DATA SCIENCE JOURNAL

Demissie, MG, et al. 2019. Understanding Human Mobility Patterns in a Developing Country Using Mobile Phone Data. *Data Science Journal,* 18: 1, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2019-001

RESEARCH PAPER

Understanding Human Mobility Patterns in a Developing Country Using Mobile Phone Data

Merkebe Getachew Demissie[1], Santi Phithakkitnukoon[2], Lina Kattan[1] and Ali Farhan[1]

[1] Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, CA

[2] Department of Computer Engineering, Excellence Center in Infrastructure Technology and Transportation Engineering (ExCITE), Faculty of Engineering, Chiang Mai University, Chiang Mai, TH

Corresponding authors: Merkebe Getachew Demissie (merkebe.demissie@ucalgary.ca), Santi Phithakkitnukoon (santi@eng.cmu.ac.th)

This study demonstrates the use of mobile phone data to derive country-wide mobility patterns. We identified significant locations of users such as home, work, and other based on a combined measure of frequency, duration, time, and day of mobile phone interactions. Consecutive mobile phone records of users are used to identify stay and pass-by locations. A stay location is where users spend a significant amount of their time measured through their mobile phone usage. Trips are constructed for each user between two consecutive stay locations in a day and then categorized by purpose and time of the day. Three measures of entropy are used to further understand the regularity of user's spatiotemporal mobility patterns. The results show that user's in a high entropy cluster has high percentage of non-home based trips (77%), and user's in a low entropy cluster has high percentage of commuting trips (49%), indicating high regularity. A set of doubly constrained trip distribution models is estimated. To measure travel cost, the concept of a centroid point that assumes the origins and destinations of all trips are concentrated at an arbitrary location such as the centroid of a zone is replaced by multiple origins and destinations represented by cell tower locations. Note that a cell tower location can only be used as trips origin/destination location when a stay is detected. The travel cost measured between cell tower locations has resulted in shorter trip distances and the model estimation shows less sensitivity to the distance-decay effect.

Keywords: mobile phone data; origin-destination matrix; trip distribution; human mobility; travel demand; developing country

1. Introduction

Travel demand modeling involves analysis of how much trip is generated, where these trips go, by which mode and on which routes. Except in few occasions, people travel to satisfy needs such as work, leisure, etc. or to perform some activity at a location which is not nearby. In order to understand travel demand, transport planners must understand the spatiotemporal distributions of these activity locations (Ortuzar and Willumsen, 2011). Travel flow estimation at different spatial and temporal scales is a continuing subject across different areas of study. The flows of people from one place to another can be grouped based on their temporal and spatial characteristics. Flows of short distance and duration are presumably commuting trips to work/school/shopping etc. and flows with long distance and duration as internal/global migration (Ortuzar and Willumsen, 2011).

Origin-destination (OD) flow is one of the key information required to provide the basis for accurate travel forecasts by a transport planning model. This information is also vital for policy making and devising travel demand management measures. Traditional methods for collecting trips origin and destination such as surveys are costly, laborious and take the time of trip makers. In recent years, mobile phone data, which includes the passively recorded spatiotemporal trajectories of large portion of the population, have emerged as promising inputs for travel demand model development (Alexander et al., 2015; Demissie, 2014; Demissie et al., 2018).

Trip distribution is one of the main stages of the traditional four-step transportation planning model. It reflects the pattern of trip making behavior in terms of number of trips between trip origins and destinations. Over the years, different types of trip distribution models have been developed. Some of the simplest model such as growth-factor model, is appropriate for short-term studies where no major change in the transportation network is foreseen. However, there are circumstances that cause changes in the transport network cost. One of the most known models suitable for long-term strategic studies is gravity model. This model responds better to changes in the trip pattern when important changes in the transport network take place (Cascetta et al., 2007; Ortuzar and Willumsen, 2011). Previous studies also applied a log-linear model, which is based on statistical estimation of flows as a function of several exploratory variables that includes the characteristics of origin and destination zones and travel cost parameter (Dennett, 2011; Flowerdew and Lovett, 1988).

The main motivation behind our study relates to the measurement problem during trip distribution model development such as those involving travel cost, which in our case is represented by travel distance. Travel cost is usually estimated by the centroid-to-centroid distance between the origin and the destination zones. This obviously is an approximation to the true average trip distance between the two zones. In addition, the centroid-to-centroid distance can lead to a zero-distance separation of the intra-zonal flow, which in reality is always positive (Kordi et al., 2012). This study proposes a method to estimate the average inter-zonal trip length based on multiple origin and destination locations within a zone. The origin and destination locations within a zone are approximated by cell tower locations, where individuals spend significant amount of their time measured through their mobile phone usage. Unlike the centroid-to-centroid trip distance, our method does not assume the origins and the destinations of trips are concentrated at any arbitrary point such as the centroid of a zone. This study improves on previous studies of the same topic and makes the following distinct contributions: (i) measure the spatial and temporal variability of users OD trips; and (ii) develop trip distribution models by applying two different approaches to estimate the travel costs that are measured in terms of trip distance.

The remainder of the paper is organized as follows: Section 2 reviews previous studies in the domain of OD estimation and trip distribution modeling using mobile phone data. Section 3 provides description of datasets and data preparation procedures. Section 4 presents detailed methods that are used to infer the OD trips. Section 5 presents three measures of entropy to understand regularity of user's OD trips. Section 6 presents the results of trip distribution models. The final section outlines the limitations of our research work and the conclusions drawn.

2. Literature review

The use of mobile phone data has been explored for the development of large scale mobility sensing since the early 2000s (Caceres et al., 2008). The data have been used to investigate various aspects of transportation issues: large-scale urban sensing (Calabrese et al., 2011a; Ratti et al., 2005); traffic parameter estimation (Bar-Gera, 2007; Demissie et al., 2013a; Demissie et al., 2013b; Liu et al., 2008); origin-destination trip estimation (Calabrese et al., 2011b; Demissie et al., 2016a; Demissie et al., 2018; White and Wells, 2002); land use inference (Demissie et al., 2015; Toole et al., 2012); travel demand estimation (Alexander et al., 2015; Colak et al., 2015; Demissie et al., 2016b; Demissie et al., 2018; Gundlegård et al., 2016; Phithakkitnukoon et al., 2017).

Previous study show mobile phone data have the advantage of updating OD flow estimates more frequently, which reduces the extensive time required to derive OD flows through traditional methods. This process can also be repeated with new datasets that can be obtained with reduced cost in contrast to data obtained through traditional surveys (Calabrese et al., 2011b). However, the validity of OD flow derived from mobile phone data is questionable, especially, when sampling and penetration rates are not adequate (Calabrese et al., 2011b; Hoteit et al., 2014). Calabrese et al. (2011b) and Csáji et al. (2013) measure the accuracy of the estimated OD flows based on how well the data fits to a gravity model. Demissie et al. (2013c) and Schneider et al. (2013) used household surveys to conform validity of mobile phone trajectories of users. Mobile phone data opens new ways to develop several trip distribution models (Wang et al., 2017; Yan et al., 2014; Yang et al., 2014).

