Analyzing Cell Phone Location Data for Urban Travel

Current Methods, Limitations, and Opportunities

Serdar Çolak, Lauren P. Alexander, Bernardo G. Alvim, Shomik R. Mehndiratta, and Marta C. González

Travelers today use technology that generates vast amounts of data at low cost. These data could supplement most outputs of regional travel demand models. New analysis tools could change how data and modeling are used in the assessment of travel demand. Recent work has shown how processed origin–destination trips, as developed by trip data providers, support travel analysis. Much less has been reported on how raw data from telecommunication providers can be processed to support such an analysis or to what extent the raw data can be treated to extract travel behavior. This paper discusses how cell phone data can be processed to inform a four-step transportation model, with a focus on the limitations and opportunities of such data. The illustrated data treatment approach uses only phone data and population density to generate trip matrices in two metropolitan areas: Boston, Massachusetts, and Rio de Janeiro, Brazil. How to label zones as home- and work-based according to frequency and time of day is detailed. By using the labels (home, work, or other) of consecutive stays, one can assign purposes to trips such as home-based work. The resulting trip pairs are expanded for the total population from census data. Comparable results with existing information reported in local surveys in Boston and existing origin–destination matrices in Rio de Janeiro are shown. The results detail a method for use of passively generated cellular data as a low-cost option for transportation planning.

Every time a cell phone is used, passive mobile monitoring generates a record with the time and approximate location of the event. When monitoring software is installed, Internet usage, GPS coordinates, and much more can be tracked. In the near future, as this information becomes more accurate, the question of whether the combination of GPS and phone data can entirely replace travel diaries will arise. How to make this possible is being studied, and the main challenge 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 do not offer the same detailed information as surveys do. Passive data cannot explain every individual's travel motives in depth or directly ask the questions that can be answered with travel diaries. However, such data contain valuable information about continuously recorded trip choices. This information can be adopted in some phases of transportation applications at low cost in any city worldwide where phone use is ubiquitous. Harvesting, interpreting, and applying these data require innovative thinking about both data acquisition and data analysis.

Applying these passive data requires innovative techniques for big data statistics and analysis. While these data mining techniques mature, a useful question to consider is how to incorporate the processed data to generate travel models. This paper focuses on the latter task. The use of call detail records (CDRs) to generate origin–destination (O-D) trips of various purposes (home, work, other) and times of day is explored. A replicable procedure is presented for processing CDRs to extract information relevant to trip generation. The results are compared with travel surveys from Boston, Massachusetts, and with independently generated O-D pairs from Rio de Janeiro, Brazil. The accuracy of the analyzed mobile phone data sets is different for the two cities. The coordinates from Rio de Janeiro are at tower resolution, and Boston coordinates come from a triangulation algorithm applied by the data provider. Furthermore, the validation sources differ. For Boston, census and travel diary survey commuting data are used, whereas for Rio de Janeiro, O-D estimates by purpose and time of day are used. Both analyses are presented here to show that the proposed method is robust to varying conditions of accuracy. The current limitations of the data to inform all the steps of a traditional transportation model are discussed. Possible avenues for future work are suggested.

Passive data alone may not be the ultimate solution to gathering detailed information needed for a complete transportation model such as mode, detailed activity types, and route assignment. Nevertheless, the substantial efficacy of passive data for O-D generation is shown here. The richness of these data comes from their ability to provide the trip choices of millions of individual users every minute and everywhere. This work has two main findings: first, evaluating the extent to which cell phone data accurately reflect daily travel, and second, developing guidelines for how to best use these data to generate origins and destinations by purpose and time of day.

LITERATURE REVIEW

In the United States, the first works that used mobile phone data for transportation applications referred to traffic monitoring. Departments of transportation (DOTs) in various states have carried out these studies in collaboration with private data providers. For example,

S. Çolak and L. P. Alexander, Department of Civil and Environmental Engineering, and M. C. González, Engineering Systems Division, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. B. G. Alvim, SCN, Quadra2 Lote A, Ed. Corporate Center, 7o. Andar, 70.712-900, Brasília–DF, Brazil. S. R. Mehndiratta, World Bank, 1818 H Street, NW, Washington, DC 20433. Corresponding author: S. Çolak, serdarc@mit.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2526, Transportation Research Board, Washington, D.C., 2015, pp. 126–135.
DOI: 10.3141/2526-14

