Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Understanding Travel Behavior from Call Detail Records

N.K.Bhagya Jeewanthi[a], Amal.S.Kumarage[a], Sirganesh Lokanathan[b]

[a]*Department of Transport and Logistics Management, University of Moratuwa, Katubadda, Moratuwa,10400, Sri Lanka*
[b]*LIRNEasia,Colombo,10400,Sri Lanka*

Abstract

The study presents a methodology to identify human mobility behavior from call detail records (CDR) of nearly 1000 anonymized users for a period of three months. The locations within Sri Lanka from which CDRs have been generated during this period by each such user were identified and classified by province. In analyzing the CDRs, made within the Western Province, the frequency of making calls from a specific location of a connected tower was considered in identifying the possible nature of the activity that the user could have been engaged in at that location. As in basic transportation planning theory, they were first classified into the potential home and non-home locations based on the time of the day these calls were generated from each such location. Thus home and work locations have been identified together random locations, considered as potential locations for shopping trips, recreational trips, and trips for visiting friends and relations. The applicability of the proposed methodology is supported by a validation against the extensive transport survey for 35,000 households in the Western Province of Sri Lanka under the CoMTrans Study

Keywords: Mobile Network Big Data; Call Detail Record; Load Balancing

1. Introduction

Human mobility patterns are critical sources of information for designing, analyzing, and enhancing transportation planning activities. If the mobility predictions can be done accurately, ongoing problems like traffic congestion, which is characterized by lower speed, longer trip times and increased vehicle queuing can be reduced by a considerable level, Farkavcova et al. (2010). Currently, roadside and household surveys are used to gather data to support transportation planning studies. As these are expensive, they have limited sample sizes and lower update frequencies and are prone to sampling biases and reporting errors, Saini et al. (2015).

Big Data can play an important role in understanding human behavioral processes. Mobile Network Big Data (MNBD) is considered as one of the most promising ICT (Information and Communication Technologies) data sources for measuring the mobility of people and having the potential to supplement traditional mobility prediction techniques much more economically and accurately.

Mobile Network Big Data (MNBD) is an ever-growing volume of data sets derived from the way people use mobile communication devices. Compared to other network related data like GPS, Call Detail Records (CDR) is often one of the largest subsets of MNBD since most telecommunication service providers automatically capture such data for billing purposes, Dash et al. (2015). Whenever a cellular transaction is made, a CDR which consists of time-stamped tower locations with caller IDs are, Iqbal et al. (2014). A CDR describes the mobile usage pattern of a particular user from which it is possible to extract information on the mobility of the user. Thus, analyzing the temporal and spatial dimensions of the CDRs of the mobile users can provide a better understanding of human mobility.

2. Literature Review

Studies had been focusing on different mobility aspects within their research areas. Trip extraction using CDR have ranged from O-D matrix estimation, Alexander et al. (2015), Saini et al. (2015), identifying meaningful places, Ahas et al. (2010), to the characterization of human mobility, Zhang (2014).

O-D matrices have been developed with different techniques. As an example in a study that was done in Mumbai, different locations visited by the users were identified and mobility traces were generated for each given user. Then the records were aggregated for all users, multiplied by a scaling factor and converted to vehicle trips to arrive at the initial OD), Saini et al. (2015). A study in Dhaka focused on generating tower to tower transient OD matrices using the trips occurring within different time periods. Corresponding nodes were converted to node-to-node transient O-D matrices of a traffic network, Iqbal et al. (2014). Another study conducted in Estonia used the number of days and the number of calls as the primary parameters to identify home and work locations, Ahas et al. (2010).

Other than these, most of the studies have focused on identifying major home and work anchor points, Saini et al. (2015), Kevin et al. (2014). As CDRs contain the time stamped-locations of users, analyzing CDRs can lead to the identifications of locations visited by users separately on weekday and weekends. Studies have used these weekday and weekend moments within given time windows have also been used as the basis for location identification. The most widely-used method for inferring home and workplaces assumes that home and workplace locations are the two locations people visit most frequently and regularly, as measured by aggregating the preferences by user locations, Leng (2013). In case of human mobility characterization, statistical models were used in identifying the variance in the number of individual’s activity locations, Jarv et al. (2013). Also, studies had been done to identify the individual spatial travel behavior, where the caller activity threshold was used along with multiple linkage analysis to identify the daily and monthly meaningful locations of users, Jarv et al. (2013). The study in Estonia introduces a model for the identification of such meaningful places for mobile users, identifying them as home and work anchor points, Ahas et al. (2010).