The problem of estimating flows between two zones is a classic problem that appears in a variety of fields such as raw material or goods distribution, flows of capital in economics, or flows of particles in Physics. One of the prior works suggested that the number of trips between two zones follows the gravity law (Ortuzar and Willumsen, 2011). A simplest version of the gravity model for the flow between two traffic analysis zones takes the following functional form in (*Eq.* 1):

$$T_{ij} = \frac{\alpha P_i P_j}{d_{ij}^2} \qquad \text{(Eq. 1)}$$

Where, p_i and p_j are the populations of zone i and zone j, d_{ij} is the distance between zone i and zone j, α is the gravity constant for trip distribution, and T_{ij} is the number of undirected trips between the two zones. The gravity law in *Eq.* 1 was further generalized and presented in the form (*Eq.* 2):

$$T_{ij} = A_i O_i B_j D_j f\left(C_{ij}\right) \quad \text{(Eq. 2)}$$

Where, the single gravity constant for trip distribution factor α is replaced by two sets of balancing factors $A_i = 1/\Sigma_j B_j D_j f(C_{ij})$ and $B_j = 1/\Sigma_i A_i O_i f(C_{ij})$, which ensure that the estimates of T_{ij}, when summed across both rows and columns of the matrix equal the known total trip ends O_i and D_j; $f(C_{ij})$ is a generalized cost function (travel cost).

Batty (1976) approached the intra-zonal trip distance estimation as a problem of finding the mean trip length within a zone. A similar approach can also be taken to measure the weighted-average inter-zonal distance. In this regard, each of the zones are divided into x origin subzones and y destination subzones. Then, the average inter-zonal trip length is calculated from (*Eq.* 3).

$$C_{ij} = \frac{\sum_{x_i}\sum_{y_j} T_{x_i y_j} C_{x_i y_j}}{\sum_{x_i}\sum_{y_j} T_{x_i y_j}},\ x \in i, y \in j \quad \text{(Eq. 3)}$$

Where, C_{ij} is the average inter-zonal trip length between zone i and zone j. $T_{x_i y_j}$ is the number of trips between origin subzone x at zone i and destination subzone y at zone j, and $C_{x_i y_j}$ is the trip distance between subzone x at zone i and subzone y at zone j.

Regardless of the recent advances on the use of mobile phone data for travel demand modeling such as OD flows of different modes (Phithakkitnukoon et al., 2017); OD estimation by purposes and time of day (Alexander et al., 2015; Colak et al., 2015; Demissie et al., 2018; Gundlegård et al., 2016; Toole et al., 2015), there is still work to be done to explore the usage of mobile phone to develop travel demand models. Undoubtedly, we are building on previous studies of the same topic. Our study aims to (i) measure the spatial and temporal variability of users OD trips; and (ii) develop trip distribution models. In this regard, however, the main contribution of this study is the introduction of the cellular network structure to represent the multiple trip origin and destination locations within a zone. We attempt to make a case for replacing the centroid-to-centroid based trip distance by cell tower-to-cell tower based trip distance to measure travel cost. Thus, the average inter-zonal travel cost that adheres to the reality can be measured. Note that cell tower locations can only be used as trips origin/destination locations when a stay is detected, where users spend significant amount of their time, which is measured through their mobile phone usage.

3. Data description and preparation

3.1. Data description

This study uses anonymized Call Detail Records (CDRs) of mobile phone users collected from the entire country of Senegal for the period of two weeks between January 7 to January 20, 2013. In 2013, Senegal had an estimated population of 13,508,715. The country is divided into 14 regions, which are further divided into 45 departments, and 123 arrondissements (districts) (Geohive, 2014). Our analysis is performed at the district level, thus there are 123 traffic analysis zones. In this study, both expressions, district and zone can be used to represent the traffic analysis zone (also inter-district trips, and inter-zonal trips provide the same meaning). The data are made available by SONATEL and Orange within the D4D Senegal Challenge framework. The mobile phone record was obtained with the granularity of cell tower coverage, where each record has anonymized unique user ID, connected cell tower location, and the corresponding timestamp of the call activity (received or made). We analyzed cellular traffic handled by 1,666 cell towers.

3.2. Stay and pass-by area detection

Consecutive mobile phone records of users are used to identify if the user stays in a particular location, engaged in some activity, or passing by the location en-route to his/her destination. Hariharan and Toyama (2004) and Zheng et al. (2009) develop a methodology that has been used to detect stay location from GPS trajectories. Zheng et al. (2009) defined a stay location as 'a geographic region where a user stayed over a certain time interval'. The detection of stay location from GPS trajectories depends on two parameters such as time and distance thresholds. Thus, a stay location would be characterized by the medoid of a group of consecutive GPS records and the user's arrival and leaving times at the stay location.

One of the key differences between the study by Zheng et al. (2009) and ours is the data. Zheng et al. (2009) used data from GPS devices that can record location information every two seconds. In our study, we use CDR where location acquired by this data is at the granularity of a cell tower (cell sector) which gives uncertainty on the exact location of a user; and CDR is sparse in time as it is only acquired when a device is engaged in a voice call or short message service. However, the CDR data used in this study remains useful for our analysis which is at a scale where the lack of a detailed level of precision is acceptable.

Because of the nature of the CDR data in our study, we cannot follow the stay location detection procedure provided by Zheng et al. (2009). Instead, we followed the distance threshold modification suggested by Demissie et al. (2018) and Wang et al. (2017) such as the 200meters distance threshold is replaced by a cell tower location. Note that more than 99% of the distances (Euclidean) between the cell towers are longer than 500 meters, so we did not merge traces from multiple cell towers to estimate a stay location. Thus, each cell tower location is used as the medoid of a group of consecutive traces of the same tower. Then, we use time (10 minutes) and distance (cell tower location) thresholds to detect a stay location. The time threshold is calculated using arrival time (first connection time to cell tower) and leaving time (last connection time to cell tower) at a given cell tower location. A detailed discussion of the stay detection procedure can be found in Demissie et al. (2018) and Wang et al. (2017).

In total, 44.4 million mobile phone connections that are obtained from 319,508 users for a period of two weeks are analyzed. After consecutive traces with time duration of less than 10 minutes are eliminated, the data points are reduced to 7.5 million stay locations (the number of times where the time duration between group of consecutive traces/calls are more than 10 minutes). **Figure 1** shows the frequency of number of stays per user. The average stay per user is 19.11, with first, second, and third quartiles of 9, 14, and 23 over two weeks period.

3.3. Significant locations detection

We identified the most significant locations visited by each user such as users' home, work and "other" locations. Then, the trip made to these locations are connected to activity types of work or home or other. To identify the significant locations, a combined measure of frequency, duration, time and day of mobile phone calls is used. For each user, a home district is identified based on the aforementioned criteria during the night-time (10pm–7am) (Phithakkitnukoon et al., 2012). To identify the work location, calls on Saturday and Sunday are filtered out to avoid the bias from typical non-working days. Then, a work district for each user is estimated based on the aforementioned criteria during night-time (8am–7pm). The remaining locations are labeled as "other".

