senseable city lab:.::

Transportation Research Part C 26 (2013) 301–313

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Understanding individual mobility patterns from urban sensing data: A mobile phone trace example

Francesco Calabrese [a,d], Mi Diao [b,e,*], Giusy Di Lorenzo [a,d], Joseph Ferreira Jr. [c], Carlo Ratti [a,c]

[a] SENSEable City Lab, Massachusetts Institute of Technology, USA
[b] Department of Real Estate, National University of Singapore, Singapore
[c] Department of Urban Studies and Planning, Massachusetts Institute of Technology, USA
[d] IBM Research, Dublin, Ireland
[e] Institute of Real Estate Studies, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form 22 September 2012
Accepted 24 September 2012

Keywords:
Mobility analysis
Mobile phone traces
Vehicle Kilometers Traveled (VKT)

ABSTRACT

Large-scale urban sensing data such as mobile phone traces are emerging as an important data source for urban modeling. This study represents a first step towards building a methodology whereby mobile phone data can be more usefully applied to transportation research. In this paper, we present techniques to extract useful mobility information from the mobile phone traces of millions of users to investigate individual mobility patterns within a metropolitan area. The mobile-phone-based mobility measures are compared to mobility measures computed using odometer readings from the annual safety inspections of all private vehicles in the region to check the validity of mobile phone data in characterizing individual mobility and to identify the differences between individual mobility and vehicular mobility. The empirical results can help us understand the intra-urban variation of mobility and the non-vehicular component of overall mobility. More importantly, this study suggests that mobile phone trace data represent a reasonable proxy for individual mobility and show enormous potential as an alternative and more frequently updatable data source and a compliment to the conventional travel surveys in mobility study.

1. Introduction

The transportation sector plays an important role in global sustainability. In 2004, it accounts for 22% of primary energy use and 27% of CO_2 emissions all over the world (de Ia Rue du Can and Price, 2008). Individual mobility consumes about two thirds of the total transportation energy use.[1] Understanding the intra-urban variation of individual mobility is important for policy makers to perform "what if" analysis of the environmental consequences of alternative development scenarios and land use controls, and develop regional growth strategies towards a more sustainable future.

The majority of empirical studies on mobility rely on travel surveys, because they provide detailed descriptions of demographics, place of residence, and travel attributes at an individual or household level to support modeling. However, travel surveys are not without limitations:

* Corresponding author at: Department of Real Estate, National University of Singapore, Singapore. Tel.: +65 65163418.
E-mail address: rstdm@nus.edu.sg (M. Diao).

[1] https://www.iea.org/impagr/cip/archived_bulletins/issue_no23.htm.

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.09.009

- The usage of small sample due to the high expense of travel surveys. For example, the National Household Travel Survey "add-on" program is among the largest available travel surveys, but it involves at most 10,000 observations for participating states and a few thousand observations for individual metropolitan areas. As a result, there are not many respondents included in any one neighborhood, which limits the efforts to adequately understand travel patterns for small areas (Handy, 1996).
- The limitation in spatial and temporal scales of the collected datasets. Privacy concerns often limit the geographic specificity with which trip origins and destinations can be revealed, thus spatial effects at fine-grained scales cannot be identified. The surveys normally only collect the travel diary of households within 1–2 days due to concerns in respondent burden. Therefore, complete household activity schedules cannot be observed and some important mobility patterns, such as intra-week and seasonal variations, are also neglected.
- The low update frequency. Survey data are normally updated every 5–10 years, which limit the responsiveness of related urban policies in addressing the rapid metropolitan growth and socio-economic, demographic, infrastructure and travel behavior changes that may have occurred or are projected to occur in the foreseeable future.

During the last two decades, we have seen an explosion in the deployment of pervasive systems like cellular networks, GPS devices, and WiFi hotspots that allow us to capture massive amounts of real-time data related to people and cities (Reades et al., 2007; Gonzalez et al., 2008; Wang et al., 2009). The usage of these datasets could enable researchers to better understand the laws governing people's movements and improve the efficiency and responsiveness of urban policies. This study aims to explore the potential value and challenges of these novel datasets in urban modeling, using the mobile phone trace data collected by mobile network operators as an example.

Compared to travel survey data, the mobile phone trace data provide researchers new opportunities to examine individual mobility from an alternative perspective with their lower collection cost, larger sample size, higher update frequency, and broader spatial and temporal coverage. The mobile phone locations are routinely collected by operators for network management purposes, therefore, the datasets are theoretically available at no cost to researchers. The datasets allow for studying individual mobility of millions of people across the metropolitan area over a longer time period compared to a few thousand households' movements within 1–2 days usually collected through travel surveys. They are updated on a real time basis, which could lead to more reliable and trackable urban performance indicators and support more prompt policy responses to emerging urban issues.

Meanwhile, the mobile phone trace data also have significant drawbacks for transportation research: (1) socio-economic and demographic attributes are not available due to privacy concerns, which are indispensable to calibrate models at disaggregate level to explore the underlying behavior mechanism of individual/household mobility choice; (2) mobile phone users might not represent a random sample of the population. The results need careful analyses to be properly interpreted; and (3) the datasets are not primarily designed for modeling purposes and are often not in an easy-to-use format, which restricts the usefulness of raw data without intensive processing.

This study represents a first step towards building a methodology to utilize mobile phone data for transportation research. In this study, we use mobile phone traces from about one million users in the Boston Metropolitan Areas, Massachusetts, USA over 3 months to characterize individual mobility and understand its spatial patterns within a metropolitan area.

To address the lack of socio-economic and demographic attributes in mobile phone data, we aggregate mobility measures generated from individual mobile phone traces to block groups, the most disaggregate level of census geography at which socio-economic and demographic information is available, and associate the aggregate mobility measures with census data. Given the aggregate nature of our analysis, we raise two cautions at the outset. First, the underlying behavior mechanism cannot be identified by this study. The second caution concerns the ecological fallacy, which is the fallacy related to inferring the nature of individuals based solely upon aggregate statistics collected for the group. Therefore, it should be noted that what we find in this study is the general spatial patterns of mobility within the metropolitan area and their relationships to neighborhood characteristics, but not how individuals' characteristics and built environment influence their own travel behavior.

Nonetheless, using aggregate data collected for a long time period could help screen the idiosyncratic factors at the individual level, identify the underlying trends, and explain the variation in intra-urban mobility patterns. As Yang (2008) and Yang and Ferreira (2008) demonstrate, even without individual preference, urban spatial structure alone could explain a significant portion of the variation in commuting distance.

Similar analytic approaches that involve data aggregation have been adopted by many previous studies in mobility research. Lindsey et al. (2011) explore the relationship between Vehicle Kilometers Traveled (VKT) and urban form characteristics at grid cell level. Yang (2008) examines the relationship between excess commuting distance and urban spatial structure at census tract level. Holtzclaw et al. (2002) find that the average annual distance driven per car at the Traffic Analysis Zone (TAZ) level is a strong function of density, income, household size and public transit. Wang (2001) explains intra-urban variations of commuting time and distance at TAZ level in Columbus, Ohio using Census Transportation Planning Package (CTPP) data. Miller and Ibrahim (1998) examine the relationship between urban form and work trip VKT at traffic zone level in Toronto. These studies provide useful insights into individual mobility, but their data are mostly aggregated from travel surveys. Therefore they have similar data-related shortcomings, such as high data collection expense and low update frequency.