the Virginia DOT in collaboration with AirSage created automatic signaling from cell phones and showed fluctuations of traffic speed on a map (*1*). Similarly, a project in Maryland in collaboration with Delcan Corporation inferred traffic along main highways (*2*). In a validation exercise, researchers in North Carolina used 1 month of data to calculate travel times of monitored devices. The extracted travel times and volume delay function for 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 works, such as those by Sohn (*4*) and Akin and Sisiopiku (*5*), performed O-D matrix calculations with simulations of mobile phone data. The main focus of those works was evaluation of the efficacy of the techniques related to metering frequency and numbers of localizations necessary to achieve accurate traffic estimates. These approaches to collecting and analyzing data rely on continuous or close monitoring. Accuracy is gained at the cost of fewer 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, that is, CDRs, can support the models of travel behaviors. Most of the literature working with CDRs addresses data mining techniques used to discover attributes of the data. For example, Ratti et al. (*6*) created mobile landscapes of Graz, Austria, visualizing the dynamic of a town in real time from hundreds of thousands of mobile phone users (*7*). The researchers measured call density (measured with Erlang) and origins and destinations (by way of handovers). More elaborate attempts have mapped millions of consecutive calls from mobile phone users to the roads and compared the results with average car volumes in a given period measured with cameras (*8*) or with the travel times of cars by using a Bureau of Public Roads function (*9*). These works do not compare with travel diaries of trip purpose. Recent advances in the area compared estimated vehicle traffic in roadways from mobile phone data with 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, finding good agreement between these processed AirSage intertrack O-D pairs and the results of the regional model. The results of Huntsinger and Donnelly represent an advance in linking CDRs with travel demand models (*10*). However, the results relied on processed origins and destinations: how these are obtained is not detailed, and less is known about the comparisons of O-D pairs by activity purpose and time of day.

This paper advances in that direction, detailing a method for use of CDRs to extract origins and destinations by purpose and time of day. The presented results are validated against surveys and existing origins and destinations available from the local DOT. The robustness of the method is shown through comparison of the results for two types of phone data sets and in two cities. Each case study is validated with various available sources of information. The discussion ends with limitations of the information provided by the passive data and future directions for overcoming these issues.

DATA

The data sets studied in this work were from the metropolitan areas of Rio de Janeiro and Boston. General information about the two cities is given in Table 1.

Each record of these data sets contains an anonymous user ID, the geographical location in latitude and longitude, and the time at the instance of the phone activity, which includes calls made and text messages sent. For Boston, the coordinates of the records are estimated by the service provider (AirSage) according to a standard triangulation algorithm whose accuracy corresponds to an average of 200 to 300 m. The data for Rio de Janeiro, however, are at tower resolution; 1,421 mobile phone towers are scattered across the state. This work builds on the comparison of two resolutions and provides a framework that can support both the generation of O-D trips by day obtained from such passive data.

TABLE 1 CDR Data and Demographic Information

General Preliminary Information	Rio de Janeiro	Boston
Number of calls (millions)	1,046	8,000
Number of users (millions)	2.8	2.0
Spatial resolution	Static latitude–longitude pairs	Triangulated latitude–longitude pairs
Duration (months)	5	2
Population (millions)	12.6	4.5
Area (km^2)	43,600	4,600

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 is a schematic example of transforming daily call records to daily trips. First stays are detected, and then trips that occur between these stay locations are detected. To generate an activity type for a specific stay, all locations are first labeled by the frequency of calls and time of day. Then the stay can be labeled as home, work, or other. From the inferred activities, the observed trips in each day can be counted. Each stage of this procedure is detailed here.

The sample subject of Figure 1 generates three home-based work (HBW) trips, four no-home-base (NHB) trips, and three home-base-other (HBO) trips in the 3 days of observation. If the call activity of this user were expanded to represent 1 day of trips by 27 subjects (which is the average expansion factor of Boston tracks in the data), this user would generate 27 HBW trips, 36 NHB trips, and 27 HBO trips for the population of 27 he or she represents. These trips would be distributed across the day according to the observation time of the stays and the departure times assigned. This procedure generates 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.

Stay Detection

CDR data inherently contain noise because of tower-to-tower call balancing performed by the mobile phone service providers,