Different techniques have been used to reduce the errors in locating CDR. Considering the movement state identification, the basic objective is to identify whether the user had stayed at the location or whether it is a pass-by or transient point. The user can make calls while staying at trip ends like home or at their office. Also, they can make calls while traveling to these trips ends. The towers to which they connect while traveling becomes the transient locations. A time threshold is imposed in most of the studies to clarify this moment state, Calabrese et al. (2011), Mellegard et al. (2011), Maldeniya et al. (2015). As an example, a threshold of 10 min for the lower boundary and 1-hour upper boundary is used and also the records should be more than a 1km apart, Saini et al. (2015). If the sequence of records fulfills this criterion it is identified as a stay location and a moment between two stays is identified as a trip. But in general, the method works well for high-frequency data. This will cause most of the data points to be labeled as pass-by. So the locations were also associated with previously visited points to be more precise. The purpose of 10 min lower boundary is to minimize the wrong identification of trips due to stationary individuals connecting to different neighboring towers, Maldeniya et al. (2015) while the other target is to identify hidden visit inferences occurred due to the difference in observed trip ends and the actual trip ends. The 1-hour upper bound potentially reduce the chances of hidden visits occurring during observed displacements, Zhao et al. (2014).

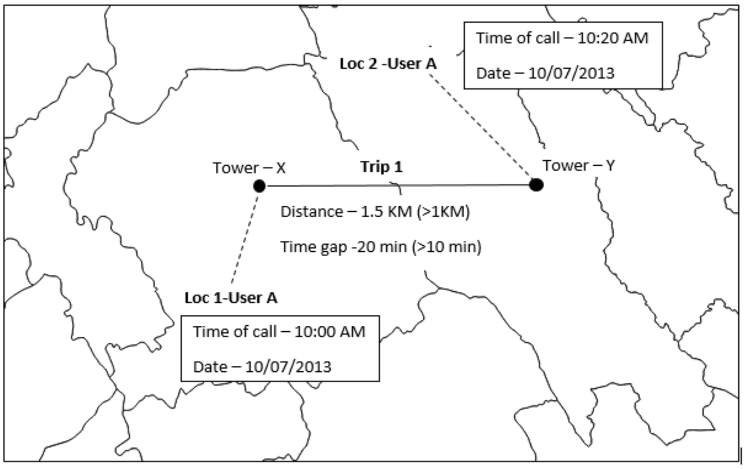

Fig. 1. Illustration of a trip

Fig. 1 above illustrates an example of the concept of a trip used in previous studies. User A had made a call at Location 1 (Loc 1) by connecting tower X at 10:00 AM and he had made a call at Location 2 (Loc 2) by connecting to tower Y at 10: 20 AM. Studies have introduced this as a trip since the towers are more than 1KM apart and the call gap exceed the boundary of 10 minutes.

In summary, different aspects related to human mobility have been addressed by the previous research including identification of significant places, generation of O-D matrices, profiling of users, etc. Although a large number of studies have used CDRs to understand mobility and different travel aspects, the methods still lack a conceptual integration with transport theory. Literature also shows that the majority of the studies carried out in the Sri Lankan context receives the most attention in identifying the trip numbers rather than evaluating the human behavior. The main two sources of error in using CDR due to data sparsity and localization error resulting from load sharing effect are addressed to some extent, Calabrese et al. (2011), Iqbal et al. (2013). The current study tries to fill the gap of evaluating human behavior through a preliminary analysis of CDR while further reducing the errors created by load sharing and data sparsity by analyzing the behavior of the most frequent stays.

3. Introduction

The study uses 3 months (May, June, and July) of Call Detail Records for nearly 1000 randomly selected SIMs from a mobile operator in Sri Lanka. Data was provided for this research by LIRNEasia - a regional ICT policy and regulation think-tank. The data is completely pseudonymized by the operator, where the phone numbers have been replaced by a unique computer generated identifier.

The main data source used for the purpose of validation is obtained from the transport data collected from household surveys from within the Western Province conducted as part of the CoMTrans study, JICA (2014). The household survey data provide transportation information, including social-demographic records, travel time, trip purposes, travel modes, etc., which can be used for validating the results from CDR trip analysis.

4. Methodology

This analysis begins with the study of the appearance of different users and the consistency of such appearance within the Western Province. This is followed by the study of the regularity of user appearance within particular zones also called stays. The study of the regularity of such stays over time is used to identify potential stays such as home and non-home locations. The analysis excluded callers who did not demonstrate regular mobile usage patterns while the remaining CDR were processed to determine the user's with regular appearances within the province in a meaningful manner. The locations or stays of each user were identified by considering the appearance threshold within particular time categories. The connectivity between these locations was analyzed further to derive significant travel patterns.

