



Origin–destination trips by purpose and time of day inferred from mobile phone data



Lauren Alexander ^{a,*}, Shan Jiang ^b, Mikel Murga ^a, Marta C. González ^a

^a Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

^b Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States

ARTICLE INFO

Article history:

Received 1 June 2014

Received in revised form 18 January 2015

Accepted 17 February 2015

Available online 9 March 2015

Keywords:

Mobile phone data

Data mining

Human mobility

Trip production and attraction

Trip distribution

Travel surveys

ABSTRACT

In this work, we present methods to estimate average daily origin–destination trips from triangulated mobile phone records of millions of anonymized users. These records are first converted into clustered locations at which users engage in activities for an observed duration. These locations are inferred to be *home*, *work*, or *other* depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we probabilistically infer departure time using survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day. These trips are multiplied by expansion factors based on the population of a user's *home* Census Tract and divided by the number of days on which we observed the user, distilling average daily trips. Aggregating individuals' daily trips by Census Tract pair, hour of the day, and trip purpose results in trip matrices that form the basis for much of the analysis and modeling that inform transportation planning and investments. The applicability of the proposed methodology is supported by validation against the temporal and spatial distributions of trips reported in local and national surveys.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The ubiquity of cell phones, along with rapid advancement in mobile technology, has made them increasingly effective sensors of our daily whereabouts (Lane et al., 2010). Call detail records (CDRs) from mobile phones contain time-stamped coordinates of anonymized customers, thereby providing rich spatiotemporal information about human mobility patterns. Since CDRs are automatically collected by cell phone carriers for billing purposes, this data can be gathered more frequently and economically than travel survey data collected once (or twice) a decade for transportation planning purposes. Additionally, mobile phone data offers digital footprints at a scale and resolution that may not be captured by surveys that typically record one day of travel diaries per household.

Despite these advantages, mobile phone data lacks information typically available from travel surveys about a respondent (e.g. age or income) or his/her trip (e.g. purpose or mode) (Richardson et al., 1995; Stopher and Greaves, 2007; Hu and Reuscher, 2004). Furthermore, CDRs contain traces of a user at approximated locations when his/her phone communicates

* Corresponding author.

with a cell phone tower, providing an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces as well as understanding its limitations.

It has been demonstrated that CDR data can be used to infer origin–destination (OD) trips using microsimulation and limited traffic count data (Iqbal et al., 2014). At the level of the individual, daily trip chains/trajectories constructed from mobile phone data are consistent with household surveys (Jiang et al., 2013; Schneider et al., 2013). Further, road usage inferred from the CDR data has been validated against GPS speed data (Wang et al., 2012) and highway assignment results from a travel demand model (Huntsinger and Donnelly, 2014).

There is still work to be done to explore the usage of phone data to generate trip distributions of different modes, purposes, and times of day. As a step in that direction, this research proposes a methodology to extract OD trips by purpose and time of day from CDR data. This segmentation captures distinct trip-making patterns pertinent for transportation planning applications. Moreover, other than CDR data, the techniques presented in this paper rely only upon nationally-available survey data to allow transferability of the methodology to other study areas in the US.

Extensive research has been conducted into OD estimation, as these trips provide the basis for transportation feasibility and impact studies. Conventional OD estimation approaches rely on surveys and/or travel demand models to provide trip matrices. Often, such trip matrices are calibrated or updated using traffic counts and estimation techniques such as maximum likelihood, generalized least squares, and optimization (Spiess, 1987; Cascetta, 1984; Bell, 1991; Yang et al., 1992). This research provides a realistic, cost-effective alternative to these traditional OD data sources and estimation approaches. By presenting a systematic and replicable procedure to extract data relevant to the transportation community, we hope this work will help to facilitate the use of mobile phone data in practice.

In this paper, we demonstrate methods to analyze mobile phone records for the Boston metropolitan area. In Section 2, we present an overview of the data and the methods developed to produce OD trips by purpose and time of day. In Section 3, we summarize and validate our results against independent data sources for the study area, including the US Census and household travel surveys. Based on these findings, we conclude with a discussion of the limitations and applications of CDR data in the context of transportation planning and modeling.

2. Data and methods

2.1. CDR data

The studied dataset contains more than 8 billion anonymized mobile phone records (from several carriers) from roughly 2 million users in the Boston metropolitan area over a period of two months in the Spring of 2010. Although the CDR data spans 60 days, the data provider reindexed the anonymous user IDs for most of the users after the 17th day of the dataset. Effectively, we observe some users for at most 17 days, some users for at most 43 days, and still others for up to 60 days.

Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200–300 m. In typical mobile phone data sets, locations are represented by cell towers rather than triangulated coordinates and therefore have a lower spatial resolution; however, the method proposed here is expected to hold for such cases (Song et al., 2010a; Wang et al., 2012).

2.2. Stay extraction

The first step to reliably infer activities and trips from CDR data is to filter out noise resulting from (1) tower-to-tower call balancing performed by the mobile service provider, creating the appearance of false movements, and (2) inexact signal triangulation. Furthermore, we wish to distinguish users' stationary stay locations (when/where users engage in an activity) from their moving pass-by locations (when/where users are en-route to activities). To do so, we develop a method based in the work of Hariharan and Toyama (2004) for processing GPS traces. The spatial and temporal filtering methods are discussed below and illustrated in Fig. 1.