Figure 2a shows the frequency distribution of stay times at the workplace. Stay time at the workplace is measured based on the differences between the arrival and departure times of the users at their workplace. The measurement is done for the period of two weeks for the users with inferred home and workplace locations. The average time people spent at their workplace is 5 hours and 38 minutes.

Figure 2b shows the distribution of the number of hour(s) people spend away from their workplace. The time spent away from the workplace is considered as the gap time between departure time from the workplace on a given day and arrival time at the workplace in the next working day. The average time spent

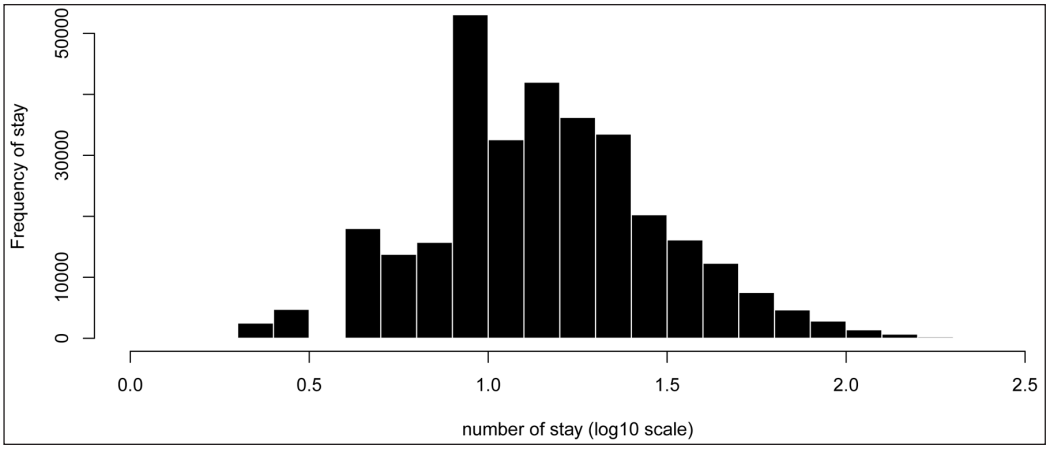

Figure 1: Frequency of stays.

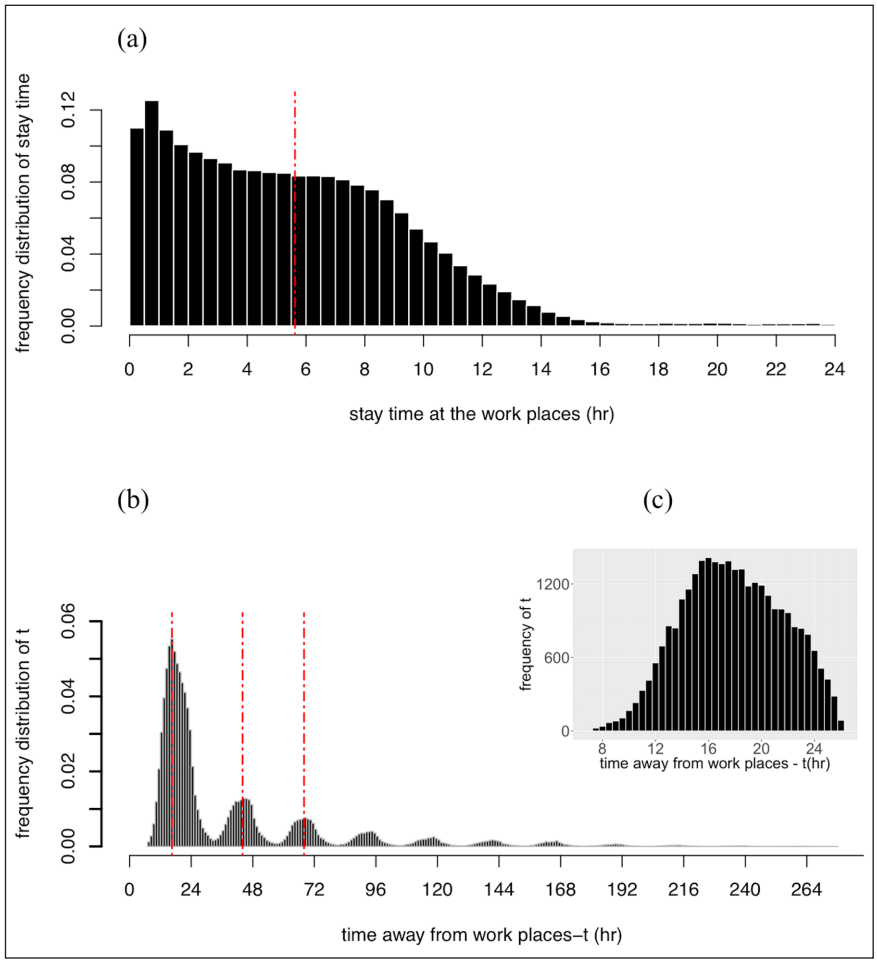

Figure 2: (a) Stay time at the workplace; **(b)** Time away from the workplace; and **(c)** Time away from the workplace between the first two working days.

away from the workplace is 39 hours and 30 minutes, with a first, second, and third quartiles of 16 hours and 42 minutes, 22 hours, and 46 hours and 42 minutes, respectively. It can be seen that 63.4 percent of the observations are concentrated in the first peak (people arriving at workplace after being way for 32 hours or less). The average time of the first peak is 16 hours, which essentially makes sense for the majority of people who work for 8 hours per day and away for the remaining 16 hours. However, only 11% of these data are observed on Monday and the rest of the observations are equally distributed on the remaining working days (Tuesday to Friday). The second peak contains 15.8% of the total observations (people arriving at their workplace after being away for 32 to 57 hours), and 31% of the data is observed on Monday. The third peak contains 10% of the total observation (people arriving at their workplace after being away for 58 to 81 hours). Interestingly, 46% of the data on the third peak is recorded on Monday, with an average time of 68 hours. The third peak may reflect on the people who are away from their workplaces since Friday evening because of the weekends.

The analysis presented in **Figure 2b** shows that the average time spent away from the workplace is 39 h and 30 min. That seems considerably high and in contrast with other findings with lower values (BLS, 2017). One of the reasons could be missing data, where some users might not use mobile phone during their trip and as a result the time spent away from workplace could not be measured between successive working days (this can happen because of lack of either the departure time or the arrival time records at the workplace, which consequently results large time interval). **Figure 2c** shows frequency of the number of hour(s) people spend away from their workplace between the first two working days, where there would not be missing working day. The average time spent away from the workplace is close to the reality, which is 17 hours and 51minutes, with first, second, and third quartiles of 15 hours, 17 hours and 40 minutes, and 20 hours and 40 minutes, respectively. However, in all the cases, the timestamps associated with the departure and arrival times at the workplace location are based on the mobile phone usage rather than the actual arrival and departure times of the user. Thus, the measured times are only approximation of the actual times.

