# ANALYZING CELL PHONE LOCATION DATA FOR URBAN TRAVEL: CURRENT METHODS, LIMITATIONS AND OPPORTUNITIES

**Serdar Çolak**
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
77 Mass. Ave., Cambridge, MA 02139
Tel: 857-891-3655; Email: serdarc@mit.edu

**Lauren P. Alexander**
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
77 Mass. Ave., Cambridge, MA 02139
Tel: 512-695-8513; Email: lpalex@mit.edu

**Bernardo Guatimosim Alvim**
SCN, Quadra2 Lote A, Ed. Corporate Center, 7o. Andar,
70.712-900, Brasília - DF, Brazil
Tel: +55 61 3329-8646; Email: balvim@worldbank.org

**Shomik R. Mehndiretta**
The World Bank
1818 H Street NW, Washington, DC 20433
Tel: 202-473-9980; Email: smehndiratta@worldbank.org

**Marta C. Gonzalez, Corresponding Author**
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
Engineering Systems Division, Massachusetts Institute of Technology
77 Mass. Ave., Cambridge, MA 02139
Tel: 857-928-4546; Email: martag@mit.edu

Word count: 5228 words text + 8 tables/figures x 250 words = 7,228 words

Submission Date: 08/01/2014

## ABSTRACT

Travelers today utilize technology, which generates vast amounts of data at low costs. In essence, these data have the potential to supplement most of the outputs from regional travel demand models. Creating new analysis tools may shift the paradigm of how we approach data and modeling when assessing travel demand. Recent work has shown how processed origin-destination trips, as developed by trip data providers, supports travel analysis. Much less is reported on how raw data from telecommunication providers can be processed to support such analysis or to what extent the raw data can be treated to extract travel behavior. This work discusses how cell phone data can be processed to inform a four-step transportation model, focusing specifically on the limitations and opportunities of such data. We show a data treatment pipeline that uses only phone data and population density to generate trip matrices in two Metropolitan areas: Boston and Rio de Janeiro. We detail how to label zones as home and work based on frequency and the time of day. Based on the labels (home, work, or other) of consecutive stays we can assign purposes to trips such as home-based-work. The resulting number of trips pairs are expanded using the total population from census data. We show comparable results with existing information reported in local surveys in Boston and existing origin-destination matrices in Rio de Janeiro. Our results detail a method to use passively generated cellular data as a low cost option for transportation planning.


*Keywords*: Mobile phone data, data mining, human mobility, trip production and attraction, trip distribution, travel surveys.

# INTRODUCTION

Every time we use our phones, passive mobile monitoring generates a record with the time and approximate location of the event. With monitoring software installed, it is possible to track Internet usage, GPS coordinates, and much more. In the near future as this information becomes more accurate, the question of whether the combination of GPS and phone data entirely replace travel diaries will arise. We are in the early days of figuring out how to make this possible, and the main challenge today for transportation modelers is to find methods that will extract meaningful information while adapting to the current opportunities and limitations of the data. Passive data does not offer the same detailed information as surveys do. We cannot learn about every individual's travel motifs in depth or directly ask the passive data the same questions we can answer with travel diaries. However, it contains valuable information regarding continuously recorded trip choices. This information can be adopted in some phases of transportation applications at low costs in any city worldwide where phone usage is ubiquitous. The challenge of harvesting, interpreting and applying these data requires innovative thinking on both the data acquisition and data analysis aspects of the discussion.

In order to leverage this passive data one has to apply innovative techniques for big data statistics and analysis. While these data mining techniques mature, another useful question to consider is: how do we incorporate the processed data to generate travel models? This paper focuses on the latter task. We explore the usage of call detail records (CDR) to generate origin-destination (OD) trips of different purposes (home, work, other) and times of day. We present a replicable procedure to process CDRs in order to extract information relevant to trip generation. The results are compared with travel surveys in Boston and with independently generated ODs in Rio de Janeiro. The analyzed mobile phone data sets have different accuracy in each city; in Rio we have coordinates at tower resolution while in Boston coordinates come from a triangulation algorithm applied by the data provider. Furthermore, the validation sources differ; in Boston we use census and travel diary survey commuting data while in Rio we have OD estimates by purpose and time of the day. We present both analyses here to show that the method proposed in this paper is robust to different conditions of accuracy. We discuss the current limitations of the data to inform all the steps of a traditional transportation model and the possible avenues for future work.

Passive data alone may not be the ultimate solution to gather detailed information needed for a complete transportation model such as mode, detailed activity types and route assignment. Nevertheless, we show here, its substantial efficacy in OD generation. The richness of these data comes from its ability to provide trip choices of millions of individual users every minute and everywhere. The main findings of this work are twofold: firstly, evaluating the extent to which cell phone data accurately reflects daily travel and secondly, developing guidelines on how to best utilize these data to generate ODs by purpose and time of the day.

## Literature Review

In the US the first works using mobile phone data for transportation applications refer to traffic monitoring. Departments of Transportation in different states carry out these studies in collaboration with private data providers. For example in Virginia, VDOT in collaboration with AirSage [1] created automatic signaling from the phones and showed fluctuations of traffic speed on a map. Similarly, a project in Maryland in collaboration with Delcan Corp. [2] infer traffic

along main highways. In a validation exercise, researchers in North Carolina used one month of data to calculate travel times of monitored devises. The extracted travel times and volume delay function over 800 centerline miles of roadway compared very well with the results of a regional travel model [3] yielding significant cost savings over traditional methods. Other set of works, like the ones lead by Sohn [4] and Akin and Sisiopiku [5] perform OD matrix calculations using simulations of mobile phone data. The main focus of these works was the evaluations of the efficacy of the techniques related to metering frequency and numbers of localizations necessary to achieve accurate traffic estimates. These approaches to collect and analyze the data rely on continuous or close monitoring to achieve this goal. Accuracy is gained at cost of less individuals being tracked in more detail.