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), \dots, d_i(n_i))$ be the observed data for a given anonymous user i , where $d_i(k) = (t(k), x(k), y(k))'$ for $k = 1, \dots, n_i$, and $t(k)$, $x(k)$, and $y(k)$ are the time, longitude, and latitude of the k -th observation of user i . First, we extract points $d_i(k)$ that are spatially close (i.e. within roaming distance of 300 m) to their subsequent observations, say, $d_i(k+1), d_i(k+2), \dots, d_i(k+m)$. To reduce the jumps in the location sequence of the mobile phone data, we assume that $d_i(k), \dots, d_i(k+m)$ are observed when user i is at a specific location, i.e., the medoid of the set of locations $(x_i(k), y_i(k))', \dots, (x_i(k+m), y_i(k+m))'$, which is denoted by

$$\text{Med}((x_i(k), y_i(k))', \dots, (x_i(k+m), y_i(k+m))').$$

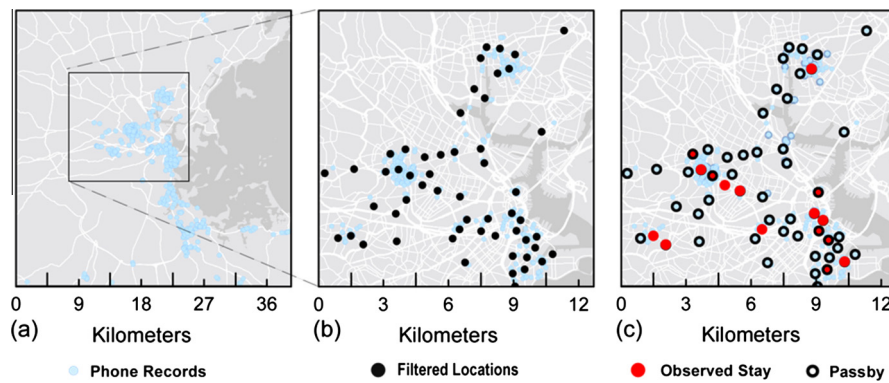


Fig. 1. Extracting stay and pass-by areas from the phone data for an anonymous user in the 2-month period.

This treatment respects the time order at first, to ignore noisy jumps in estimated location, but then disregards time ordering to apply the *agglomerative clustering algorithm* (Hariharan and Toyama, 2004) to consolidate points that are close in space but may be far apart in time. The points to be consolidated together form a cluster whose diameter is required to be no more than a certain threshold (set as 500 m). Again we modify the observation locations to the corresponding medoids of the clusters (see Fig. 1(a) and (b)).

Next, we impose the time duration criterion on the clean data, and extract the stay locations whose durations exceed a certain threshold (set as 10 min). In the example presented in the figure we extract 31 distinct stay locations from the 1776 phone records in the two-month period of the exhibited anonymous user (see Fig. 1(c)). The rest of the points are called pass-by points, at which we do not observe any lengthy stays. Note that it is possible that the user stays in some of these pass-by locations as well as locations that we do not observe. In these cases, information about time and location is totally or partially latent to us as we do not observe it from the phone records. However, all the stay locations frequently visited by the user ought to be extracted from the mobile phone data, if the observation period is long enough. As such, the pass-bys are filtered out and the stays are assumed to be trip origins or destinations, between which trips are made. Analysis of the pass-by points is out of the scope of the present work, in which we focus on simple trip chains with origins and destinations labeled as: home, work, or other.

2.3. Activity inference

Trips are induced by the need or desire to engage in activities (Pinjari and Bhat, 2011) and therefore understanding patterns and types of activities is crucial in estimating travel demand. It has been demonstrated that human mobility patterns are characterized by regularity with frequent returns to previously visited locations (Song et al., 2010b; Song et al., 2010a; Hasan et al., 2013). Due to this predictability, we are able to reasonably infer stay activities for users' most visited locations (i.e. home and work).

Accordingly, our first task is to label the stay regions in order to assign trip purpose. For each user, the stay extraction process detailed above results in a timestamp and duration for each observed visit to a stay location. For this study, we assign an activity type of either *home*, *work*, or *other* to each users' stay locations. Future research can expand the *other* designation to activity types such as school, shopping, recreation and social, using land use information.

Each user's *home* location is identified as the stay with the most visits on (i) weekends and (ii) weekdays between 7 pm and 8 am, representing the time windows in which we expect users to spend substantial amounts of their time at home. In addition to inferring trip purpose, the *home* stay location of each user is used to filter out users with too few data points and expand the data from phone users to study area population, as summarized in Section 2.4.

A *work* location is identified as the stay (not previously labeled as *home*) to which the user travels the maximum total distance from *home*, $\max(d \cdot n)$, where n is the total number of visits to a given stay on weekdays between 8 am and 7 pm and d is the distance between the latitude-longitude coordinates of the *home* stay and the given stay using plane approximation. This assumption is based on the rationale and historical evidence (Levinson and Kumar, 1994; Schafer, 2000) that for a given frequency of visits, longer distance trips are more likely to be work trips than shorter distance trips, which are more likely to be for non-work purposes (i.e. to the nearby grocery store).

If the user visits the identified *work* stay less than 8 times ($n < 8$; once a week, on average) or the distance is less than 0.5 km ($d < 0.5$), then the activity of the stay region is identified as *other* rather than *work*. In effect, not all users are assigned a *work* stay, accounting for the fact that not all users commute to a job. Subsequently, all the remaining stay locations not identified as *home* or *work* are designated as *other*. These classification assumptions serve to avoid falsely identifying a location as work that is either not visited frequently enough or close enough to a user's home that it could reflect signal noise rather than a distinct location.

We acknowledge that under these simple assumptions we may misidentify users' *true* home and work locations and, by extension, their trip purposes. However, based on comparisons with census data (presented below) this procedure give us very good estimates of the distribution of home and work locations and home-work flows in our study region. Note that these assumptions are related to the duration and spatial resolution of this dataset, and it may be necessary to adjust them for applications of other datasets.

2.4. Data filtering and expansion

For users with too few stay locations, the CDR data may not fully represent their travel patterns. Accordingly, users with fewer than 8 (one per week, on average) visits to designated *home* stays are filtered out. This filter serves the additional purpose of ensuring with a reasonable degree of certainty that the designated stay is the user's home, a key assumption in our method of upscaling users to population. Note that this filtering process necessarily excludes visitors, for whom a home location is not observed in the studied dataset. Future research could look at extracting visitor trips from CDR data using an assumption other than home location to upscale these trips.