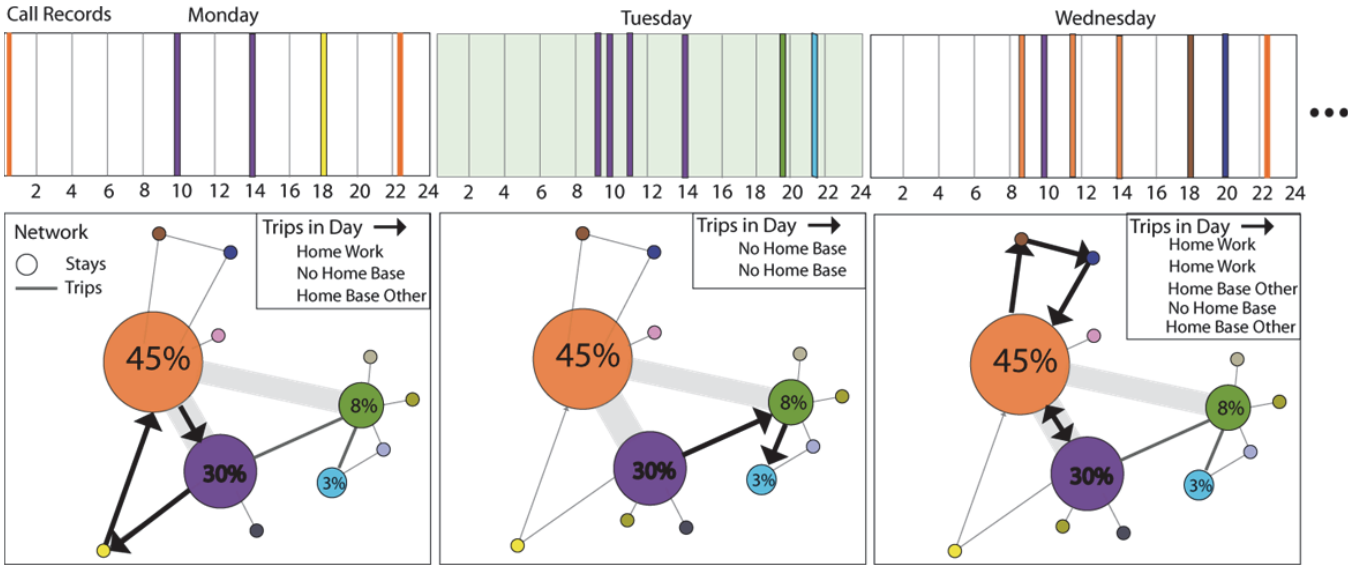

FIGURE 1 Schematic example of phone records converted to daily trips for one mobile phone user. Activities are inferred in stay locations, and daily trips are measured by time of day between stays.

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 jumps and similar such discrepancies in the triangulated Boston data (*11, 12*). 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, only the tower closest to the user's location can be known, so the estimate of a user's position is known only up to the Voronoi cell for that tower. Because of the discrete nature of these data, the call sequence simplification is carried out by joining sequences of calls made from a set of towers within a certain distance, followed by joining the sequence of calls made from the same tower. To address issues of temporal resolution stays are counted only if the user is known to be in that location for at least 10 min in both cities.

Stay extraction comprises a series of steps carried out to remove noisy data, referred to as pass-by points. As a consequence of the CDR's passive nature, the user releases location information only when she interacts with her phone. Consequently, it is possible that the user does stay in some of these removed pass-by locations, or even visits other locations that cannot be observed because of lack of phone interaction. The essence of this analysis, however, is that in a period that is long enough, the periodicities and the regularities of a user's travel patterns will emerge.

Activity Inference

Human mobility patterns 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 typically are categorized by purpose, such as HBW, HBO, or NHB.

For successful extraction of purpose for every trip, simple tags are assigned to locations. For every user, his most visited location on weekdays from 7 p.m. to 8 a.m. and on weekends is classified as his home. Users with too little activity from their home locations are filtered out of the analysis. Next, the user's workplace is assigned, defined as the location aside from home that the user visits second most frequently during the complement of the home time period on weekdays. Users with too few calls from their assigned workplace are excluded. All stays made from other locations are classified as other, because this level of data resolution does not allow for a distinction between types of other locations, such as school or shopping. After each stay has been labeled with a purpose, the resulting trips obtained from stay locations are assigned purpose pairs, such as HBW, HBO, or NHB. The origins and destinations obtained in this way are then classified according to their purpose pairs.

These methods are not definite solutions for perfectly estimating user home and work locations; on the contrary, they are simple and straightforward and could lead to incorrect assignment of home and work locations, resulting in misclassification. However, with increased spatial and temporal granularities of data and the inclusion of refined geographic information system data 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 weekday 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 from 3 a.m. of the previous day to 3 a.m. of the next. The 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 users travel from home to work.

An important part of this procedure is assigning the raw trip a departure time. CDR data are passive and are generated only 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, a departure time estimation procedure is used. For Boston data, call distributions from the National Household Travel Survey are used on which to base a specific time within the time range of the user's two consecutive calls. Because such surveys could not be accessed for Rio de Janeiro, this weighting scheme is carried out with the overall call activity. This is a simple assumption that is not entirely accurate, but it 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. Although all distributions look qualitatively similar, there are unique differences: for example, users in Rio de Janeiro make HBW trips, on average, a couple of hours later than Bostonians.

Every raw trip is associated with the purpose, that is, HBW, HBO, or NHB; an origin; a destination; a departure time; and whether the user made the trip on a day on which she was observed at work (that is, a workday).

Data Expansion

Before expansion, additional selection criteria are applied for further analysis. This filtering eliminates users that have too many or too few total calls or insufficient home or work calls. After this procedure, roughly 300,000 users in Boston and 500,000 users in Rio de Janeiro remained.

