson, S., Huff, J., 1982. Assessing day-to-day variability in complex travel patterns. Transp. Res. Rec.: J. Transp. Res. Board 891, 18–24.
Hanson, S., Huff, J., 1986. Classification issues in the analysis of complex travel behavior. Transportation 13, 271–293.
Hato, E., Shinji, I., Mitani, T., 2006. Development of MoALs (Mobile Activity Loggers supported by GPSphones) for travel behavior analysis. In: Proceedings of the 85th Annual Meeting of the Transportation Research Board. Washington D.C..
Huang, L., Li, Q., Yue, Y., 2010. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. ACM, New York, pp. 27–30.
Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., Smoreda, Z., 2013. Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. Geogr. Inform. Sci. Heart Europe Lect. Notes Geoinform. Cartogr., 247–265
Iqbal, M.S., Choudhury, C.F., Wang, P., Gonzalez, M., 2014. Development of origin-destination matrices using mobile phone call data. Transp. Res. Part C 40, 63–74.
Jiang, B., Yin, J., Zhao, S., 2009. Characterizing the human mobility pattern in a large street network. Phys. Rev. E 80 (021136), 1–11.
Jiang, S., Fiore, G.A., Yang, Y., Ferreia, J., Frazzoli, E., Gonzalez, M.C., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp. Chicago, IL, pp. 1–9.
Joh, C.H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity-travel pattern similarity: a multi-dimensional alignment method. Transp. Res. Part B 36, 385–403.
Jou, R.-C., Mahmassani, H.S., 1997. Comparative analysis of day-to-day trip-chaining behavior of urban commuters in two cities. Transp. Res. Rec.: J. Transp. Res. Board 1607, 163–170.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263.
Kang, C., Liu, Y., Ma, X., Wu, L., 2012a. Towards estimating urban population distributions from mobile call data. J. Urban Technol. 19 (4), 3–21.
Kang, C., Ma, X., Tong, D., Liu, Y., 2012b. Intra-urban human mobility patterns: an urban morphology perspective. Physica A: Stat. Mech. Appl. 391 (4), 1702–1717.
Kang, H., Scott, D.M., 2010. Exploring day-to-day variability in time use for household members. Transp. Res. Part A 44, 609–619.
Kitamura, R., 1988. An evaluation of activity-based travel analysis. Transportation 15, 9–34.
Kitamura, R., Chen, C., Narayanan, R., 1998. Traveler destination choice behavior: effects of time of day, activity duration, and home location. Transp. Res. Rec. 1645, 76–81.
Kitamura, R., Chen, C., Pendyala, R., 1997. Generation of synthetic activity-travel patterns. Transp. Res. Rec. 1607, 154–162.
Kitamura, R., Chen, C., Pendyala, R., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation 27 (1), 25–51.
Larijani, A.N., Olteanu-Raimond, A., Perret, J., Bredif, M., Ziemlicki, C., 2015. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. Transp. Res. Procedia 6, 64–78.
Lee, J.-K., Hou, J.C., 2006. Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application. In: Proceedings of the 7th ACM International Symposium on Mobile Ad hoc Networking and Computing. ACM, pp. 85–96.
Liu, Y., Kang, C., Gao, S., Xiao, Y., Tian, Y., 2012. Understanding intra-urban trip patterns from taxi trajectory data. J. Geogr. Syst. 14 (4), 463–483.
Liu, Y., Kang, C., Wang, F., 2014. Towards big data-driven human mobility patterns and models. Geomat. Inform. Sci. Wuhan Univ. 39 (6), 660–666.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: a new approach to understanding our socio-economic environments. Ann. Assoc. Am. Geogr. 105 (3), 512–530.
Long, Y., Thill, J.C., 2015. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. Comput. Environ. Urban Syst. 53, 19–35.
Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proc. Natl. Acad. Sci. 109 (29), 11576–11581.
Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L., 2013. Approaching the limit of predictability in human mobility. Sci. Rep. 3, 2923.
Ma, J., Goulias, K.G., 1997. A dynamic analysis of person and household activity and travel patterns using data from the first two waves in the Puget Sound Transportation Panel. Transportation 24, 309–331.
Ma, J., Yuan, F., Joshi, C., Li, H., Bauer, T., 2012. A new framework for development of time-varying O–D matrices based on cellular phone data. In: 4th TRB Innovations in Travel Modeling (ITM) Conference. Tampa, FL.
Ma, X., Wu, Y., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C 36 (1–12).
McNally, M.G., 2000. The four step model. In: Hensher, D., Button, K. (Eds.), Handbook of Transport Modeling. Pergamon, Amsterdam, pp. 35–52.