Work schedule such as stay time at the workplace and time spent away from the workplace only provide limited information for transportation planning and operation. The timing of when these working hours occur, what time employee arrive and depart from their workplace provide more insightful information, but mobile phone data have limitation in terms of providing actual arrival/departure times.

4. Identifying trip-makers' origins and destinations

For each user, the consecutive traces associated with the stay are arranged along the date and time of the day. A trip can be identified if the trip-maker has more than one stay location within 24 hours (one-day) period where midnight is taken as the transition time from one day to the next. It is assumed that a trip is made between two consecutive stay locations. Thus, the first interaction time at location i and the last interaction time at location i + 1 should be within a period of one day.

OD trips are categorized by time of the day (24 hours) and purpose. Home-based work trips (HBW) are trips between a person's home and workplace. Home-based other trips (HBO) are trips between a person's home and other destinations which are not for the purpose of working. A non-home based trip (NHB) is a trip that neither begins or ends at a person's home regardless of the purpose of the trip. The relative share of average weekday trips for HBW is 19.5%, HBO 13.6%, and NHB 66.9%. Alexander et al., (2015) also found a similar proportion of HBW trips. However, the proportion of HBO and NHB trips differ significantly. A previous study in the same region also suggested that most of the inter-departmental trips in Senegal are irregular trips instead of commuting trips (Wang et al., 2017). At this point our research does not point a reason for the high percentage of NHB trips and we suggest further investigation in future work.

Figure 3a shows country-wide daily average hourly OD flows. **Figure 3b** shows spatial distribution of inter-district (inter-zonal) OD trips derived from sample users for the period of two weeks. Most of the inter-district OD flows are made between districts in the region of Dakar, Thies, Saint-Louis, and Diourbel, where these regions have economic, social, and political significance.

5. Measuring regularity of trips made by users

A number of studies have explored the use of mobile phone data for OD estimation. One of the cornerstones of these studies is the precise inference of activity locations (Ahas et al., 2010; Gonzalez et al., 2008). Besides home and work locations, people spend a significant amount of their time in other locations (Ahas et al., 2010). González et al., (2008) mined the trajectory of phone users for six months and showed that people spend their time at few locations and most travelers show a high degree of temporal and spatial regularity. Previous studies by Song et al., (2010) and Qin et al., (2012) introduced the concept of entropy to measure predictability of human mobility. This study also apply entropy with the focus of further understanding the spatial and temporal variability of users OD trips.

Three entropy measures are calculated for each user's mobility pattern: (i) Entropy 1: $H1_x = \log_2 N_x$, where N_x is the number of unique locations visited by the user x. $H1_x$ is used to understand the degree of irregularity of a user assuming each visited location has equal probability; (ii) Entropy 2: $H2_x = -\Sigma_{y=1}^{N_x} p_x(y) \log_2 p_x(y)$, where $p_x(y)$ is the probability of visiting location y by the user x. The probability depends on the frequency of previous visits. $H1_x$ and $H2_x$ are solely based on user's spatial pattern, which do not capture the time of location visitation. To incorporate the time parameter and the sense of OD trips (visited locations pair), a joint entropy of user's mobility is introduced; (iii) Entropy 3: $H3_x = -\Sigma_o \Sigma_d \Sigma_t p(o, d, t) \log_2 p(o, d, t)$, where o

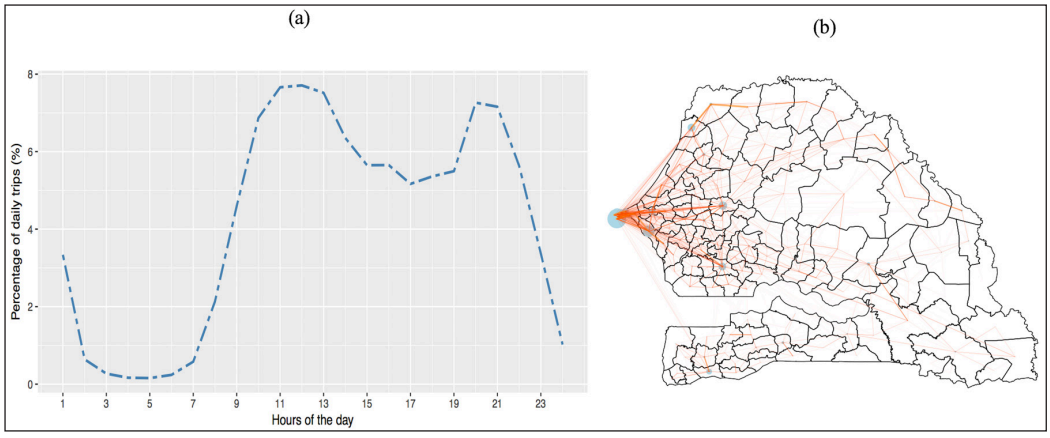

Figure 3: (a) Daily average hourly OD flows **(b)** Inter-district OD flows.

and d are spatial parameters representing the origin and destination of a trip, respectively. The origin and destination of a trip can be home (H), work (W), or other (O). The temporal parameter is represented by t, where t_1 (9am to 2pm), t_2 (2pm to 7pm), t_3 (7pm to 10pm), and t_4 (10pm to 9am). p (o, d, t) is the joint probability of o, d, and t. At this level, trips made from/to locations other than home and work locations are not explicitly modeled and classified under 'other' location. If "other" is considered as one location, a user can make 28 trip types: {(H, W, t_1), (H, O, t_1), (W, H, t_1), (W, O, t_1), (O, H, t_1), (O, W, t_1), (O, O, t_1), ..., (O, H, t_4), (O, W, t_4), (O, O, t_4)}. However, based on the two weeks data, by average a user visited 4.85 distinct locations. Thus, there can be more than 28 trip combinations.

To calculate the entropy values, 243,928 users with inferred home and workplace locations are selected. **Figure 4(a–c)** shows the distribution of the three entropy measures. The value of $H1_x$ and $H2_x$ measure user's regularity in terms of visited locations. In fact, $pH1_x$ peaks at $H1_x = 1.86$, suggesting that on average it took around four locations to identify user's randomly chosen next location ($2^{1.86} = 3.63$). The value of $H3_x$ ranges between 0 and 12.7. The users with an entropy value of 0 are regarded as highly regular. These users travel between the same origins and destinations and within the same time interval, daily. On the other hand, the users whose entropy are high, travel between different origins and destinations at different time intervals.

To further understand user's travel pattern, k-means clustering method is used to categorize the users into groups based on their entropy value ($H3_x$). A plot of the within groups sum of squares by the number of clusters is used to determine the appropriate number of clusters. The resulting number of clusters is three. **Figure 4d** shows the proportion of OD trips in each cluster. The entropy values ($H3_x$) range between 5.5 and 12.7; 3.2 and 5.5; 0 and 3.2 in the high entropy cluster, in the moderate entropy cluster, and in the low entropy cluster, respectively. The result shows 35% of the users are in the high entropy cluster with 77% of their trips are non-home based. Users in the low entropy cluster account for 21% and 49% of their trips are commuting, indicating a high regularity. The rest of the users are in the moderate entropy cluster and account for 44% of all.