To validate the representativeness of mobile phone users, we compare mobility measures generated from mobile phone traces with mobility measures computed using odometer readings from annual safety inspections of all private vehicles registered in the Boston Metropolitan Area. By combining these two datasets, it is possible to accurately characterize both individual mobility (the movement of individuals regardless of transportation modes, measured by individual total daily trip length, computed using the mobile phone trace data), and vehicular mobility (the movement of vehicles, measured by vehicular total daily trip length, computed using the vehicle safety inspection data), and investigate how each of them varies across the region.

The paper is structured as follows: Section 2 describes the two datasets considered: mobile phone traces and vehicle odometer readings. Section 3 presents the methodology defined for the mobility analysis and its application to both datasets. Finally discussion and conclusion are given in Sections 4 and 5.

2. Datasets

In this section we introduce the mobile phone trace data and vehicle safety inspection data, and describe the methods used to extract mobility statistics from both datasets for people in the Boston Metropolitan Area.

2.1. Mobile phone trace data

The mobile phone trace data used in this study consist of anonymous location estimations collected by AirSage[2] from about 1 million mobile phones in East Massachusetts generated each time a device connects to the cellular network, including:

- when a call is placed or received (both at the beginning and end of a call);
- when a short message is sent or received; and
- when the user connects to the internet (e.g. through web browsers, or through email programs that periodically check the mail server).

In the remainder of the paper we call these events network connections. These events represent a superset of the ones contained in the Call Details Records, previously considered in Gonzalez et al. (2008). The location estimations not only consist of ids of the mobile phone towers that the mobile phones are connected to, but an estimation of their positions generated through triangulation by means of the AirSage's Wireless Signal Extraction technology.

Each location measurement $m_i \in M$ is characterized by a position p_{m_i} expressed in latitude and longitude and a timestamp t_{m_i}. An example is shown in Fig. 1a, where the location measurements are connected into a sequence $\{m_1 \rightarrow m_2 \rightarrow \cdots \rightarrow m_n\}$ according to their time series.

In order to infer trips from these measurements, we first characterize the individual calling activity and verify whether that is frequent enough to monitor the user's movement over time with a fine enough resolution. The available dataset consists of 829 million anonymous location estimations – latitude and longitude. For each user we measure the inter-event time i.e. the time interval between two consecutive network connections. The average inter-event time measured for the whole population is 260 min, much lower than the one found in Gonzalez et al. (2008). Since the distribution of inter-event times for an individual spans over several temporal scales, we further characterize each calling activity distribution by its first and third quartile and the median. Fig. 1b shows the distribution of the first and third quartile and the median for the whole population. The arithmetic average of the medians is 84 min (the geometric average of the medians is 10.3 min), which results in values small enough to detect changes in location where the user stops for as long as 1.5 h.[3]

From a spatial point of view, mobile phone location data has a greater uncertainty range than GPS data, with an average of 320 m and median of 220 m as reported by AirSage based on internal and independent tests. Moreover, some peak errors appear when the user is connected to the network not using the closest mobile phone tower. In these cases it can appear that the user travels several kilometers in just a few seconds. To avoid this problem, we apply a low-pass filter to the data and resample every 10 min, following the approach proposed and evaluated in Rome (Calabrese and Ratti, 2006; Calabrese et al., 2011a).

2.1.1. Determining trips

To explore the statistical properties of individual mobility, we estimate the trips that people make and the length of those trips.

A first approach for the trip's length estimation is to consider a trip to be a path between user's positions at consecutive network connections and then calculate the length as the distance between those points. This approach was used in Gonzalez et al. (2008) but fine-grained spatial resolution could not be reached in their study because only the mobile phone tower location for each network connection was available.

[2] http://www.airsage.com/.

[3] There are cases where trips are underestimated because of the low number of network connection generated, but these cases represent outliers in the dataset. Meanwhile these "outliers" are equally distributed over the population, thus do not influence the empirical study.

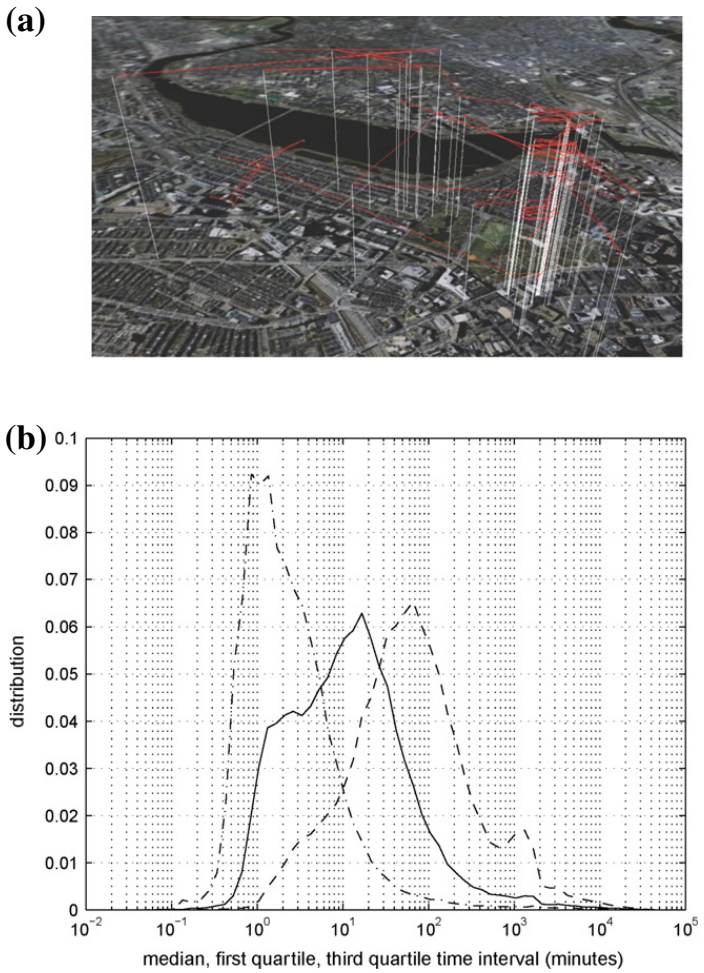

Fig. 1. Mobile phone locations data collected for the Boston Metropolitan Area: (a) Example of location measurements collected for a mobile phone user. *Z*-axis represents the time. (b) Characterization of individual calling activity for the whole population: median (solid line), first quartile (dash-dotted line) and third quartile (dashed line) of individual inter-event time.

The drawback of this approach is that we can detect several very short trips due to localization errors and users making consecutive network connections in the same area. Since these fictitious trips could drastically modify the trip length estimation, we propose a second approach for which we manipulate the data applying a methodology which generalizes what was used for GPS traces, see Ye et al. (2009) and Krumm (2006).

The methodology is composed of the following steps.

- We infer measurement series $M_s = \{m_q, m_{q+1}, \ldots, m_z\} \in M$ where the user makes network connections over a certain time interval $\Delta T = t_{m_z} - t_{m_q} > 0$ into an area within the radius ΔS, i.e.

$$\max \mathit{distance}(p_{m_i}, p_{m_j}) < \Delta S \quad \forall\, q \leqslant i, \quad j \leqslant z$$

A lower bound on the spatial resolution has been defined as 1 km, to take into account the localization errors estimated by AirSage.