Miller, E., Hunt, J.D., Abraham, J.E., Salvini, P.A., 2004. Microsimulating urban systems. Comput. Environ. Urban Syst. 28 (1–2), 9–44.
Moiseeva, A., Timmermans, H., Choi, J., Joh, C.H., 2014. Sequence alignment analysis of activity-travel pattern's variability using eight weeks' diary data. Transp. Res. Rec. 2412, 49–56.
Mokhtarian, P.L., Cao, X., 2008. Examining the impacts of residential self-selection on travel behavior: a focus on methodologies. Transp. Res. Part B: Methodol. 42 (3), 204–228.
Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A tale of many cities: universal patterns in human mobility. PLoS ONE 7 (5), e37027.
Ortuzar, J.d.D., Willumsen, L.G., 2011. Modeling Transport. John Wiley & Sons.
Pas, E., 1985. State of the art and research opportunities in travel demand: another perspective. Transp. Res. Part A 19A, 460–464.
Pas, E.I., 1983. A flexible and integrated methodology for analytical classification of daily travel-activity behavior. Transp. Sci. 17 (4), 405–429.
Pas, E.I., 1987. Intrapersonal variability and model goodness-of-fit. Transp. Res. Part A 21A (6), 431–438.
Pas, E.I., 1988. Weekly travel-activity behavior. Transportation 15, 89–109.
Pas, E.I., Koppelman, F.S., 1987. An examination of the determinants of day-to-day variability in individuals' urban travel behavior. Transportation 14, 3–20.
Pas, E.I., Sundar, S., 1995. Intrapersonal variability in daily urban travel behavior: some additional evidence. Transportation 22, 135–150.
Patel, S.N., Kientz, J.A., Hayes, G.R., Bhat, S., Abowd, G.D., 2006. Farther than you may think: an empirical investigation of the proximity of users to their mobile phones. In: UbiComp 2006: Ubiquitous Computing, pp. 123–140.
Pelletier, M.-P., Trepanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C 19 (4), 557–568.
Phithakkitnukoon, S., Horanont, T., Lorenzo, G.D., Shibasaki, R., Ratti, C., 2010. Activity-aware Map: Identifying Human Daily Activity Pattern using Mobile Phone Data. MIT-Senseable City Lab, Boston.
Phithakkitnukoon, S., Smoreda, Z., Olivier, P., 2012. Social-geography of human mobility: a study using longitudinal mobile phone data. PLoS ONE 7 (6), e39253.
Ramos, G.M., Daamen, W., Hoogendoorn, S.P., 2011. Expected utility theory, prospect theory, and regret theory compared for prediction of route choice behavior. Transp. Res. Rec. 2230, 19–28.
Ramos, G.M., Daamen, W., Hoogendoorn, S.P., 2013. Modeling travelers' heterogeneous route choice behavior as prospect maximizers. J. Choice Model. 6 (17–33).
Ramos, G.M., Daamen, W., Hoogendoorn, S.P., 2014. A state-of-the-art review: developments in utility theory, prospect theory and regret theory to investigate travellers' behaviour in situations involving travel time uncertainty. Transp. Rev. 34 (1), 46–67.
Ranjan, G., Zang, H., Zhang, Z.-L., Bolot, J., 2012. Are call detail records biased for sampling human mobility? Mobile Comput. Commun. Rev. 16 (3), 33–44.

Rasouli, S., Timmermans, H., 2015. Bounded Rational Choice Behavior: Applications in Transport. Emerald Group Publishing Limited, United Kingdom.
Roorda, M.J., Ruiz, T., 2008. Long- and short-term dynamics in activity scheduling: a structural equations approach. Transp. Res. Part A 42, 545–562.
Schlaich, J., Otterstatter, T., Friedrich, M., 2010. Generating trajectories from traces. In: 89th Annual Meeting of the Transportation Research Board. Washington, D.C..
Schlich, R., Axhausen, K.W., 2003. Habitual travel behaviour: evidence from a six-week travel diary. Transportation 30, 13–36.
Schlich, R., Schönfelder, S., Hanson, S., Axhausen, K.W., 2004. Structures of leisure travel: temporal and spatial variability. Transp. Rev. 24 (2), 219–237.
Schneider, C.M., Belik, V., Couronne, T., Smoreda, Z., Gonzalez, M., 2013. Unraveling daily human mobility motifs. J. Roy. Soc. Interface 10 (84), 20130246.