After this filtering, 335,795 users remain in the Boston CDR dataset. This sample size is an order of magnitude larger than in most household travel surveys, and should increase given longer periods of observation. To upscale these users to total population of the study region, the number of *home* stays were aggregated to the 974 Census Tracts in the study area. An expansion factor was then calculated for each Tract as the ratio of the 2010 Census population and the number of residents identified in the CDR data. For the 10 Census Tracts with fewer than 10 CDR residents, the scaling factor is set to 0 to ensure that we do not overweight users that are not representative of a given Census Tract. The 1st, 2nd, and 3rd quartiles of the expansion factors are 9.4, 14.2, and 25.1, respectively, as illustrated by the tight probability distribution of expansion factors in Fig. 2a. The spatial distribution illustrated in Fig. 2b suggests that the Tracts in the western portion of the study area tend to be more heavily weighted. CDR data for a period greater than 60 days would likely have lower expansion factors and an improved spatial distribution of users, however, we show that already this limited data set gives reasonable results.

2.5. Trip estimation

With stays for each user designated by activity type and expansion factors to upscale users to population, average daily origin–destination trips can be constructed by time of day and purpose—home-based work (HBW), home-based other (HBO), and non-home based (NHB). This segmentation allows us to capture distinct trip-making patterns and is consistent with segmentation in the trip distribution stage of trip-based travel demand models.

Since the timestamp and duration associated with each stay reflect the *observed* (based on phone usage) rather than *true* arrival time and duration of a user, we infer trip departure time using probability density functions to account for this uncertainty. The publicly-available 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011), filtered for respondents residing in a consolidated metropolitan statistical area (CMSA) or MSA with populations greater than or equal to 3 million, is a reasonable source as it approximates temporal travel patterns of major US cities comparable to Boston, while allowing for transferability of this methodology to other US cities. Using this departure time data, we generate six hourly distributions for weekdays and weekends and the following trip purposes: HBW, HBO, and NHB.

For each user, it is assumed that a trip is made between two consecutive stays ($i, i + 1$) occurring within a 24 h period beginning and ending at 3 am. The trip occurs at a point in time spanned by the range $[s_i + \delta_i, s_{i+1}]$, where s is the observed arrival time and δ is the observed duration of a stay. The departure hour is randomly generated in this time window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose identified from the origin and destination stay activities (HBW, HBO, NHB).

Furthermore, it is presumed that a user starts and ends each 24 h period at home such that if a user is not recorded at his/her *home* stay for the first (last) record of the 24 h period, his/her first (last) trip begins (ends) at his/her *home* stay. The first (last) trips are assumed to occur at point in time spanned by the range $[3AM, s_{i+1}]$ ($[s_i + \delta_i, 3AM]$), where s is the observed arrival time and δ is the observed duration of a stay. As before, the departure hour is randomly generated in this window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose based on the destination (origin) stay activity (HBW, HBO).

Through this process, we construct trips on all days we observe each user. The frequency of weekday observations per user is illustrated in Fig. 3. The distribution of total weekday trips per user is shown in Fig. 3a, with first, second, and third quartiles of 33, 58, and 96 trips, respectively. The reindexing of anonymous user IDs mentioned previously in Section 2.1 is evident in the two peaks of the distribution of the number of weekday days we observe each user, as seen in Fig. 3b. Despite this reindexing, we achieve a sufficiently large number of observation days per person, with first, second, and third quartiles of 11, 17, and 21 days, respectively. Dividing each user's total weekday trips by his/her total weekday days, we get the distribution of average weekday trips shown in Fig. 3c. The distribution has a long tail, however, the first, second, and third quartiles are 2.6, 3.2, and 4.3 average trips per weekday, respectively, demonstrating that the vast majority of users have a reasonably small number of daily trips.

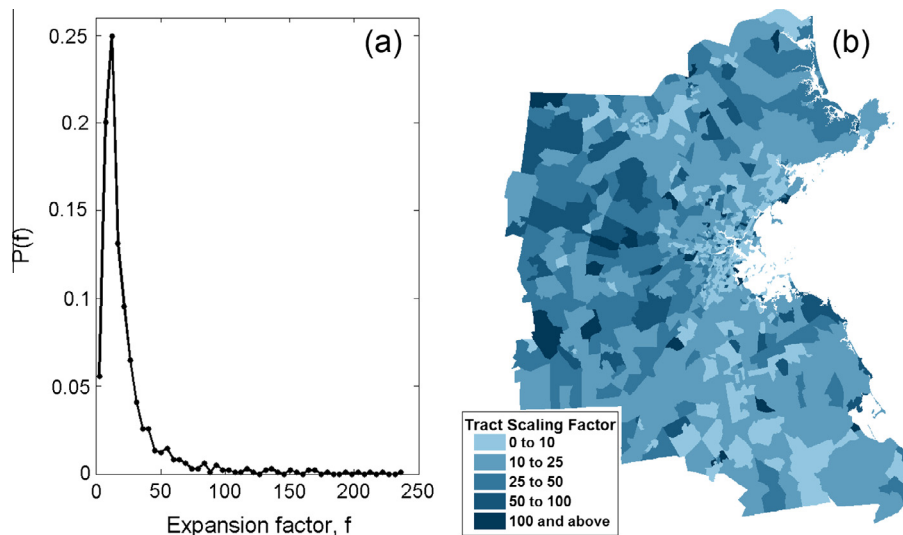


Fig. 2. (a) Probability distribution of Census Tract expansion factors. (b) Thematic map showing the spatial distribution of Census Tract expansion factors.