- The points $M_s = \{m_q, m_{q+1}, \ldots, m_z\} \in M$ are fused together so that a single geographic region $p_s = (z-q)^{-1} \sum_{i=q}^{i=z} p_{m_i}$ (centroid of the points) can be regarded as a virtual location characterized by a group of consecutive location measurements.
- Once the virtual locations are detected, we evaluate the trips as paths between a user's positions at consecutive virtual locations.

As a first analysis we study the trip length distribution (see Fig. 2a), showing that trips range from 1 to 300 km. The distribution is well approximated by $P(x) = (x + 14.6)^{-0.78} \exp(-x/60)$ with $R^2 = 0.98$, which confirms what was found in Gonzalez et al. (2008). The slightly different coefficients found in this case could be attributed to the different built environment in Europe and US, see Liu et al. (2009).

To check the plausibility of our segmentation of the trajectory in trips, we compute the same statistic computed in Krumm (2006) about the number of trips to see if the results are reasonable. The distribution of trips per day over the whole

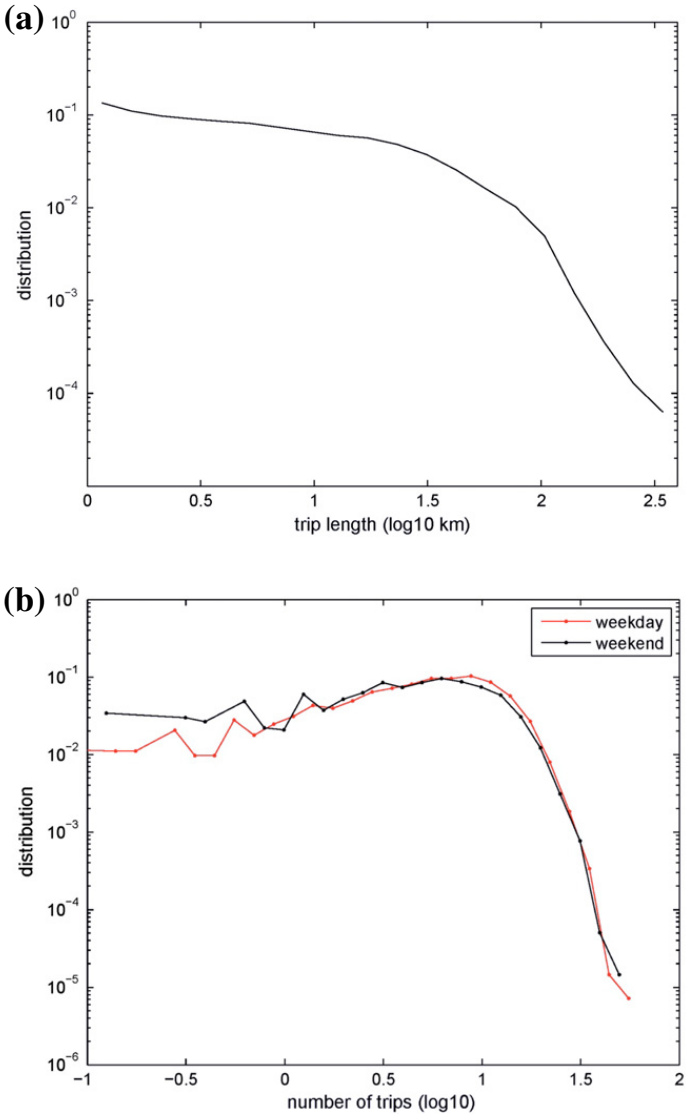

Fig. 2. Statistics on the detected trips: (a) Trip length distribution. (b) Number of trips distributions.

population is shown in Fig. 2b, separating weekday and weekend trips. We obtain an average of 5.0 one-way trips per day during the weekday, and 4.5 during the weekend. This number is reasonable when compared to the US National Household Travel Survey which evaluated this number to be between 4.2 during weekdays and 3.9 during weekends. Finally, aggregating trips at census tract and county levels, it was shown that the mobile-phone-measured Origin–Destination flows correlate well with the US Census estimates, as described in Calabrese et al. (2011b).

2.1.2. Determining the home location

As the next step in our data analysis, we are interested in estimating the home location of each mobile phone user. To detect the home location, first we divide the study area into 500 m by 500 m grid cells. For each cell we evaluate the number of nights the user connects to the network in the time interval 6 pm–8am while in that cell, and select as home location the cell with the greatest value.

The accuracy of the home location estimation is evaluated by a measure of repetitiveness, i.e. dividing the above value by the number of nights when the user connects to the network, and ranges from 0 to 1. The cumulative distribution of the accuracy for the whole population is shown in Fig. 3a, resulting in more than 40% of the people having an estimated home location with accuracy greater than 0.5, which implies that they have been detected at the estimated home location at least half of the monitored days.

Based on the area covered by the mobile phone data, we select eight counties in east Massachusetts (Middlesex, Suffolk, Essex, Worcester, Norfolk, Bristol, Plymouth, and Barnstable) with an approximate population of 5.5 million people.[4] To sim-

[4] To match the spatial coverage of the vehicle safety inspection data, only mobile phone users living in the Boston Metropolitan Area are included in Section 3: individual mobility analysis.

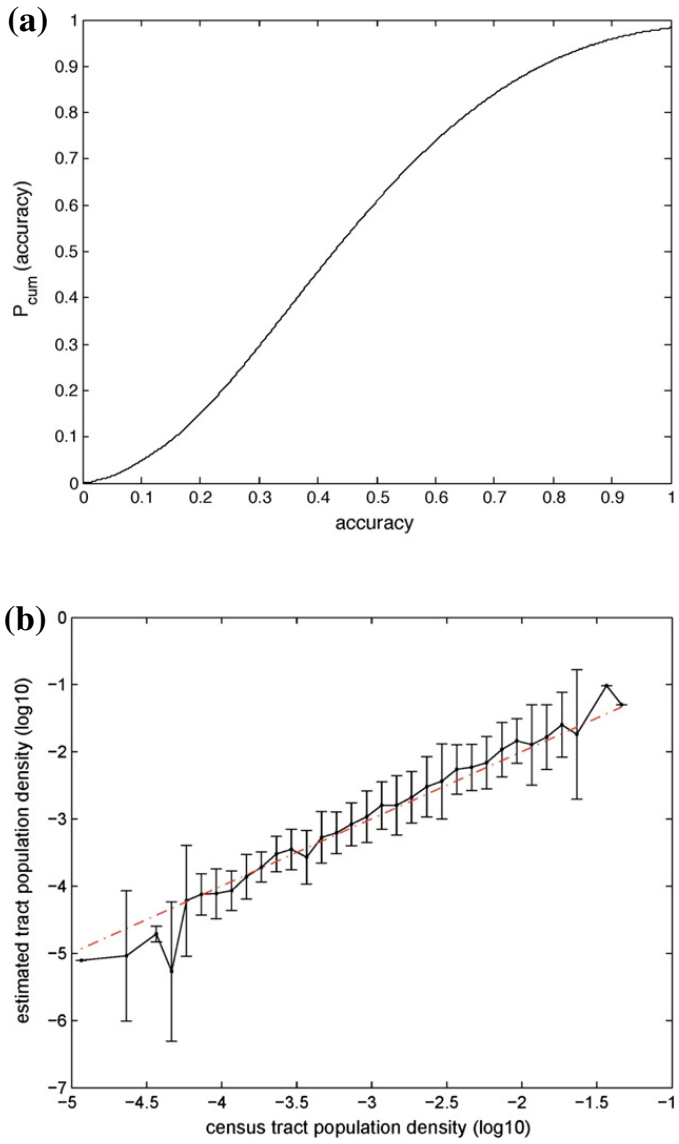

Fig. 3. Repetitiveness of home location estimation and comparison with 2000 Census population estimates: (a) Home location estimation accuracy: cumulative distribution. (b) Population density comparison. Mobile phone population multiplied by 23.2 (i.e. 4.3% sample). Error bars represent the standard error.