4.1. Methodological framework

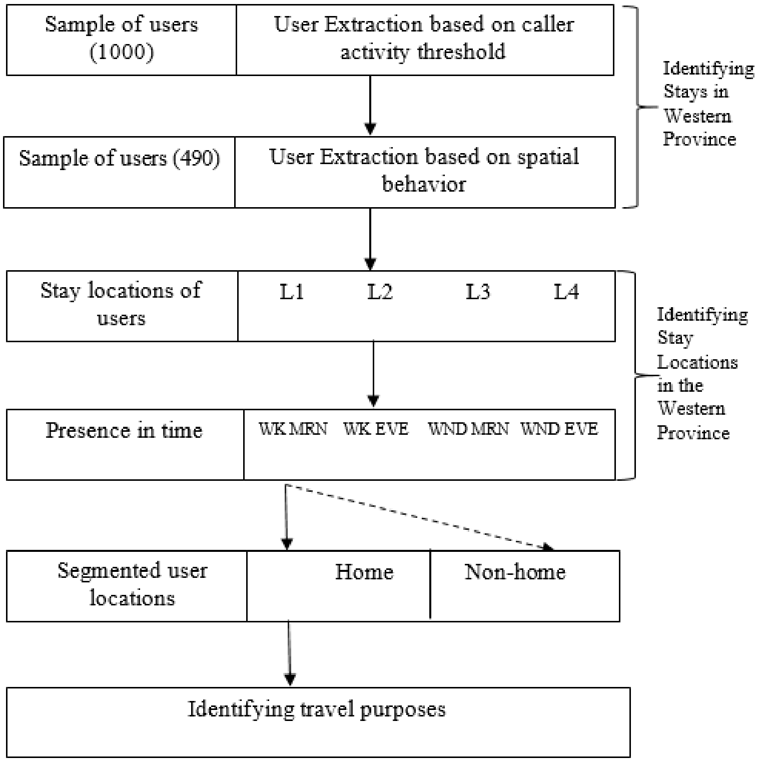
Sample of users (1000)
User Extraction based on caller activity threshold
Sample of users (490)
User Extraction based on spatial behavior
Identifying Stays in Western Province
Stay locations of users
L1
L2
L3
L4
Presence in time
WK MRN
WK EVE
WND MRN
WND EVE
Identifying Stay Locations in the Western Province
Segmented user locations
Home
Non-home
Identifying travel purposes

Fig. 2. Methodological framework of the study.

4.2. Value creation with Call Detail Records

Below is an example which the authors may find useful. CDRs, being the most widely available type of data in the interest of the transportation community, contain the time-stamped locations of the user which is the exact requirement for identifying trips. As an example, after processing the CDR data a single entry takes the following format.

Table 1. A sample of Data Record- Single user data after pre-processing.

Incoming/Outgoing	User	Cell ID	Date	Time
I	A	23452	23/07/2013	09:30
I	A	23452	23/07/2013	10:30
I	A	11432	23/07/2013	18:30

Table 1 can be interpreted, such that user A took a call at cell Id 23452 on 23/07/2013. Which means that this user had been to, stayed or passed by this particular location on this date at 9.30 am. User A had also made a call by connecting to the same cell ID at 10:30 AM. There are three possible explanations for this scenario;

- The user might have moved to different locations during that time of 30 minutes and come back to his original location
- He might have moved, but only within the coverage area of base station 23452
- He may have stayed at the same location without moving

Considering the last entry in Table 1, the user is observed to have moved to a different location by 6:30 PM. In order to create the value from these entries for transport studies, it is important to analyze whether these are random points visited by the user or a pass by point or whether this is a significant location in the particular user's daily life. The current study focuses on identifying the above aspects through systematic analysis.

4.3. Identifying stays in Western Province

To identify the stays within the Western Province, the initial sample was processed by removing the noise in the data. Due to the dynamic nature of the records, it is difficult to define the exact threshold (intensity of caller activities that a certain user should exceed to be qualified for further analysis) of caller activities to extract users with credible data for further analysis. Accordingly, users who have records continuously over all the three months were selected. For example, Table 2 indicates the dispersion of the call activities of a single user. This particular user does not have even a single record in the 3rd month of the data set, such users were initially eliminated from the main sample of users which left 764 users for further analysis.