Next the choice of O-D resolution is considered. In Boston the choice is between town boundaries (164) and census tracts (974), and in Rio de Janeiro the choice is between subdistricts (118) and TAZs (730). The choices here result in the creation of origins and destinations at various spatial resolutions, which affect the resulting O-D correlations. The difference in choice between Rio de Janeiro and Boston can be attributed to the granularity of the data in the respective cities.

A count of the number of residents is used to determine the sample size for each of the areas in the 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 offer little information about the users; therefore, traditional methods like stratification (*16*) that use statistical decision making to ensure healthier sampling are not applicable. However, the sample sizes are generally larger. In addition, since mobile phone carriers already store these data for other purposes, the additional cost of gathering the data is negligible, and use of the data for transport planning is a by-product.

So the selected sample can be used to represent the whole population of the area, 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 the expansion factor. Typically in this upscaling procedure the standard is to use models such as iterative proportional fitting (*17*), whose functions take into account not only the origin of the trip but also the destination and other parameters. Figure 3 shows the expansion factor distributions for both cities. For Rio de Janeiro, of the two options for spatial resolution, zones appear

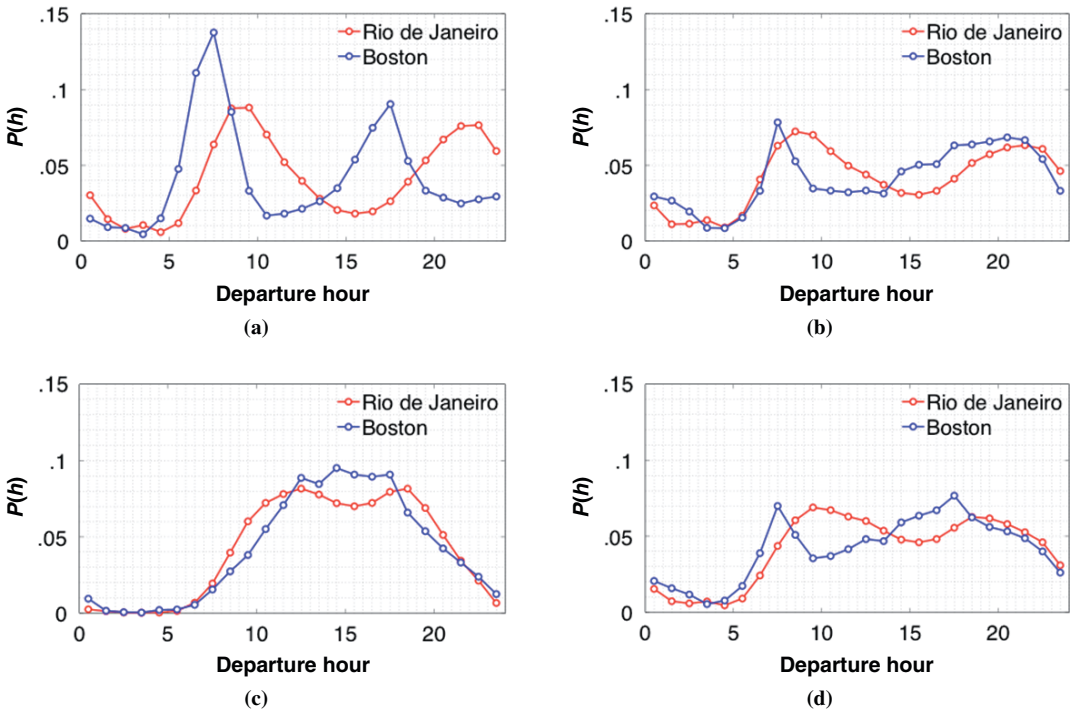

FIGURE 2 Trip departure hours of typical weekday categorized by trip purpose: (*a*) HBW, (*b*) HBO, (*c*) NHB, and (*d*) all ($P(h)$ = probability distribution of h, the departure hour).