Less is known about how the massive amount of information hidden in several months' of anonymized mobile phone bills; also known as call detailed records (CDRs), can support the models of travel behaviors. Most of the literature working with CDRs deals with data mining techniques to discover attributes of the data. For example in [6,7], the authors created 'Mobile Landscapes' of Graz (Austria), visualizing the whole dynamic of a town in real time from hundreds of thousands of mobile phone users. The researchers measured call density (measured in Erlang) and ODs (by way of 'handovers'). More elaborated attempts have mapped millions of consecutive calls from mobile phone users to the roads and compared these results against average car volumes in a given time period measured via cameras [8] or against the travel times of cars, via a BPR function [9]. These works do not compare with travel diaries of trip purpose. Recent advances in the area, compared estimated vehicle traffic in roadways using mobile phone data against the values of a travel demand model calibrated via surveys [10]. The authors obtained trips from 600,000 individual users. AirSage provided the track-to-track trip information. The researchers disaggregated to traffic analysis zones (TAZs) to create trip tables and assigned them to roads. They found good agreement between these processed AirSage's inter-track ODs and the results of the regional model. The results in [10] present an advance in linking CDRs with travel demand models. However, these results rely on processed ODs how these are obtained is not detailed and less is known about the comparisons of ODs by activity purpose and time of the day.

Here we advance in that direction, detailing a method of how CDRs can be used to extract ODs by purpose and time of the day. The presented results are validated against surveys and existing ODs available from the local DOT. We show the robustness of the method comparing the results over two types of phone data sets and in two cities. We validate each case of study against different sources of information available. We end the discussion with limitations of the information provided by the passive data and future directions to overcome these issues.

## DATA

The datasets studied in this work are from the metropolitan area of Rio de Janeiro, Brazil and for Boston, USA. General information about the two cities are summarized in Table 1.

Each record of these datasets contains an anonymous user ID, the geographical location in the form of the latitude and the longitude, and the time at the instance of the phone activity- which includes made calls and sent text messages. For Boston, the coordinates of the records are estimated by the service provider (AirSage) based on a standard triangulation algorithm whose

accuracy corresponds to an average of 200 to 300 meters. The data for Rio de Janeiro, on the other hand, is at tower resolution, with 1421 mobile phone towers scattered across the whole state. This work aims to build on the comparison between two resolutions as well as providing a framework that can support both in the generation of origin-destination trips by day period obtained from such passive data.

**TABLE 1 A summary of the CDR data and the demographic information of the two metropolitan areas analyzed in this study, Rio de Janeiro and Boston**

| | Rio de Janeiro | Boston |
|---|---|---|
| # of Calls (millions) | 1046 | 8000 |
| # of Users (millions) | 2.8 | 2.0 |
| Spatial Resolution | Static Lat/Lon pairs | Triangulated Lat/Lon pairs |
| Duration (months) | 5 | 2 |
| Population (millions) | 6.3 | 4.5 |
| Area ($km^2$) | 4557 | 12105 |

## METHODOLOGY

This section explains how the time stamped call records can be converted into individual trajectories with labeled locations, which are then used to generate trip types for each user. The results are then expanded to account for the difference between the number of phone users and the population distribution in the cities considered.

Figure 1 shows a schematic example of transforming daily call records to daily trips. We first detect *stays*, and then *trips* that occur between these stay locations. To generate an activity type for a specific stay, all locations are first labeled by the frequency of calls and time of the day there. As a result, the stay can be labeled as *home*, *work*, or *other*. Based on the inferred activities we can count the observed trips of in each day. We detail each stage of this procedure.

Note that the sample subject of Fig.1 generates 3 HW trips, 4 NHB and 3 HBO trips in the 3 days of observation. Assuming the call activity of this user was expanded to represent 1 day of trips of 27 subjects (which is the average expansion factor of Boston tracks in our data), this user would generate 27 HBW trips, 36 NHB trips and 27 HBO trips for this population of 27 he/she represents. These trips would be distributed across the day based on the observation time of the stays and the departure times assigned. This procedure aims to generate a representative sample of trips from phone users to account for choices of the total population. Thus, the larger the market share and the more phone activity, the better the trips' reflection of the choices of the entire population. The rising ubiquity of phone usage coupled with improvements in localization accuracy means the estimates obtained from passive devices will only improve.