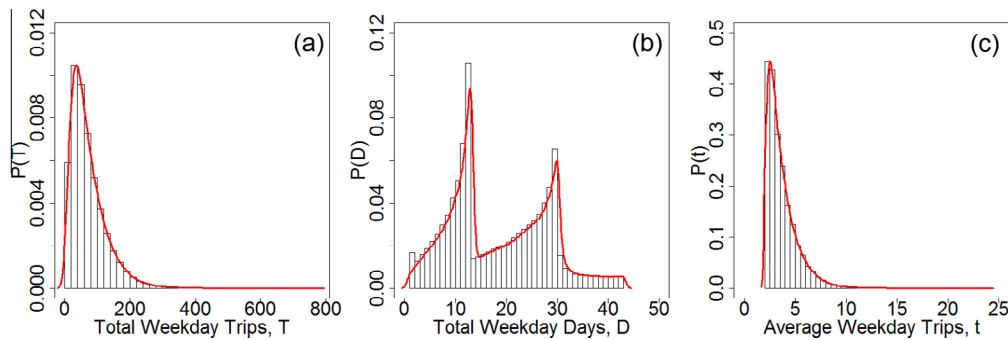


Fig. 3. Frequency of weekday observations per user. (a) Probability distribution of total weekday trips per user. (b) Probability distribution of total weekday days per user. (c) Probability distribution of average weekday trips per user.

In order to obtain average daily OD trips, each users trips are multiplied by the expansion factors described in Section 2.4 for the user's *home* Census Tract and divided by the number of days from which we constructed the user's trips. For users assigned a *work* stay, weekday trips are only constructed on days in which the user is observed at his/her *work* stay to ensure we capture representative weekdays of commuters. Unlike traditional travel surveys which ask a respondent details about one or a few recent days, this method has the advantage of capturing many days per user and thus variations in his/her daily travel behavior. Lastly, each user's average daily trips are aggregated into Census Tract pair trip matrices by day type (weekday, weekend), purpose (HBW, HBO, NHB), and hour of departure.

3. Results and validation

3.1. Productions and attractions

Accurately extracting and upscaling users' stays is crucial to trip generation. Due to the regularity of human behavior (Song et al., 2010a; Song et al., 2010b; Hasan et al., 2013), we are able to infer users' *home* and (if applicable) *work* stay locations from CDR data. For this dataset, we find that we can reasonably represent the spatial distribution of home and work locations when aggregated to the 164 study area cities and towns (MassGIS, 2014). Refer to Section 3.2 below for more information on the impact of aggregation level on accuracy. Fig. 4a shows a comparison of home locations by town from 2010 Census data and the raw and upscaled CDR data.

As we would expect since Tract population was used to upscale the data, the number of residents in each town is almost identical to that of the upscaled CDR data. However, the slope of a best-fit line through the raw CDR data is close to 1, which speaks to the fact that the overall distribution of raw CDR users is fairly representative and a simple factoring

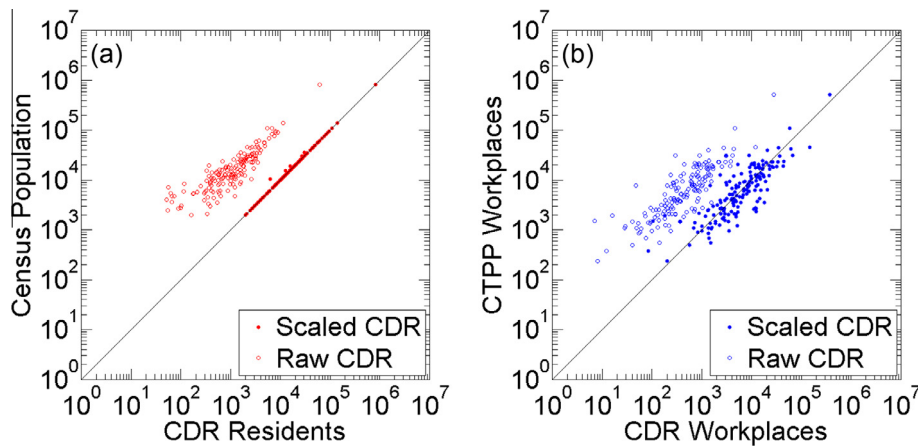


Fig. 4. (a) CDR residents vs. 2010 Census population by town before and after population expansion. (b) CDR vs. Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013) workers by town before and after population expansion.

method is in fact appropriate to expand the phone users to population. Similarly, Fig. 4b shows a comparison of work locations aggregated by town. As with the raw CDR data on the home-end, the distribution of raw workplaces is fairly consistent with the 2006–2010 Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013) data (slope approximately 1), and the upscaling method adjusts well for the difference in magnitude. This strong correlation is noteworthy considering that each users' *home* and *work* locations were scaled based on their *home* location only.

3.2. Trip distribution

With the establishment of reasonable distributions of trip productions and attractions, we next validate the distribution of trips using two local surveys. The 1991 Boston Household Travel Survey (BHTS) contains information on 39,300 trips made by 3737 households (Boston Metropolitan Planning Organization, 1991), while the 2010/2011 Massachusetts Travel Survey (MHTS) contains data on 153,099 trips made by 32,739 people (NUSTATS, 2012). We find that the CDR trips compare well with trips from these data sources by time of day and purpose. Fig. 5 illustrates the distributions of hourly departure times for (a) HBW, (b) HBO, (c) NHB, and (d) total average weekday trips. Note that we also benchmark against the NHTS departure time distributions, which were used to infer departure time for the CDR trips. Accordingly, differences between each of the hourly NHTS and CDR distributions reflect the observed arrival and duration times of CDR stays.

Most notably, there are consistently more CDR trips in the late night hours than that of the surveys. While this may be due to a slight mismatch between the frequency of calling and trip-making throughout the day, it may also highlight an advantage of CDR data to capture late night trips not typically reported in survey responses of an average day. Regardless, most transportation planning applications focus on trips in the morning and evening peak periods, when congestion is most prevalent, and for which we compare well. Similar trends are evident for average weekday trip shares segmented by key time periods, as presented in Table 1.