plify the analysis, we extract traces for one quarter of the users per county to validate the home locations distribution. We compare it with data from the US 2000 Census at the census tract level. In the selected 8 counties, we have 1171 distinct census tracts, with a population ranging from 70 to 12,000 people (on average 4705), and an area ranging from 0.08 to 203 km^2 (on average 10.8 km^2). The distribution of mobile phone users' estimated home locations matches quite well with the population distribution, as shown in Fig. 3b, corresponding to about 4.3% of the population being monitored.

2.2. Vehicle safety inspection data

This study uses a second unique dataset, the annual vehicle safety inspection records from the Registry of Motor Vehicles (RMV) to estimate annual kilometers traveled for every private passenger vehicle registered in the Boston Metropolitan Area. Safety inspection is mandated annually beginning within 1 week of registering a new or used vehicle. The safety inspection utilizes computing equipment that records a vehicle identification number (VIN) and an odometer reading and transmits this data electronically to the RMV where it can be associated with the residential street address of the vehicle's owner. We get access to this dataset through the research collaboration of the MIT Urban Information Systems Group with MassGIS. Mass-

GIS compared the two recent vehicle safety inspection records for all private passenger vehicles, calculated the odometer reading difference, and pro-rated it based upon the time period between inspection records so as to reflect the estimated annual kilometers traveled. MassGIS then geocoded each vehicle to the owner's address using GIS tools.[5] Overall, 2.47 million private passenger vehicles are included in this dataset. To summarize, for each vehicle the dataset provides the following information: vehicle identifier, annual VKT, home longitude and latitude. For privacy reasons, neither the owner name nor owner address was available for our research. The XY locations are street centerline locations that are estimated by MassGIS to be proximate to the home address using MassGIS address matching tools.

For the initial 2.47 million vehicles, 2.10 million (84.9%) have credible odometer readings. For the remaining 0.37 million vehicles, we know their locations of garaging but do not have reliable odometer readings, either because the reported reading was determined to be in error or because two readings sufficiently far apart were not available, for example, for a brand new vehicle. While this dataset lacks individual trip details, it does provide a very high percentage sample of total VKT in the region. Furthermore, unlike travel surveys, this dataset does not depend on the subjects' willingness or ability to remember and report their driving habits, thus providing a more reliable estimate of VKT. The broad coverage and relatively high reliability of safety inspection records make it a good choice to validate the mobile phone data and explore vehicular mobility patterns.

3. Individual mobility analysis

In this study, we analyze mobility represented by two measures: (1) individual mobility measured by the average daily total trip length that mobile phone users make, and (2) vehicular mobility measured by the average daily VKT per vehicle. To understand intra-urban mobility patterns, we map each mobile phone user and vehicle owner to an estimated home zone, compute the zonal average daily trip length for mobile phone users and vehicles respectively, and relate them to the built-environment and demographic characteristics of the home zone. The census block group is selected as the home zone to take into account the uncertainties in the mobile phone location data, which is in the same order of magnitude as the radii of block groups. A higher aggregation level such as census tracts could reduce the potential misclassification of home zone, but in the meantime mask the variations within the zone.

Vehicles and mobile phone users seem to be distributed over population with similar patterns as shown in Fig. 4a,[6] at least for block groups with more than 500 people. Fig. 4b compares the two mobility measures. There seems to be a linear relationship between the two estimates for areas with an average vehicle total trip length of up to 65 km per day (the red line has the formula $y = 0.328x$, constant term equals 0). This shows that the average daily individual total trip length is approximately 1/3 of the one measured for vehicles. It has to be noted that the individual total trip length measured using mobile phone trace data corresponds to the sum of the Euclidean distance between places, while the VKT per day measured using the odometer readings takes into account the length of the real paths followed from place to place. Using Euclidean distance to measure individual mobility could bring some downwards bias. But, Boarnet and Chalermpong (2001) indicate that urban road distance is linearly related to straight-line distance. Due to the relatively high road density in Metro Boston, measuring the Euclidean distance should not introduce large errors. Moreover, as we are interested in comparing the distances traveled by residents living in different areas, we can assume that the measurement method affects the measures in a similar manner, thus limiting the potential bias. Others have used a similar approach to estimate VKT (Chatman, 2008).

Fig. 5 shows the spatial distribution of daily trip lengths computed using the two datasets. Fig. 5b is plotted using quantile classification method with five categories. To make the two maps comparable, the cutting points of Fig. 5a equal the corresponding points in Fig. 5b multiplied by 0.328. The two maps display similar spatial patterns, with some notable differences. To understand the intra-urban spatial pattern of individual and vehicular mobility we compare the two measures as they change with respect to the built environment.[7] Based on literature, we compute built-environment variables at block group level along four dimensions: density, land use mix, street network layout, and accessibility (see Diao, 2010; Diao and Ferreira, 2010).

- Population density: In this study, we compute the density measure as 2000 census block group population divided by area of residential land.
- Land use mix: We use an entropy-type land use mix indicators in this study. A value of 0 means that the land is exclusively dedicated to a single use, while a value of 1 suggests equal mixing of the 5 land uses, including single family, multi-family, commercial, industrial, and recreation and open space.

[5] People may change locations. MassGIS addresses this problem as follows: (1) for VINs with same license plate number but different owner addresses in consecutive years, the pro-rated annual mileage is associated with the later address; (2) VINs with different plate numbers, suggesting a change of ownership during the observation period, are excluded from the study.

[6] Fig. 4a shows the sample means and standard errors of the number of vehicles and mobile phone users for block groups that have been categorized into bins based on block group population. Since the frequency of large, medium and small block groups varies considerably, the error bars are uneven.

[7] Note that CTPP data could also be used to relate aggregate commuting patterns with built environment factors, as many researchers did before, e.g., Yang (2008) and Wang (2001). However, (1) CTPP data is only available for commuting travels, while non-work trips are a significant part of individual mobility. (2) CTPP provides self-reported commuting time but does not report commuting distances. Researchers have to apply a network analysis to compute the commute distance, which is not necessarily the real distance traveled. (3) CTPP data are updated every 10 years. Therefore, the urban trend within decade cannot be identified.