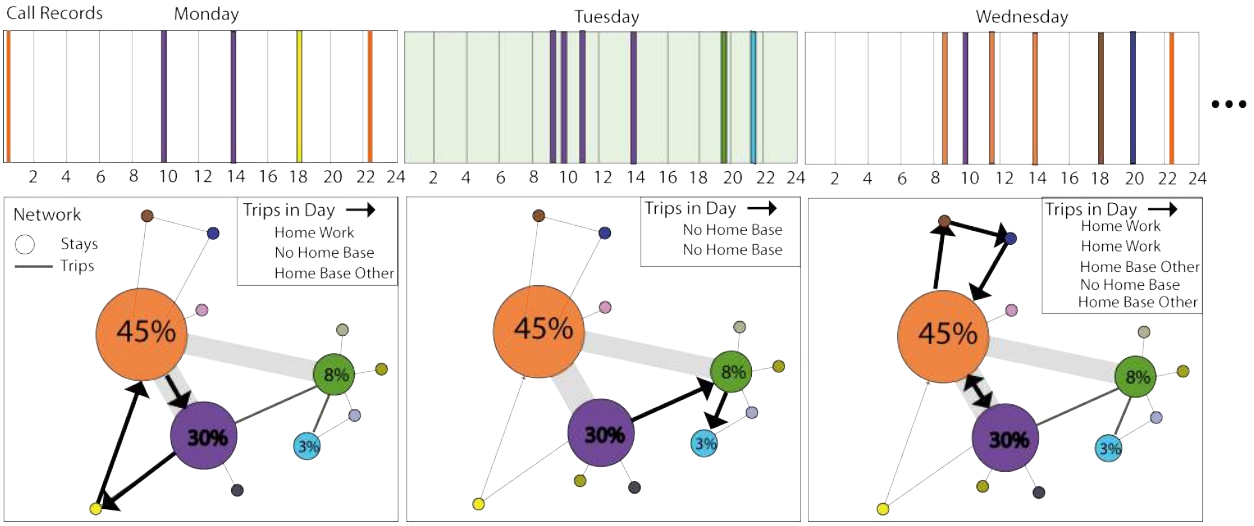


**FIGURE 1. Schematic example of phone records converted in daily trips for one mobile phone user. Activities are inferred in stay locations and daily trips are measured by time of the day between these stays.**

## Stay Detection

CDR data inherently contains noise due to tower-to-tower call balancing performed by the mobile phone service providers, which creates signal jumps that do not represent actual movement. A procedure for GPS traces is applied to the CDR data to correct for these jump and similar such discrepancies in the triangulated Boston data [14-15]. This method simplifies a sequence of calls that are within a specified proximity to the medoid of all such calls.

An analogous process, with minor differences, is applied to Rio de Janeiro's tower-based data. At this resolution one can only know the closest tower to the user's actual location, so the estimate of a user's position is only known up to the Voronoi cell for that tower. Due to the discrete nature of this data, the aforementioned call sequence simplification is carried out by joining sequences of calls made from a sets of towers within a certain distance, followed by joining the sequence of calls made from the same tower. To address issues of temporal resolution we only stays if the users is known to be in that location for at least 10 minutes in both cities.

In summary, stay extraction comprises a series of steps carried out in order remove noisy data, referred to as pass-by-points. Here we should note a consequence of the CDR's passive nature: the user only releases location information when he/she interacts with his/her phone. Consequently, it is possible that the user does indeed stay in some of these removed pass-by locations, or even visits other locations that we simply cannot observe due to lack of phone interaction. The essence of this analysis, however, is that in a long enough time period, the periodicities and the regularities of a user's travel patterns will emerge.

## Activity Inference

Human mobility patterns as captured from mobile phone data exhibit regularity and frequent returns to previously visited sites [13-15]. This behavior can be integrated in transportation

planning, as travel surveys typically focus primarily on home and work locations, and trips are typically categorized by purposes such as home-based-work (HBW), home-based-other (HBO), or non-home-based (NHB).

To successfully extract purposes for every trip, we begin by assigning simple tags to locations. For every user, his/her most visited location on weekdays from 7pm-8am and on weekends is classified as that his/her home. Users with too little activity from their home locations are filtered out of the analysis. This is followed by assigning the user's workplace, which we define to be the location (that is not home) that the user visits second most during the complement of the home time period on weekdays. Similarly, users with too few calls from their assigned workplace are excluded. Stays made from other locations are all classified under *other*; as this level of data resolution does not allow for a distinction between types of other locations, such as school, shopping, etc. Once each stay is labeled with a purpose, then the resulting trips obtained from stay locations can be assigned purpose pairs, such as HBW or HBO or NHB. The ODs obtained in this way will then be classified in terms of their purpose pairs.

These methods are by no means definite solutions for perfectly estimating user home and work locations; on the contrary, they are very simple and straightforward and might lead in some cases to incorrect assignment of home and work locations, and consequently result in misclassification. However, with increased spatial and temporal granularities of data and the inclusion of refined GIS information and demographic data, these methods can and should be replaced by more sophisticated algorithms.

**Trip Generation**

After all the calls have been assigned one of the three location tags (*home*, *work* or *other*), the next step in the procedure is to go through the time-ordered stay sequence for every user. Two consecutive weekdays calls constitute a *raw trip* if they are both weekday calls, are not from the same location and are in the same *effective day*, which spans 3am of the previous day to the 3am of the next. Our method assumes that users typically travel from their home location at the beginning of an effective day and travel back to their home at the end. Therefore if a user's last call of the day is not from the home location, a raw trip is added to home. Similarly, if a user's first call of the day is made from a location other than home, a trip is added to ensure user's travel from his/her home to work.

An important part of this procedure is assigning the raw trip a departure time. As stated before, CDR data is passive and is only generated when users choose to interact with their phones. Therefore the assumption that users start their trip at the exact time they make the call from the origin is flawed. To account for this we introduce a departure time estimation procedure. For Boston, we use call distributions from the National Household Travel Survey (NHTS), to choose a specific time within the time range of the user's two consecutive calls. Due to our lack of access to such surveys in Rio de Janeiro, we carry this weighting scheme using the overall call activity. This is a simple assumption that is not entirely accurate, but yields better results than assuming the call time and trip departure time are concurrent. Figure 2 depicts the distributions for the departure times for the two cities broken down by the purpose of the trip. It can be noted that while all distributions look qualitatively similar, there are unique differences: users from Rio de Janeiro make HBW trips, on average, a couple hours later than Bostonians.

As an overview, every raw trip is associated with a purpose of ‘HBW’, ‘HBO’ or ‘NHB’, an origin, a destination, a departure time and whether the user made it on a day where he was observed at work (making it a *workday*).