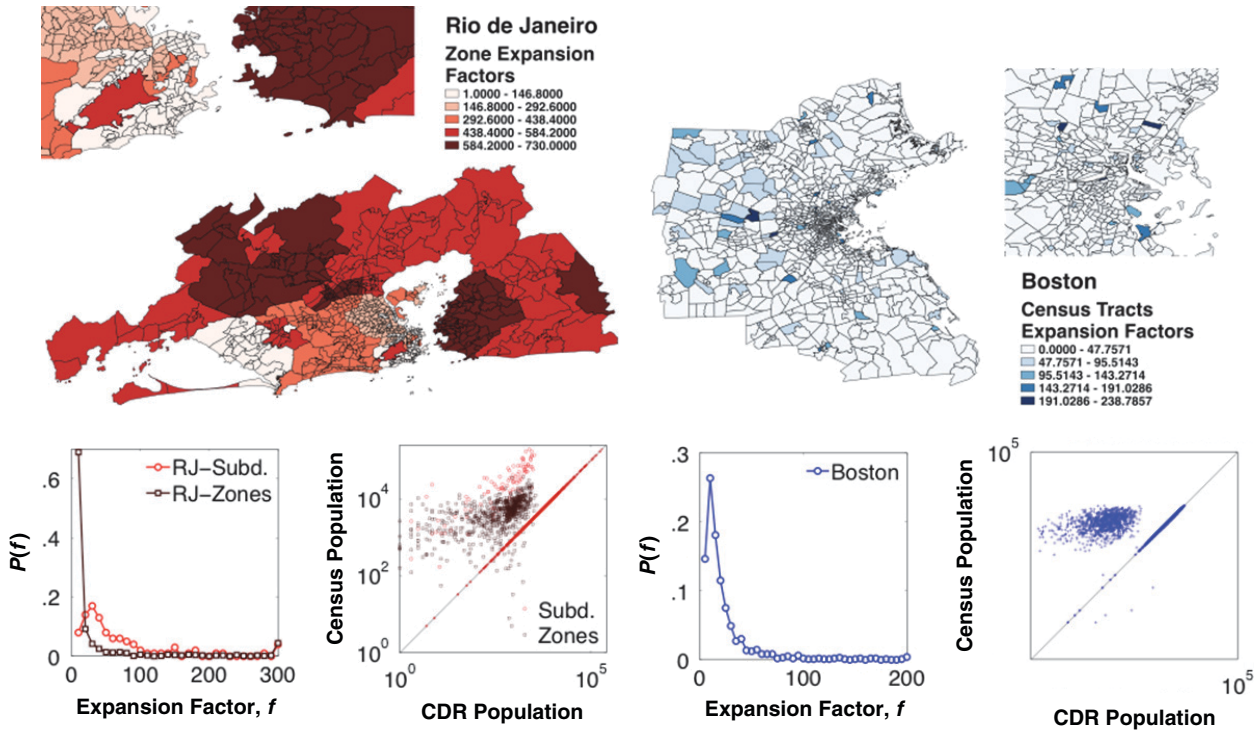

FIGURE 3 Maps depicting expansion factors in both cities in specified resolutions and distributions of expansion factors, and comparison of CDR population and how it is scaled up to census population ($P(f)$ = probability distribution of f, the expansion factor).

to achieve lower expansion factors on average compared with the subdistrict level. This effect is attributed to the irregularities of the subdistrict sizes and populations compared with those of the zones. In Boston, the expansion factors are smaller and, more significant, are more evenly distributed across the metropolitan area. The CDR data for Boston are triangulated and thus almost continuous in space, which allows for a more accurate assessment of home locations.

Algorithm

The procedure and the methods outlined so far in this section are summarized as an algorithm in Equation Box 1. 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 or 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 1 day apart constitute a raw trip. Such trips are assigned purposes according to the purposes of the two stays and are mapped to origins and destinations according to the 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 she has been observed at work in that day of the trip and by her total number of weekdays otherwise. This magnitude is then added to the O-D table with the appropriate origin, destination, purpose, and time period, and the final average weekday O-D pairs are determined.

RESULTS

This section tests the accuracy of O-D information obtained from CDR data with the method, the method's limitations, and how spatial resolution influences accuracy. The steps of the traditional four-step model are used.

For validation, the results were compared with the origins and destinations obtained from Census Transportation Planning Package results for Boston for 2006 to 2010 (*18*) and the Rio de Janeiro transportation plan for 2013. The comparisons are confined to morning home-to-work commuting flows, because information about trips by other purposes and times of day is not available from either of these data sources.

Trip Generation

In accordance with the traditional four-step model, the results analysis begins by comparing trip generations: the total numbers of trip productions and attractions in both cities. Figure 4 exhibits very

EQUATION BOX 1 Algorithm 1. Estimating O-D 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 0.
    if $n_{min} < u.numCalls < n_{max}$, and $u.homecalls > minhomecalls$,
    and $u.workcalls > minworkcalls$, and $u.home \neq u.work$ then
      $u.selected \leftarrow$ 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.