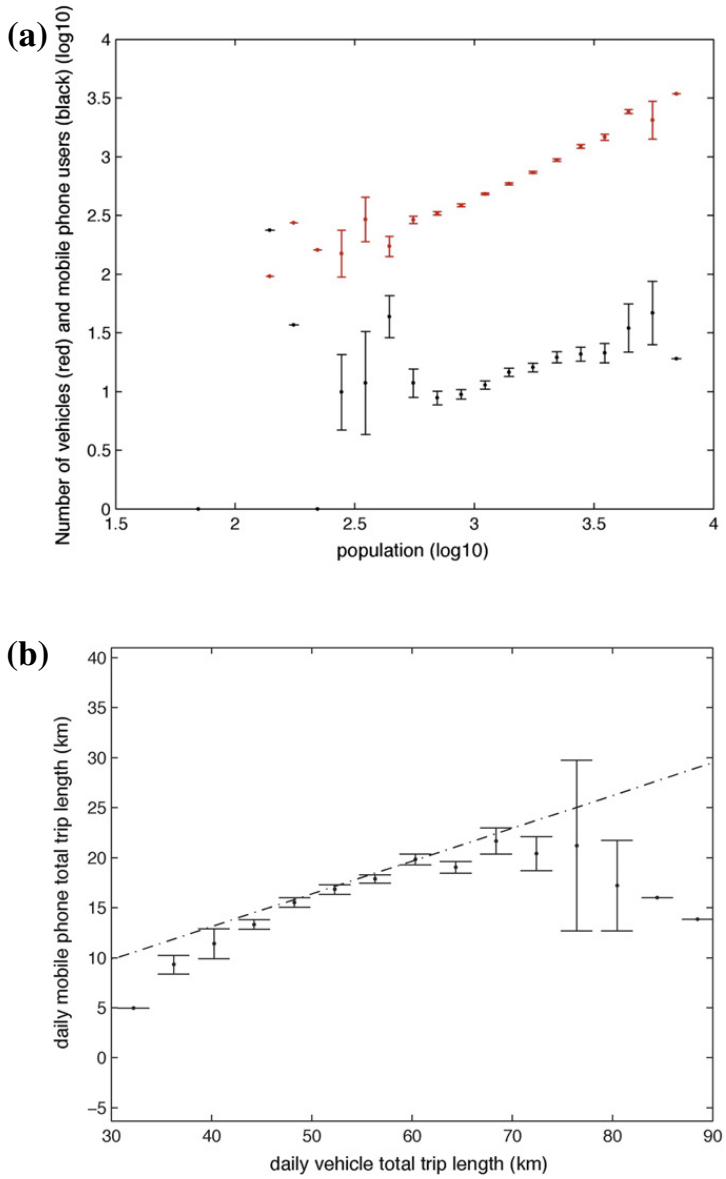

Fig. 4. Comparison between mobile phone and vehicle data. (a) Number of cars (red) and mobile phone users (black) as functions of block group population. Error bars represent the standard error of the means. (b) Comparison between average daily individual and vehicle total trip lengths at the census block group level. Error bars show the standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Intersection density: To show the differences between urban "sprawl" and a traditional block pattern, we compute intersection density within the block group as an indicator of the street network layout.
- Accessibility: Four accessibility measures are computed in this study.
  - *Job accessibility:* Job accessibility is computed at the TAZ level using the following formula:

$$A_i = \sum_j O_j f(C_{ij}), \quad f(C_{ij}) = \exp(-\beta C_{ij})$$

where O_j is the number of jobs in TAZ j; $f(C_{ij})$ is an impedance function; C_{ij} is the network distance between TAZ i and j; β is set to 0.1, based on Zhang's calibration using an Activity-Travel Survey conducted by the Central Transportation Planning Staff for the Boston region (Zhang, 2005). All block groups are assigned the job accessibility of the TAZ that they are associated with.

  - *Distance to non-work destinations:* The distance to non-work destinations indicator is computed by MassGIS. MassGIS utilized Dun & Bradstreet database to locate all non-work destinations in Metro Boston and classified them into 29 categories. The indicator is a weighted average of the minimal distances to these 29 categories of non-work activities, where the weight is the number of trips to a category of destinations divided by the total number of trips based on the 2000 National Household Travel Survey.

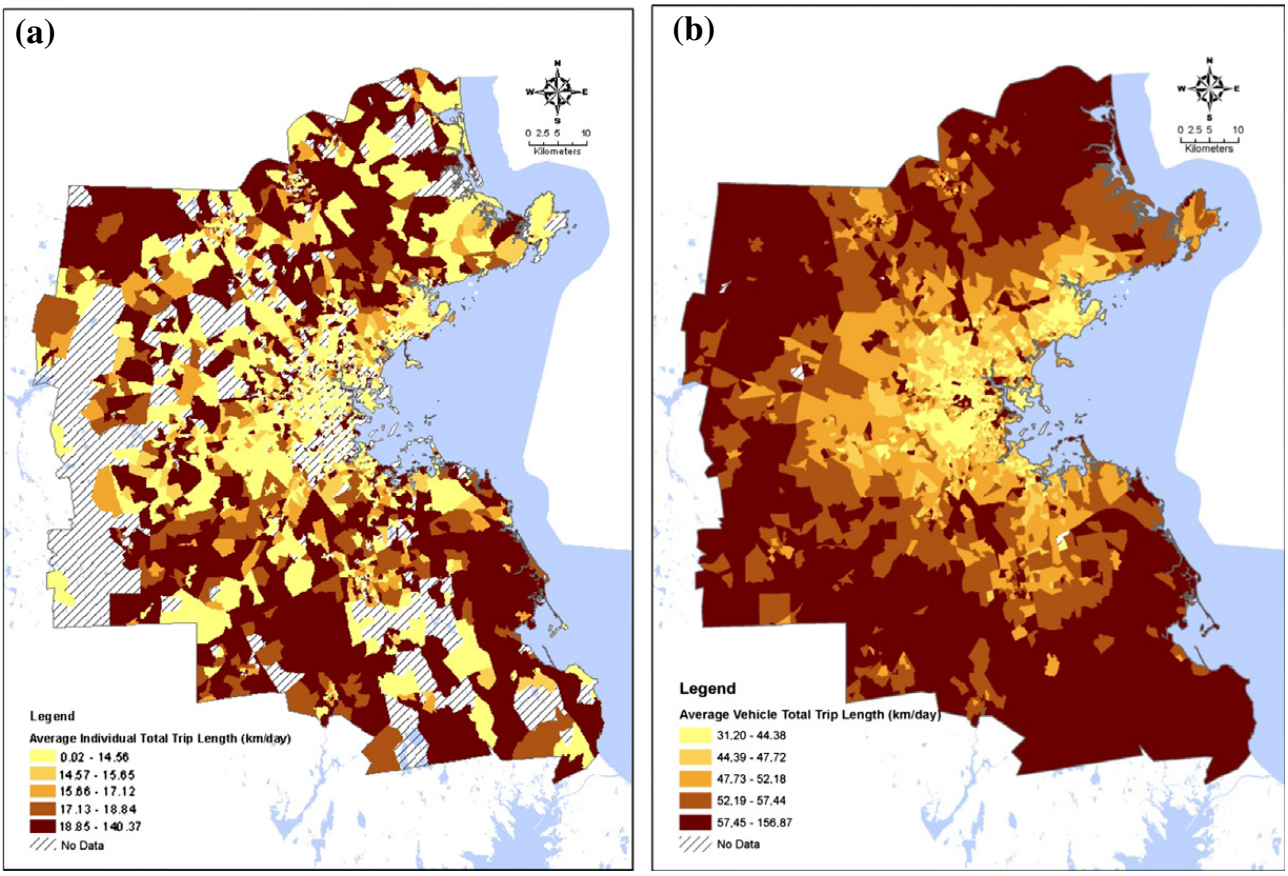

Fig. 5. Spatial distribution of daily mobility at block group level. (a) Individual mobility measured by average individual total trip length. (b) Vehicular mobility measured by average vehicle total trip length.