## Data Expansion

Post filtering described in section 2.2, we apply additional selection criteria to be involved in further analysis. This filtering consists of eliminating users with too many or too few calls in total or insufficient number of home or work calls. After this procedure, roughly 300,000 users in Boston and 500,000 users in Rio de Janeiro remain.

Next we consider the choice of OD resolution. In Boston we choose between town boundaries (164) and census tracts (974), whereas in Rio our choice is between subdistricts (118) and TAZs (730). The choices here result in the creation of ODs at different spatial resolutions, which affect the resulting OD correlations. The difference in choice between Rio de Janeiro and Boston can be attributed to the granularity of the data in the respective cities.

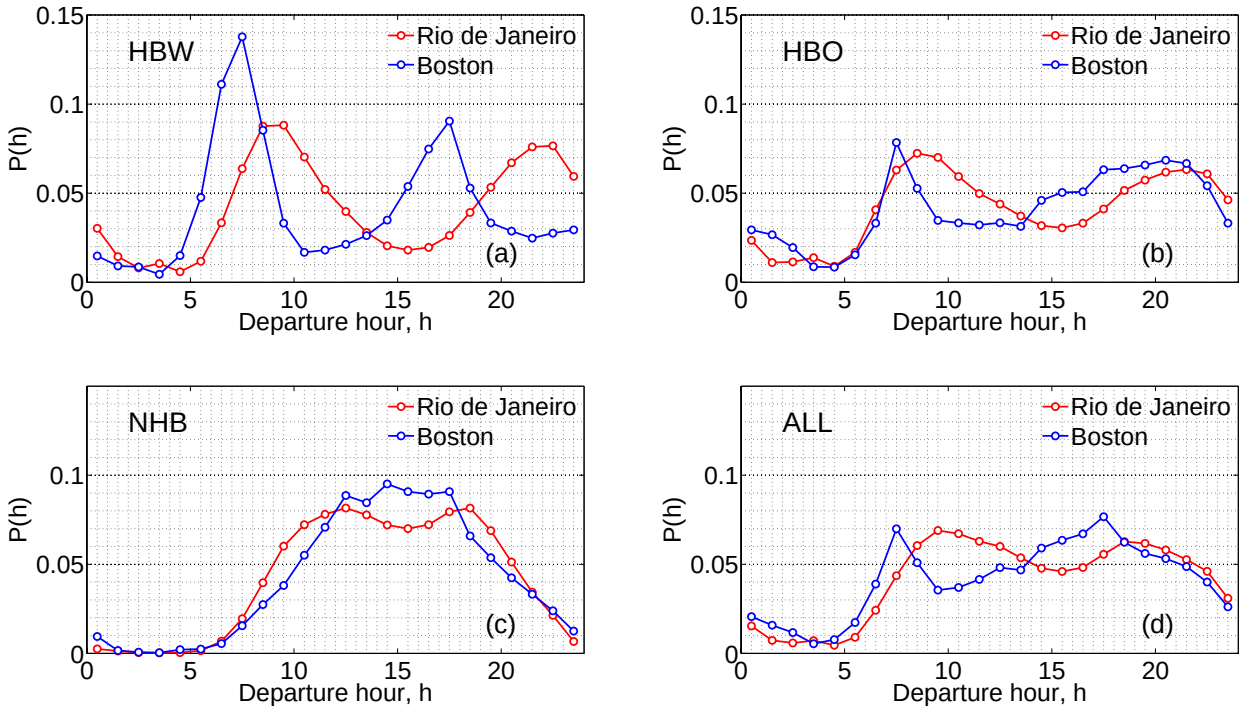


**FIGURE 2. Trip departure hours of a typical weekday for the two areas categorized by trip purpose.**

We count the number of residents to determine the sample size for each of the areas in our analysis. These numbers are qualitatively similar to the number of people surveyed in each zone in the travel surveys, as they will be used as a representation for the population of that zone. Unlike surveys however, mobile phone data offers very little information about the users; therefore traditional methods like stratification [16] that use statistical decision making to ensure healthier sampling are not applicable. On the other hand, the sample sizes are generally larger. In

addition, since mobile phone carriers already store this data for other purposes, the additional cost of gathering this data is negligible, and its use for transport planning is a byproduct.

To use the selected sample to represent the whole population of the area, we propose the following: the actual population of each polygon is divided by the number of users who have been classified as that zone's residents from the CDR data to obtain what we'll refer to as the *expansion factor*. Typically in this upscaling procedure the standard is to use models like IPF [17] that utilize functions that take into account not only the origin of the trip but also the destination and other parameters. Figure 3 exhibits the expansion factor distributions at both cities. For Rio de Janeiro, between the two options of spatial resolution, zones appear to achieve lower expansion factors on average when compared to the subdistrict level. This can be attributed to the irregularities of the subdistrict sizes and populations compared to the zones. In Boston, the expansion factors are smaller in general and more importantly, more evenly distributed across the metropolitan area. This is also because the CDR data for Boston is triangulated and thus almost continuous in space, which allows for a more accurate assessment of home locations.