Furthermore, the relative share of average weekday trips for each trip purpose is comparable for the CDR and survey data. Table 1 shows that the shares of HBW, HBO, and NHB CDR trips are within the ranges of trip purpose shares across all three surveys. This again suggests that our inferences of *home*, *work*, and *other* activities, as well as their relative prevalence in the data set, seem reasonable.

To draw comparisons on the magnitude of daily CDR trips, we MHTS data, which includes weights to expand respondents to population estimated from the 2006–2010 American Community Survey (NUSTATS, 2012). Table 2 shows a comparison of average weekday trips by purpose and period of the day for the CDR trips and weighted MHTS trips. The survey reports more daily trips than we observe in the CDR data, with most of the difference coming from the NHB trip segment. Still, the total CDR and MHTS trips imply reasonable numbers of average weekday trips per person – 3.50 and 4.24, respectively.

Lastly, Table 2 presents a comparison of the spatial distribution of daily CDR and MHTS trips at the Tract-pair and town-pair level. The correlation coefficients of the trip matrices improve significantly with aggregation to the 164 study area cities and towns. In particular, the HBW and AM correlations at the Tract-pair level see the largest improvement. This may be indicative of the role of the size of Tracts, which are considerably smaller in downtown Boston where many of the morning commute trips end. We discuss the relationship between aggregation level and correlation in more detail in Section 3.3 below.

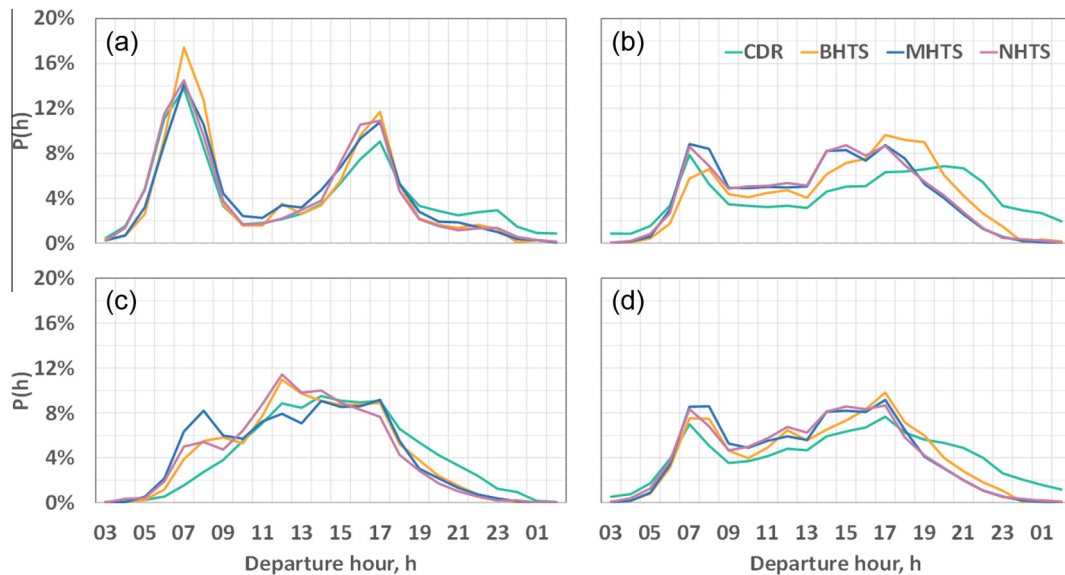


Fig. 5. Distribution of average weekday hourly departure time from CDR data, 1991 Boston Household Travel Survey (BHTS) (Boston Metropolitan Planning Organization, 1991), the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012), and 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011) for (a) home-based work trips, (b) home-based other trips, (c) non-home based trips, and (d) all trips.

Table 1

Average weekday trip shares by purpose and period from CDR data, 1991 Boston Household Travel Survey (BHTS) (Boston Metropolitan Planning Organization, 1991), the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012) and the 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation Federal Highway Administration, 2011).

Source (%)	HBW (%)	HBO (%)	NHB (%)	Morning 6a–9a (%)	Mid-day 9a–3p (%)	Evening 3p–7p (%)	Rest-of-day 7p–6a (%)
CDR	18	51	31	16	27	27	30
BHTS	20	48	32	18	32	33	17
MHTS	12	49	39	21	34	33	12
NHTS	14	55	30	19	37	31	13

Table 2

Average daily trips by purpose and period from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) (NUSTATS, 2012), as well as the correlation coefficients of CDR and MHTS Tract-pair and Town-pair trips.

	HBW	HBO	NHB	AM 6a–9a	MD 9a–3p	PM 3p–7p	RD 7p–6a	Total
CDR Trips (in millions)	2.81	7.84	4.73	2.46	4.12	4.15	4.65	15.37
MHTS Trips (in millions)	2.14	8.99	7.18	3.99	6.24	6.06	2.31	18.61
Tract-pair correlation	0.30	0.64	0.58	0.42	0.65	0.54	0.40	0.58
Town-pair correlation	0.96	0.97	0.98	0.97	0.98	0.97	0.96	0.98

3.3. Home-work flows

Commuting trips represent a key travel market and source of daily roadway congestion, and accurately representing these trips is an important step in validating trips estimated from CDR data. Accordingly, we next compare with flows between people's home and work locations, as reported by the 2006–2010 Census Transportation Planning Products (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013). Distinct from the average daily HBW trips compared in Section 3.2, these flows simply link home and work, ignoring that people's daily trip chains may in fact include work trips to/from locations other than home.