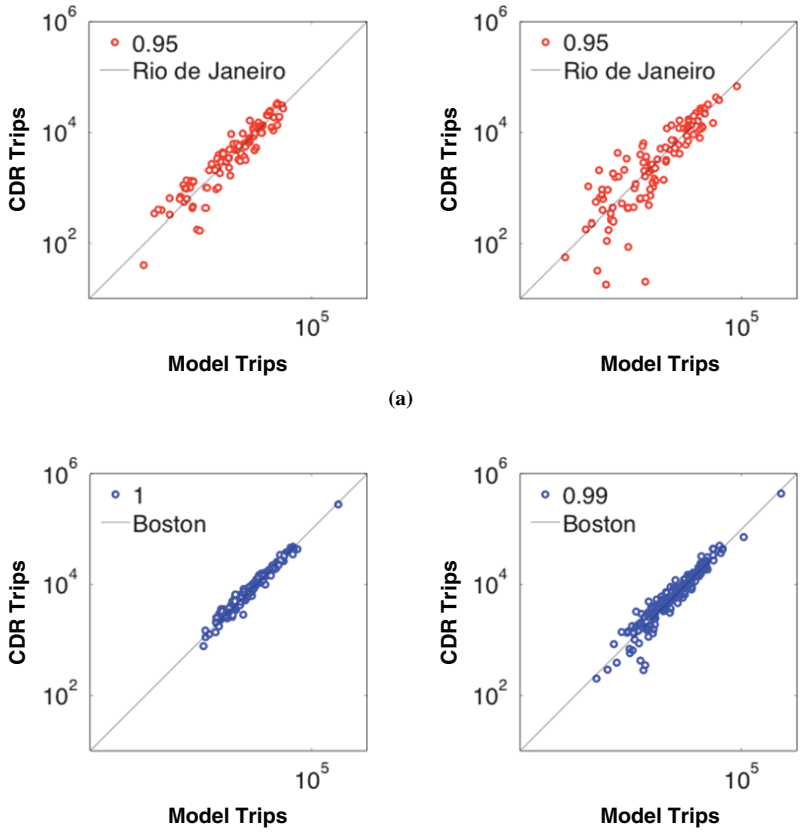

FIGURE 4 Trip production and attraction comparisons, between subdistricts in (a) Rio de Janeiro and (b) between towns in Boston metropolitan area.

high correlation between the CDR and survey data, which almost reaches ρ = 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 that the procedure can provide production and attraction data.

Trip Distribution by Time and Purpose

The next step compares trip distributions. Figure 5 assesses the results. Comparison of the HBW trips for each O-D pair in the morning peak shows strong correlations for both intertown and intratown trips, reaching ρ = .84 for Rio de Janeiro and ρ = .99 in Boston. Figure 5 also illustrates spatially the flow distribution of the model and the CDR origins and destinations for both cities with color-coded and width-adjusted lines between O-D pairs whose flow values exceed 0.10% of the total study area trips. The figure shows that CDR data capture the flow distribution of that of the model O-D pairs, the majority of the flows concentrating toward downtown Boston and downtown and across the bay in Rio de Janeiro.

Finally, correlations and total trip counts are compared by purpose and time of day. Table 2 compares the findings and the model results. For Rio de Janeiro, the O-D generation procedure is carried out in three distinct cases. Case 1 applies to home detection and O-D generation at the larger subdistrict level. Case 2 does both at the smaller zone level. Case 3 detects homes and initially generates origins and destinations at the zone level, which are then aggregated to subdistrict-to-subdistrict origins and destinations. In both cities at the higher resolution, at the tract level in Boston and at the zone level (Case 2) in Rio de Janeiro, the correlations are weak. When resolution is smaller, corresponding to towns in Boston and subdistricts in Rio de Janeiro, the correlations are .96 and .83, respectively. At this point, the CDR data are good only at generating origins and destinations at a certain resolution, because, especially in Rio, there are not enough CDR users in certain areas, inflating 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 origins and destinations are aggregated to the larger level (Case 3 in Rio de Janeiro). Therefore, for better results, home detection at the finest resolution available is important, but final aggregation of the origins and destinations may be necessary for more representative O-D information.

A significant shortcoming of this procedure is the mismatch of 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 appears to be the simplicity of the expansion procedure used here. Thus a more elaborate procedure for trip distribution is needed that takes into account more than just the ratio of CDR users to the actual population at the origin and the average number of daily trips. More attention is needed for the configuration of daily trip chains used per user (*15*), the number of daily trips counted per user, and other factors that could affect the representative power of a single raw trip.