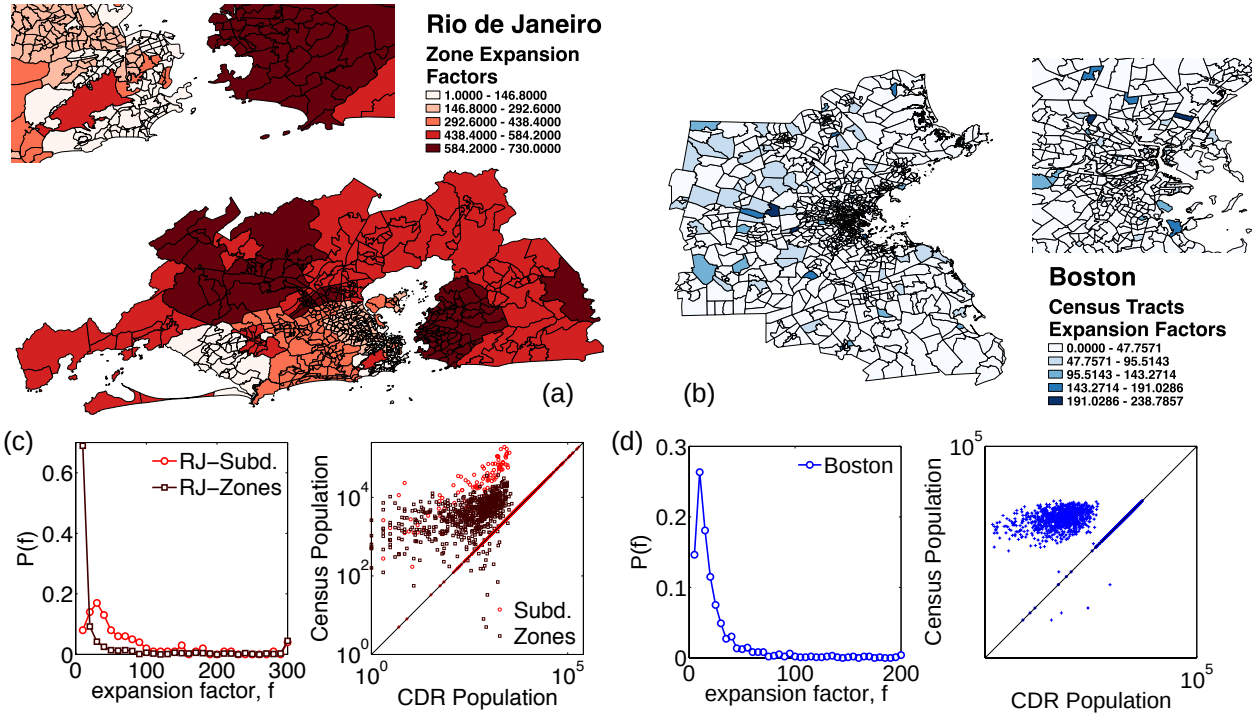


**FIGURE 3. Maps depicting the expansion factors in both cities in the specified resolutions and the distributions of the expansion factors, and comparison between the CDR population and how it's scaled up to the census population.**

## Algorithm

The procedure and the methods outlined so far in this section are summarized in an algorithm form in Figure 4. Once stays are extracted from every user's raw call data, home and work locations are determined. All unique stays are labeled with a purpose of *home*, *work* or *other*. Then a user is selected if his number of calls is within acceptable bounds and he has enough calls

from home and work. For every selected user, the number of CDR residents in the polygon corresponding to his/her home location is incremented. In addition, the number of weekdays and workdays on which a user has been observed is stored.

Then, raw trips are extracted from each user's stays. Two consecutive stays that are not both from home or both from work and are either in the same effective day or only one day apart constitute a raw trip. Such trips are assigned purposes based on the purposes of the two stays, and mapped to origins and destinations based on locations of these stays. Departure time is chosen from the time range between two stays according to preset distributions of trip departure times.

Finally, all raw trips are assigned a magnitude equal to the expansion factor divided by that user's total number of workdays if he/she has been observed at work in that day of the trip, and by his/her total number of weekdays otherwise. This magnitude is then added to the OD table with the appropriate origin, destination, purpose and time period to obtain the final average weekday ODs.

## RESULTS

In this section, our goal is to test the accuracy of origin-destination information obtained from CDR data using our methodology, its limitations, and how spatial resolution influences accuracy. In doing so we will proceed in parallel with the steps of the traditional four-step model.

For validation, we compared our results to the ODs obtained from 2006-2010 Census Transportation Planning Products (CTPP) [18] results in Boston and the Transportation plan in Rio in 2013. Our comparisons are confined to morning home-work commuting flows as information about trips by other purposes and times of day is not available in either of these data sources.

### Trip Generation

As in accordance with the flow of the traditional four-step model, we begin analyzing our results by comparing trip generations: the total numbers of trip productions and attractions in both cities. Figure 5 exhibits very high correlation between the CDR and survey data, which almost reaches $\rho = 1$ in Boston. For both cities the trip productions and attractions lie along the $y = x$ line, validating the strength of the CDR data and our procedure to provide production and attraction data.

**Algorithm 1** Estimating OD Matrices from CDRs

```
POP[o] = 0 for each location o
N[o] = 0 for each location o
OD(o, d, p, t) = 0 for origin o, destination d, purpose p and period t

{Detecting home and work locations, assigning labels, selecting users}
for all users u do
  u.stays = vector of stays of u sorted by time
  u.home = most visited location on weekday nights and weekends
  u.work = most visited location on weekday work hours
  for all stays s in u.stays do
    set s.label as either H, W or O.
    if n_min < u.numCalls < n_max, and u.homecalls > minhomecalls,
    and u.workcalls > minworkcalls, and u.home ≠ u.work then
      u.selected ← true
    end if
  end for
  if u.selected = true then
    N[u.home] + +
  end if
  calculate u.weekdays, unique weekdays user has been observed
  calculate u.workdays, unique weekdays user has been observed at work
end for

{Generating raw trips}
rawtrips = set of all raw trips
for all users u | u.selected = true do
  for i = 2 to i = length(u.stays) do
    s0 = u.stays[i − 1] and s1 = u.stays[i]
    if s0 and s1 are in the same effective day then
      create trip and trip.user = u
      trip.o = s0.location and trip.d = s1.location
      set trip.purpose based on s0.label and s1.label
      set trip.workday = true if user was observed at work in this day
      set trip.departure based on overall trip departure knowledge
      add trip to rawtrips
    else
      create ntrip, mtrip and mtrip.user = ntrip.user = u
      ntrip.o = s0.location and ntrip.d = u.home
      mtrip.o = u.home and mtrip.d = s1.location
      set ntrip.purpose based on s0.label and H
      set mtrip.purpose based on H and s1.label
      set trip.workday for both days and trip.departure for both trips
      add ntrip, mtrip to rawtrips
    end if
  end for
end for

{Trip expansion}
for all rawtrips r do
  u = r.user
  if u.workdays > 0 and r.workday = true then
    f = POP[o]/N[u.home]/u.workdays
  else
    f = POP[o]/N[u.home]/u.weekdays
  end if
  OD(r.o, r.d, r.purpose, r.departure)+ = f
end for

* inPolygon(b) returns the polygon from which the call was made.
```