Schwanen, T., Kwan, M.P., Ren, F., 2008. How fixed is fixed? Gendered rigidity of space time constraints and geographies. Geoforum 39 (6), 2109–2121.
Schwartz, S.H., 1977. Normative influence on altruism. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, New York, pp. 221–279.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, New York.
Shi, L., Chi, G., Liu, X., Liu, Y., 2015. Human mobility patterns in different communities: a mobile phone data based social network approach. Ann. GIS 21 (1), 15–26.
Simma, A., Axhausen, K., 2001. Successive days, related travel behaviour? Arbeitsbericht Verkehrs- und Raumplanung, 62.
Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010. Modelling the scaling properties of human mobility. Nat. Phys. 6, 818–823.
Spinsanti, L., Celli, F., Renso, C., 2010. Where you stop is who you are: understanding people's activities by places visited. In: Proceedings of the Workshop on Behavior Monitoring and Interpretation (BMI 2010). Kuala Lumpur.
Srivastava, G., Schoenfelder, S., 2003. On the Temporal Variation of Human Activity Spaces. ETH Zürich, Institute für Verkehrsplanung und Transportsysteme (IVT), Zürich.
Stern, P.C., 2000. Toward a coherent theory of environmentally significant behavior. J. Soc. Issues 56, 407–424.
Stopher, P., Clifford, E., Zhang, J., FitzGerald, C., 2008b. Deducing mode and purpose from GPS data. Working Paper of the Austrian Key Centre in Transport and Logistics. Sydney, Australia, University of Sydney.
Stopher, P., FitzGerald, C., Zhang, J., 2008b. Search for a global positioning system device to measure person travel. Transp. Res. Part C 16, 350–369.
Stopher, P., FitzGerald, C., Zhang, J., Bretin, T., 2007. Analysis of 28-day global positioning system panel survey. Transp. Res. Rec.: J. Transp. Res. Board 2014, 17–26.
Susilo, Y.O., Axhausen, K.W., 2014. Stability in individual daily activity-travel-location patterns: a study using the Herfindahl-Hirschman Index. Transportation 41, 995–1011.
Susilo, Y.O., Kitamura, R., 2005. Analysis of day-to-day variability in an individual's action space. Transp. Res. Rec.: J. Transp. Res. Board 1902, 124–133.
Tettamanti, T., Demeter, H., Varga, I., 2012. Route choice estimation based on cellular signaling data. Acta Polytech. Hungarica 9 (4), 207–220.
TMIP, 2013. Household Surveys at a Glance. Federal Highway Administration, Washington, D.C..
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: travel demand estimation using big data resources. Transp. Res. Part C 58, 162–177.
Tversky, A., Kahneman, D., 1986. Rational choice and the framing of decisions. J. Bus. 59 (S4), S251.
Wang, M., 2014. Understanding Activity Location Choice with Mobile Phone Data. Civil and Environmental Engineering (Ph.D). Seattle, University of Washington, Seattle.
Wang, H., Calabrese, F., DiLorenzo, G., Ratti, C., 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: Proceedings of 13th International IEEE Conference on ITS. Funchal, Portugal, pp. 318–323.
Wang, M., Chen, C., 2012. Attitudes, mode switching behavior, and the built environment: a longitudinal study in the Puget Sound Region. Transp. Res. Part A 46, 1594–1607.
Wang, J., Wei, D., He, K., Gong, H., Wang, P., 2014. Encapsulating urban traffic rhythms into road networks. Sci. Rep. 4, 4141.
Wang, P., Hunter, T., Bayen, A., Schechtner, K., Gonzalez, M.C., 2012. Understanding road usage patterns in urban areas. Sci. Rep. 2, 1001.
Weiner, E., 1999. Urban Transportation Planning in the United States: An Historical Overview. Praeger, Santa Barbara, CA.
Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: 80th Annual Meeting of the Transportation Research Board. Washington, D.C., p. 24.
Xie, K., Deng, K., Zhou, X., 2009. From trajectories to activities: a spatio-temporal join approach. In: Proceedings of the 2009 International Workshop on Location Based Social Networks. ACM, New York, pp. 25–32.
Yanez, M.F., Mansilla, P., Ortúzar, J.D., 2010. The Santiago panel: measuring the effects of implementing Transantiago. Transportation 37, 125–149.
Ye, Y., Zheng, Y., Chen, Y., Feng, J., Xie, X., 2009. Mining individual life pattern based on location history. In: Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. IEEE.
Zandlebergen, P., 2009. Accuracy of iPhone locations: a comparison of assisted GPS, WiFi and cellular positioning. Trans. GIS 13 (s1), 5–26.