– *Distance to subway stations and highway exits:* The distance to subway stations and highway exits indicators compute the Euclidean distance to the nearest subway station and highway exit respectively.

Fig. 6 plots the relationships between the built-environment variables and mobility measures. Although very simple bivariate analyses, these graphs reveal that most built-environment indicators are correlated with individual mobility and vehicular mobility in similar ways. Interestingly, distance to non-work destinations exhibits the largest differences in pattern and the vehicular mobility measure is more strongly correlated with built-environment variables than the individual mobility measure.

We then build multi-variate regression models to predict daily mobility with built-environment indicators, controlling for the impact of demographic characteristics from the 2000 US Census. Two models on individual mobility and vehicular mobility are calibrated in our study. Descriptive statistics of variables are shown in Table 1. Estimation results are presented in Table 2. Overall, our models can explain 49.40% of the variation in individual mobility and 56.48% of the variation in vehicular mobility. The mobile-phone-based mobility measure generally correlates with built-environment factors in similar ways as the odometer-reading-based mobility measure, which further confirms the validity of mobile phone trace data in representing mobility.

The estimation results provide some useful insights into the intra-urban spatial patterns of mobility. We find that the spatial distribution of activities has significant impact on total trip length. Both job accessibility and distance to non-work destinations are negatively associated with total trip length, and they are the two most important factors that explain the variations in individual and vehicular total trip lengths, as reflected by their higher standardized coefficients compared to other variables. One standard deviation increase in job accessibility is associated with 0.57 and 0.39 standard deviations decrease in individual and vehicular total trip length respectively, while one standard deviation closer to non-work destinations is associated with 0.17 and 0.48 standard deviations decreases in individual total trip length and vehicular total trip length respectively.

The distance to subway stations variable has a negative and significant coefficient in the individual mobility model, while its coefficient in the vehicular mobility model is insignificant. The results suggest that living close to a subway station can actually increase the total distance traveled, but not necessarily reduce the usage of each car.[8] Proximity to highway exits is

[8] In a separate analysis, we find that distance to subway station is negatively correlated with household car ownership level, which could mean that transit accessibility influences the total VKT mainly through reducing car ownership level.

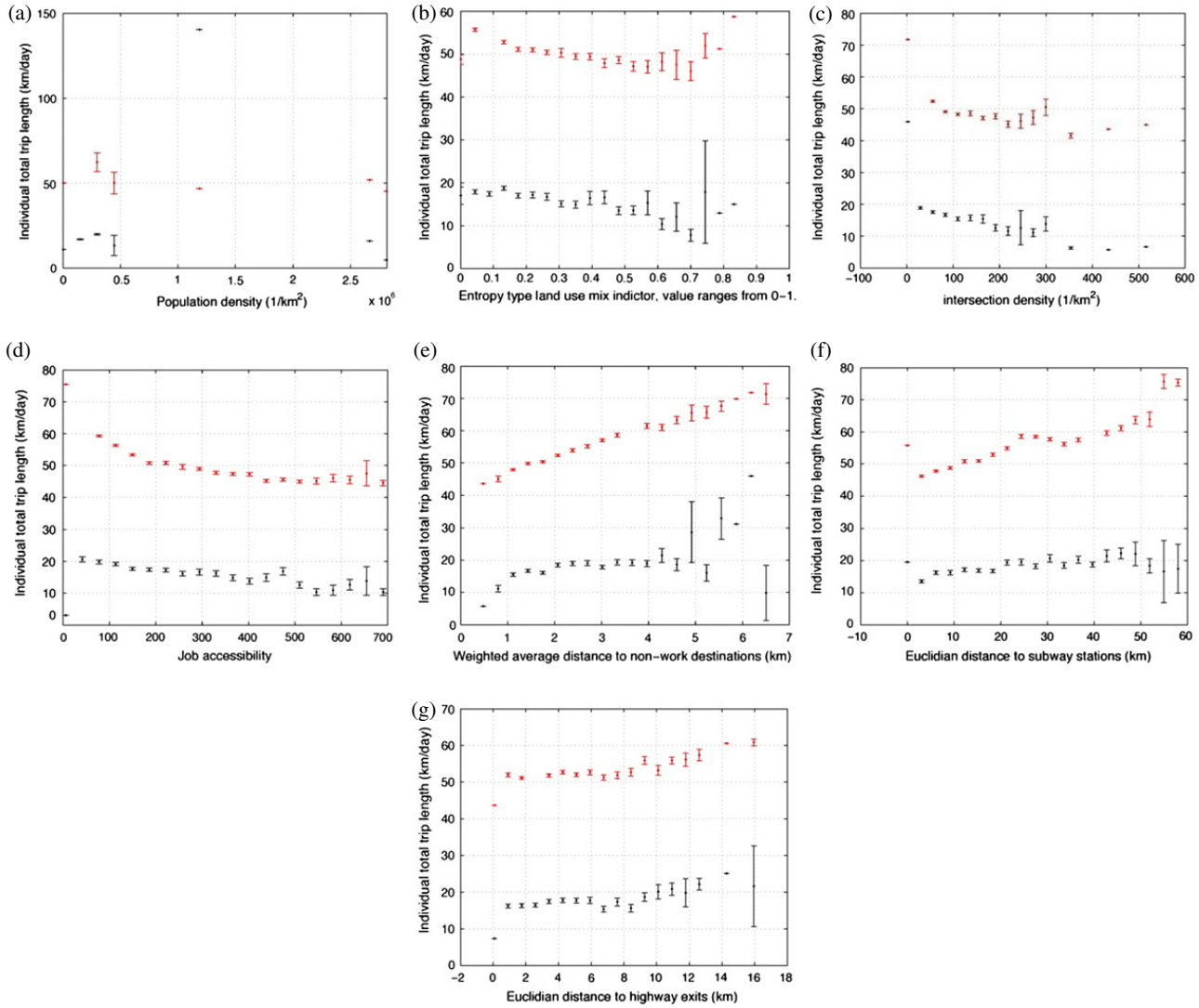

Fig. 6. Daily mobility as a function of different built environment factors. Mobile phone users (black) and vehicles (red). Error bars represent the standard error of the mean. (a) Population density, (b) land use mix, (c) intersection density, (d) job accessibility, (e) distance to non-work destinations, (f) distance to subway stations, and (g) distance to highway exits. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Descriptive statistics.