**FIGURE 4. Algorithm outlining the OD generation procedure.**

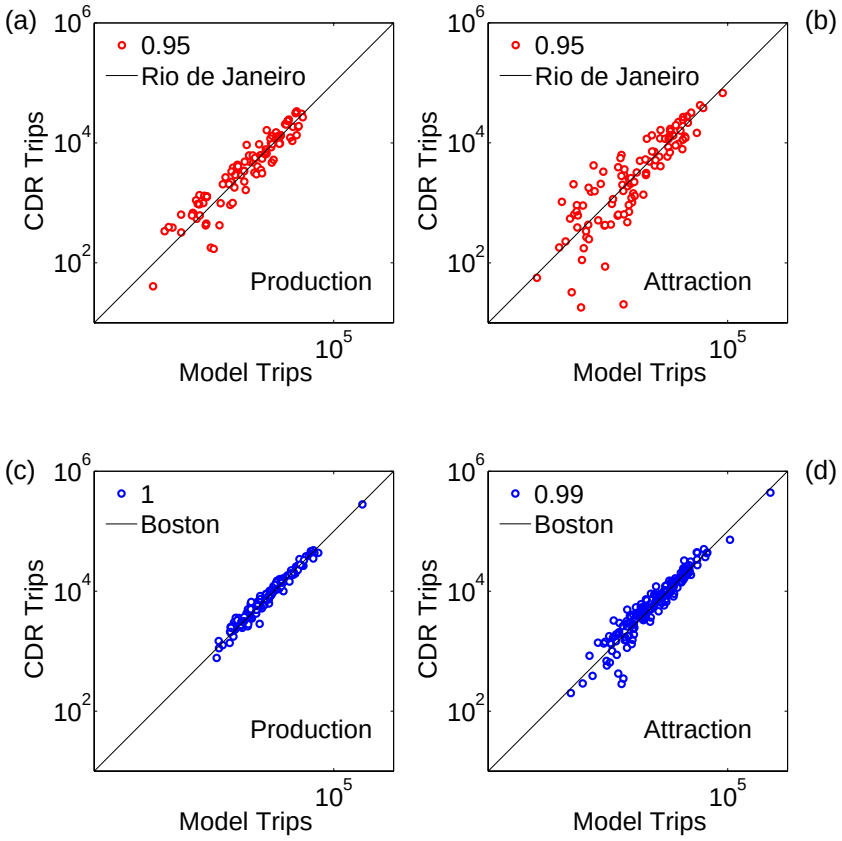


**FIGURE 5. Trip production and attraction comparisons of inter-subdistricts in Rio de Janeiro and inter-towns in Boston**

## Trip Distribution by Time and Purpose

In the next step we compare trip distributions. Figure 6 assesses the results. Comparison of the HBW trips for each origin-destination pair in the morning peak shows strong correlations for both inter-town and intra-town trips, reaching $\rho = 0.84$ for Rio de Janeiro and $\rho = 0.99$ in Boston. Figure 6 also illustrates spatially the flow distribution of the model and the CDR ODs for both cities by mapping color-coded and width adjusted lines between OD pairs whose flow values exceed 0.10% of the total study area trips. By visual inspection, it can be said that CDR data manages to capture the flow distribution of that of the model ODs, with majority of the flows concentrated towards downtown in Boston and downtown and across the bay in Rio de Janeiro.

Finally, we set out to compare correlations and total trip counts by purpose and time of day. Table 2 shows how our findings compare with the model results. For Rio, the OD generation procedure is carried out in three distinct cases. Case 1 applies to home detection and OD generation at the larger subdistrict level. Case 2 does both at the smaller zone level. Case 3 detects homes and initially generates ODs at the zone level, which are then aggregated to subdistrict-to-subdistrict ODs. In both cities at the higher resolution, at tract level in Boston and at zone level (Case 2) in Rio, the correlations are weak. When resolution is smaller, corresponding to towns in Boston and subdistricts in Rio, the correlations are 0.96 and 0.83, respectively. At this point it can be noted that the CDR data is only good at generating ODs at a

certain resolution. This can be attributed to that fact that, especially in Rio, there are not enough CDR users in certain areas, which consequently inflates the expansion factors and estimated trips. The best results are obtained when home detection is carried at the smaller zone level but then the final ODs are aggregated to the larger level (Case 3 in Rio). Therefore it can be argued that for better results, home detection at the finest resolution available is important, but final aggregation of the ODs may be necessary for more representative OD information.

One significant shortcoming of this procedure is the mismatch on number of total daily trips with transportation models of the city. For both cities, with the exception of HBW trips, the total number of daily trips differs significantly from those estimated in the models. The main reason for this issue seems to be the simplicity of the expansion procedure used here. This reflects the need for a more elaborate procedure for trip distribution that takes into account more than just the ratio of CDR users to the actual population at the origin and their average number of daily trips. More attention must be paid to the configuration of daily trip chains used per user [15], the number of daily trips counted per user and other factors that might affect the representative power of a single raw trip.