6. Trip distribution modeling

The modeling framework set out in this study focuses on inter-zonal trips between origin zone *i* and destination zone *j*. Note that trips within the same zone (intra-zonal trips) are not considered in the model development. Log-linear models have been applied to the analysis of values contained within contingency tables. Previous studies by Dennett (2012) and Demissie et al. (2018) showed how the doubly constrained, multiplicative model in Eq. 2 can be equivalent with statistical (additive) log-linear model. Using a similar notation to Dennett (2012), the additive version of a log-linear model can be expressed as (*Eq.* 4):

$$T_{ij} = \tau\tau_i^O\tau_j^D\tau_{ij}^{OD} \qquad \text{(Eq. 4)}$$

where τ is the overall component representing the level of inter-zonal flows, τ_i^O and τ_j^D are the raw and column 'main effects' represented by categorical variables with i and j are both 123 levels, τ_{ij}^{OD} is the interaction component representing the physical or social separation between the origin and destination zones with i = 1, ..., n * j = 1, ..., n parameters, where, n = 123 (i.e., 15,006 in our case, where the 123 intra-zonal

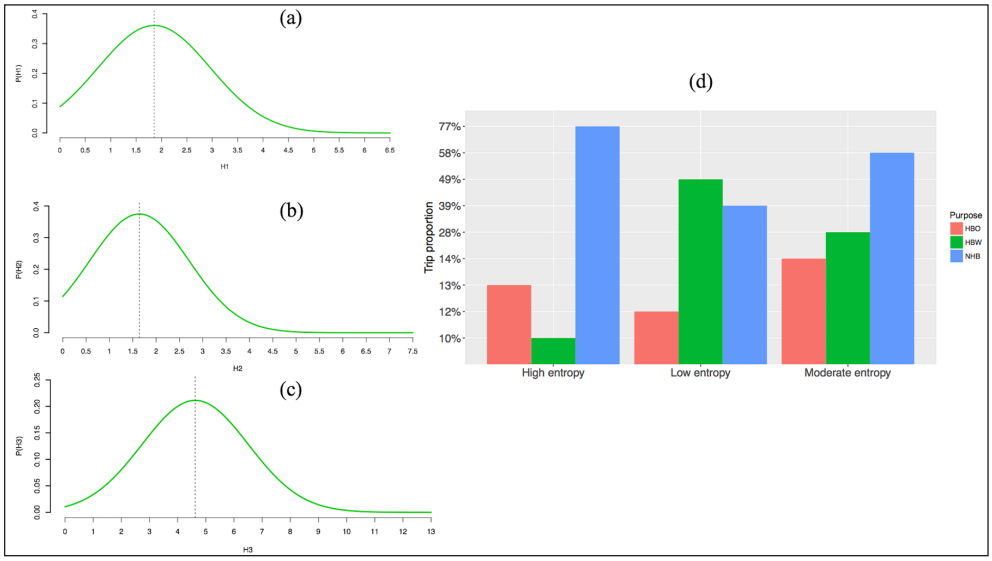

Figure 4: The distribution of the entropy **(a)** H1, **(b)** H2, **(c)** H3, and **(d)** Number of OD trips in each cluster.

interactions are not considered). By taking the natural logarithm, the multiplicative component model in *Eq.* 4 can be expressed as a log-linear (additive) model as follows (*Eq.* 5):

$$\ln\left(T_{ij}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_{ij}^{OD} \quad \text{(Eq. 5)}$$

6.1. Trip expansion

Once the sample OD flows are detected (Section 4); the next step is to expand them in order to represent the mobility behavior of the total population. There are no available model outputs or comprehensive travel survey data (land use, number of employees, floor area, socioeconomic characteristics, etc.) in Senegal that can be used to develop trip generation models. A simplified method is developed to produce the following two information: (i) total daily person trips originating from each district (T_{Oi}); and (ii) total daily person trips destined to each district (T_{Dj}). To obtain the total daily person trips originating in each district, first, average number of daily trips made by each sampled user is calculated (t_{user}). Then, the total daily trips generated by sampled users in each district (t_{total_i}) is obtained by summing t_{user} of sampled residents of each district, *i*. The number of sampled residents are based on the number of "Homes" detected in each district, *i*. Then, an expansion factor (f_i) is developed for each district as the ratio between the total number of population of each district (age group ≥ 5 years) and the number of sampled users identified as residents of that district based on the CDR data. Finally, the total daily person trips originating from each district is obtained by multiplying t_{total_i} and f_i. To obtain T_{oi}, we use census information near the timespan of the CDR data as a secondary source of data describing the population which can be related to the sample. There is no accurate secondary source of data describing the total attractions in each district. We assume the model for total daily person trips originating from each district is reasonable. Thus, we control the number of total daily person trips so that the number of person trip origins equals the number of person trip destinations.

6.2. Trip distance

Two approaches are used to estimate the average inter-district trip distances. Approach 1: the average inter-district trip distance is based on the Euclidian distance measured between the centroids of the origin and destination districts. Approach 2: Eq. 3 is used to measure the average inter-district trip distance. **Figure 5a** shows the daily average inter-district trip distances measured based on approach 1 and approach 2, respectively. The daily average inter-district trip distance measured based on approach 1 assumes trips between two zones have the same origin and destination points (the centroid), which is unrealistic because the origin and destination of trips are usually scattered within a zone and influenced by time of day, trip purpose, and other district attributes (Bharat and Larsen, 2011). The daily average inter-district trip distance measured based on approach 2 is relatively shorter.

6.3. Results of trip distribution models

Two doubly constrained log-linear models are estimated using the average daily inter-district OD flows derived from sample users and expanded to the general population based on census data of the region (in Section 6.1): (i) Model 1 – trip distance is obtained based on Approach 1; and (ii) Model 2 – trip distance is

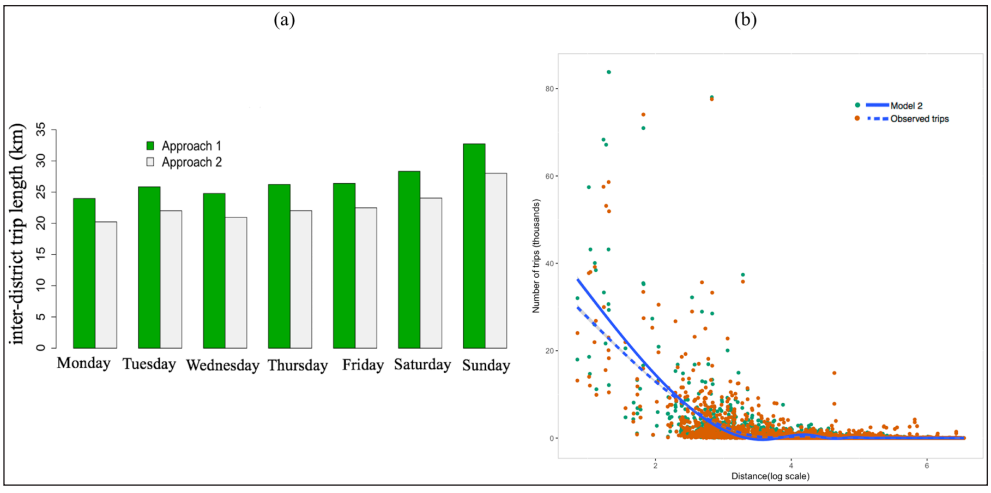

Figure 5: (a) Average inter-district trip distances **(b)** Observed vs. estimated inter-district OD trips (Model 2).

obtained based on Approach 2. The inverse power (C_{ij}^{-n}) cost function is used as it was the one that provided a better fit (n is parameter for cost function).