Variable	Obs.	Mean	Std. dev.	Min	Max
Ln(average individual trip length – individual mobility)	1101	2.771	0.344	1.348	4.219
Ln(average vehicle trip length – vehicular mobility)	1101	3.949	0.150	3.440	4.729
Population density (1 m/km^2)	1101	0.008	0.079	0.000	2.584
Land use entropy	1101	0.171	0.140	0.000	0.750
Intersection density (100/km^2)	1101	0.648	0.553	0.027	5.160
Distance to subway station (km)	1101	16.541	12.952	0.060	57.072
Distance to highway exit (km)	1101	3.419	2.351	0.171	15.502
Distance to non-work destinations (km)	1101	1.860	0.948	0.483	5.520
Job accessibility (10 k)	1101	20.962	16.767	0.589	69.072
Pct. of population below poverty level	1101	0.078	0.098	0.000	0.840
Pct. of owner-occupied units	1101	0.646	0.271	0.000	1.000
Pct. of population with 13+ years of education	1101	0.858	0.120	0.340	1.000
Pct. of population that is white	1101	0.852	0.163	0.030	1.000
Pct. of population under 5	1101	0.072	0.033	0.000	0.300
Pct. of population 16 years and over in labor force	1101	0.671	0.099	0.000	0.910
Median household income (10 k$)	1101	6.040	2.501	0.000	20.000

Table 2
Regression results for individual and vehicular daily total trip length.

Variable	Individual mobility model			Vehicular mobility model		
	Unstd. coeffi.	Std. coeffi.	*t*-Stat.	Unstd. coeffi.	Std. coeffi.	*t*-Stat.
Constant	3.0727		28.930**	3.9384		92.090**
Built-environment characteristics						
Population density (1 m/km^2)	0.0323	0.0075	0.340	−0.0072	−0.0038	−0.190
Land use entropy	−0.0946	−0.0386	−1.240	0.0186	0.0175	0.610
Intersection density (100/km^2)	−0.0120	−0.0193	−0.590	−0.0204	−0.0758	−2.510*
Distance to subway station (km)	−0.0037	−0.1405	−3.230**	0.0004	0.0304	0.750
Distance to highway exit (km)	−0.0071	−0.0484	−1.990*	−0.0137	−0.2164	−9.580**
Distance to non-work destinations (km)	0.0617	0.1704	4.010**	0.0749	0.4763	12.090**
Job accessibility (10 k)	−0.0118	−0.5747	−12.280**	−0.0035	−0.3917	−9.030**
Demographic characteristics						
Median household income (10 k$)	−0.0228	−0.1649	−4.450**	−0.0009	−0.0154	−0.450
Pct. of population below poverty level	−0.3285	−0.0922	−2.550*	0.1187	0.0767	2.290*
Pct. of owner-occupied units	0.1481	0.1170	2.630**	−0.0548	−0.0997	−2.410*
Pct. of population with 13+ yrs of education	−0.2472	−0.0862	−2.150*	−0.1018	−0.0817	−2.200*
Pct. of population that is white	0.1196	0.0567	1.610	−0.0625	−0.0683	−2.100*
Pct. of population under 5	0.3788	0.0367	1.470	0.1523	0.0339	1.460
Pct. of population 16+ yrs old in labor force	0.1362	0.0383	1.450	0.2332	0.1509	6.160**
No. of observations	1101			1101		
Adjusted *R*-squared	0.4940			0.5648		

* Coefficient significant at the 0.05 level.
** Coefficient significant at the 0.01 level.

significantly associated with both total trip length measures. In particular, the distance to highway exits variable is the third most important factor to predict vehicular total trip length, after distance to non-work destinations and job accessibility. Presumably, those living close to highways and subway stations use them to travel further than they might otherwise.

Increasing intersection density is significantly associated with decrease in vehicular total trip length, but its effect on individual total trip length is insignificant, which may be attributed to the fact that our measure of individual total trip length only counts for the Euclidian distance between locations. After controlling for other factors, the impacts of population density on individual and vehicular total trip lengths are both insignificant, which confirms previous findings that density itself has relatively little impact on travel (see Ewing, 1995; Kockelman, 1997), and that other factors associated with density, such as regional accessibility, actually have far greater impacts on travel. Land use mix effects are also insignificant in both models.

It should be noted that the associations found in this study do not necessarily mean causality since residential self-selection may confound the association between the built environment and travel behavior (see Mokhtarian and Cao, 2008).

4. Discussion and limitations

By combining mobile phone traces and odometer readings from annual vehicle safety inspections, our empirical results offer some useful insights into intra-urban mobility patterns. We find that accessibility to work and non-work destinations are the two most important factors in explaining the regional variations in individual and vehicular mobility, while the impacts of population density and land use mix on both mobility measures are insignificant. A well-connected street network is negatively associated with daily vehicular total trip length. These results suggest that regional planning policies dealing with accessibility may significantly reduce the overall travel even more than local policies.

Moreover, comparing the two trip length models we find that (1) closeness to highway exits is related to higher individual and vehicular total trip lengths; and (2) proximity to subway stations correlates with individuals who travel longer distances every day, but the usage of each car is not significantly different from other areas. That suggests that being closer to public transportation and highways is associated with increased propensity for people to move for longer distances. This new finding confirms the importance of transportation infrastructure (both transit and highway networks) for allowing people to access places in the region, but also suggests that the more the system is pervasive (i.e. close to people) the more people will use it to perform more and longer trips.

This study also demonstrates pervasive datasets such as mobile phone traces and vehicle safety inspection records provide rich information to support urban modeling and metropolitan planning, which can serve as a useful alternative data source and a compliment to the traditional travel surveys in mobility research. Meanwhile, some related limitations should also be addressed when applying these datasets in mobility analysis. One common shortcoming for both datasets is the lack of information on mobile phone users or vehicle owners, which limit their usefulness in modeling at an individual or household level to identify the underlying behavior mechanism.

For the mobile phone data, a crucial factor to take into account is the localization error, which could limit the minimum size of the spatial units that can be considered and lead to errors in statistical analysis. In this study, the mobile phone location data from AirSage have a localization error of zero mean and a mean absolute error of 320 m. Given the uncertainties in mobile phone locations, there might be cases in which we assign mobile phone users into wrong home grid cells and/or into wrong home block groups (especially in dense urban areas where block groups are small in size), which could further lead to inaccurate zonal mobility measures. Similarly, vehicles are associated with block groups by geocoding their address to an approximate street location. Since the street centerlines define the boundaries of block group, the address matching can also assign a vehicle to the wrong (neighboring) block group. However, given the fact that localization errors can be assumed independent and identically distributed with zero mean, and the large number of mobile phone users tracked, the misclassification of home zones for some individuals should not introduce large biases in the zonal mobility measures and regression results produced in Section 3. Indeed, Fig. 3 shows that the allocation of mobile phone users into home grid cells lead to a density pattern similar to census data at the census tract level. Figs. 4 and 5 show that the vehicle and mobile phone based mobility measures match each other pretty well at the block group level, which implies that no significant bias are introduced by the proposed allocation approach. Meanwhile, the accessibility measures, which are the most important factors in explaining mobility variations as identified by this study, are not strongly affected by the misclassification of home zones, because the distance between the centroids of two adjacent block groups in dense urban areas is very small.