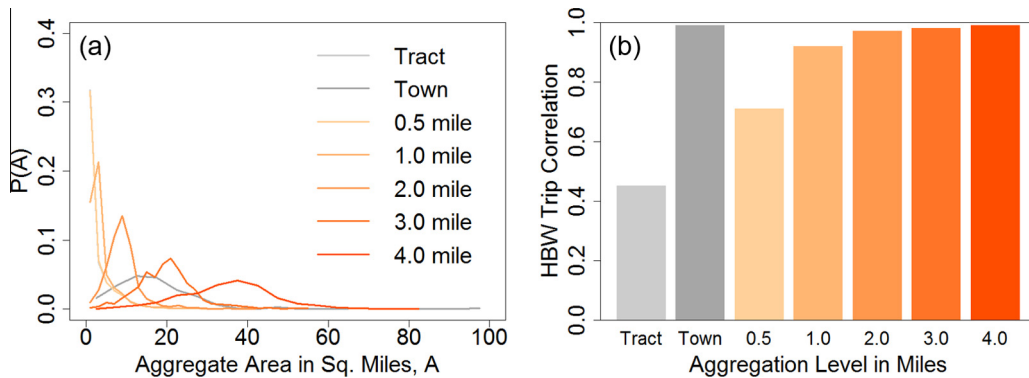
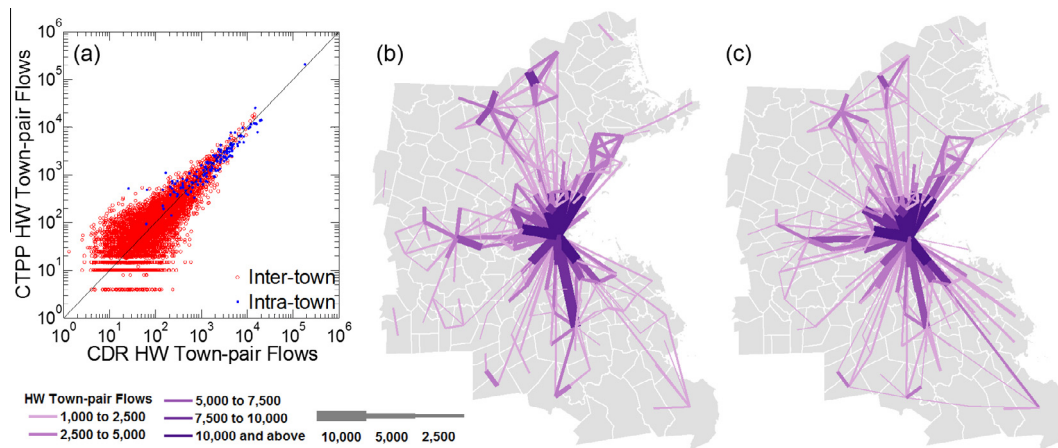
Table 3 summarizes statistics that support the comparison of CDR and CTPP home-work (HW) flows. In addition to the total magnitude of trips, the similarities between the percentages of inter-tract and inter-town flows and average trip length give a high-level indication that the distributions of HW flows are similar.

At the flow level, we find that the correlation between CDR and CTPP HW Tract-to-Tract and town-to-town flows is 0.45 and 0.99, respectively, indicating that the level of aggregation of trips has a significant impact on accuracy. We demonstrate that as we gradually increase average aggregation size using variably-sized buffers around each origin and destination Tract

Table 3

Comparison of average weekday HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows.

Source	Daily HBW trips (millions)	Inter-Tract share (%)	Inter-town share (%)	Average trip length (miles)
CDR	2.11	94	68	9.67
Census	2.10	90	68	10.72

**Fig. 6.** (a) Probability density distributions of aggregation area size by designated areas (Tract or towns) and variable buffers. (b) Correlation between HBW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows corresponding to the aggregation levels in (a).**Fig. 7.** (a) Intra-town and inter-town pair daily HW CDR flows and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows. (b) Spatial distribution of daily inter-town HW CDR flows (>1000). (c) Spatial distribution of daily inter-town HW 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows (>1000).

(Fig. 6a), the correlation between CDR and CTPP HW trips increases as well (Fig. 6b). We find that using small aggregation buffers has the most significant impacts on correlation, while having minimal influence on average aggregation size (as illustrated by the fact that the distribution for the 0.5 mile buffer obscures that of the Tract-level aggregation in Fig. 6b). In effect, using a 0.5 mile buffer aggregates the small, dense Tracts (i.e. in the city center) and results in a notable improvement in accuracy. In the absence of meaningful districts or communities to which to aggregate, this can inform suitable distance thresholds for trip clustering to overcome limitations of sparse data and/or spatial inaccuracy.

We further investigate comparisons of the data sets using town-pairs flows. Fig. 7a shows the CDR and CTPP HW flows for all of the intra-town and inter-town pairs, which have correlations of 0.99 and 0.95, respectively. It is evident from Fig. 7a that town pairs with many trips validate better than those pairs with few trips, especially those with fewer than about 500 daily trips. This trend is likely due to sparsity in data for these smaller markets. Fig. 7b and c illustrate spatially the HW flow distribution for key markets (inter-tract pairs with greater than 1000 daily trips) for the CDR and Census data, respectively. Inspecting the figure, it is evident that the CDR data captures very similar patterns to that of the CTPP commuting data, with the majority of flows directed in and out of Boston as well as a few shorter distance markets in the suburban towns.

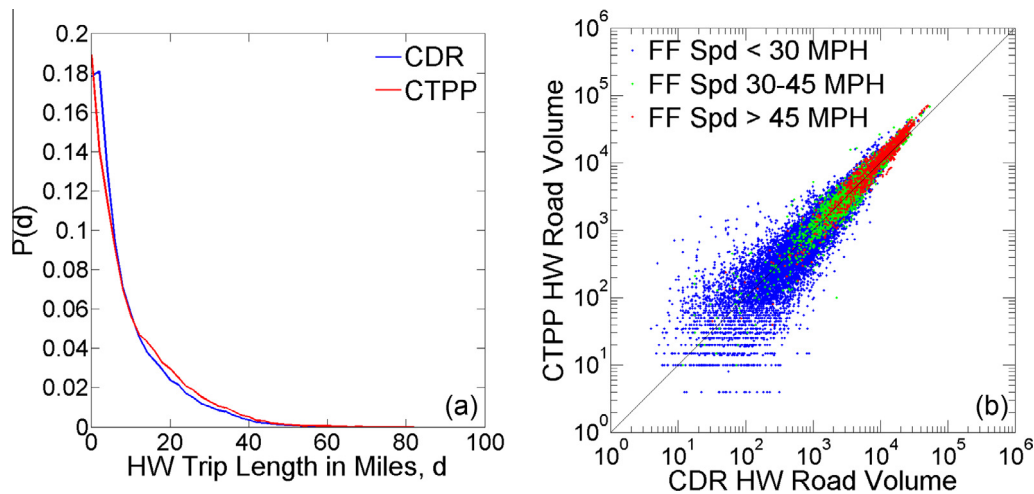


Fig. 8. (a) Trip length distribution of daily HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows. (b) Road segment volumes for daily HW CDR and 2006–2010 CTPP (U.S. Department of Transportation Federal Highway Administration, 2013) flows by free flow speed.