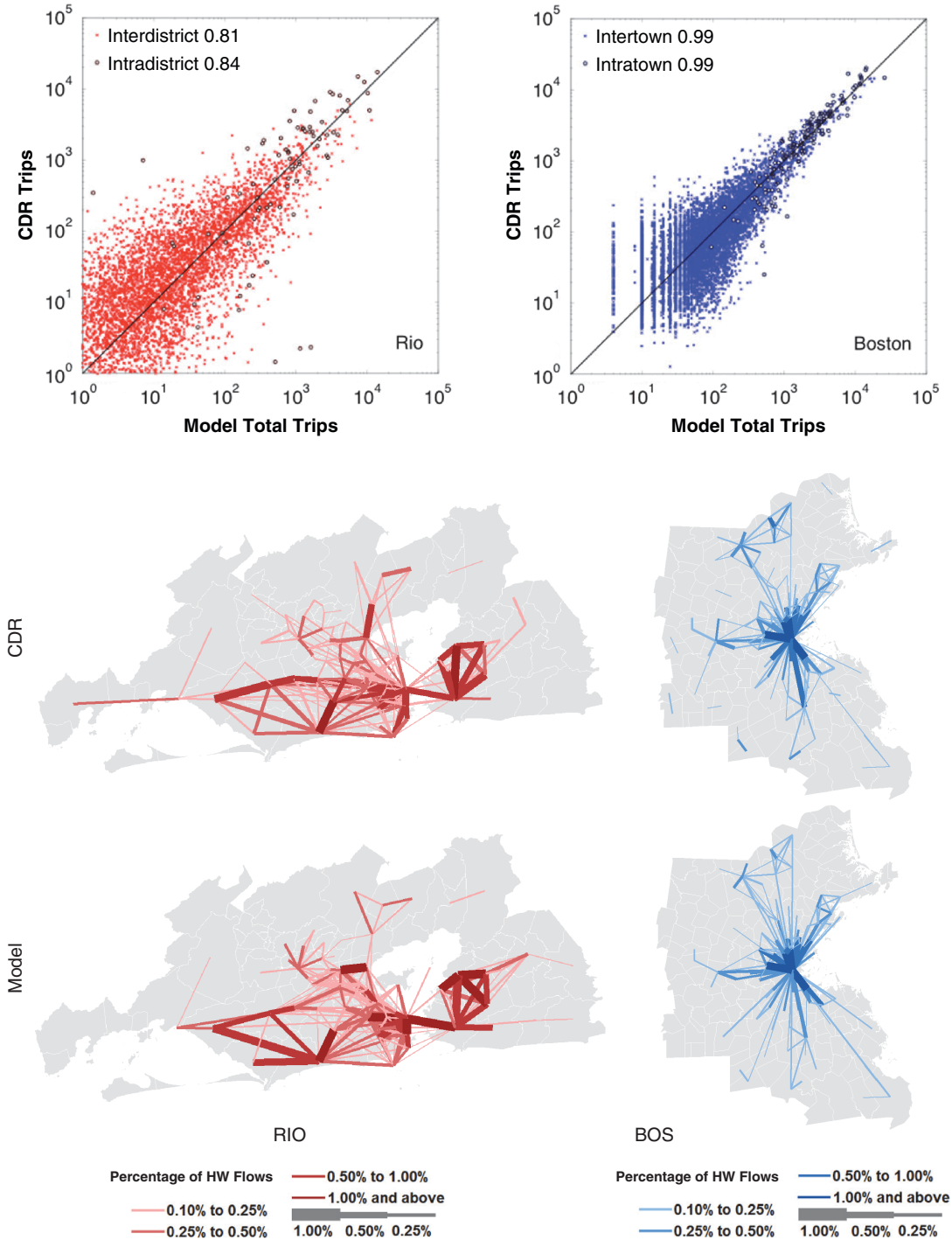

FIGURE 5 Assessment of trip distributions by comparison of inter- and intratown (subdistrict) O-D pairs and spatial illustration of flows (>0.10% of total trips) (RIO = Rio de Janeiro; BOS = Boston).

TABLE 2 Trip Distributions by Purpose and Period

Trip Distribution Comparison	HBW	NHB+HBO	a.m.	MD	p.m.	Total
Rio de Janeiro						
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	.43					
Case 2, correlation	.36					
Case 3, correlation	.83					
Boston						
CDR trips (millions)	2.81	12.57	2.46	4.12	4.15	10.73
NHTS trips (millions)	2.14	16.17	3.99	6.24	6.06	16.29
Tract pair correlation	.3	0.61	0.42	0.65	0.54	0.58
Town pair correlation	.96	0.98	0.97	0.98	0.97	0.98

NOTE: NHTS = National Household Travel Survey; a.m. = morning; MD = midday; p.m. = evening. Blank cells indicate that resolution of model O-D pairs was not suitable for validation.

SUMMARY AND CONCLUSIONS

This work extracted O-D information for two cities, Boston and Rio de Janeiro, from large passive mobile phone data sets, consisting of billions of geotagged records of mobile phone calls made by millions of users. A portable method available for immediate use in other cities was proposed, and its applicability in the two subject cities was demonstrated. A step-by-step formulation in pseudocode was provided so the algorithm would be easily deployable.

As for any passive data set, CDRs require considerable preprocessing to distill relevant information. First a standard procedure was applied to remove noise and extract stays from the call data. Then home and work locations of users who had enough calls were determined, and their stays were labeled as home, work, or other. In the next step, raw trips were extracted through analysis of consecutive stays for each user, followed by application of proper expansion factors to raw trips to get representative O-D trips. The method was then validated with models created for the subject cities.

The results suggest that CDR data could accurately estimate production and attraction of trips in addition to trip distributions. Trip production and attraction of home–work trips produced very strong results for intertown trips in Boston and interdistrict trips in Rio de Janeiro. Trip correlations are significantly higher when aggregated to larger polygons, indicating that CDR data are not representative of the population for inferring trips between smaller census tracts or TAZs. Only when trips among larger zones were calculated, such as those of the subdistricts in Rio de Janeiro and towns in Boston, were good results obtained. These results were measured as correlations with origins and destinations 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 O-D pairs to the larger subdistricts or towns yielded high correlations. The magnitude of total trips remains a concern that should be addressed when only phone data and distribution of population are used to estimate origins and destinations.

The demonstrated method uses only CDR data and population distribution to produce O-D pairs by purpose and time of day. These O-D pairs were compared with those produced by existing O-D models that use travel surveys. The new method yielded excellent correlations of home production and attraction for HBW trips and good correlations in interzone trips when the zones contained enough users. However, total numbers of trips were larger than those estimated by the models of the subject cities. Nevertheless, CDR data may overcome the problem of stated preferences that is inherent in surveys.

Future directions include building on this methodology to better augment survey information and completing a four-step model, which will encompass additional complexities, such as modal split and route traffic assignment. CDR data, treated carefully, can be a fertile source for learning about patterns of urban mobility and finding better ways to harness these data.

ACKNOWLEDGMENTS

This work was partially funded by collaborations between BMW and the Massachusetts Institute of Technology (MIT), under the supervision of Mark Leach; by the World Bank and the MIT Human Mobility and Networks Lab; and by the Center for Complex Engineering Systems at King Abdulaziz City for Science and Technology, under the codirection of Anas Alfaris. The work of Serdar Çolak was supported by a UPS Center for Transportation and Logistics graduate research fellowship, and the work of Marta C. González was supported by the Department of Civil and Environmental Engineering, MIT. The authors thank Alexandre Evsukoff, Pedro Bittencourt, and Pu Wang for technical support and AirSage and the Rio de Janeiro City Hall for support and data.

REFERENCES

1. *The Future of Transportation Studies: A Comparative Review.* AirSage, Atlanta, Ga., 2013. http://airsage.com/Contact-Us/White-Paper.
2. Delcan, Markham, Ontario, Canada. http://delcan.com/markets-and-services/services/category/its-technology-systems-integration.
3. Ward, K. Using Cell Phone Technology to Collect Travel Data. Presented at TRB Planning 355 Applications Conference, Reno, Nev., 2011.
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, South Korea, 2004.

5. Akin, D., and V. P. Sisiopiku. Estimating Origin–Destination Matrices Using Location Information from Cell Phones. *Proc., 49th Annual North American Meetings of the Regional Science Association International,* San Juan, P.R., 2002.
6. Ratti, C., R. M. Pulselli, S. Williams, and D. Frenchman. Mobile Landscapes: Using Location Data from Cell-Phones for Urban Analysis. *Environment and Planning B: Planning and Design,* Vol. 33, No. 5, 2006, pp. 727–748.
7. *Mobile Landscape, Graz in Real Time.* SENSEable City Laboratory, Massachusetts Institute of Technology. http://senseable.mit.edu/graz.
8. Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. Development of Origin–Destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C,* Vol. 40, 2014, pp. 63–74.
9. Wang, P., T. Hunter, A. Bayen, K. Schechtner, and M. C. González. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports,* Vol. 2, No. 1001, 2012.
10. Huntsinger, L. F., and R. Donnelly. Reconciliation of Regional Travel Model and Passive Device Tracking Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
11. Hariharan, R., and K. Toyama. Project Lachesis: Parsing and Modeling Location Histories. In *Geographic Information Science,* Lecture Notes in Computer Science, No. 3234, Springer, New York, 2004, pp. 106–124.
12. Jiang, S., Y. Yang, G. Fiore, J. Ferreira, E. Frazzoli, and M. C. González. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. *Proc., ACM SIGKDD International Workshop on Urban Computing,* 2013.
13. Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science,* Vol. 327, 2010, pp. 1018–1021.
14. Song, C., T. Koren, P. Wang, and A.-L. Barabási. Modelling the Scaling Properties of Human Mobility. *Nature Physics,* Vol. 6, 2010, pp. 818–823.
15. Schneider, C. M., V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling Daily Human Mobility Motifs. *Interface,* Vol. 10, No. 84, 2013.
16. Ben-Akiva, M. E., and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand,* Vol. 9, MIT Press, Cambridge, Mass., 1985.
17. Pukelsheim, F., and B. Simeone. On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis. University of Augsburg, Augsburg, Bavaria, Germany, 2009. http://opus.bibliothek.uniaugsburg.de/opus4/frontdoor/index/index/docId/1229.
18. Census Tract Flows. FHWA, U.S. Department of Transportation. http://www.fhwa.dot.gov/planning/censusissues/ctpp/data products/2006-2010 tractflows/index.cfm.

The Standing Committee on Travel Survey Methods peer-reviewed this paper.