We use the glm () function implemented in in R to fit the log-linear model in Eq. 5 to the data (R Core Team, 2013). The use of 123-district level zoning system resulted a set of 15,006 inter-district flows. The models also use 123 categorical variables related to the origins and another set of 123 categorical variables related to the destinations (one of the variables from each group is used as a reference category). A trip distance variable is also used to represent the travel cost. The estimation made by generalized linear model (GLM) for this model and the upcoming model are on the basis that the Poisson assumption is satisfied. Thus, the predicted value of OD trips is assumed to follow a Poisson distribution with a mean that is logarithmically linked to a linear combination of the origin and destination categorical variables and the distance variable.

Eq. 6 shows the estimation of Model 1, where $\widehat{T_{ij}}$ is the estimated flow between the districts i and j. Analysis of the origin-specific (orig) and destination-specific (dest) variables was carried out to understand their relative importance against the reference category $orig_1$ and $dest_1$, respectively. The coefficients of the origin categorical variables vary from 4.53 (Mbane), which is one of the districts that generates a high number of daily trips to –1.54 (Rufisque), which is one of the districts that generates a low number of daily trips. The coefficients of the destination categorical variables go from 0.77 (Dakar Plateau) to –4.11 (Fongolimbi), which is one of the districts that attracts few daily trips. The coefficient of distance variable is negative and significant.

$$\widehat{T_{ij}} = \exp(13.58 + 1.43orig_2 + 0.31orig_3 + 1.15orig_4 + \ldots + 4.12orig_{123} + 0.24dest_2 - 0.09dest_3 + 0.77dest_4 + \ldots - 1.37dest_{123} - 2.32lnC_{ij}) \quad \text{(Eq. 6)}$$

Eq. 7 shows the estimation of Model 2.

$$\widehat{T_{ij}} = \exp(12.41 + 1.23orig_2 + 0.38orig_3 + 1.50orig_4 + \ldots + 2.71orig_{123} + 0.08dest_2 - 0.06dest_3 + 1.14dest_4 + \ldots - 1.76dest_{123} - 1.98lnC_{ij}) \quad \text{(Eq. 7)}$$

Figure 5b shows comparison of the observed and estimated inter-district trips of Model 2. A more conventional R-Squared (R^2) value is also used to compare the observed trips against the models outputs and we found 0.78, and 0.91 for Model 1, and Model 2, respectively.

6.4. Model summary

The trip distance (or travel cost) (lnC_{ij}) parameters of the two doubly constrained log-linear models are shown in *Eq.* 6 and *Eq.* 7. The similar negative term corresponding to the trip distance parameters in both models resembles the decrement in the level of origin/destination interaction as the trip distance increases. However, there is a difference in the value of the trip distance parameters. The average inter-district trip distance parameter for Model 1 is –2.322. On the other hand, the parameter for the average inter-district trip distance for Model 2 is larger (–1.983). In the case of Model 1, the average inter-district trip distance is obtained based on centroid-to-centroid distance, which is a function of the geometry of the origin and destination districts. Thus, in spite of the fundamental attributes that influence trips, inter-district trips are only influenced by the shape of the origin and destination zones. As a result, the average inter-district trip distance is large and Model 1 is highly sensitive to the distance-decay effect, so people disutility of distance may be overestimated. In the case of Model 2, the length of inter-district trip is obtained based on *Eq.* 3. In this case, the concept of a centroid point is replaced by multiple trip origin and destination points within a zone. These multiple origin and destination points are locations, where individuals spend significant amount of their time (measured through their mobile phone usage). Unlike the centroid-to-centroid trip distance, this method reflects more of the reality as it does not assume the origins and destinations of all trips are concentrated at any arbitrary location such as the centroid. Because of its larger trip distance parameter, Model 2 generates more inter-district trips when compared with the Model 1. This indicates that trip making is influenced by trip distance and people may prefer making shorter trips.

6.5. Orientation ratio

Additional analysis is done to check the reasonableness of the estimated OD flows using Orientation Ratio (OR). This ratio is a simple indicator to show the tendency of trips moving from a given production area to the attraction area (FHWA, 2010). It can be calculated by $OR_{ij} = (T_{ij}/D_j)/(O_i/\Sigma_{i=1}^{123}\Sigma_{j=1}^{123}T_{ij})$, where, OR_{ij} is orientation

ratio between trip production district i and trip attraction district j, T_{ij} is number of trips from i to j, D_j destination total of district j, O_i is origin total of district i and there are 123 districts in the study area. **Figure 6a** and **Figure 6b** show an example of estimated and observed orientation ratios to understand the propensity of trips from all districts towards the capital city of Senegal, Dakar (marked by the circle), which has a total of 10 districts. The analysis is done based on estimated OD flows of Model 2. The figure demonstrates high matching patterns between the observed and estimated orientation ratios in most parts of the country.

7. Conclusions

In the developing countries, it is a challenging task to obtain mobility data because of the limited budget available to conduct large scale mobility surveys. In this study, we used CDR data obtained through the D4D Senegal challenge to analyze country-wide mobility patterns of people. Our main focus of the present study was exploring the potential of CDR data to detect the origin and destination flows and measure the spatial and temporal variability of user's OD trips. Then, we developed two trip distribution models. In this regard, we attempted to make a case for replacing the centroid-to-centroid based trip distance by cell tower-to-cell tower based trip distance to measure travel cost. Thus, the average inter-zonal travel cost that adheres to the reality can be measured.

The extracted trips are categorized by purpose and time of the day. The results show the NHB trips are considerably high and in contrast with previous findings from another region of the world (Alexander et al., 2015). At this point our research does not point a reason for the high percentage of NHB trips and we suggest further investigation in future work. Three entropy measures are applied with the focus of understanding the spatial and temporal variability of individuals OD trips. The result shows that user's in the high entropy cluster has a high percentage of non-home based trips (77%), and user's in the low entropy cluster has a high percentage of commuting trips (49%), indicating high regularity.