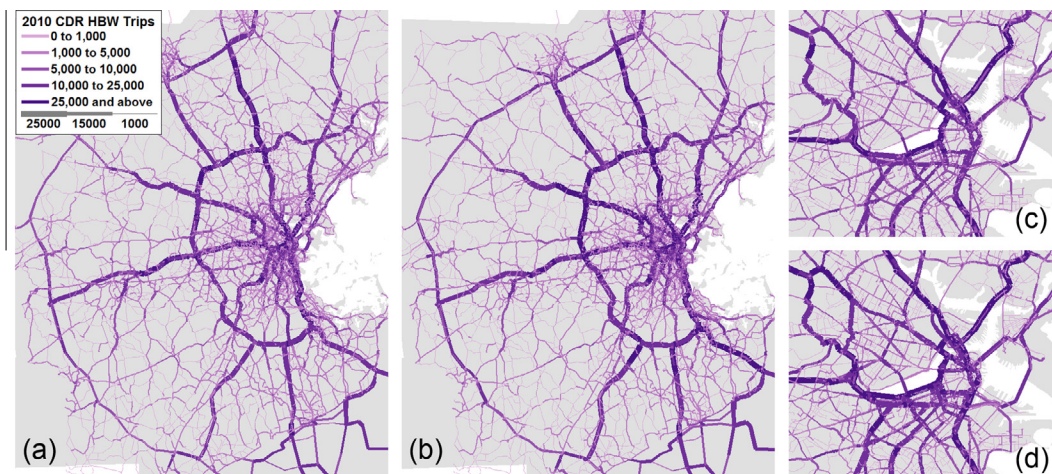


Fig. 9. Road segment volumes of HBW trips (a) from CDR data in the Boston metro area and (c) downtown Boston, and (b) from 2006–2010 CTPP data (U.S. Department of Transportation Federal Highway Administration, 2013) in the Boston metro area, and (d) downtown Boston. Flows are generated from Tract-to-Tract ODs in TransCAD (Caliper, 2009) using all-or-nothing highway assignment minimizing travel time.

Assigning the Tract-to-Tract trips to roads offers another valuable spatial comparison as it considers potential paths of OD trips and has important implications for planning applications. Although it is not representative of a meaningful traffic scenario, we assign all daily HW flows (irrespective of time of day or mode) to a road network for this comparison. Traffic assignment also allows us to estimate and compare trip length distributions across the two datasets. Fig. 8a illustrates that the trip length distributions are indeed very similar, consistent with our findings of comparable trip distributions.

Fig. 8b illustrates strong correlation between CDR and CTPP road segment volumes by free-flow speed, which serves as a proxy for major and minor arterials. With correlations of 0.97, 0.98, and 0.95 for all segments with free flow speeds greater than 45 MPH, between 30 and 45 MPH, and less than 30 MPH, respectively, it is evident that the roadway volumes estimated from CDR and CTPP data are very similar, especially for major roads. The lower correlation on more minor road segments follows the finding observed in Fig. 7a, in which Tract-pairs having few daily trips have the lowest correlation, since minor roads typically serve these smaller markets. Spatially, Fig. 9 illustrates these correlations for the greater metropolitan area and downtown. Although differences in road segment volumes are virtually indistinguishable visually (in Fig. 9), the CTPP

commuting trips result in slightly higher road segment volumes on major roads experiencing the highest volumes (in Fig. 8b).

Despite the lower correlations observed for Tract-pairs and road segments with lower flows and volumes, respectively, these markets have minimal impact on the network as a whole. Further, surveys are susceptible to inaccurate sampling and/or upscaling due to infrequency or scarcity of trips in these minor markets. Accordingly, the *ground truth* number of trips for Tract-pairs with few trips is unknown and comparisons between survey and CDR trips reflect this uncertainty and noise.

4. Conclusions

In this paper, we detailed steps necessary to extract average daily origin–destination trips by purpose and time of day from mobile phone call detail records (CDRs). The proposed techniques were applied to CDRs in the Boston metropolitan area and validated against local and national surveys. The methods are transferable to other study areas and could be reproducible by researchers and practitioners using mobile phone and census data.

Emphasizing the importance of data preprocessing, much of the methods serve to filter out noise and extract accurate travel patterns representative of the study area. While this processing reduces the immensity of the CDR data, we are left with a sample size that is an order of magnitude larger than most household travel surveys. Further, we observe many days per user, allowing us to capture variation in daily behavior, including weekends, not typically reported household travel surveys.

We find that the size of the areas used to aggregate trips is a very important factor in how well the CDR and survey data compare. We observe significantly higher trip correlation when aggregating origins and destinations to 164 cities and towns rather than the 974 Census Tracts in the study area. This improvement in accuracy is seemingly an effect of aggregating small Census Tracts (i.e. in the city center), for which CDR data may not have a sufficiently-large sample size or the necessary spatial accuracy. In general, aggregating trip origins and destinations to areas greater than 1 square mile produces agreement with survey data. As mobile phone providers collect more dense data such as GPS traces or wifi access points, spatial and temporal data sparsity will decrease, and accordingly, aggregation size can decrease relative to a given level of precision. Although we can reasonably represent average daily activity and trip patterns with CDRs, data limitations preclude its use in applications requiring richer data such as real-time, dynamic OD estimation.