Table 2. User behavior illustration_ no records in all three months.

User ID	Cell ID	Date	Time
A	23452	23/05/2013	09:30
A	23452	23/05/2013	10:30
⋮			⋮
A	23452	04/06/2013	11:20
A	15432	02/06/213	16:25

As the next step, each user's stays within the Western Province during the considered period were identified. Only the Moments within the Western Province was taken by considering the possibility of the validation with the existing data. The total no. of days the users had visited the Western Province were cumulated as in Table 3

Table 3. Example_ Cumulating the total no. of days within the Western Province

User ID	Cell ID	Date	Time	Geographical Location (Province)	
A	23452	23/07/2013	09:30	Western Province	Day 1
A	23452	23/07/2013	10:30	Western Province	
A	11432	23/07/2013	18:30	Other	
A	23452	24/07/2013	08:45	Western Province	Day 2
A	23452	24/07/2013	11:50	Western Province	
A	44512	25/07/2013	07:45	Other	

As an example, User A had created 5 entries within three days (Made 5 calls within the three days), but in the actual scenario, the user had present at the Western Province only for two days. The total no. of days that each user had appeared within the Western Province were calculated in a similar manner. Out of them, the users who have appeared at least a single day within the Western Province were considered for further analysis.

Based on that criteria, 49% of users were qualified for further analysis. This can also be extrapolated to be understood that 49% of all mobile callers from within the country would visit the Western Province at least once during a three month. Remaining 51% have their regular movements outside the boundary of the Western Province. Fig 3 illustrates the summary of the user identification based on the total no. of stays within the Western Province of Sri Lanka.

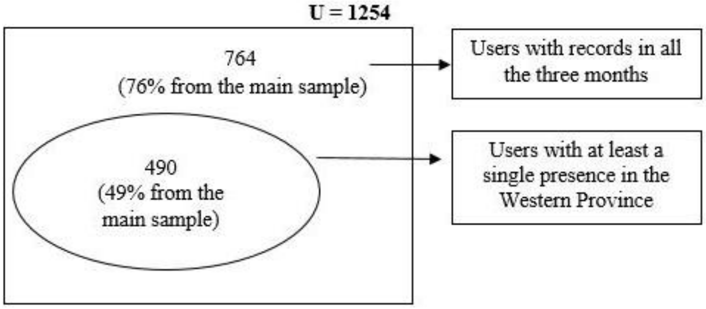
U = 1254
764
(76% from the main sample)
Users with records in all the three months
490
(49% from the main sample)
Users with at least a single presence in the Western Province

Fig. 3. Initial user identification based on the nature of the records

4.4. Identifying home and non-home locations

Home and other non-home locations of users were decided based on the total number of appearances of each user in each zone. Due to the differences in mobile usage, it is difficult to define the exact trip purpose at this preliminary stage. Therefore, the study assumed that, a caller who appeared to be found in a zone more than 75% of the considered time window, as a regular visitor to that specific zone within the considered time period. The most frequent zones during evenings were considered as the home location while the regularly visited zones during the mornings were considered as non-home locations. Fig 4. Below shows typical caller stays within different zones where locations A and B are identified to be in the vicinity of the home since 92% of the stays were at night or in the weekend while locations A and B appear to be the non-home location since they were found to be visited during 81%, 76% of the time respectively during weekday mornings. However, many callers did not demonstrate clear home or non-home locations. These could be those who do not have regular trips away from homes or those who have both their home and work or school located in the same zone. Also itinerant persons such as salesmen, transport workers, tradesmen, part-time workers or those doing shift work etc. would not demonstrate trip patterns as assumed in this identification.

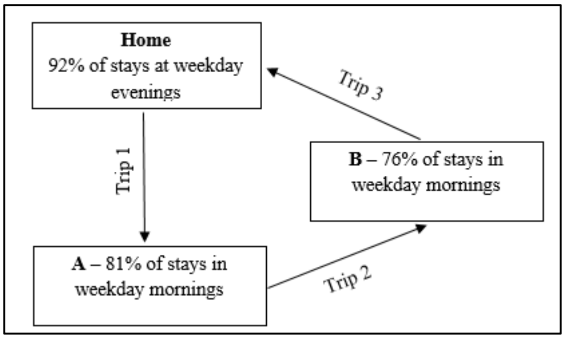
Home
92% of stays at weekday evenings
Trip 3
Trip 1
B – 76% of stays in weekday mornings
A – 81% of stays in weekday mornings
Trip 2