Other elements that can affect the statistical results include: (1) the market share of the mobile phone operator from which the dataset is obtained; (2) the potential non-randomness of the mobile phone users; (3) calling plans which can limit the number of samples acquired at each hour or day; and (4) number of devices that each person carries. Moreover, due to the fact that the considered dataset is event-driven (location measurements available only when the device makes network connections) the connection patterns of users are affecting the possibility to capture more or less trips.

Nonetheless, the analyses performed on the inter-event time, the spatial distribution of mobile phone users, and the comparison with vehicle odometer reading data confirm that: mobile phone users have fairly frequent calling activities so that their location changes can be tracked through mobile phone traces; mobile phone data have reasonable reliability in determining trips and home locations; and mobile-phone-based mobility measures have similar spatial distribution patterns as the odometer-reading-based mobility measures. Therefore, mobile phone traces represent a reasonable proxy for individual mobility, whose quality and representativeness could be further improved in the future as the penetration rate of smart phones keeps increasing.

We recognize two limitations that arise while comparing the two datasets:

- The individual total trip length computed using mobile phone data measures individual spatial movement, regardless of the mode, while the vehicular total trip length reflects the usage of each vehicle, regardless of how many people drive it;
- Individual total trip length is measured as sum of the Euclidean distances among stop points, so it does not take into account the real (most probably longer) path taken by a person. The vehicular total trip length instead measures this since it is based on odometer readings.

Addressing these limitations will be part of our future work.

5. Conclusions

This study focuses on exploring the potential values and limitations of applying large-scale urban sensing data to transportation research, using the mobile phone trace data as an example. In this study, we provide algorithms that can extract mobility information from mobile phone trace data, such as trips and home locations at 500 m by 500 m grid cell level, which could lead to useful mobility statistics for transportation research. By integrating these statistics into statistical analyses, our study shows that mobile phone traces represent a reasonable proxy for individual mobility and can provide useful insights into intra-urban mobility patterns and the non-vehicular part of the overall mobility when combined with the vehicle safety inspection data.

While further studies still need to be performed, especially to take into account the user's calling patterns, this study seems to show enormous potential in applying mobile phone trace data to mobility study. In the future, we plan to extend the current study along multiple directions:

(1) Performing longitudinal studies to monitor the temporal change in individual mobility as well as the interconnection between individual mobility and the built environment;
(2) Using Geographically Weighted Regression to examine spatial differences in the relationship between the two mobility measures and the built-environment factors.
(3) Utilizing mobile phone trace data in other aspects of mobility research, such as identifying individual activity pattern for activity-based modeling, and generating real-time Origin–Destination matrix for travel demand modeling.

Acknowledgements

Acknowledgments go to AirSage and MassGIS for providing us with the data used in the study. The authors thank the MIT CCES-KACST program, Audi Volkswagen, BBVA, Ericsson, GE, Ferrovial and all the companies that generously support the SENSEable City Lab Consortium. We also acknowledge the helpful comments of the anonymous referees and the partial support of University Transportation Center (Region One) Grant MITR21-4, and the Singapore National Research Foundation (NRF) through the Future Urban Mobility program at the Singapore-MIT Alliance for Research and Technology (SMART).

References

Boarnet, M., Chalermpong, S., 2001. New highways, house prices, and urban development: a case study of toll roads in Orange County, California. Housing Policy Debate 12, 575–605.

Calabrese, F., Ratti, C., 2006. Real time Rome. Networks and Communications Studies 20 (3–4), 247–258.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011a. Real-time urban monitoring using cell phones: a case study in Rome. IEEE Transactions on Intelligent Transportation Systems 12 (1), 141–151.

Calabrese, F., Di Lorenzo, G, Liu, L., Ratti, C., 2011b. Estimating Origin–Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. IEEE Pervasive Computing, 99.

Chatman, D.G., 2008. Deconstructing development density: quality, quantity and price effects on household non-work travel. Transportation Research Part A 42, 1008–1030.

de Ia Rue du Can, S., Price, L., 2008. Sectoral trends in global energy use and greenhouse gas emissions. Energy Policy 36 (4), 1386–1403.

Diao, M., 2010. Sustainable Metropolitan Growth Strategies: Exploring the Role of the Built Environment. Ph.D. Thesis, Massachusetts Institute of Technology.

Diao, M., Ferreira, J., 2010. Vehicle miles traveled and the built environment: evidence from vehicle safety inspection data in the Boston metropolitan area. In: Paper presented at the 12th World Conference on Transportation Research, Lisbon.

Ewing, R., 1995. Beyond density, mode choice, and single-purpose trips. Transportation Quarterly 49 (4), 15–24.

Gonzalez, M., Hidalgo, C., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.

Handy, S., 1996. Methodologies for exploring the link between urban form and travel behavior. Transportation Research Part D 1 (2), 151–165.

Holtzclaw, J., Clear, R., Dittmar, H., Goldstein, D., Hass, P., 2002. Location efficiency: neighborhood and socioeconomic characteristics determine auto ownership and use – studies in Chicago, Los Angeles and San Francisco. Transportation Planning and Technology 25, 1–27.

Kockelman, K.M., 1997. Travel behavior as a function of accessibility, land use mixing, and land use balance: evidence from the San Francisco Bay Area. Transportation Research Record: Journal of the Transportation Research Board 1607, 117–125.

Krumm, J., 2006. Real time destination prediction based on efficient routes. In: Society of Automotive Engineers (SAE) 2006 World Congress. pp. 2006–01–0811.

Lindsey, M., Schofer, J., Durango-Cohen, P., Gray, K., 2011. The effect of residential location on vehicle miles of travel, energy consumption and greenhouse gas emissions: Chicago case study. Transportation Research Part D 16, 1–9.

Liu, L., Calabrese, F., Biderman, A., Ratti, C., 2009. The law of inhabitant travel distance distribution. In: European Conference on Complex Systems. Warwick, UK.

Miller, E., Ibrahim, A., 1998. Urban form and vehicular travel: some empirical findings. Transportation Research Record: Journal of the Transportation Research Board 1617, 18–27.

Mokhtarian, P.L., Cao, X., 2008. Examining the impacts of residential self-selection on travel behavior: a focus on methodologies. Transportation Research Part B 42 (3), 204–228 (a Tribute to the Career of Frank Koppelman).

Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C., 2007. Cellular census: explorations in urban data collection. IEEE Pervasive Computing 6 (3), 30–38.

Wang, P., Gonzalez, M., Hidalgo, C., Barabasi, A.-L., 2009. Understanding the spreading patterns of mobile phone viruses. Science 324 (5930), 1071–1076.

Wang, F., 2001. Explaining intra-urban variations of commuting by job accessibility and worker characteristics. Environment and Planning B 28, 169–182.

Yang, J., 2008. Policy implications of excess commuting: examining the impacts of changes in US metropolitan spatial structure. Urban Studies 45 (2), 391–405.

Yang, J., Ferreira, J., 2008. Choices vs. choice sets: a commuting spectrum method for representing job-housing possibilities. Environment and Planning B 35 (2), 364–378.

Ye, Y., Zheng, Y., Chen, Y., Feng, J., Xie, X., 2009. Mining individual life pattern based on location history. In: IEEE International Conference on Mobile Data Management, pp. 1–10.

Zhang, M., 2005. Exploring the relationship between urban form and nonwork travel through time use analysis. Landscape and Urban Planning 73 (2–3), 244–261.