Aggregating to towns results in similar distributions of upscaled home and work locations inferred from the CDR data and the home- and workplace-based tabulations from the 2006–2010 US Census Transportation Planning Package (CTPP) (U.S. Department of Transportation Federal Highway Administration, 2013). Additionally, our inferred distributions of trips by hour of the day and purpose are comparable with the 1991 Boston Household Travel Survey (Boston Metropolitan Planning Organization, 1991), 2010/2011 Massachusetts Travel Survey (NUSTATS, 2012), and the 2009 National Household Travel Survey (U.S. Department of Transportation Federal Highway Administration, 2011) (filtered for trips in MSAs and CSAs with populations greater than 3 million). Finally, the spatial distribution of home-work flows is highly correlated with that of the CTPP, a well-established nation-wide source for Tract-to-Tract commuting data.

In validating OD trips by purpose and time of day, we demonstrate that CDR data can be effectively used to represent distinct mobility patterns across market segments typically relevant to transportation planning applications. In particular, CDR data can be used to augment or complement traditional survey data, which provides detailed information about a respondent and his/her trip but is more costly and onerous to collect. Transportation models rely heavily on survey data for inputs, calibration, and validation, and CDR data can be a valuable new resource. Furthermore, the outputs of our proposed methodology are analogous to the outputs of the trip generation and distribution steps of traditional four-step travel demand models. In areas where public transportation is significant, OD matrices developed from CDRs can be post-processed to obtain mode-specific trip tables, equivalent to the mode split step. As such, CDR data can be very useful for planning applications and/or study areas where running such a model is either not feasible or not necessary.

In addition to average daily origin–destination trips, mobile phone data captures individuals' daily trip chains and is therefore well-suited for activity-based models, especially if land use information can be used to infer activity types beyond home, work, and other. Future steps for analyzing this data include traffic assignment of vehicle trips inferred from CDRs by time of day, allowing us to explore how these data sets help to improve existing urban trip models and applications related to mitigating congestion.

Acknowledgements

This work was partially funded by the MIT-Accenture alliance, the BMW-MIT collaboration under the supervision of PI Mark Leach,¹ the Austrian Institute-HuMNet collaboration agreement under the supervision of PI Dietmar Bauer² and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris.³ We thank Yingxiang Yang and Peter Widhalm for technical support.

¹ mark.leach@bmw.de.

² Dietmar.Bauer@ait.ac.at.

³ anas@mit.edu.

References

- Bell, M.G.H., 1991. The estimation of origin-destination matrices by constrained generalised least square. *Transport. Res. Part B: Methodol.* 25, 13–22.
- Boston Metropolitan Planning Organization, 1991. 1991 Boston Household Travel Survey. <http://www.surveyarchive.org/Boston/Boston_91.zip>.
- Caliper, 2009. TransCAD Transportation Planning Software. <<http://www.caliper.com/tcovu.htm>>.
- Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. *Transport. Res. Part B: Methodol.* 18, 289–299.
- Hariharan, R., Toyama, K., 2004. Project lachesis: parsing and modeling location histories. *Geogr. Inform. Sci.*, 106–124.
- Hasan, S., Schneider, C., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. *J. Stat. Phys.* 151.
- Hu, P.S., Reuscher, T.R., 2004. Summary of Travel Trends: 2001 National Household Travel Survey. Technical Report. U.S. Department of Transportation Federal Highway Administration. <<http://nhts.ornl.gov/2001/pub/stt.pdf>>.
- Huntsinger, L.F., Donnelly, R., 2014. Reconciliation of regional travel model and passive device tracking data. In: *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*.
- Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin-destination matrices using mobile phone call data. *Transport. Res. C* 40, 63–74.
- Jiang, S., Yang, Y., Fiore, G., Jr., J.F., Frazzoli, E., González, M., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*.
- Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. *IEEE Commun. Mag.* 48, 140–150.
- Levinson, D.M., Kumar, A., 1994. The rational locator: why travel times have remained stable. *J. Am. Plan. Assoc.* 60, 319–332.
- MassGIS, 2014. Community Boundaries. <<http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/towns.html>>.
- NUSTATS, 2012. Massachusetts Department of Transportation: 2010/2011 Massachusetts Travel Survey.
- Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. *Handbook Transp. Econ.* 1, 1–36.
- Richardson, A., Ampt, E.S., Meyburg, A.H., 1995. *Survey Methods for Transport Planning*. Eucalyptus Press, Melbourne.
- Schafer, A., 2000. Regularities in travel demand: an international perspective. *J. Transport. Stat.* 3, 1–31.
- Schneider, C., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unraveling daily human mobility motifs. *J. Roy. Soc. Interface* 10.
- Song, C., Koren, T., Wang, P., Barabási, A.L., 2010a. Modelling the scaling properties of human mobility. *Nature Phys.* 6, 818–823.
- Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010b. Limits of predictability in human mobility. *Science* 327, 1018–1021.
- Spiess, H., 1987. A maximum likelihood model for estimating origin-destination matrices. *Transport. Res. Part B: Methodol.* 21, 395–412.
- Stopher, P., Greaves, S., 2007. Household travel surveys: where are we going? *Transport. Res. Part A: Policy Practice* 41, 367–381.
- U.S. Department of Transportation Federal Highway Administration, 2011. 2009 National Household Travel Survey. <<http://nhts.ornl.gov/download.shtml>>.
- U.S. Department of Transportation Federal Highway Administration, 2013. CTPP 2006–2010 Census Tract Flows. <http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm>.
- Wang, P., Hunter, T., Bayan, A.M., Schectner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. *Sci. Rep.* 2.
- Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transport. Res. Part B: Methodol.* 26, 417–434.