Fig. 4. Example of the moment between home and non-home locations

4.5. Identifying travel purposes

Considering the regularity and the connectivity of the regular stays, three user categories were identified as low mobility users, commuters, and expatriates. When the most frequently used zone within the mornings and the evenings were approximately equal, such callers were considered as low mobility users indicating that they did not leave the cell very often. These could be people who rarely left their homes or did in fact, leave regularly, but did not leave the zone. Commuters are those whose show frequent stays during the day of the weekdays in a zone different from that of the night time and weekends. These are identified as being regular commuters. In the case of those who were found to have three different frequent locations between the daytimes of weekdays, the night of weekdays and weekends were identified as expatriates who stayed during the week in temporary locations and regularly went to the permanent residence over weekends. Further, it was possible to identify those who were expatriates in the Western Province and those who were residing permanently in Western province. Results from the above caller types are shown in Table 4.

Table 4. Summary of Caller Types (by % of Users)

	Low mobility users	Commuters	Expatriates
% of users from 490 sample	63%	20%	17%
% of users with primary stay locations within the Western Province	52%	58%	Weekdays in the Western Province – 40% Weekends in the Western Province – 60%

4.6. Validation with household data

The research concludes its findings by validating the CDR results with the results from the Household Visit Survey conducted in the Western Province. Table 5 shows the validation which compares the percentage of those clearly distinguished work and home locations within the Western Province, where 26% of the users from the CDR analysis were found to be working within the Western Province, whereas the HVS survey indicated this as 27.9%.

Table 5. Validation with HVS data

Criteria	HVS Data	CDR data – (sample of 1000 Caller IDS)
% of users with work and home location within the Western Province	= (1624632/5821710)*100% = 27.9%	= (58/223) *100% = 26%

5. Conclusion

The main outcome of this research is the formulation of a methodology to understand travel attributes of users within the Western Province using mobile phone CDR. The study also validates the home-work distribution using traditional household survey data proving the possibility of using Call Detail Records to predict travel behavior. The development of the methodology is demonstrated using randomly selected 1254 IDs.

The research also identifies several limitations of using such data primarily arising from load balancing of mobile phone towers causing mobile devices to oscillate between alternative towers indicating a physical movement of a user even though the device and the user remain stationary. This results in the recording of false movements. A user who habitually makes several calls e- route between stays may also be indicating to have many stays in zones which are actually transit zones. The pursuit of further analysis using larger CDR sample sizes will lead to better results.

6. References

Ahas, R., Silm, S., Saluveer, E. & Tiru, M., 2010 Using mobile positioning data to model locations meaningful to users of mobile phones.Journal of Urban Technology, 17, pp. 3-27

Alexander, L., Jiang, S., Murga, M. & González, M., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transport Research Part C-Emerging Technologies, 58, pp.240-250

Calabrese F, Di Lorenzo, L., Liu, C., Ratti, 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10, 36-44

Dash, M., Kiat, K.K., Decraene, J., Yap, G.E., 2015. CDR-To-MoVis: Developing A Mobility Visualization System From CDR Data. IEEE 31st International Conference on Data Engineering in Seoul South Korea

Farkavcová, Guenther, E., Greschner, V., 2010. Decision making for transportation systems as a support for sustainable stewardship.Management Research Review 33, 317-339

Iqbal, M., Charisma, F., Choudhury, Wangb, P., & Gonzalez, M. (2013). Development of origin-destination matrices using mobile phone call data. Transport Research Part C-Emerging Technologies,40, pp.63-74

Jarv, O., Ahas, R. & Witlox, F., 2013. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transport Research Part C-Emerging Technologies, 38, pp.122-135

JICA, 2014. Urban Transport system development project for Colombo metropolitan region and Suburbs.

Kevin, Kung, S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data

Leng, Y., 2013. Urban Computing using Call Detail Records: Mobility Pattern Mining, Next-location Prediction, and Location Recommendation.

Maldeniya, D., Kumarage, A.S., Lokanathan, s, 2015. Where did you come from? Where did you go? Robust policy-relevant evidence from mobile network big data.

Mellegard, E.S., Zahoor, M.M., 2011. Origin/Destination-estimation Using Celluar Network Data. 11th IEEE International Conference on Data Mining Workshops

Saini, T., Barot, K., Sinha, A., Gogineni, R., Krishnan, R., & Venkata. (2015). Estimating Origin-Destination Matrix using Telecom Network Data.

Zhang, Y., 2014. User Mobility from the View of Cellular Data Networks. IEEE Conference on Computer Communications