Two approaches were presented to estimate the inter-zonal trip distance by analysing Senegal's CDR data at 123-district level traffic analysis zone system. The centroid-to-centroid distance produced relatively larger average trip distance. As a result, the estimated model is highly sensitive to the distance-decay effect. In the second approach, the origin and destination locations within a zone are approximated by the locations of cell towers, where users spend a significant amount of their time (measured through their mobile phone usage). This has resulted in a shorter trip distance and the model shows less sensitivity to the distance-decay effect. We also detected stay time at the workplace and time spent away from the workplace, which only provided part of the information required for the transportation planning and operation. The timing of when these working hours occur, what time employee arrive and depart from their workplace provide more insightful information, but mobile phone data have limitation in terms of providing actual arrival/departure times.

One of the aims of our study is to provide transport planners in the developing countries with an option that can be considered in the absence of detailed transport data for transport planning. The results can also be used to support decisions regarding the inter-district public transport planning as well as major national and regional road network development projects. In our analysis, residence of sample users is used to obtain the expansion factor required to expand the results to a general population. However, future studies should incorporate detailed profile such as socio-economic and demographic of sample users to properly represent the composition of sample data.

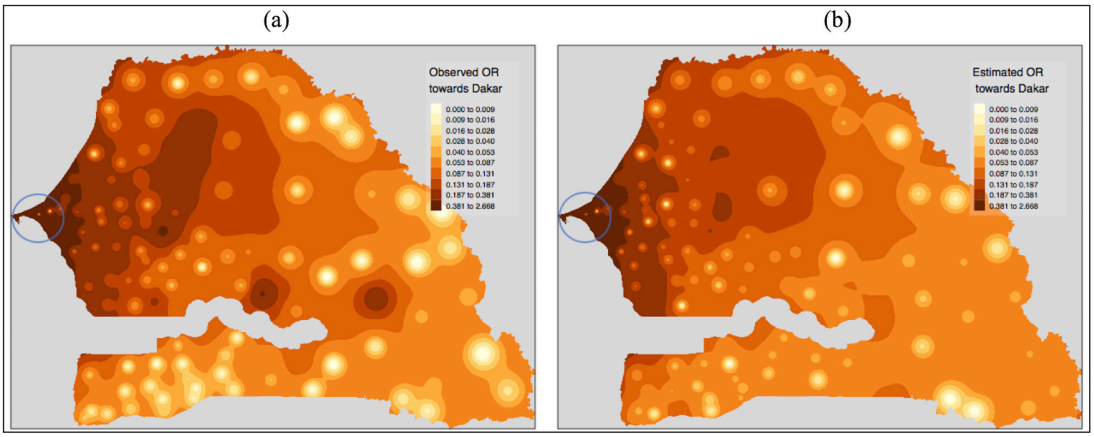

Figure 6: (a) Orientation ratios (observed trips) **(b)** Orientation ratios (estimated trips).

The presented results are not validated against ground truth because of lack of mobility data, where Senegal does not have traffic counting in a regular basis or has no priori travel demand data from previous surveys that tend to be costly, labour intensive and time disruptive to the trip makers. Future studies should also apply CDR data from a longer period in order to capture seasonality of travel demand and improve the representativeness of the extracted movements and flows. In our study, the choice of 10 min threshold value to categorize traces into stay or pass-by is arbitrary. In reality, the threshold value should be area specific. Future studies should consider cell tower service area coverage, traffic and pedestrian congestions to determine area specific threshold values.

Data Accessibility Statement

The dataset used in this study was obtained through the framework of Data for Development (D4D) Senegal Challenge. The researchers are not allowed to share the data directly. However, the D4D Senegal Challenge data is available on request to other researchers for academic, non-commercial purposes by D4D organizers.

Acknowledgements

This work was supported by the Eyes High Postdoctoral Fellowship at the University of Calgary. This research work was partially supported by Chiang Mai University. The authors would like to thank the Data for Development Senegal Challenge for providing the mobile phone data.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

MGD designed the study and processed the data. All authors analyzed the results and wrote the manuscript. All authors have read and approved the final manuscript.

References

Ahas, R, Silm, S, Järv, O, Saluveer, E and **Tiru, M.** 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1): 3–27. DOI: https://doi.org/10.1080/10630731003597306

Alexander, L, Jiang, M, Murga, M and **Gonzalez, M.** 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58: 240–250. DOI: https://doi.org/10.1016/j.trc.2015.02.018

Bar-Gera, H. 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies*, 15(6): 380–391. DOI: https://doi.org/10.1016/j.trc.2007.06.003

Batty, M. 1976. Urban modelling: algorithms, calibrations, predictions. London: Syndics of Cambridge University Press.

Bharat, PB and **Larsen, O.** 2011. Are intrazonal trips ignorable? *Transport Policy*, 18(1): 13–22. DOI: https://doi.org/10.1016/j.tranpol.2010.04.004

BLS – Bureau of labor statistics. 2017. American time use survey – 2017 results. U.S. Department of Labour. USDL-18-1058. Available at: https://www.bls.gov/news.release/pdf/atus.pdf [Accessed 21.11.2018].

Caceres, N, Wideberg, J and **Benitez, F.** 2008. Review of Traffic Data Estimations Extracted from Cellular Networks. *IET Intelligent Transport Systems*, 2(3): 179–192. DOI: https://doi.org/10.1049/iet-its:20080003

Calabrese, F, Colonna, M, Lovisolo, P, Parata, D and **Ratti, C.** 2011a. Real-time urban monitoring using cellphones: A case study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1): 141–151. DOI: https://doi.org/10.1109/TITS.2010.2074196

Calabrese, F, Lorenzo, G, Liu, L and **Ratti, C.** 2011b. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE*, 10(4): 36–44. DOI: https://doi.org/10.1109/MPRV.2011.41

Cascetta, E, Pagliara, F and **Papola, A.** 2007. Alternative approaches to trip distribution modelling: a retrospective review and suggestions for combining different approaches. *Regional Science*, 86(4): 597–620. DOI: https://doi.org/10.1111/j.1435-5957.2007.00135.x

Çolak, L, Alexander, B, Alvim, S, Mehndiretta, M and **Gonzalez, M.** 2015. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. *Transportation Research Board Annual meeting.* Washington, D.C.

Csáji, B, Browet, A, Traag, V, Delvenne, J, Huens, E, Dooren, P, Smoreda, Z and **Blondel, V.** 2013. Exploring the mobility of mobile phone users. *Physica A Statistical Mechanics and its Applications*, 392(6): 1459–1473. DOI: https://doi.org/10.1016/j.physa.2012.11.040

Demissie, MD. 2014. Combining datasets from multiple sources for urban and transportation planning: Emphasis on cellular network data, Ph.D. dissertation, Dept. Civil Eng., Coimbra Univ., Coimbra, Portugal.