**TABLE 2 Trip Distributions by purpose and by day period for the two cities**.

| **RIO DE JANEIRO** | HBW | NHB+HBO | AM | MD | PM | Total |
|---|---|---|---|---|---|---|
| CDR Trips (millions) | 2.24 | 7.40 | 2.25 | 4.60 | 2.79 | 9.64 |
| Model Trips (millions) | 2.06 | 1.67 | 1.31 | 1.19 | 1.24 | 3.74 |
| Case 1 - Correlation | 0.43 | Model ODs are not suitable in resolution for validation. | | | | |
| Case 2 - Correlation | 0.36 | | | | | |
| Case 3 - Correlation | 0.83 | | | | | |
| **BOSTON** | HBW | NHB+HBO | AM | MD | PM | Total |
| CDR Trips (millions) | 2.81 | 12.57 | 2.46 | 4.12 | 4.15 | 10.73 |
| MHTS Trips (millions) | 2.14 | 16.17 | 3.99 | 6.24 | 6.06 | 16.29 |
| Tract Pair Correlation | 0.3 | 0.61 | 0.42 | 0.65 | 0.54 | 0.58 |
| Town Pair Correlation | 0.96 | 0.98 | 0.97 | 0.98 | 0.97 | 0.98 |

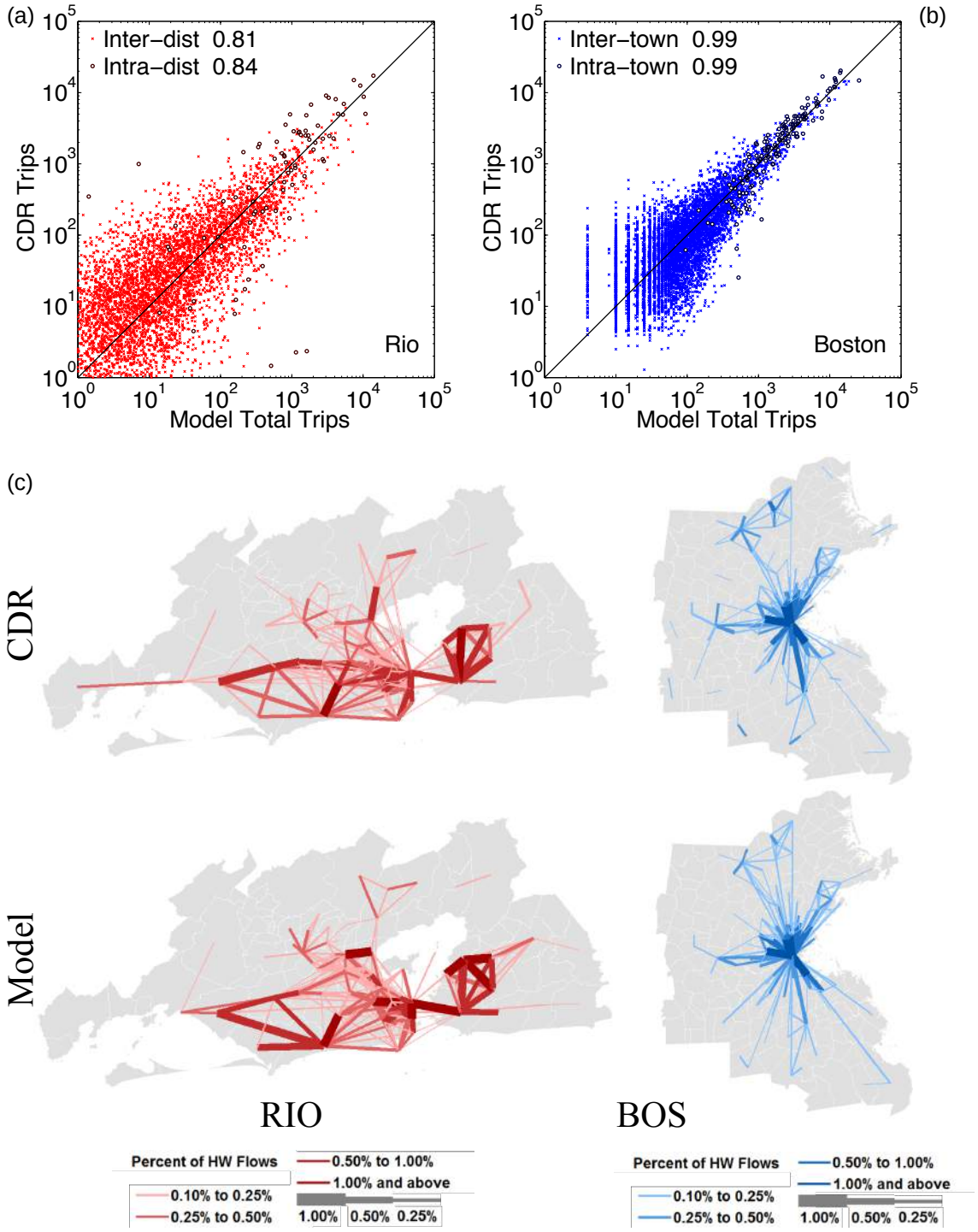


**FIGURE 6. Assessment of trip distributions by comparison of inter and intra town (subdistrict) OD pairs and a spatial illustration of the flows (>0.10% of total trips)**

## SUMMARY AND CONCLUSIONS

In this work we set out to extract origin destination information for two cities, Boston, USA and Rio de Janeiro, Brazil, from large passive mobile phone data sets, consisting of billions of geo-tagged mobile phone call records, made by millions of users. We proposed a portable methodology available for immediate use in other cities and demonstrated its applicability in the two subject cities. To make our proposed algorithm easily deployable we provide a step by step formulation in pseudo code.

As is the case with any passive dataset, CDRs require considerable preprocessing to distill relevant information. We first applied a standard procedure to remove noise and extract stays from the call data. We then determined home and work locations of users with enough calls, and labeled their stays as home, work or other. In the next step we extracted raw trips by analyzing consecutive stays for each user followed by applying proper expansion factors to raw trips to get representative origin destination trips. We finalized our work by validating against models created for the subject cities.

Our results suggest that CDR data holds promise in accurately estimating production and attraction of trips in addition to trip distributions. Trip production and attraction of HW trips produce very strong results for inter town trips in Boston and inter district trips in Rio. A key point to stress, is that trip correlations are significantly higher when aggregated to larger polygons, indicating that CDR data is not representative of the population when inferring trips between smaller census tracts or travel analysis zones (TAZs). Only when calculating trips among larger zones, like those of the sub districts in Rio and towns in Boston, good results were obtained. These results were measured as correlations with ODs generated by models by time and purpose.

In both cities carrying out the home detection procedure at the smaller zones (TAZ or census tracks) and then aggregating the resulting ODs to the larger sub-districts or towns yielded high correlations. The magnitude of total trips still remains an issue that should be addressed when using only phone data and distribution of population to estimate ODs.

In conclusion, we demonstrated a method that uses CDR data and population distribution alone to produce ODs by purpose and time of the day. We then compared our ODs to those produced by existing ODs models utilizing travel surveys. The method yields excellent correlations of home production and attraction for home-work trips and good correlations in inter-zone trips when the zones contain enough users. Yet, total numbers of trips are larger than estimated by the models of the subject cities. Nevertheless, CDR data may overcome the issue of stated preferences that are inherent in surveys.

Future directions include building on this methodology to better augment survey information, and complete a 4-step model, which encompasses additional complexities such as modal split and route traffic assignment. CDR data treated carefully can be a fertile source to learn about patterns of urban mobility and finding better way to harness this data will continue to be a rich field of research.

## ACKNOWLEDGMENTS

This work was partially funded by the BMW-MIT collaboration under the supervision of PI Mark Leach, the World Bank-HuMNet and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris. We thank Alexandre Evsukoff, Pedro Bittencourt, and Pu Wang for technical support, AirSage and the Rio City Hall for the support and the data they have provided. Our work was also supported, in part, by the UPS Center for Transportation and Logistics Graduate Research Fellowship awarded to Serdar Colak and by the Departmet of Civil and Environmental Engineering at MIT through the Winslow Career development chair to Marta C. Gonzalez.

## REFERENCES

1. http://airsage.com/Contact-Us/White-Paper/
2. http://delcan.com/markets-and-services/services/category/its-technology-systems-integration
3. Ward, K., *Using Cell Phone Technology to Collect Travel Data*, in *TRB Planning* 355 *Applications Conference*. 2011: Reno, NV.
4. Sohn, K. *Dynamic estimation of origin–destination flows using cell phones as probes.* SDI 2004-R-04, Department of Urban Transportation, Seoul Development Institute, Korea, 2004.
5. Akin, D., and Sisiopiku, V.P. *Estimating origin–destination matrices using location information from cell phones*. Proc. 49th Annual North American Meetings of The Regional Science Association Int, Puerto Rico, 2002
6. Ratti, C., Pulselli, R.M., Williams, S., and Frenchman, D. *Mobile landscapes: using location data from cell-phones for urban analysis*, Environ. Plan. B – Plan. Des., 33, (5), pp. 727–748.
7. SENSEable City Laboratory, MIT: *Mobile landscape, Graz in real time*, Available at: http://senseable.mit.edu/graz/
8. Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. *Development of origin–destination matrices using mobile phone call data.* Transportation Research Part C: Emerging Technologies, *40*, 63-74, 2014.
9. Wang, P., Hunter, T., Bayen, A., Schechtner K., and Gonzalez, M.C. *Understanding road usage patterns in urban areas.* Scientific Reports, Vol. 2, Number 1001, 2012.
10. Huntsinger, L. and Donnelly. *Reconciliation of Regional Travel Model and Passive Device Tracking Data.* 93rd Annual Meeting of the 27 Transportation Research Board, Washington, DC, January 2014.
11. R. Hariharan and K. Toyama. *Project Lachesis: Parsing and modeling Location Histories*. Geographic Information Science, Vol. 3234, pages 106-124.
12. Jiang, S., Yang, Y., Fiore,G., Ferreira Jr., J., Frazzoli, E., and Gonzalez, M.C.. *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, 2013.
13. Song, C., Qu, Z., Blumm, N., Barabási, A-L. *Limits of Predictability in Human Mobility*. Science 327, 1018-1021 (2010).
14. Song, C., Koren, T., Wang, P., Barabási, A-L. *Modelling the scaling properties of human mobility.* Nature Physics 7, 713- (2010).

15. Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). *Unravelling daily human mobility motifs. Journal of The Royal Society Interface*, Vol. *10*, No. 84, 20130246.
16. Ben-Akiva, Moshe E., and Steven R. Lerman. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press, 1985.
17. Pukelsheim, F. and Simeone, B. *On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis*. http://opus.bibliothek.uniaugsburg.de/opus4/frontdoor/index/index/docId/1229
18. U.S. Department of Transportation Federal Highway Administration. CTPP 2006-2010 Census Tract Flows. http://www.fhwa.dot.gov/planning/census issues/ctpp/data products/2006- 2010 tract flows/index.cfm.