Demissie, MG, Correia, G and **Bento, C.** 2013a. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. *Transp. Res. C, Emerging Technol*, 32: 76–78. DOI: https://doi.org/10.1016/j.trc.2013.03.010

Demissie, MG, Correia, GH and **Bento, C.** 2013b. Traffic volume estimation through cellular networks handover information. *13th World Conference on Transportation Research*. Rio de Janeiro, Brazil.

Demissie, MG, Correia, GH and **Bento, C.** 2013c. Exploring cellular network handover information for urban mobility analysis. *Journal of Transport Geography*, 31: 164–170. DOI: https://doi.org/10.1016/j.jtrangeo.2013.06.016

Demissie, MG, Correia, GH and **Bento, C.** 2015. Analysis of the pattern and intensity of urban activities through aggregate cellphone usage. *Transportmetrica A: Transport Science*, 11(6): 502–524. DOI: https://doi.org/10.1080/23249935.2015.1019591

Demissie, MG, Phithakkitnukoon, S and **Kattan, L.** 2018. Trip Distribution Modeling Using Mobile Phone Data: Emphasis on Intra-Zonal Trips. *IEEE Trans. Intell. Transp. Syst*. Oct. 2018. DOI: https://doi.org/10.1109/TITS.2018.2868468

Demissie, MG, Phithakkitnukoon, S, Sukhvibul, T, Antunes, F and **Bento, C.** 2016a. Inferring origin-destination flows using mobile phone data: A case study of Senegal. *International conference on electrical engineering/electronics, computer, telecommunications and information technology*. Chiang Mai, Thailand. DOI: https://doi.org/10.1109/ECTICon.2016.7561328

Demissie, MG, Phithakkitnukoon, S, Sukhvibul, T, Antunes, F, Gomes, R and **Bento, C.** 2016b. Inferring passenger travel demand to improve urban mobility in developing countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems*, 17(9): 2466–2478. DOI: https://doi.org/10.1109/TITS.2016.2521830

Dennett. 2012. Estimating flows between geographical locations: "Get me started in" spatial interaction modelling. Centre for Advanced Spatial Analysis, University College, London.

FHWA. 2010. Travel model validation and reasonableness checking manual second edition. Cambridge Systematics, Inc.

Flowerdew, R and **Lovett, A.** 1988. Fitting constrained poisson regression models to interurban migration flows. *Geographical Analysis*, 20(4). DOI: https://doi.org/10.1111/j.1538-4632.1988.tb00184.x

Geohive. 2014. http://www.geohive.com/cntry/senegal%20ext.aspx. Accessed Jun. 01, 2017.

González, M, Hidalgo, C and **Barabási, A.** 2008. Understanding individual human mobility patterns. *Nature*, 453: 779–782. DOI: https://doi.org/10.1038/nature06958

Gundlegård, D, Rydergren, C, Breyer, N and **Rajna, B.** 2016. Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95: 29–42. DOI: https://doi.org/10.1016/j.comcom.2016.04.015

Hariharan, R and **Toyama, K.** 2004. Project lachesis: parsing and modeling location histories. *Geogr. Inform. Sci*, 106–124.

Hoteit, S, Secci, S, Sobolevsky, C, Ratti, C and **Pujolle, G.** 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64: 296–307. DOI: https://doi.org/10.1016/j.comnet.2014.02.011

Kordi, M, Kaiser, C and **Fotheringham, A.** 2012. A possible solution for the centroid-to-centroid and intra-zonal trip length problems. *AGILE'2012 International Conference on Geographic Information Science. Avignon.*

Liu, HX, Danczyk, A, Brewer, R and **Starr, R.** 2008. Evaluation of cellphone traffic data in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*, 2086: 1–7. DOI: https://doi.org/10.3141/2086-01

Ortuzar, JD and **Willumsen, LG.** 2011. Modelling transport. Wiley. DOI: https://doi.org/10.1002/9781119993308

Phithakkitnukoon, S, Smoreda, Z and **Olivier, P.** 2012. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. *PLoS ONE*, 7(6). DOI: https://doi.org/10.1371/journal.pone.0039253

Phithakkitnukoon, S, Sukhvibul, T, Demissie, M, Smoreda, Z, Natwichai, J and **Bento, C.** 2017. Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science*, 6(11). DOI: https://doi.org/10.1140/epjds/s13688-017-0108-6

Qin, S, Verkasalo, H, Mohtaschemi, M, Hartonen, T and **Alava, M.** 2012. Patterns, Entropy, and Predictability of Human Mobility and Life. *PloS one*, 7(12). DOI: https://doi.org/10.1371/journal.pone.0051353

Ratti, C, Sevtsuk, A, Huang, S and **Pailer, R.** 2005. Mobile landscapes: Graz in real time. *Symposium on LBS & TeleCartography*. November 28–30, 2005.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Schneider, C, Belik, V, Couronné, T, Smoreda, Z and **González, M.** 2013. Unraveling daily human mobility motifs. *J. Roy. Soc. Interface*, 10.

Song, C, Qu, Z, Blumm, N and **Barabási, A.** 2010. Limits of Predictability in Human Mobility. *Science*, 327(5968): 1018–1021. DOI: https://doi.org/10.1126/science.1177170

Toole, J, Colak, S, Sturt, B, Alexander, L, Evsukoff, A and **Gonzalez, M.** 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58: 161–428. DOI: https://doi.org/10.1016/j.trc.2015.04.022

Toole, J, Ulm, M, González, M and **Bauer, D.** 2012. Inferring land use from mobile phone activity. *ACM SIGKDD international workshop on urban computing*. Beijing, China. DOI: https://doi.org/10.1145/2346496.2346498

Wang, Y, Correia, G, Romph, E and **Santos, BF.** 2017. Road network design in a developing country using mobile phone data: An application to Senegal. *IEEE Intelligent Transportation Systems Magazine*. DOI: https://doi.org/10.1109/MITS.2018.2879168

White, J and **Wells, I.** 2002. Extracting origin destination information from mobile phone data. *International Conference on Road Transportation and Control*, 30–34. London. DOI: https://doi.org/10.1049/cp:20020200

Yan, X, Zhao, C, Fan, Y, Di, Z and **Wang, W.** 2014. Universal predictability of mobility pattern in cities. *J R Soc Interface*, 11(100). DOI: https://doi.org/10.1098/rsif.2014.0834

Yang, Y, Herrera, C, Eagle, N and **Gonzalez, M.** 2014. Limits of Predictability in Commuting Flows in the Absence of Data for Calibration. *Scientific reports*, 4(5662).

Zheng, Y, Zhang, L, Xie, X and **Ma, W.** 2009. Mining interesting locations and travel sequences from GPS trajectories. *Proceedings of the 18th international conference on World wide web. ACM*. DOI: https://doi.org/10.1145/1526709.1526816

How to cite this article: Demissie, MG, Phithakkitnukoon, S, Kattan, L and Farhan, A. 2019. Understanding Human Mobility Patterns in a Developing Country Using Mobile Phone Data. *Data Science Journal,* 18: 1, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2019-001

Submitted: 31 August 2018 **Accepted:** 30 November 2018 **Published:** 03 January 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS