DEVELOPING A TRIP DISTRIBUTION MODEL FOR IDENTIFIED MOBILITY GROUPS USING BIG DATA

Buddhi Ayesha Rathnayaka

188006L

Thesis/Dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

February 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis/dissertation under my supervision.

Name of the Supervisor: Dr. Charith Chitraranjan
Signature of the Supervisor: Date:

Name of the Supervisor: Dr. Amal Shehan Perera
Signature of the Supervisor: Date:

Name of the Supervisor: Prof. Amal S. Kumarage
Signature of the Supervisor: Date:

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisors Dr. Amal Shehan Perera, Dr. Charith Chitraranjan and Prof. Amal S. Kumarage for the continuous support and guidance I received in every aspect while completing this research. I could not have imagined having a better set of advisors for my research. I am highly grateful for their insightful guidance and encouragement provided throughout my research work.

I would also like to thank my progress review committee, Dr. Lochandaka Ranatunge, and Dr. Dilum Bandara for their valuable insights and guidance. Their advice helped me to improve the state of my research work.

A very special gratitude goes to Sriganesh Lokanathan (Team leader, LIRNEasia), and I also thank Lasantha Fernando (Researchers - LIRNEasia). Further, I would like to extend my gratitude towards LIRNEasia organizations for all the support given during the study period.

This work was funded by a Senate Research Committee (SRC) Grant of the University of Moratuwa. Further, I would like to thank LK Domain Registry for the support given for the conference participation via travel grants.

I am sincerely thankful to all the staff members and the non-academic staff of the Department of Computer Science and Engineering. Finally, I would like to thank my fellow lab mates in Data Science and Engineering Research Hub of University of Moratuwa and all of my friends and family. Not just academic work, I was able to enjoy my time while involving different activities and outings. It helped me to keep my life balanced. I thank my parents who have given me a very fortunate life and always believing, trusting and supporting me.

ABSTRACT

The need for frequent transportation planning has become a key factor since people started becoming more mobile making urban traffic patterns more complex. The primary source for analysing such travel behavior is through manual surveys. These surveys are expensive, time consuming and often are outdated by the time the survey is completed for analysis. To overcome these issues, Mobile Network Big Data (MNBD) which concerns large data sets can be used over such traditional data collection processes. Call Detail Records (CDR) which is a subset of MNBD is readily available as most of the telecommunication service providers maintain CDR. Thus, analyzing CDR leads to an efficient identification of human behavior and location.

However, many researches on CDRs have been done focusing to identify travel patterns in order to understand human mobility behavior. Relatively high percentage of sparse data and other scenarios like the Load Sharing Effect (LSE) causes difficulties in identifying precise location of the user when using CDR data. Existing approaches for identifying precise user location patterns have certain constraints. Past researches utilizing CDRs have used primary approaches in recognizing load sharing effects and have given minimum consideration to the transmission power of the respective cell towers when localizing the users. Furthermore, these studies have neglected the differences in mobility behavior of different segment of users and taken the entire community of users as a single cluster.

In this research, a novel methodology to overcome these limitations is introduced for locating users from CDRs by dividing the users into distinct clusters for identifying the model parameters and through enhanced identification of load sharing effects by taking the transmission power into consideration. Further, this study contributes to the transport sector by identifying secondary activities from CDR data, without limiting to the primary activity recognition. This research uses approximately 4 billion CDR data points, voluntarily collected mobile data and manually collected travel survey data to find techniques to overcome the existing limitations and validate the results.

Proposed dynamic filtering algorithm for load shared records identification showed a significant improvement on accuracy over previous predefined speed based filtering methods. Further, we found that, IO-HMM outperforms standard HMM results on activity recognition.

Keywords: Travel Demand Modeling with Mobile Network Big Data, User Localization Based on CDRs, Activities identification from CDRs.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CA	Call Activity
CAC	Call Activity Count
CDR	Call Detail Records
CLS	Cordon Line Surveys
DSD	Divisional Secretariat Divisions
HBW	Home-based Work
HMM	Hidden Markov Model
HVS	Household Visit Survey
IO-HMM	Input/Output Hidden Markov Model
JICA	Japan International Cooperation Agency
LSE	Load Sharing Effect
LSR	Load Sharing Records
MNBD	Mobile Network Big Data
O-D	Origin-Destination
PCA	Principal Component Analysis
SLS	Screen Line Survey
SVM	Support Vector Machine
TGS	Trip Generation Survey
TSS	Travel Speed Survey
VLR	Visitor Location Register

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Travel is the one of the most common activity of people living in urban areas. Only a limited number of cities in the world have the facility of water-based traveling whereas traveling by air is considered to be inappropriate for urban travel. Thus, the means of travel available for urban passenger transportation are mainly land-based, which include private transportation (Walking and private motor vehicles) and various public transportation services [1]. As urban travel in Sri Lanka is purely based on land, planning and implementing of proper transport initiatives should be given significant attention.

Mobility patterns of humans are a critical information source when designing, analyzing and enhancing transport planning activities. A set of mobility rich data is essential in order for these transport activities to become a continuous process with progression. At present, these input data are gathered through a manually carried out roadside survey or by household surveys, which are done occasionally. These approaches ought to be expensive, time-consuming, and outdated by the time of conducting the analysis using the gathered data. Because of that, more and more innovative and modern data extraction methods have become an essential necessity and in demand, in order to analyze human mobility.

The higher engagement of modern computing technologies, specially mobile devices being used by a substantial portion of entire population has made the possibility of capturing large-scale spatio-temporal data related to human mobility [2]. With the introduction of Big Data and it's deep understanding capability of social practices, extensive penetration of mobile devices has proven to be a vital source of extracting insights on human mobility.

Mobile Network Big Data (MNBD) is known as complex, large-volume, growing data sets retrieved from the manner people interact with the communication devices [3]. These Mobile Network Big Data is capable of using the telecommunication infrastructure which contains location information to locate mobile devices with respective to time with a certain accuracy [4]. While data collected via mobile phone networks, GPS and Bluetooth have been using in a wide range of mobile-based applications to power functionalities such as location navigation, services, tracking, etc., the most common type of MNBD used in transport-related studies are the Call Detail Records (CDR) data which is collected by the mobile phone carriers for customer usage tracking and billing purposes.

This study focuses on understanding the behavioral patterns of mobile device users within the specified study area, i.e. the Western Province of Sri Lanka. In order to validate the CDR findings for model development, attributes of CDR data were transformed into behavioral elements and home visit survey data. Validation of CDR data confirms the outstanding relationship with the actual human mobility abiding by the basic concepts of transportation.

Mobility analysis provides outputs on par with several intermediate and final results of the traditional forecasting framework. The section below drafts the basic concepts of transportation along with the traditional framework, which acts as the foundation for deriving the CDR outputs.

1.1 Transportation Forecasting

Transportation forecasting is the process of estimating the total number of people or vehicles that will use the given facility in the future [5]. The majority of the travel demand models are based on the conventional transport forecasting technique known as the "Four-step Model". The process initiates with the following three key indicators known as destination choice, mode choice, and route choice as shown in Figure 1.1.

Trip generation is the analytical process that provides the relationship between the urban activities and travel [6]. It depends on factors such as land use and socio-economic characteristics of the particular area whereas conventional trips are a production of a particular zone and attracted to other specified land uses. Trips ending from the origins or destinations are the ones that are originated or

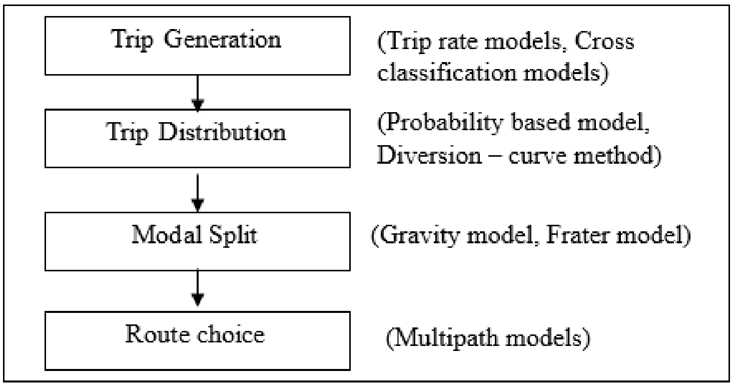

Trip Generation
(Trip rate models, Cross classification models)
Trip Distribution
(Probability based model, Diversion – curve method)
Modal Split
(Gravity model, Frater model)
Route choice
(Multipath models)

Figure 1.1: Four-step model

Source: 2015 Regional Travel Demand Model

terminated within each zone. The primary stage of the model is to predict the number of trips that will commence and terminate from each travel analysis zone within a particular region during a day. With the help of the above-mentioned surveys, the total number of trip generations can be calibrated. The second part is about identifying to where the trips are destined for or in other words, to estimate the interchange volumes between each pair of zones [1], i.e. trip productions generated from the trip generation phase in each zone I is distributed among the trip attracting zones J. Trip volume of the zone J would depend on its relative attractiveness. Usually, residential areas act as the trip generation points whereas the non-residential areas become the attractors. Figure 1.2 below demonstrates the trip distribution of residential and non-residential zones. When people travel from the residential areas to work in the morning, the residential area becomes the production zone and the workplace becomes the non–residential area of attraction. This situation takes place vice-versa during the peak hours of the evening.

Based on the origin and the destinations of the trips, movements can be categorized into four categories as shown in Table 1.2. Movements were segmented paying special attention to the origin and destinations in the context of the study area and the traffic analysis zones. When we categorize the trips under the topics given below, it makes the process of identifying the overall behavior of the users

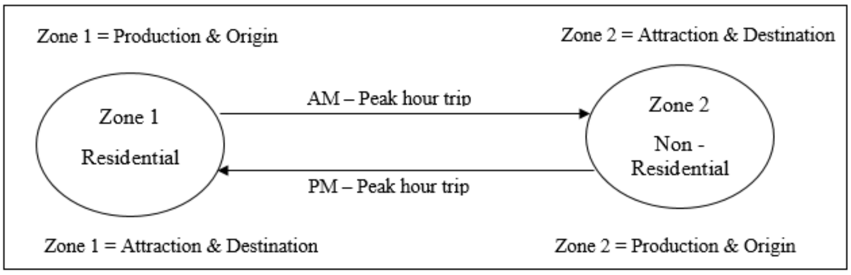

Zone 1 = Production & Origin
Zone 2 = Attraction & Destination
Zone 1
Residential
AM – Peak hour trip
PM – Peak hour trip
Zone 2
Non - Residential
Zone 1 = Attraction & Destination
Zone 2 = Production & Origin

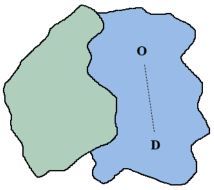
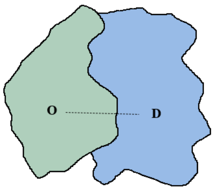
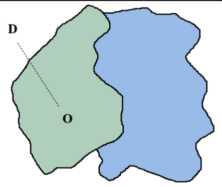
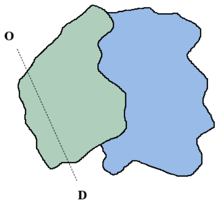
Figure 1.2: Characteristics of residential and non-residential zones

much easier, such as whether they are traveling within a specified zone or to any other areas and whether the specified zone is a transit zone, etc.

The third stage of the model is the model split which determines which vehicle will be utilized while it travels from one zone to another or within the same zone, mainly as auto, transit, bicycle, walking, etc. The complexity of this stage is mainly based on the availability of transit modes in that area. A variety of choices is available for travelers starting from the mode of transport such as bus, subway, commuter rail, ferry as well as whether they would like to drive alone or carpool. The choices made by the travelers will be undertaken and assigned to a transportation network while determining the route or path to be taken when traveling from one zone to another. The supply provided by the road and transit systems will interact with the travel demand of the traffic zone. Determining an individual's path based on the minimal travel time and the congestion that can arise from several vehicles using the same route can be taken as one such example.

Travel information is a requirement to model the trips using the above mentioned techniques. Even though surveys are used as the source for data aggregation at the moment, it is considered to be an expensive and time-consuming process in transportation model development. These surveys study when, where and how people travel within the areas. A number of transport surveys such as Home Visit Surveys (HVS), Cordon Line Surveys (CLS), Screen Line Survey (SLS), Trip Generation Survey (TGS), Travel Speed Survey (TSS) were conducted to obtain data within this context. The primary aim of this study is to take the initial steps in supplementing these surveys with CDR data by establishing a relationship between CDR and traditional data.

Table 1.1: Types of trip movements

Movement Type	**Description**	**Illustration**
Intra-zonal trips	Trips which have both ends within the same zone in the study area	
Inter-zonal trips	Trips which have both ends within the same study area, but in different zones	
External trips	Trips which have one end within the study area and the other end outside the study area	
Transit trips	Trips which have both ends outside the study area, but travel through the study area	

1.2 Introduction to Mobile Network Big Data

Big data is an information technology term that refers to high volume, velocity, and variety of data [7]. The biggest feature of Big Data is the high volume with rising quantities. The diversity and generation methods of the data create the structural difference in Big Data. Some of the diversified big data sources are stock market fluctuation data, online trading data, credit card usage data, mobile positioning data and administrative data such as digitized medical records, insurance records, and tax records. Mobile Positioning Data or Mobile Network

Big Data is one of the main categories of big data generated by the use of mobile phones creating the technical ability to locate devices in time and space using available network infrastructure with a certain accuracy. Data generated through mobile devices such as Wi-Fi networks, GPS, Bluetooth involve in a variety of applications including location services, navigation, etc.

These mobile positioning data are introduced in different names based on their purposes and collection techniques. This study uses CDR data which is known as one of the most common types of big data. Every mobile network carrier collects call detail records generated during incoming voice calls, outgoing voice calls and text messages which is an event-driven data for billing purposes. While CDR generates the exact time-stamp of the user's locations above mentioned unique identifier links each phone owner with their Caller Activities (CA). In order to protect the privacy and the privacy of every customer, a unique computer-generated identifier will be used to replace every phone number. Table 1.2 contains data of a single CDR.

Table 1.2: CDR data parameters

Name	**Description**
Call Direction Key	1. Incoming call 2. Outgoing call
Device Name	Ex:- MICROMAXX103
A Number	The anonymized identifier for a particular phone number
Other Number	The anonymized identifier for the other phone number
Cell ID	Each cell corresponds particular base station which has a latitude and longitude position
Call Time	Formatted as:- YYYY-MM-DD:HH:MM:SS
Call Duration	Measured in seconds

All the data will be recorded in the form of a data array each time a call is made. Based on mobile phone recordings and interests of travel patterns in transportation, the processing of the CDR leads to the collection of mobility rich information of human insights.

1.3 Purpose of the Research

Travel has become more complicated as humans have become more itinerant, generating the requirement to commence a continuous process to plan transport initiatives. Hence, the necessity of frequently updated sets of data arises with the analysis of mobility aspects. The existing mechanism has the capacity beyond the requirement. A handful of researchers are interested in using CDR, which is a subset of mobile network big data to analyze human mobility behaviour. Nevertheless, there are several issues with existing CDR data analysis methods. This study will be focusing on two significant areas namely, "Load sharing effect identification" and "Activity pattern recognition". The ultimate motive behind this study is to utilize the combination of MNBD and traditional transport forecasting approaches, which include trip generation and distribution steps to recognize human mobility behaviour.

1.4 Research Gap

A high number of transport studies have been carried out in different countries using CDR, including Sri Lanka . Different methodologies have been adopted for a range of transport studies, including O-D matrix estimation [8], [9] identifying meaningful places [10] to the characterization of human mobility [11] which will be discussed in detail within literature review. Machine learning-based algorithms perform as the foundation of the study even though the existing studies present comprehensive mechanisms. Deficiencies shown in the studies abiding the fundamental transport concepts pave the way to several conceptual issues, as discussed below. Another noticeable barrier would be the ignorance of load sharing effect by a majority of researchers; especially in the local context due to mobile cell coverage.

Initially, a variety of parameters should be established with regard to sample selection, while identifying the stays, movements and etc. At the moment, there is a lack of availability in properly labeled CDR data set with socio-economic factors to analyze human mobility behaviours. Therefore, the main priority is to create a labeled CDR data set with socio-economic information.

1.5 Research Objectives

This study identifies two significant issues related to CDR data analysis, and the following research objectives were set to minimize the issues mentioned above.

- Develop a methodology to perform improved trip detection using CDRs
- Generating Machine learning models to identify mobility groups using CDRs
- Developing activity pattern recognition models and validating the derived model with existing transport data

1.6 Scope of the Research

The factor behind selecting Western province as the study area was the possibility of data validation since "house-hold visit survey" data (Ground truth data) only covers the three urban districts, including Colombo, Gampaha, and Kalutara with 47 Divisional Secretariat Divisions (DSD). Only inter trips were validated at the DSD level even though both intra and inter trips were validated at the district level. The most common and precise home-based work trips were extracted and used for primary validation purposes. Shopping, visiting friends & relatives, travelling to school and etc, were considered for further analysis.

1.7 Project Contributions

Following are the contributions of this project towards the growth of the research community in the domain of CDR data analysis for Travel Demand Modeling.

- Novel way to identify the load sharing effect on CDR data and minimize its effect on data analysis.
- Study area lacks labeled CDR data. A mobile app developed is used to collect labeled CDR data and socio-economic data.

- Introduction of a novel methodology for User Localization using Call Detail Records data.

- Publication of a research paper for the "*20th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL) 2019*" at The University of Manchester, United Kingdom (Paper title: "User Localization based on Call Detail Records").

Chapter 2

LITERATURE SURVEY

The analysis of human mobility patterns for the use of empirical data collection has been an active area of research in many occasions [12]. Most of the literature studies had focused on working with data obtained via wireless tracking devices. With the emergence of wireless communication methods which essentially has changed the way people communicate, work, and engage in day-to-day activities around the world. Hence, mobile phone data which is the most common wireless communication method in the world can be employed to derive the spatio-temporal information of anonymous phone users' whereabouts in order to conduct an analysis of their social mobility patterns [13]. This chapter gives an overview of the studies in the transportation field using mobile data, followed by the introduction to digital data used in transport analysis and a discussion on different aspects of transport studies covered along with the methodologies. This chapter also addresses the issues with the existing techniques of using CDR and the identified research gap in a detailed manner.

2.1 Introduction to Digital Data in Transport Analysis

The current generation of transport models exploits the advantage of the growing availability of mobile phone use data, GPS tracks, transactions data and online activity patterns of users to evaluate and predict on behaviors of people moving around a specific city and to the depth of how they interact with different places.

Traveling while a GPS-enabled smart mobile phone on-board and also embedded with in-vehicle systems in some cases has become the norm which enables moving vehicle data to be more prevalent and accessible as ever [14]. With GPS data, studies employ probabilistic models to predict future vehicle destinations alongside Markov models to find the most-likely and probabilistic next location based on a subset of previous locations [15]. Without using different modes other than personnel vehicles some studies have focused on developing multi-modal destination prediction models [16].

Transaction data have been used for the travel prediction apart from GPS data. Smart-card payment systems such as "Smartrip" in Washington and "Suica" in Tokyo are implemented in some cities which generates considerably valuable travel data while being an essential part of the present-day public transport fare collection systems. In addition to that, a study has proposed and instigated a method to acquire an Origin-Destination matrix from smart-card data for a multi-modal public transport system based on the estimated alighting stop for trip segments [17].

CDRs which contains location and time stamp information for consequential activity destinations are not as structured as traditional travel survey data, or as accurate as GPS data which provides higher frequency and accuracy [18], but out of all these, CDR data which is a byproduct of the billing process of mobile phone connections are the most economically and readily available information type generated which can be used for the travel predictions. The large volume of data and long observation period of mobile phone data can be used to infer human footprints on an unprecedented scale even Though such data are often sparse in space and time [19]. Depending on the location positioning technology employed by service carriers, CDRs can reasonably represent spatio-temporal information of mobile phone users' movements at cellular-tower or much finer-grained level [13].

Primarily there are two types of mobile positioning data categorized based on their collection method as "active positioning data" and "passive positioning data" [20]. Both of these types can be observed in real time as well as historically, but in general, a specific target request is made to locate the device in active positioning while in passive positioning historical data is collected without a request [21]. Visitor Location Register (VLR) and CDR are two main inclusions under passive positioning data. VLR data generates a record whenever a subscriber changes the serving base station. But in contrast, CDR is generated only when a phone call is made. Therefore, VLR represents human mobility with a higher resolution than the CDR. But due to the large volume involved, associated storage and the processing cost, analyzing VLR is difficult at the current stage. Passive positioning data in the form of CDR is the most common and universally captured by all mobile connection providers when all factors are considered [22]. Compared to other network related data like GPS, CDR is easily available as most of the telecommunication service providers maintain such data [23]. Including the aspects of economic activities and urban planning, etc. CDR had been used in a wide range of studies in relation to transportation.

2.2 CDR and Human Mobility

CDR have been used considerably in previous studies to explore the knowledge of individual human mobility. The movement of an individual is governed by his or her daily routing which includes frequent places of visit, which is described as the activity space of that particular person [13]. Individual trajectories which reflect movement of an individual in space over time can be used to characterize the activity space [24]. Significant amount of research has been focused on individual activity space to identify different attributes of movement. Analysis on travel or activity space can be mainly studied under two main sections as trip based analysis and activity based analysis.

2.2.1 Trip Based Analysis Using CDR

When a user's mobile phone corresponds with a cell phone tower, CDRs in relation to that specific stream of connection contain a traces of the user with estimated locations providing an approximate and incomplete picture of the user's daily trip-making [8]. Using micro-simulation and limited traffic count data it has been recognized that CDR data can be used to infer O-D of trips [3].

For instances in the road networks of Boston and San Francisco, Wang et al. developed a technique to generate tower-based transient O-D matrices for different time periods and convert them to node-to-node transient ODs [25]. In that research, Wang et al. have created transient O-D matrix by observing the people's movement from one destination to another.

In order to learn about the spread of the travel demands in a time period of a day and gather movements over the total observational period, the researches have divided the day into four periods as follows, Morning period: 6am to 10am, Noon & Afternoon period: 10am to 4pm, Evening period: 4pm to 8pm, Night period: 8pm to 6am. A movement is when it is observed that the same mobile phone user has been in two distinct zones in a time span of one hour (Please refer to the latter part of the literature review chapter for error mitigation topic). Nodes considered as the intersections and links considered as the road segments, it can define a road network. Every movement in the O-D matrix is assigned to the road network by an incremental traffic assignment.

A variety of techniques have been used to develop O-D matrices. In a study of Mumbai CDR, data were filtered to obtain data from the given day type, and data from all the days corresponding to the day-type is combined to generate mobility traces for each given user by superimposing the location of all activities for each user. The records are then aggregated for all users, multiplied by a scaling factor and converted to vehicle trips to arrive at the initial O-D [9]. Apart from that, the majority of the studies have developed algorithms to assign mobile phone towers extracted from CDR to traffic nodes [26]. Inter-relation between mobile O-D and traffic O-D have not been explored in detail as most of these studies have been focused on computational issues.

Coming to the identification of meaningful places, a model had developed to identify home and work locations. The study was conducted in Estonia and the process uses the number of days and the number of calls as the primary inputs [10]. Other than that, most of the studies focus on identifying major home and work anchor points [9], [27]. Weekday, Weekend movements with time window was taken 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 the most frequently, measured by aggregating the preferences by user locations [28]. In the case of human mobility characterization, statistical models have been used in identifying the variance in the number of individual's activity locations [29]. Also, studies had been carried out to identify the individual spatial travel behavior, there the caller activity threshold was used along with multiple linkage analysis to identify the daily and monthly meaningful locations to users [29].

Table 2.1 and Table 2.2 indicates the different methodologies followed by different researchers within their corresponding scope of analysis. These studies can be categorized into two sections as follows;

1. Identifying Significant Locations.

2. O-D Matrix Estimations.

Based on the behavioral structure, stay locations have been assigned as home or work in order to identify the significant location. In order to identify the common structures, Principal Component Analysis (PCA) is one of the main techniques used. Usually, PCA is a technique used to emphasize variation and to bring out robust patterns in a dataset [30]. Significant points have been based considering the frequently used cells by month (Monthly activity locations) or day (Daily activity locations). Eventually, identifying the stays and generating tower to tower or geographical location based matrices is the general technique used to derive O-D matrices.

2.2.2 Use of CDR in Sri Lankan Context

In Sri Lanka, mobile penetration has increased from 87% - 123% between 2011 to 2016 [33] suggesting that almost every person is using mobile for their daily activities. Additionally, some people may be using more than one SIM and some may not be using a SIM at all. Since the mobile device is used by a majority of the population, it is reasonable to state that the population of mobile users would be represented by the travel patterns of a random sample of users.

CDR in the Sri Lankan context has been used in a wide range of studies including economic, urban planning, transportation and etc. When combined with other geographical and various demographic data, CDR data can be used to gain awareness of human mobility.

Table 2.1: Identifying significant locations using CDR in transport aspects

Attribute	Name of the Research	Methodology Followed
Identifying Significant Locations	Urban Computing using Call Detail Records: Mobility Pattern Mining, Next-location Prediction and Location Recommendation [28].	- Capture when, how often and how long each user appears at each user location. - Identify the common behavioural patterns and daily routines from the identified appearances with Principal Component Analysis (PCA). - Identify home and workplaces based on the similarity in the present patterns.
	Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone Call Detail Records [29].	- Assign meaningful monthly and daily activity locations using mathematical linkage analysis technique. - Total number of caller activities made within locations on considered days or months is used as the main influencing criteria for identifying locations.
	A Hierarchical Approach for Identifying User Activity Patterns from Mobile Phone Call Detail Records [31].	- Finding the usual stay locations of the users from one-month CDR data. - Work time considered as 9am to 5pm while other period as off hours. Distinguished two main groups as regular and irregular workers. - Identified the distance of travel to work and some other random locations within these two groups.

The research on quantifying urban economic activity using cell phone data [34], study the connection between urban economic activity and commuting flows using fine spatial and temporal variations. CDR has also been used to understand communities using communication patterns by applying several community detection algorithms [35]. Land use classification being an important concept in

Table 2.2: O-D matrix estimations using CDR in transport aspects

Attribute	Name of the Research	Methodology Followed
O-D Matrix Estimations	Estimating Origin-Destination Flows Using Mobile Phone Location Data [32].	- Trip determination All consecutive points for which the radius <1km combined together such that the centroid becomes a virtual location. - Once the virtual locations are determined, the trips and stops were evaluated as paths among users' positions at consecutive virtual locations. - The geographical area under analysis is separated into regions to derive the O-D matrices.
	Development of Origin–Destination Matrices Using Mobile Phone Call Data [3].	- Generate tower-to-tower transient O-D matrices using and trips occurring within different time periods. - Converted the corresponding traffic network modes to node-to-node transient O-D matrices. - Derive the true O-D matrices by scaling up these node-to-node transient O-D matrices using traffic counts.
	Origin-Destination Trips by Purpose and Time of Day Inferred From Mobile Phone Data [8].	- The records were first converted into clustered locations at which users engage in activities for an observed time period, to be work, home, or other depending on frequency of observations based on triangulated mobile data. - Probabilistically infer departure time based on survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day.

urban planning is one other aspect involved with CDR in the Sri Lankan context [36]. This explores the potential of leveraging massive amounts of human mobile usage to derive conclusions at spatio-temporal activities of the masses and provide a useful measure for activity based classification of land use. Apart from the above, CDR is also used in disaster management and disaster resilient development aspects [37].

Coming into the state of art in transportation with MNBD in Sri Lanka, one of the main studies is an origin-destination matrix estimation with CDR [38] which uses different techniques like stay based approach, frequency-based approach and transient-based approach in deriving OD-matrices using CDR [39]. A stay is defined as a continuous series of records in the stay-based approach, such that;

- The distance between the cell towers connected should be less than 1 km.
- The entire series of records should span a period of more than 10 minutes
- Two series of records are separated by a time interval, as a means to T $<$ 1 hour

A trip by the stay-based approach was introduced with the movement between two stays. In the transient-based approach, the movement is considered as a trip if the cell tower utilized for each call is different and records are separated by a time interval of 10 minutes $<$ T $<$ 1 hour.

Apart from that, the daily sequences of an individual over a period of 1 year helps in identifying the frequent locations using the frequent trip-based approach. The movement between the most frequently visited locations was calculated as trips to generate O-D pairs. Nonetheless, there are variations in the results of the traditional transport surveys and MNBD estimations which have to be understood before MNBD can be used to replace traditional transport forecasts.

2.2.3 Limitations of Using CDR in Travel Predictions

There are issues in using big data due to its high volume and complexity. Though CDR can be obtained more frequently and economically from the telecommunication service providers, there are several limitations that need to be addressed when using them in mobility estimations.

A critical limitation of CDR data for mobility analysis is data sparsity [40]. Data entry is generated only when communication activity is initiated [31]. A user may be monitored to move from zone B to zone E, but user's initial origin (O) and final destination (D) may actually be located in zone A and zone F. In such cases, segments of the trip information are overlooked in the CDR [3]. Trip estimation based on sparsely and irregularly distributed data can be incomplete. Sparsity also causes movement state identification, making it challenging to identify whether the user had stayed at the location or it is a pass by point.

Apart from that using CDR we cannot recognize the exact location of the user, but the location is recorded as the geographic position of the cell tower used by the network [31]. The user can be anywhere within the coverage area. This is further referred to as the localization error, where it requires different algorithms to identify the sequence of locations or frequent locations visited by users. The "density of a cell tower" and the "density of cell towers varies with the area" are the dependents of the spatial granularity of the location data. The coverage area of the cell towers varies depending on the socio-economic features including population density [41].

Another limitation of CDR data is that there can be changes in the tower although there is no actual displacement of the user which is described as the effect of load sharing [3]. This can happen when the mobile service provider tries to balance cell tower traffic among adjacent cell towers which allows them to optimally serve the traffic at that instance.

There are perceptible concerns with the three approaches described in the above as in the stay-based approach where a trip is defined as a movement between two stays, the transient locations or random locations that cannot be captured. Long distance trips are also considered as a collection of small trips since they cannot be identified properly. In the frequent- based approach, Ad-hoc movements are ignored as the location sequence is only considered frequent, if it occurs on at least 10% of the daily sequences of the individual. As a collection, all these restrictions create different types of errors in identifying the location of users, identifying the stays of users, identifying pass by points.

2.2.4 Techniques Used in Minimizing the Limitations of CDR

To interpret the most accurate results, most of the studies have used a variety of methods to reduce the errors of CDR. One primary factor considered by most of the studies is the location identification which identifies the locations visited by the users and the results were obtained through various techniques. The majority had produced a sequence of visited locations using different algorithms [41], [42], [43] which is referred by the term trajectory smoothing. For instance, the sequence of CDR extracted within a certain time threshold was filtered using time, speed or by assigning a center location for close by points, for such algorithms assist in identifying the common or the significant location sequences of users. Locations have been identified clusters based in some studies depending on their spatial distribution without considering the time factor. Finally, the points of locations were consolidated even though they were visited on different days [44], [45]. The most frequently visited locations of an individual's routine can be understood by this methodology.

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 point. A time threshold is imposed in most of the studies to clarify this movement state [26], [32], [38]. 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 1km apart [9]. If the sequence of records fulfills this criterion, it is identified as a stay location and a movement 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. Moreover, the 10-minute lower boundary minimizes the wrong identification of trips due to stationary individuals connecting to different neighboring towers [38].

A distinction exists between "observed end-locations" and "hidden end-location" with the help of the data sparsity, which occurs due to differences in observed trip ends and the actual trip ends. One study had explicitly addressed this issue directly based on that a hidden location occurs when a significant amount of time is elapsed between cell transitions. They have set a 1-hour upper bound that potentially reduces the chances of hidden visits occurring during observed displacements [40]. The selection of the threshold is empirical as they tend to overestimate the existence of hidden visits with sparse data. This issue is paid less attention in many of the studies with total awareness that it has serious insinuations on trip detection which results in the wrong extraction of O-D pairs.

2.2.5 Research Gap Identification

The literature review indicates that the CDR can be used in a wide range of applications in a transport context, including O-D matrix estimation, identifying meaningful places to the characterization of human mobility. Even though the theoretical involvement of the method is not up to its standards, a large body of literature is available regarding the use of CDR in transportation aspects.

The basic trip concepts like trip purposes, intra-inter trips are comparatively low with regards to previous studies. There is always a defined purpose for travel and it is entwined with these concepts verifying that these ideas should be prioritized in the analysis work. Heuristic approaches and arbitrary criteria are used in the existing studies in setting up different location and time boundaries which depend on the socio-economic background of the users. The socio-economic criterion should be adjusted to the Sri Lankan context and included in the analysis

This research attempts to derive travel using CDR by focusing on the purpose of home-based work. In addition the study area lacks studies with labeled CDR data. Therefore this study primarily focuses on labeling the CDR data and deriving meaningful travel estimations from CDR data. Also the load sharing effect which is an embedded issue in CDR data has not been thoroughly studied in previous studies. Since, analyzing the load sharing effect is also a primary target in the current study. This study further attempts to validate the methodology proven by comparing the CDR data with the existing transport data.

Chapter 3

METHODOLOGY

This chapter provides a detailed analysis of the methodology used in this research. Figure 3.1 depicts the overall approach, and it contains two distinct steps of model creation and model validation. The model creation step involves developing different models using CDR and mobile app data. The following subsections of this chapter will discuss the actions of each component presented in Figure 3.1.

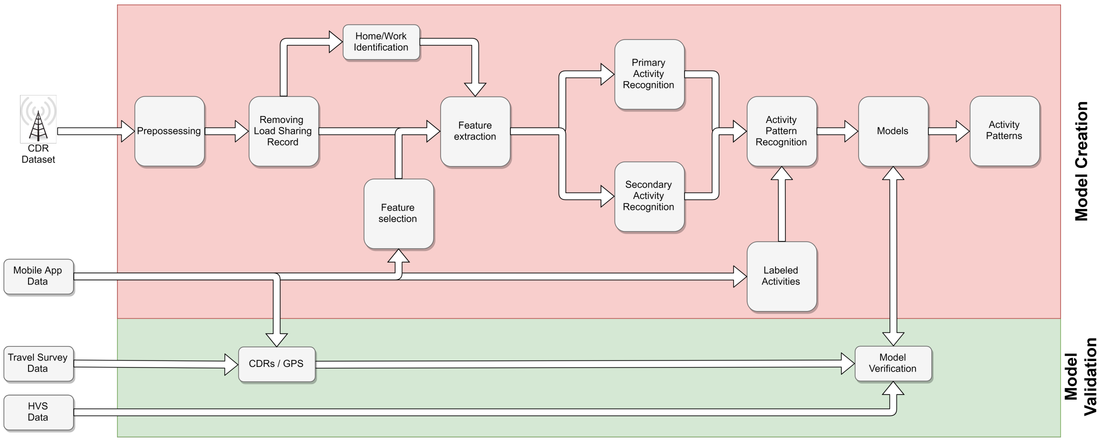


Figure 3.1: Flow diagram for the overall methodology

3.1 Data

The study primarily uses three data sources as in analysis namely CDR data, Voluntarily collected mobile app data and HVS data. Detailed description on each of the used data sources are explained as follows.

3.1.1 Call Detail Records

The study utilizes CDRs of nearly 10 million SIM cards from mobile operators in Sri Lanka. Data was offered for this research by LIRNEasia - a regional ICT policy and regulation think-tank. Data is entirely pseudonymized by the operator, wherever the phone numbers appear.

3.1.2 Voluntarily Collected Mobile App Data

This research occupies GPS and CDR data collected voluntarily from an unbiased sample population of mobile phone users in a pseudonymized manner. To observe user behaviours through cell phone localization and activity, we developed a mobile application which can run on android devices. This mobile application gathers various data periodically and sends them to a central database. Table 3.1, 3.2, 3.3 shows the type of data collected and its contribution to the study. Table 3.4 shows the total list of attributes of data collected through the app.

3.1.2.1 Usability of the mobile app

Users can download the mobile app from the Google play store (Figure 3.2). After the installation of the app, users can open the app and select the preferred language where the future functioning of the app will be based (Figure 3.3). Next, an embedded video will be automatically cast, which will provide users with a sound knowledge with respect to the usage of the app.

After the introductory video, a registration window (Figure 3.4) will appear, so that users are required to fill up the general questions. Income is marked as an optional question, as it is a highly confidential matter. The user can edit the entered details as per their requirement by clicking on the edit button in the window appeared, after registration.

Thereafter a notification will popup if the user initiated a travel by indicating the time as "Were you traveling at 9.00 am?" (Figure 3.5) Since the travel initiation is identified by the GPS mechanism, there can be instances where notifications will pop up even without an actual moment. This is due to the GPS error in certain locations. The user can provide the answer by swiping the notification. If the user was not traveling he can select the "I wasn't" button in the window (Figure 3.6), and if the user was actually traveling he can fill up the other questions on origin, destination, pass by destinations and the mode of travel. Users are allowed to select multiple pass by destinations and modes of travel.

Table 3.1: Data collected at the user registration interface

Data collected	**Contribution to the study**
Gender	The study seeks to determine different travel behavioral structures based on their caller activity patterns. Socio-demographic data collected as in the left column were identified as independent measures of travel choice behavior, which are presentable as questions to be filled within a minimum time consumption. Analyzing these socio-demographic data along with the CDR data will generate different user profiles based on their travel participation.
Age	
Occupation	
Average income	
Home address	Home based work trips are the most conveniently and accurately identifiable from the CDR data. This fact is clear from the state of art. But in previous researchers the level of validation goes only up to inter, intra district level and inter DSD level where intra DSD level remains undone due to the accuracy issues. But this gap can be filled by the analysis CDR along with GPS data. Collection of work–home locations from the users will facilitate in the validation of sample results related to work–home distribution up to the intra DSD level.
Work address	
Mobile number	Mobile number will be taken initially for the unique identification, which does not have any backward tracing. It will be pseudonomized internally during the analysis.

Special Remarks:

- Since GPS locations will be collected from every user with a frequency of 10 minutes, it is mandatory to switch on GPS for the proper functioning of the app. If the GPS mode is not activated in a user, initially a notification will pop up after 10 min of app installation to activate GPS (Figure 3.7). If the GPS mode is still inactive the notification will pop up with after 1hr and the next reminder after 2hrs. App is designed to send notifications automatically at 7.00 am and

Table 3.2: Data collected with the initiation of travel

Data collected	Contribution to the study
Nature of origin, Destination, Pass by location	Travel purpose identification from CDR data is a main objective of the study as it is one of the primary attribute of travel. Data like nature of origin, destination will support in identifying trip purposes and analyzing the changes of caller behavior patterns with the purposes on different users will facilitate in modeling solutions to the primary problem of identifying travel purposes with CDR data. Additionally, Obtaining pass by destinations will support in identifying the moment state of points (Pass by or A stay locations).
Mode of travel	Travel mode identification is a primary concern in traditional sequential travel demand forecasting models. As it is expected to supplement these traditional techniques with digital data sources, it is important to identify travel modes with digital data. Obtaining the particular mode of travel and analyzing them simultaneously with CDR data will address the above fact and will facilitate in the validation.

12.00 pm after initial reminders to activate the GPS mode.

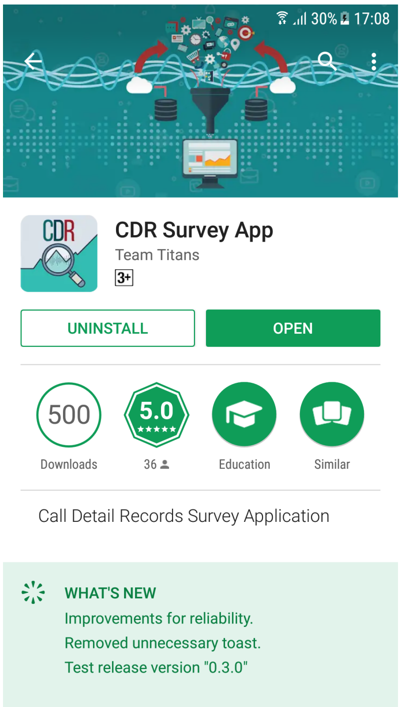
- As mentioned earlier, a notification for requesting travel details will pop up with every trip initiation. If a user does not answer for the question, the notifications will be stored in the device for 24 hours and deleted automatically.

3.1.2.2 Feasibility of the study

Following section discusses the feasibility of the mobile app in technological, legal and economic aspects.

Technological feasibility

API Level:- The app is available for Android versions above API level 17, where the working percentage is above 94%.



30% 17:08

CDR Survey App

Team Titans

3+

UNINSTALL

OPEN

500 Downloads

5.0 ★★★★★ 36

Education

Similar

Call Detail Records Survey Application

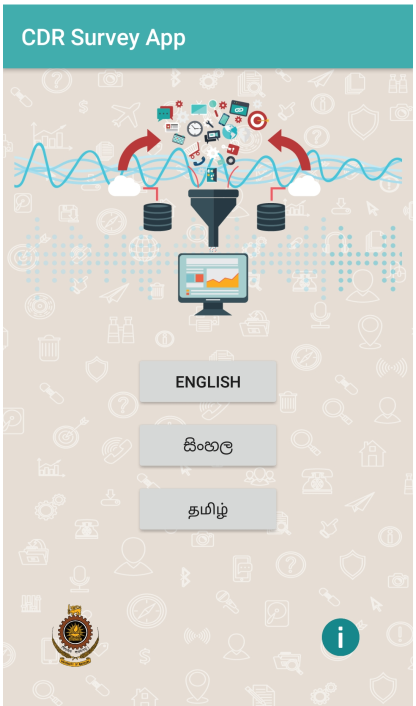
WHAT'S NEW

Improvements for reliability.

Removed unnecessary toast.

Test release version "0.3.0"

Figure 3.2: CDR mobile app listed on the Google play store



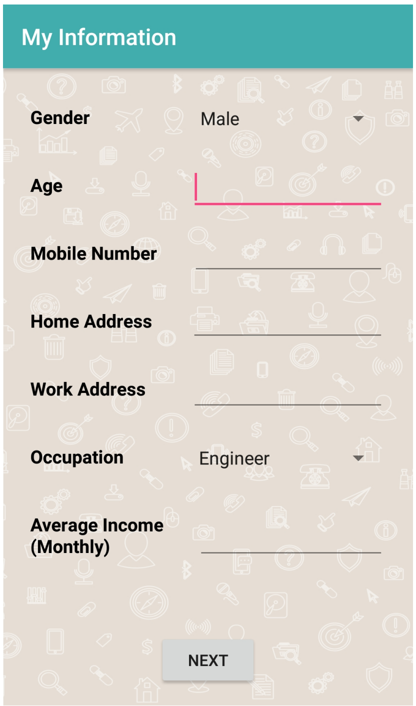
CDR Survey App

ENGLISH

සිංහල

தமிழ்

Figure 3.3: CDR mobile app supporting three languages



My Information

Gender: Male

Age:

Mobile Number:

Home Address:

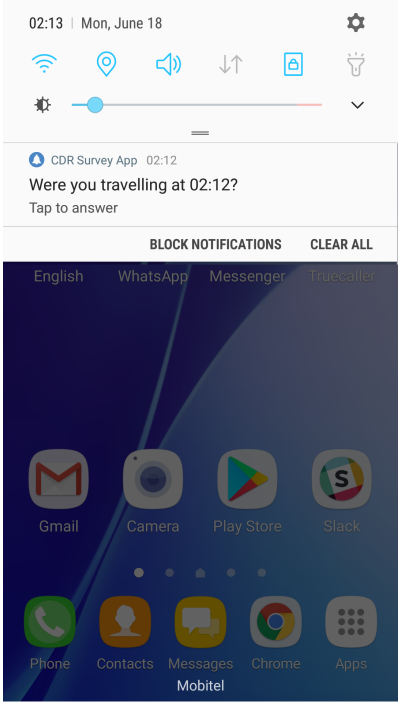
Work Address:

Occupation: Engineer

Average Income (Monthly):

NEXT

Figure 3.4: Data collection at the user registration interface



02:13 | Mon, June 18

CDR Survey App 02:12

Were you travelling at 02:12?

Tap to answer

BLOCK NOTIFICATIONS

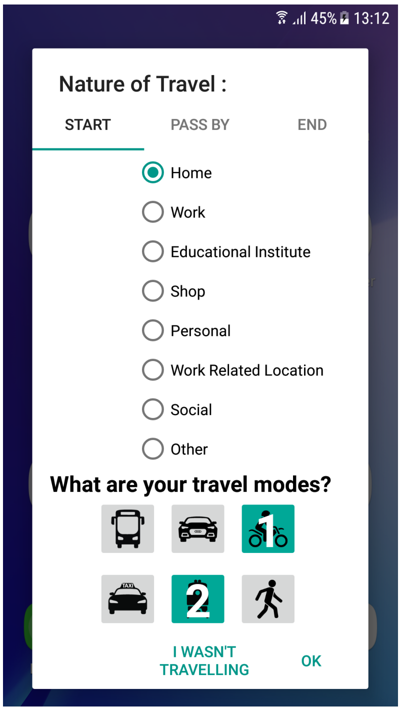
CLEAR ALL

Figure 3.5: Pop-up notifications during the initiation of travel

Data consumption:– App will consume only 3MB for the installation and 15 MB per month during the running of the app.

Table 3.3: Automatically collected data through background processing

Data collected	Contribution to the study
CDR Data	Every time a user makes a call, a CDR will be generated and stored in the main database. Identifying travel attributes using CDR data is the primary objective of the research and GPS data collection in parallel, supporting accurate location identification with temporal attributes. Further analysis of CDR along with GPS data will support in minimizing location identification errors from CDR data by modeling solutions to load sharing effect and moment state identification (Categorizing locations as pass by or stay points).
GPS Data	GPS data of the app users will be collected with a 10 min frequency.
Signal Strength Data	App collects signal strength data every 10 min

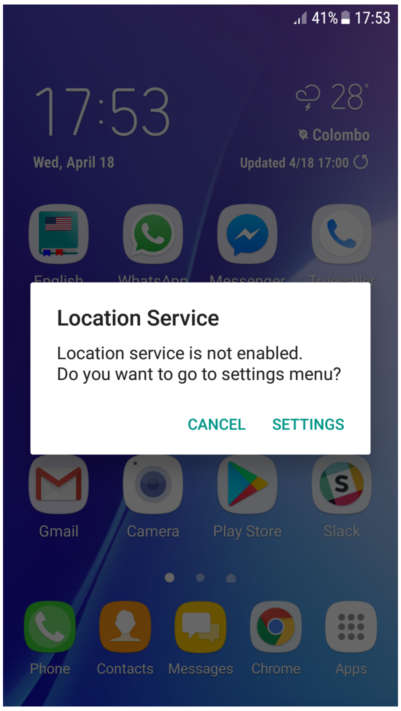

Nature of Travel :

START PASS BY END

● Home
○ Work
○ Educational Institute
○ Shop
○ Personal
○ Work Related Location
○ Social
○ Other

What are your travel modes?

Figure 3.6: Data collection with the initiation of travel


Location Service

Location service is not enabled.
Do you want to go to settings menu?

CANCEL SETTINGS

Figure 3.7: Permission request to use the GPS navigation

Table 3.4: Data collected through the mobile app

Type	Attributes
Survey Questions	1. Gender 2. Age 3. Home Address 4. Work Address 5. Mobile Number 6. Occupation 7. Average Income (Monthly) (Optional)
Popup Questions	1. Are you travelling? 2. What is your purpose of travel? (To home, To work, Private matters, Education, Business, Shopping, Other) 3. What is your mode of travel? (Bus, Private Motor Vehicle, Motorbike, Taxi, Railway, Walking)
Call Record	1. Caller ID (Hashed using SHA-256 Algorithm) 2. Receiver ID (Hashed using SHA-256 Algorithm) 3. Cell ID (With Signal Strength Levels) 4. Call Duration (in seconds) 5. Timestamp 6. GPS Location
SMS Record	1. User ID (Hashed using SHA-256 Algorithm) 2. Receiver ID (Hashed using SHA-256 Algorithm) 3. Cell ID (With Signal Strength Levels) 4. Timestamp 5. GPS Location
Data Record	1. User ID (Hashed using SHA-256 Algorithm) 2. Cell ID (With Signal Strength Levels) 3. Data Volume (in bytes) 4. Timestamp 5. GPS Location

Battery life of the mobile phone:– Only 1% of the battery will be consumed by the app when the mobile device is fully charged.

Economic feasibility

Installation of a Mobile app is a general activity performed by a smartphone user and mainly the cost is incurred for the data consumption. It is expected to distribute the app with the support of mobile service providers like Dialog by giving data packages (Internet) as incentives in order to motivate the users to install the app.

Legal feasibility

User can view the privacy policy after the installation and data is only gathered after user's consent.

Data were gathered for fifteen months from more than 700 users starting from 21-03-2018 to 21-06-2019. Some statistics of the voluntary collected mobile app survey data;

- Total number of users:- 538
- Total number of active users:- 137
- Total number of users with caller activities:- 381
- Total number of users with GPS records:- 462

80% of the participants of our mobile app dataset consists of people who are related to the education sector (e.g. Students, researchers, lecturers). We used algorithms and techniques which sensitive to the above class imbalance problem in our research.

3.1.3 Household Visit Survey (HVS) Data

Although, surveys act as the primary data source for current transport studies, in order to supplement the traditional data collection techniques with the CDR data, the CDR findings should be validated with the existing data. This study uses the household visit survey (HVS) data [46] to validate its discoveries. Since understanding the process conducted in surveys is an initial requirement, this chapter gives an overview to the household visit survey data, data collection methodologies and attributes used from the survey for the purpose of validation.

3.1.3.1 Introduction to Household Visit Survey

Transport demand in Western Province has increased remarkably over the past few years [46]. Several numbers of transport development projects and plans were carried out to cope with the anticipating transport demand and problems along with a number of surveys to obtain reliable transport data. Out of different types of surveys, Household Visit Survey (HVS) conducted by Ministry of Transport with the technical support of Japan International Cooperation Agency (JICA) is used for the validation of the current study since it is one of the largest and comprehensive transport surveys carried out in Sri Lanka. HVS covers the boundaries within the Western Province which includes Colombo, Kalutara, and Gampaha districts consisting of 2496 Grama Niladhari Divisions and 47 Divisional Secretariat Divisions. Due to the hesitation of people to participate in these kinds of surveys, the original sampling size of 3% had to be extended up to 4%.

3.1.3.2 Trip purpose composition

There were 7 purposes for the household survey questionnaire. The total number of trips made by each household was recorded under these seven purposes for the convenience of analysis. Table 3.5 indicates the seven primary trip purposes and their compositions in the survey result and briefly explains each of the considered purposes.

The survey estimated that around 10 million trips were made on a weekday within the Western Province and among these, around 2 million trips were counted as Home-based Work (HBW) trips. The current study focuses on these 2 million home-based work trips to match with the CDR data due to its regularity in frequency, time and space.

3.1.3.3 Distribution of HBW trips

As stated earlier, the HVS data estimation was based on a randomly selected weekday within the Western Province. Identified trips can be distributed among the districts within the study area (Colombo, Gampaha, Kalutara) based on their

Table 3.5: Description of trip purposes

Purpose	**Description**
Work	Commutes performed from home place towards the workplace and returning to the home location
Business	Trips from the workplace to business destinations and the return trip from the business destination to work location.
Shopping	Commute towards the stores and return to the origin.
Educations	Commutes towards any type of learning establishments and return trip to origin.
Private matter	Other matters like meeting relations, movements for medication, recreation and etc., including the return trip
Other	Any movement beyond the above categories.

Source: Urban Transport Development project for CMR and Suburbs

origins and destinations as in Table 3.6.

Table 3.6: Distribution of home-based work trips

Home/Work	**Colombo**	**Gampaha**	**Kalutara**
Colombo	45%	2%	1%
Gampaha	10%	28%	0%
Kalutara	4%	0%	10%

3.2 Preprocessing

CDR data obtained from mobile operators is from the whole country, and the validation data set only covers the traveling patterns of the Western province population. Hence, initial preprocessing tasked with filtering out Western province users. Further, data captured in public holidays were removed since the usual travel behavior of the people is changing.

In addition to this, there are some other preprocessing steps that are used in different stages of the model creation pipeline. These steps will be discussed in the respective subsections.

3.3 Load Sharing Effect on Call Detail Record Data

As stated above CDR data has the possibility of giving rich information and insight on human mobility at a large scale, As previously mentioned, CDR data has a great potential of providing rich information and insights on human mobility at a large scale, but inherits with some bottlenecks and ultimately cause for low accuracy of derived information. Frequency switching of mobile phone users among cell towers even without actual displacement has been identified as one of the major barrier, which is indicated as the Load Sharing Effect [3]. This happens due to the frequent balancing of call traffic among neighboring towers to provide a good quality of service. As an example, Figure 3.8 depicts the general demonstration of cell towers. Here, the user has the opportunity of connecting to any tower. However, the connecting tower is decided by the telephone operator based on the total number of calls each tower can handle. Hence, the user's call might get connected to X1 at one time and X2 next time, without any changes in the A's location.

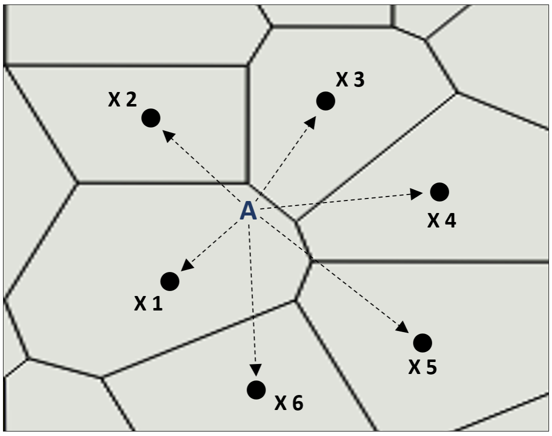


Figure 3.8: Load sharing effect

The signal gained from each tower varies without an actual displacement. Consequently, this affects on various transportation related assessments such as trip counts between particular origin and destination (it may exceed the actual count), behavioral patterns (it may be identified inaccurately). Therefore, it is essential to fix aforesaid existing issues of using CDR data in transpiration modeling and forecasting.

3.3.1 Load Sharing Effect Identification

The main objective of this section is to determine the LSRs. Previous studies has followed two different approaches to avoid the load sharing effect (Table 3.7). The first method is trajectory smoothing that will be used to spot jumps in a sequence of locations. The second approach is based on clustering.

Table 3.7: Methodology to minimize the Load sharing effect

Approach	Description	Filtering techniques
Trajectory Smoothing	Consider the sequence of CDRs within a certain time threshold and apply filtering techniques to reduce "jumps" in the location sequence.	• Speed based filtering [42] • Time-weighted smoothing [41] • Assigning a single medoid location to every record in the sequence if they are close by[44]
Spatial Clustering	Ignores the ordering or the temporal distribution of CDRs and clusters data points based on their spatial distribution only.	• Agglomerative clustering[44],[47] • Leader clustering [41], [43]

Most of the previous work is not applicable for Sri Lankan context. Therefore we adopt a new methodology to minimize and remove Load sharing effects.

3.3.1.1 Minimizing LSE clustering-based approach

Generally there are clusters composed of cells with high frequency and cells with lesser frequency around it. Cells appeared within Weekday evenings and weekday mornings will be clustered based on their distance and the frequency. In this method, we consider the ratio of caller activity frequency and the distance between cells. The clusters will be based on equation 3.1 ratio.

$$Cluster\ \alpha = \frac{\text{Frequency of caller activities}}{\text{Distance between cells}} \tag{3.1}$$

We applied this method to western province CDR dataset, as shown in Figure 3.9. One of the main drawbacks is that we can not identify individual Load shared record using this method.

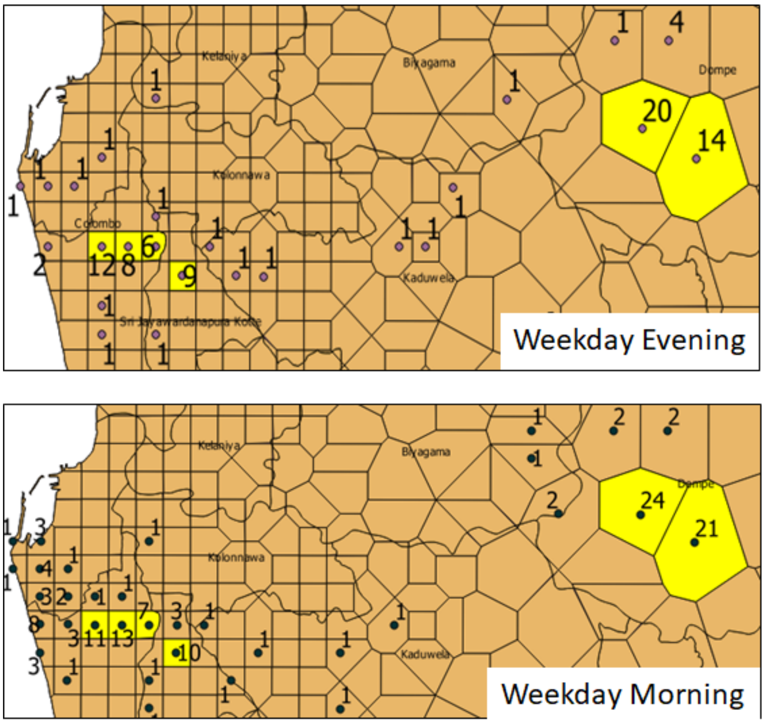


Figure 3.9: Clustering-based approach for western province

3.3.1.2 Minimizing LSE speed-based approach

Initial attempt was to employ a speed based filtering method. In speed based filtering, Trajectory data within a predefined time interval were observed any movement which are irrational based on the travel speed or travel distance are marked as load shared record CDRs. Subsequently, the speed of each user is calculated based upon two consecutive CDRs. This method tries to spot the jumps in location sequence.

The speed limit was set to 120 kmph based upon the findings in previous studies. But the results were not accurate as expected. In this setup, recall is very low even though the precision is high. In other words, most of the records which are determined as load shared records from this setup, are actually load shared records. Nevertheless, this approach was unable to identify lots of load shared records.

The study suggests a new approach to recognize load shared records. Main disadvantage of a predefined speed based filtering approach is that speed of the vehicle is not always constant. For instance, average vehicle speed is low in the morning, 7am to 9am compared to evening, 8.30pm to 10pm. The speed in one of the main corridors to Colombo is 6kmph in peak hours in general, but speed can goes up to 100 kmph in off peak hours. Vehicle speed also heavily influenced by the geographical areas. Vehicle speeds are higher in remote areas compared to urban areas. This study used dynamic speed-time filtering approach combined with road segment type to address the Load Sharing Effect. Speed limit data was obtained from previous transportation O-D surveys.

The speed limits is based on the time. Thus, the time frames were declared for our study as, 07:00 to 09:00, 09:00 to 12:00, 12:00 to 13:00, 13:00 to 16:30, 16:30 to 19:00, 19:00 to 22:00 and 22:00 to 07:00.

Then, the study area is divided based on the Divisional secretariat divisions (DSD). Average speeds were calculated based on distinct time windows. CDR were chosen upon the θ values which represent in kmph, $\theta = [0, 5, 10, \ldots, 200]$.

Initially, this approach was implemented using sample CDR data and later, this was applied nearly to 4 billion user records using Scala language on Apache Spark framework.

Spark job for identifies Load sharing record based on the dynamic speed threshold

- Remove low-frequency users.
- Filter Western Province users.
- Filter consecutive CDR records within the time frame.
- Calculate the speed of the user base on 2 consecutive CDR and add a speed filter.
- Remove identified Load sharing CDR records.
- Identify user's most frequent appearance based on the time window

- Calculate the mobility speeds of each user using those separated CDRs.

- Apply an arbitrary speed limit (θ) and filter the records which were affected by load sharing effect.

Then the records were evaluated and labeled as Load shared records using GPS data which represent actual movements. Based on this result, the same process was carried out for the main dataset.

3.4 User Localization

The discussion of this section is based on the estimating the individual user locations. CDR data were clustered based on their spatial behavior to identify distinct areas of stay (Stay location clusters). DBSCAN is employed as the clustering algorithm in this study. Significant number of past studies have only relied on data related to call frequency. This does not output accurate results nevertheless. Therefore, we present a novel method to lessen user location error. A particular cell tower is assigned to a particular user, based on the location of the user, call traffic and the signal strength from the cell tower. Three factors that were considered to assign locations for a particular user are listed below.

- Load sharing records.

- Signal strength of the particular cell tower.

- Number of days appeared in the particular cell tower.

For instance, consider the scenario bellow. Let's suppose that user X is connected to cell towers A, B, C and D with equal number of caller activities. If the weights are depends on the caller activities, the user location should be at the centroid of the recognized four locations within the considered period (Figure 3.10).

To happen above scenario, signal strength should be equal among cells. But, the location of the centroid will move from that of the position taken from the

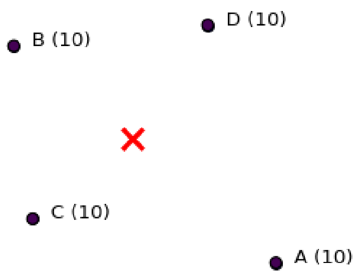


Figure 3.10: Use locations based on the number of appearances days

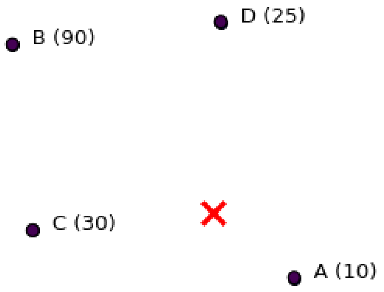


Figure 3.11: Use locations based on the signal strength

caller activity level when the signal strength data of cell towers are taken as weights Figure 3.11.

Thus, using the Load shared record frequency, the location for a particular user can be calculated. In this study area, power of signal transmission of each cell is obtained and used to rank the power of connecting users as a measurement. P_i shows the power of signal transmission of the i^{th} cell.

Using the above factors, the weight is calculated by the Equation 3.2. For a given cell i, Load sharing records, signal transmit power, and amount of appearances days are indicated by L_i, P_i, C_i, respectively for a considering user.

$$W_i = \alpha L_i + \beta \frac{1}{P_i} + \gamma C_i \tag{3.2}$$

Then, the L_i, P_i and C_i values were scaled as in Equation 3.3.

$$0 \leq L_i,\ \frac{1}{P_i},\ C_i \leq 1 \tag{3.3}$$

α, β, γ are distinct weights calculated for each user segmentation separately within the range mentioned in the Equation 3.4.

$$0 \leq \alpha,\ \beta,\ \gamma \leq 1 \tag{3.4}$$

Weighted k-means++ algorithm was employed with the weights obtained by Equation 3.2 to determine the cluster centroid in a particular time. Calculated centroids are then assigned as a user location for that corresponding time.

3.5 User Activity Pattern Recognition

The next step is user activity recognition using filtered CDR data. These data are used to identify not only primary activities, but also secondary activities. Home and Work based trips are categorized into primary activities. Other trips such as social, educational, shopping, etc are considered as secondary activities. Application of feature engineering on collected raw CDR data to identify user activities is a necessity.

3.5.1 User Profiling Based on Mobility

One setback in previous travel demand modeling research on CDR data is that all users are considered as a single set. But, factors such as occupation, age category, and etc directly affect travel behavior. Speciality of this research is collecting CDR data with socioeconomic labels through a mobile application. This subsection presents how this data is used for user profiling.

Preprocessing

Two filtering techniques are employed in data preprocessing. Typical users are removed using the first filtering technique. The typical users involve a limited caller activity levels and a limited representation within the study area (e.g. visitors, tourists). Almost all the previous studies were conducted based on call activity count (CAC). It is vital to remove the rarely visited users who are accessing the selected research area. This involves the user's spatial behavior while neglecting the CACs.

In order to filter the users, this study uses a filter based approach on spatial behavior. Analysis of the user's spatial activities is done by Shannon Entropy (3.5)

$$H = -\sum_{i=1}^{N} P_i \, log_2 \, P_i \tag{3.5}$$

In the above equation, 'P' defines the probability that activity was taken place at particular location, 'I' from the set of 'N' locations that the user visits. The number of locations that user appears in the dataset influence the behaviour measure directly. The probability of appearances of locations increases along with the entropy value when the total number of locations of appearing of a particular increases. Out of the entire dataset, 80% of the users display mostly a homogeneous behaviour when their spatial behaviours were considered.

Since the study is conducted on the Western Province of Sri Lanka, the Western Province users were considered.

3.5.2 Occupation Identification From CDR Data

Grouping all users into a single group and use of the similar model parameters for all users are the main drawbacks of previous studies. This comes to forth especially in investigating individual level mobility patterns. At the beginning, it is necessary to profile users based on their socio-economic parameters because the socio-economic parameters such as occupation, age, income and etc heavily influence the individual travel behaviour.

For occupation identification, we employed classification algorithms on different features obtained from CDR data. For classification model training, voluntarily collected mobile app data is used.

Here, work location and home location is identified from a time window based method. In this method 7AM to 4PM is used for work location identification time window and 7PM to 5AM is used for home location. After identification of work and home locations for each person, distance values were obtained for each stay location.

Based on correlation analysis, the study selected following input features.

- Average call frequency for each hour
- Average distance for each hour
- Number of cells appear in weekday mornings and weekday evenings
- Average time spend on a stay location

- Distance from home location
- Distance from work location

The study use the following set of classification algorithms to analyze the data collected from mobile app.Users are segregated into profiles using these classification algorithms.

- Neural Networks
- Support Vector Machine (SVM)
- Nearest Neighbor
- Random Forest
- Decision Table
- ZeroR

We used SVM weighted classes (SVM_{WC}) method since the dataset has some class imbalance problem. As the next step, the same models were used to classify the main CDR dataset.

3.5.3 Feature Engineering

It is necessary to have a feature analysis to identify most important features that help to identify user profiling. Following section will explain different features used.

3.5.3.1 Identifying Stay Locations

Stay location is a place where users generally stay without moving. Homes, Workplaces, Education institutes, shops, restaurants can be identified as some major stay locations. These stay locations are directly related with the user travel patterns. In CDR data, a user location means a mobile cell tower location where a user's mobile phone is connected. Load sharing effect has a massive impact on the user's stay location identification. Hence, minimizing the load sharing effect on CDR data improves the correct identification of different user locations. Following definition is used to identify user stay locations from CDR data.

If the time difference between the two records is greater than 10 minutes and the spatial displacement between two recorded locations is less than 1km, then it is defined as stay location. Further, if the time difference between the two records is greater than 10 minutes or the spatial displacement between two recorded locations is larger than 1km, it is considered as a non-stay location.

To explain this further, Figure 3.12 depicts the aggregated stay location of a randomly selected person on Mondays for the period from 1/5/2018 to 4/6/2018 (5 Monday in total). Figure 3.12 contains stay locations obtained based on both CDR and GPS data for comparison. When we consider the first 5 hours for the selected random user, it is visible that there are no recorded stay locations in CDR data. However when we consider GPS records, it is possible to identify a stay location, but from CDR data we can not capture this information because of the sparsity of data.

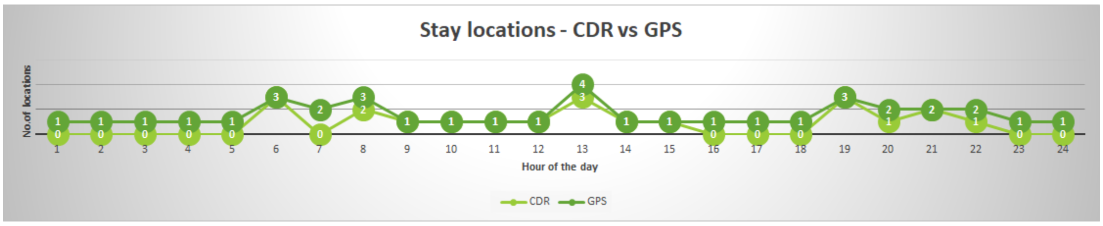


Figure 3.12: Stay locations based on CDR and GPS

This is an example of aggregated stay locations of a randomly selected user, on Mondays. Stay locations were identified by aggregating the Mondays of the considered period.

Aggregated Stay Locations Analysis

The next step was analysing stay locations for all users. For this we compared the accuracy between calculated total number of stay locations obtained from CDR and GPS data. Table 3.8 presents these accuracy details for each day of the week. Since there is a low accuracy, further investigation was done on the available data.

Table 3.8: Accuracy of stay locations identified from CDR

Day	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Accuracy	27.35%	29.40%	38.21%	26.35%	38.09%	25.03%	21.32%

We studied the aggregated accuracy percentages with caller activity levels. These results indicated that it is necessary to have at least around 200 total caller activity records to get a 50% or a higher stay location accuracy. Additionally, one other identified fact was that separate studying of weekdays and weekend data tend to improve the stay location identification.

As the following step, we try to identify a time window that contains more than 80% of caller activities in a day for each user. When we consider a person, usually CDR data is available between 6AM to 11PM for a day. However this time period can be differ from person to person. We calculated this time window for each user which contains 80% of total caller activities and used this time window to filter out stay locations. Even though this improves accuracy, when observing this data, it was difficult to identify stay locations for some individuals because there was a limited number of calls in a particular considering period of the day. For some people there were no calls on the 10th hour of the day and for another one this can be another hour of the day. Because of that, it is impossible to identify any stay location for such a person within that no call period. Hence, as a solution we assumed that person is staying in the previous stay location since there is no call record data. This method resulted in a significant accuracy improvement for stay location identification. These results are shown in Figure 3.13 and 3.14.

3.5.3.2 Travelling and non-travelling hours

Trip end identification and travelling hour identification is important for transportation. This study considers stay locations as trip end identification. Next step is the identification of the traveling hours or non-traveling hours. This is also important for trip activity identification. These travelling and non-travelling hours are defined as follows for both GPS and CDR data.

GPS:- Calculate the speed between two consecutive GPS records. If the speed exceeds 6KM per hour, consider that as a travelling hour.

CDR:- Calculate the speed between two consecutive calls. If the speed exceeds 6KM per hour, consider that as a travelling hour.

As an example, we first calculate the speed between two consecutive GPS records. If the speed exceeds 6KM per hour we considered that as a travelling hour. This same process is done for CDR dataset and compared with data obtained from GPS. Accuracy for travelling hour and non travelling hour predictions from CDR is calculated as follows (Equation 3.6 and 3.7).

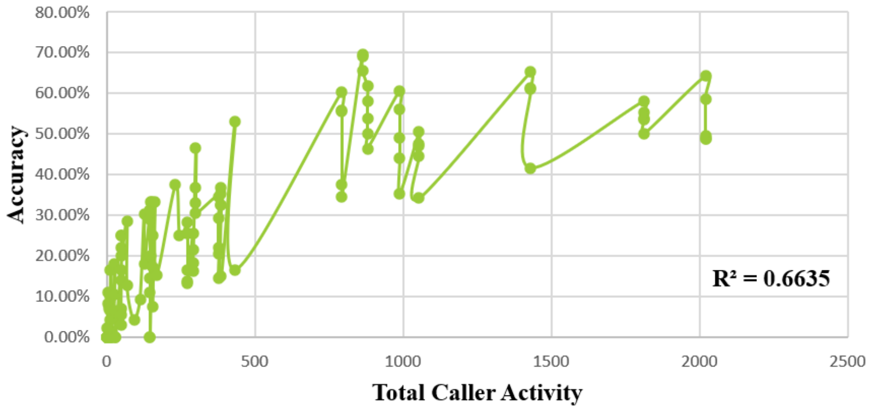


Figure 3.13: Aggregated accuracy with caller activity levels for weekdays

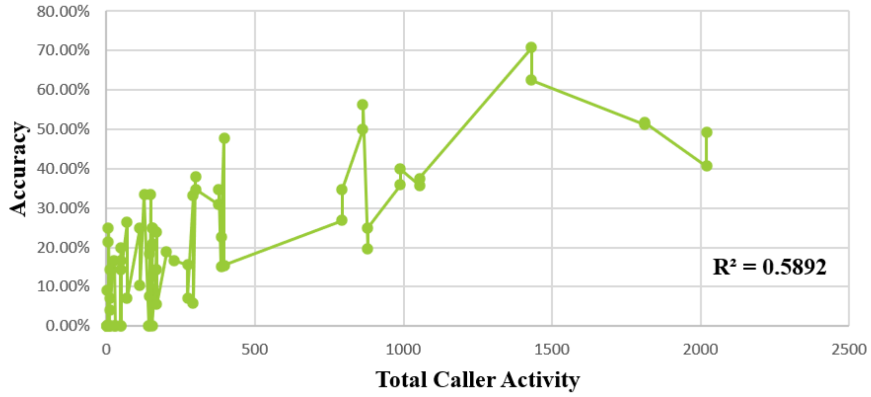


Figure 3.14: Aggregated accuracy with caller activity levels for weekends

$$\text{Accuracy of travelling hour} = \frac{\begin{array}{r}\text{Total number of accurately identified}\\ \text{travelling hours from CDR per day}\end{array}}{\begin{array}{r}\text{Total number of travelling hours}\\ \text{identified from GPS}\end{array}} \tag{3.6}$$

$$\begin{array}{r}\text{Accuracy of}\\ \text{non travelling hour}\end{array} = \frac{\begin{array}{r}\text{Total number of accurately identified}\\ \text{non travelling hours from CDR per day}\end{array}}{\begin{array}{r}\text{Total number of non travelling}\\ \text{hours identified from GPS}\end{array}} \tag{3.7}$$

Obtained results for weekdays are shown in Table 3.9 and Figure 3.15. Results for weekends are shown in Table 3.10 and Figure 3.16.

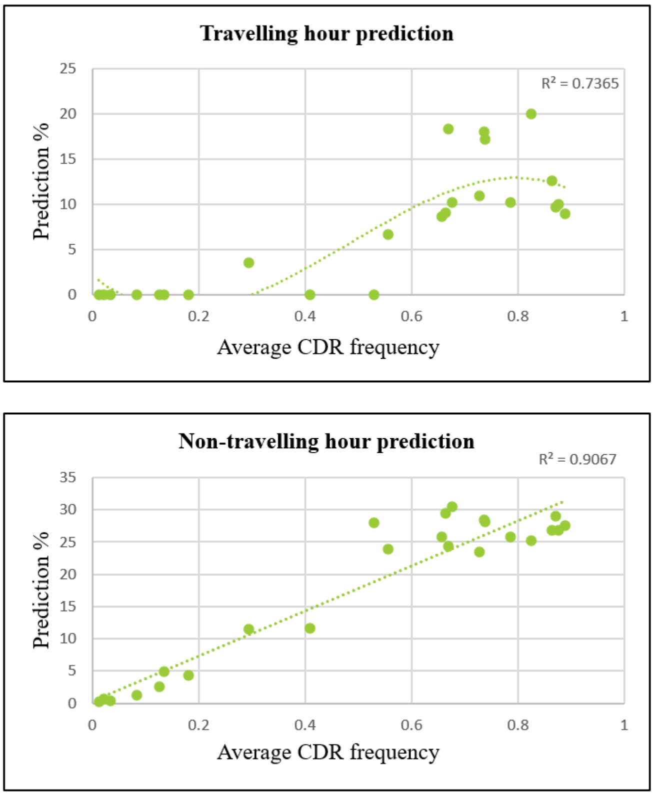


Figure 3.15: Travelling and non-travelling hour prediction accuracy with average frequency of CDR for weekdays

Accuracy of travelling and non-travelling hours identification

When we examined the results obtained, accuracy for traveling hour and non-traveling hour prediction were below 40% (Figure 3.17). This was due to not having call records for some hours and having a very limited number of calls for some people. Hence, we used the active time window method as described in section 3.5.3.1 to obtain traveling hour and non travelling hour predictions. This gave a significant accuracy increase and the results were used as input for the activity recognition model.

Average travel distance analysis

Analysis results of Household visit survey data (Figure 3.18) indicates that those who use the private and non-motorized modes travel less than people who

Table 3.9: Analysis of travelling and non-travelling hours for weekdays

Hour of the day	**Prediction (%)**		**Average CDR frequency**
	Travelling hour	**Non-travelling hour**	
0	0.00	2.56	0.13
1	0.00	0.40	0.03
2	0.00	0.20	0.01
3	0.00	1.30	0.08
4	0.00	0.69	0.02
5	20.00	25.23	0.83
6	0.00	4.38	0.18
7	3.56	11.44	0.29
8	6.67	23.95	0.56
9	10.00	26.89	0.88
10	10.25	30.54	0.68
11	18.00	28.36	0.74
12	9.71	29.02	0.87
13	9.00	27.55	0.89
14	10.25	25.82	0.79
15	0.00	27.93	0.53
16	8.67	25.77	0.66
17	12.58	26.80	0.86
18	9.05	29.40	0.66
19	17.26	28.11	0.74
20	18.33	24.28	0.67
21	11.00	23.53	0.73
22	0.00	11.70	0.41
23	0.00	4.93	0.14
Average Accuracy	**7.26**	**18.37**	

travel by public transportation means such as buses and railways. The distribution of transport mode shows that the public transport means are the major way of transportation when the distance is 4 km or longer. Further detail of trip distance by mode shows that railway users travel 25 km in average as the longest trip. Travellers by bus or cars have average travel distances of 9 km and 8 km respectively. Once the trip distance by mode is plotted with Google Activity Recognition API data, user filled data data and mobile app data (Figure 3.19) the same pattern in the curves could be observed where the public transport users indicate a higher distance of travel than the other private and non-motorized

Table 3.10: Analysis of travelling and non-travelling hours for weekends

Hour of the day	**Prediction (%)**		**Average CDR frequency**
	Travelling hour	**Non-travelling hour**	
0	0.00	0.30	0.05
1	0.00	0.92	0.02
2	0.00	0.89	0.01
3	0.00	1.92	0.01
4	0.00	0.00	0.02
5	0.00	1.92	0.02
6	16.67	5.40	0.14
7	8.33	7.81	0.25
8	0.00	17.61	0.36
9	12.50	16.17	0.33
10	4.17	17.79	0.60
11	12.50	19.08	0.81
12	12.50	24.98	0.69
13	18.06	21.36	0.44
14	3.13	18.79	0.36
15	0.00	12.01	0.36
16	10.71	19.13	0.55
17	13.10	11.14	0.44
18	19.58	17.19	0.67
19	16.67	22.64	0.49
20	10.00	19.10	0.50
21	16.67	14.17	0.35
22	0.00	8.54	0.19
23	0.00	3.77	0.11
Average Accuracy	**7.27**	**11.78**	

users.

Activity generation time window analysis

Figure 3.20 show a comprehensive hourly fluctuation of trips in between home

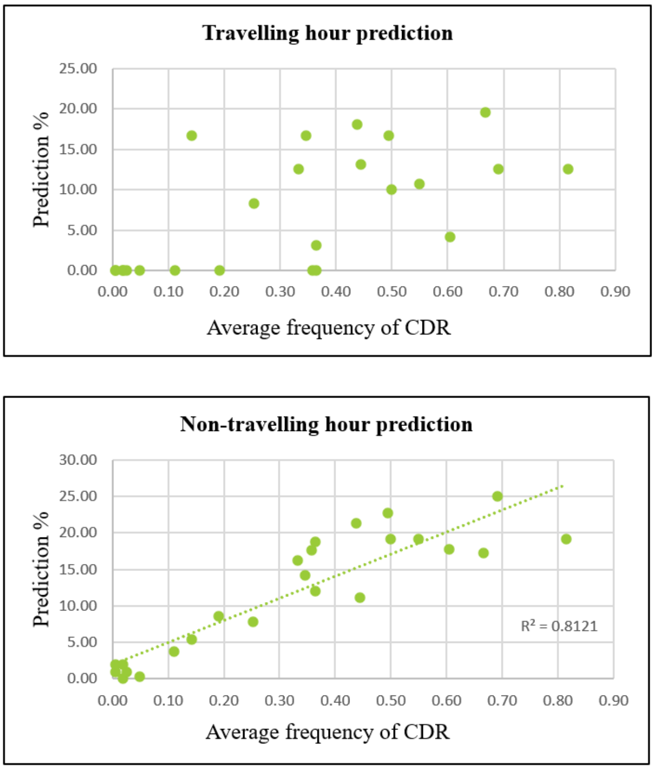


Figure 3.16: Travelling and non-travelling hour prediction accuracy with average frequency of CDR for weekends

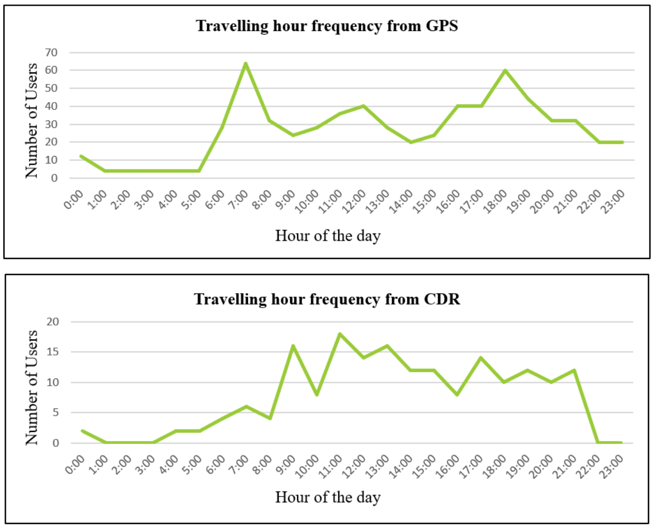
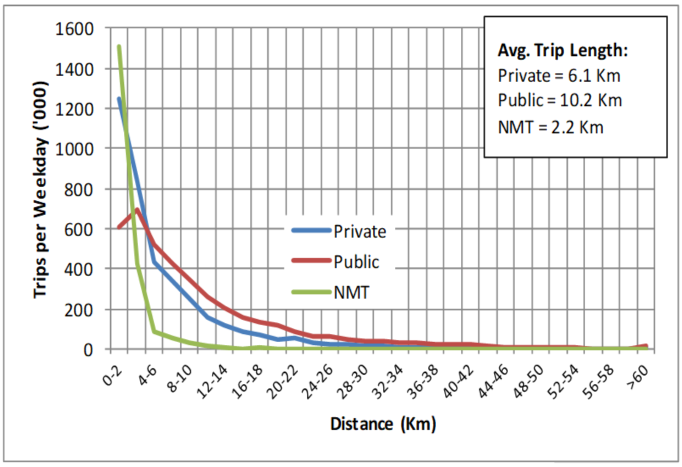


Figure 3.17: Accuracy of travelling and non-travelling hours



Source: Urban Transport System Development Project for CMR and suburbs

Figure 3.18: Average travel distance based on HVS data

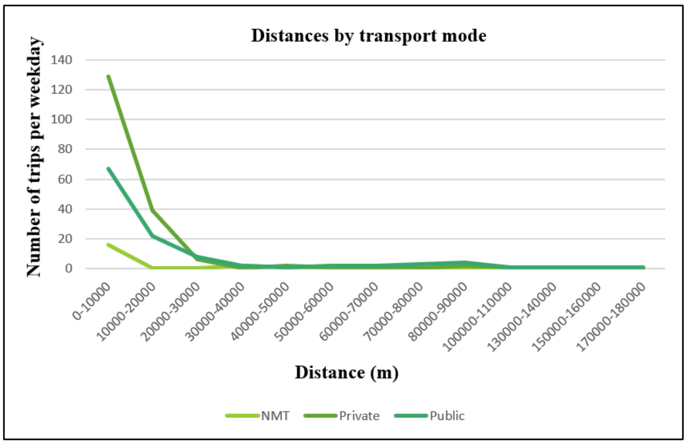


Figure 3.19: Average travel distance based on CDR data

to work, home to education and vice versa. The pattern of the fluctuations are much similar to data derived from mobile app, but the education trips present a peak which is not indicated in HVS patterns. The peak is assumed to be due to the travelling taking place into private classes.

3.5.4 Activity Recognition From CDR Data

This section involves activity recognition from features described in Section 3.5.3. Hidden Markov Models (HMMs) are used as the mainstream algorithm to recognize the activity pattern using CDR data. But, the main problem of common HMMs is the use of homogeneous transition and emission probabilities. This

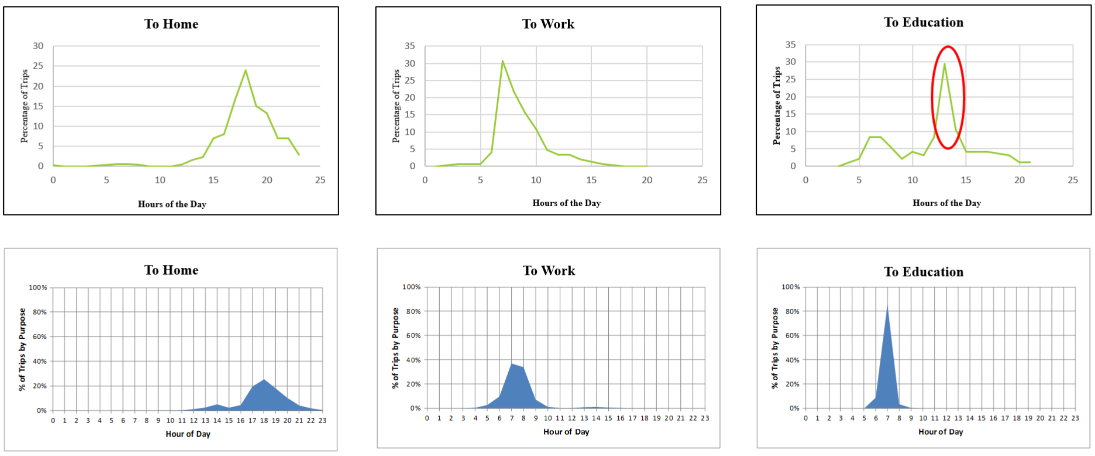


Figure 3.20: Activity generation time-window analysis for CDR and HVS data

causes drawbacks in the process of activity recognition using CDRs since current activity mostly depends mostly on previous activity, time and other contextual information. To overcome this issue, Input/Output Hidden Markov Models (IO-HMM) proposed by Yoshua Bengio et al. [48] are used to identify the activity patterns of the users. Probability distribution over a discrete state dynamical system is considered in IO-HMMs. The system is generally based on the following state space description.

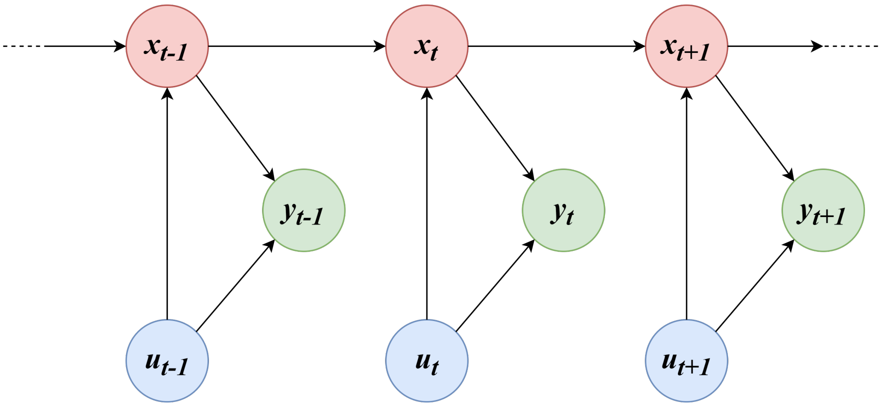


Figure 3.21: IO-HMM architecture

In Figure 3.21 u_t is defined as the input variables while x_t is the latent categorical (hidden state) variables and y_t is introduced as output variables [49].

$$x_t = f(x_{t-1},\ u_t) \tag{3.8}$$

$$y_t = g(x_t,\ u_t) \tag{3.9}$$

Where $x_t \in V = \{1, 2, \ldots, n\}$ is a discrete state, $u_t \in \mathbf{R}^m$ is the input vector at time t, and $y_t \in \mathbf{R}^r$ is the output vector. Here f and g refers to state transition function and output function respectively.

A joint probability P$(u_1^T,\ x_1^T,\ y_1^T)$ exists in the probabilistic interpretation of the IO-HMM which involves Input, State and Output random variables. Compute transition probabilities and the expected output value [49] is given by φ, $n_{i,t}$ respectively;

$$\varphi_{ij,t} = P(x_t = i \mid x_{t-1} = j, u_t) \tag{3.10}$$

$$n_{i,t} = E(y_t \mid x_t = i, u_t) \tag{3.11}$$

Forward recursion ($\alpha_{i,t}$) and Backward recursion ($\beta_{i,t}$) as follows [49];

$$\alpha_{i,t} = P(y_t \mid x_t = i, u_t) \sum_l \varphi_{il}(u_t)\alpha_{l,t-1} \tag{3.12}$$

$$\beta_{i,t} = \sum_l P(y_{t+1} \mid x_{t+1} = l, u_t)\varphi_{li}(u_{t+1})\beta_{l,t+1} \tag{3.13}$$

Transition posterior probabilities ($h_{ij,t}$) and state posterior probabilities ($g_{i,t}$) can be expressed as follows [49];

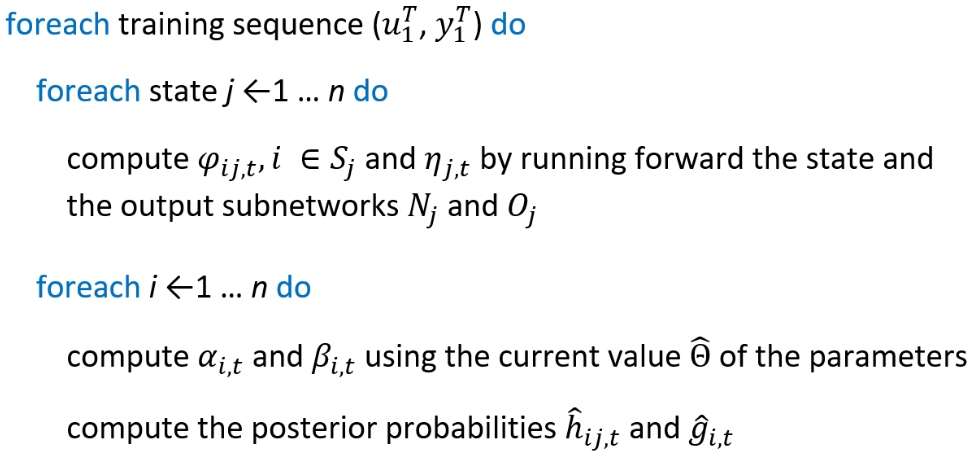
$$h_{ij,t} = \frac{P(y_t \mid x_t = i, u_t)\alpha_{j,t-1}\beta_{i,t}\varphi_{ij}(u_t)}{L} \tag{3.14}$$

$$g_{i,t} = \frac{\alpha_{i,t}\beta_{i,t}}{L} \tag{3.15}$$

Initial parameters, transition parameters, and emission parameters are the three sets of unknown parameters within the IO-HMM architecture.

Initial Probability and Transition Probability are obtained through the Multinomial logistics regression. Parameters of IO-HMM are estimated through the Expectation-Maximization (EM) approach [49]. IO-HMM is trained using the following algorithm [49]. It contains three main steps.

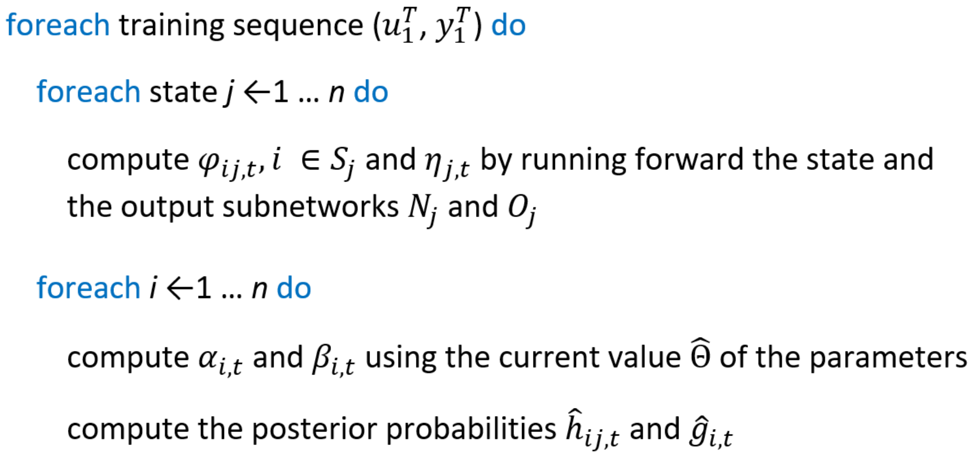
First Step:- Estimation step (Figure 3.22)



```
foreach training sequence $(u_1^T, y_1^T)$ do
  foreach state $j$ ←1 … $n$ do
    compute $\varphi_{ij,t}, i \in S_j$ and $\eta_{j,t}$ by running forward the state and
    the output subnetworks $N_j$ and $O_j$
  foreach $i$ ←1 … $n$ do
    compute $\alpha_{i,t}$ and $\beta_{i,t}$ using the current value $\widehat{\Theta}$ of the parameters
    compute the posterior probabilities $\hat{h}_{ij,t}$ and $\hat{g}_{i,t}$
```

Figure 3.22: Estimation step

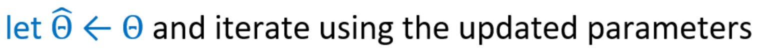
Second Step:- Maximization step (Figure 3.23)



```
foreach training sequence $(u_1^T, y_1^T)$ do
  foreach state $j$ ←1 … $n$ do
    compute $\varphi_{ij,t}, i \in S_j$ and $\eta_{j,t}$ by running forward the state and
    the output subnetworks $N_j$ and $O_j$
  foreach $i$ ←1 … $n$ do
    compute $\alpha_{i,t}$ and $\beta_{i,t}$ using the current value $\widehat{\Theta}$ of the parameters
    compute the posterior probabilities $\hat{h}_{ij,t}$ and $\hat{g}_{i,t}$
```

Figure 3.23: Maximization step

Third Step:- Updated parameters (Figure 3.24)



```
let $\widehat{\Theta} \leftarrow \Theta$ and iterate using the updated parameters
```

Figure 3.24: Parameters update step

Chapter 4

RESULTS

This chapter discusses the results generated by this study. The study used Mobile app data for user profiling and activity recognition result validation. For Load Sharing Effect identification and User localization results validation, the study used both Mobile app data and HVS data.

4.1 User Profiling

Results of the User Profiling models are listed in Table 4.1. The SVM_{WC} model outputs the highest accuracy for user profiling among the models generated in the study. Though there is a class imbalance problem in the data set since the study used the SVM_{WC} model which is sensitive to the class imbalance, the results generated had the best possible accuracy.

Table 4.1: User profiling results comparison (Precision, Recall and F1 Score)

Models	**Random Forest**			**Artificial Neural Network**			**SVM$_{WC}$**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Full-Time Employees	0.64	0.53	0.58	0.70	0.44	0.54	0.71	0.66	**0.68**
Part-Time Employees	0.45	0.39	0.42	0.55	0.40	**0.46**	0.50	0.43	**0.46**
Student	0.70	0.91	0.79	0.71	0.93	0.81	0.77	0.95	**0.85**
Housewife	0.68	0.72	0.70	0.72	0.78	0.75	0.71	0.81	**0.76**
Retired	0.32	0.21	0.26	0.36	0.24	0.28	0.31	0.29	**0.30**
Others	0.52	0.64	0.57	0.56	0.36	0.44	0.65	0.53	**0.58**

4.2 Analysis on Load Sharing Effect Identification

Coming into the load sharing identification, results were derived from the CDR data and mobile app data while they were validated from the mobile app data and HVS data.

In-order to validate results, initially we define Load shared Records. If there is a change in the last connected cell tower without an actual movement of 100m of the user, study considered it as a load shared record.

Accuracy results for corresponding speeds using mobile data are shown in the Table 4.2. Our approach shows a significant accuracy improvement, particularly in recall. The load shared records can be clearly identified when the speeds between two consecutive records are high. But there are a large number of records which are load shared but the speeds lower than required. The proposed methodology specifically captures records of such types, since the signal strength and the previous records are also taken into consideration.

Table 4.2: Load sharing record identification results comparison with mobile app data

Method	**F1 Score**								
	30	**40**	**60**	**80**	**100**	**120**	**140**	**180**	**200**
Pre-define speed base filter	0.14	0.28	0.19	0.17	0.11	0.06	0.03	0.03	0.03
Proposed Methodology	**0.54**	**0.79**	**0.61**	**0.57**	**0.51**	**0.46**	**0.38**	**0.38**	**0.38**

Further, error percentages were also calculated by validating with the HVS data for different speed limits. 40 kmph was identified as the optimum speed, such that the error is minimized. Results are shown in the Figure 4.1.

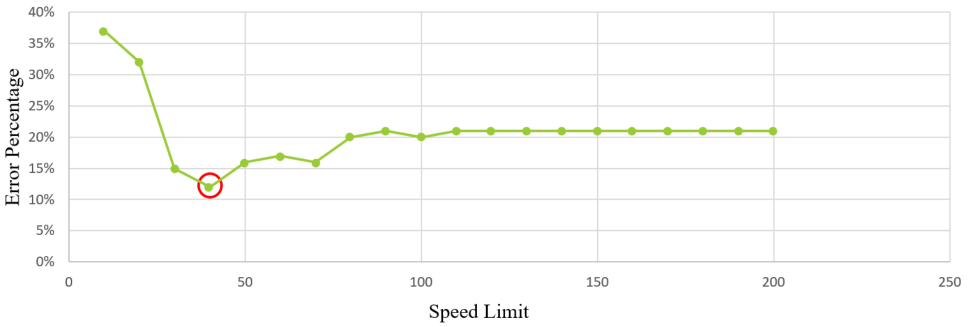


Figure 4.1: Speed vs Error after removing load sharing effect for CDR dataset

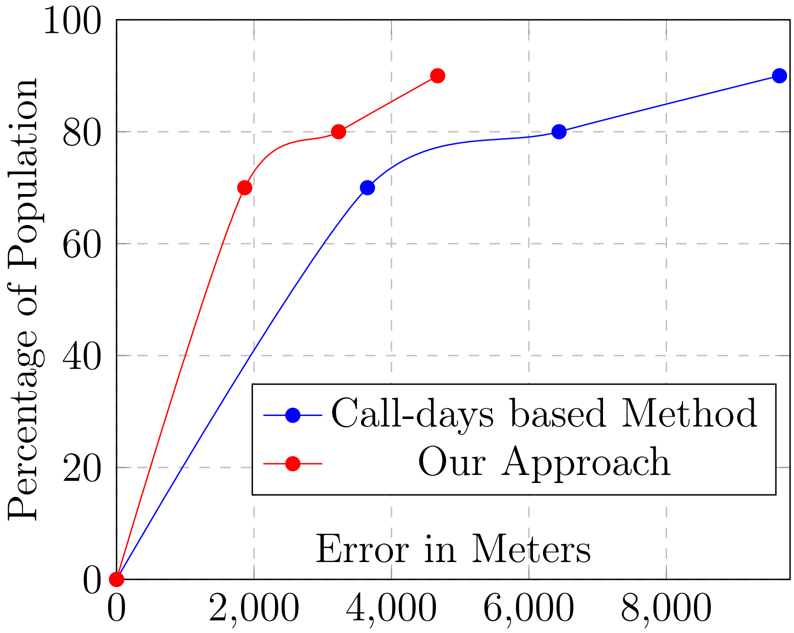


Figure 4.2: Home localization error

4.3 User Localization

Two approaches were used to evaluate the Localization results. Results were evaluated using voluntarily collected mobile app data as the first step. First, the user's home location provided by the user was taken to determine home location error and compared against the location predicted by our model. Generated results are shown in Figure 4.2. The same process was carried out to identify user work location and the Figure 4.3 shows the generated results. Our proposed methodology outperforms results of previous studies for the majority of the population. Particularly, work locations are more accurately determined than home locations. As per the analysis, 70% (468) of the users working in an urban area are observed to have their home locations in suburban areas. Due to the different cell tower densities where the urban areas have high cell tower density while suburban areas have low density, the work location identifications are more accurate compared to the home location.

Localization result evaluation was carried out as the next step, where O-D matrices [50] were created from the large CDR dataset consisting of 13 million subscribers and, the weekends and holidays were eliminated from the CDR dataset

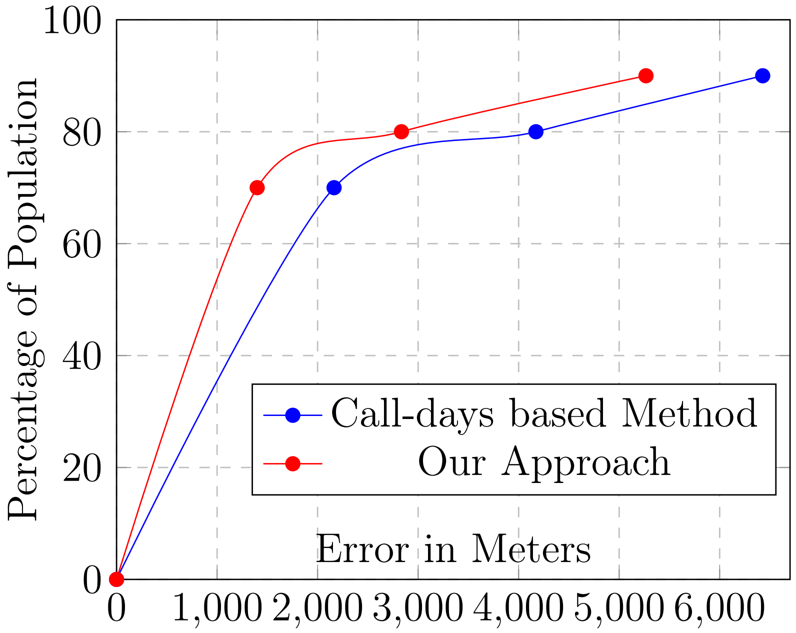


Figure 4.3: Work localization error

in the preprocessing stage. Home activities and work activities were defined between 8pm - 5am [27] and 10am to 12pm and 1pm to 4pm [44] respectively and the generated results were analysed with Household Visit Survey 0-D matrix data.

Table 4.3 depicts derived the O-D matrices from HVS data in separate levels. This research has been conducted by covering the Western province consisting of three districts namely Colombo, Kalutara and Gampaha. For instance, cell 1 indicates that 44% of users have both their residences and work places in Colombo district. Also, cell 2 indicates that 3% of users have their residences in Colombo district and work places in Gampaha district.

Table 4.4 shows the localization of users according to the number of appeared days in a particular cell tower [43]. It was derived from the CDR dataset, which consists of 4 billion CDR records of users. This same interpretation as in Table 4.3.

The user localization according to our methodology which employed signal strength, load shared records of users and number of appearance days, is showed in Table 4.5. It also has the same interpretation as Table 4.3.

O-D matrices derived from aforesaid three methods were statistically analysed with ground truth (HVS data) using Pearson's chi-square test showed in Equation 4.1.

$$\tilde{\chi}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.1}$$

χ refers to the Pearson's cumulative test statistic, where the E_i denotes the Expected value, O_i the Observation value and n is the number of cells in the table. According to the Chi-square calculation, p values of Table 4.4 and Table 4.5 are 0.64 and 0.88 respectively. Our methodology outperforms current state of the art methods and there is no statistical difference between the O-D matrix created by our method and O-D matrix generated from Household Visit Survey dataset.

Table 4.3: Home work distribution - HVS data

Home/Work		Trip Attractors		
		Colombo	Gampaha	Kalutara
Trip Generators	Colombo	44%	2%	1%
	Gampaha	10%	27%	1%
	Kalutara	4%	1%	10%

Table 4.4: Home work distribution - Call days based method

Home/Work		Trip Attractors		
		Colombo	Gampaha	Kalutara
Trip Generators	Colombo	52%	2%	2%
	Gampaha	12%	18%	1%
	Kalutara	5%	1%	7%

Table 4.5: Home work distribution - Proposed methodology

Home/Work		Trip Attractors		
		Colombo	Gampaha	Kalutara
Trip Generators	Colombo	44%	3%	2%
	Gampaha	8%	30%	1%
	Kalutara	1%	1%	10%

4.4 Activity Recognition

Results of Activity Recognition Models are depicted below in Table 4.6. As you can see, the IO-HMM model gives the highest accuracy compared to other models. Primary activities recognition shows significant accuracy compared to Secondary activities recognition. In the analysis, Home and Work trips are considered as Primary activities. While Education, Shopping, Social, Private matters and other trips are considered as Secondary activities. Table 4.7 shows the aggregated Primary and Secondary activities recognition accuracy.

Table 4.6: Activity recognition models accuracy

Activities	**Model**	
	HMM	**IO-HMM**
Home	73.9%	**78.7%**
Work	80.5%	**88.1%**
Education	72.6%	**88.2%**
Shopping	63.3%	**81.8%**
Social	55.0%	**72.4%**
Private matters	21.7%	**39.6%**
Others	46.9%	**61.5%**

Table 4.7: Primary and Secondary activities recognition accuracy

Model	**Accuracy**	
	Primary Activities	**Secondary Activities**
HMM	77.2%	51.9%
IO-HMM	**83.4%**	**68.7%**

Chapter 5

CONCLUSION AND FUTURE WORK

Since CDR has been continuously identified for a decade there is a lot of research done in the literature on different aspects including urban planning, geography etc. In this thesis, the CDR phenomenon is analyzed to create value in the transportation sector. Current chapter summarizes the findings of the study in three directions including the methodological framework of using CDR for transport initiatives and applications of CDR based findings which is introduced as the labeling the secondary activity. Final section of the chapter describes the several ways to expand the current study in future work.

5.1 Research Summary

This study focuses on different uses of CDR in transportation, including mobility pattern identification and validation of the findings. Specifically, the research questions the thesis addresses include, minimizing localization error and the labeling of secondary activity locations. Further , study develops activity pattern recognition machine learning models for different user categories, identified based on user profiling.

5.1.1 Methodological Framework of Using CDR

The first part of the study develops a methodology through the reduction of the localization error to minimize the load sharing effect. To identify different regions of stay for the user (Stay location clusters), CDRs were clustered based on their spatial behaviour. DBSCAN algorithm was used to perform the clustering. The study identifies that the assignment of a cell tower to a user depends, among others, on signal strength, user location and the call traffic load of the particular cell tower in order to minimize the localization error. Therefore, we propose a novel methodology considering the load sharing records, signal strength of the particular cell tower and number of days appeared in the particular cell tower, in order to minimize the localization error.

The localization error was evaluated in two approaches. Earlier, the results are evaluated against mobile app data. To calculate home location error, we compared the location estimated by the developed methodology with the home location given by the user. The same procedure was followed to identify user's work location. The proposed methodology outperforms results of previous studies for the majority of the population. The results of study work location identification are more accurate than the results of home location identification. O-D matrices were generated from the large (13 million subscribers) CDR dataset as the second approach to evaluate the Localization results. Home activities were defined between the time period of 8pm to 5am and work activities are from the time period of 10am to 12pm and 1pm to 4pm. Then, the study results were then analysed and compared with HVS O-D matrix data.

5.1.2 Labeling the Secondary Activities

In this section an Input-output Hidden Markov model had been trained with the CDR data for the labeling of the secondary activities. Accordingly Weekend or Weekday tagging (binary format), Travelling or non-travelling hours, Average travel distance, activity generated time window had been inserted as model inputs and the Distance between current stay location and user home, Distance between current location and user work location, Duration of the activity and Frequency of user visits to the particular location were derived as the model outputs.

5.1.3 Limitations of the Study

Use of CDR in transport context is an innovative method in Sri Lankan context, therefore very limited number of researches are available for the literature review. Most of the previous research has been in developed countries and not in developing countries like Sri Lanka. Using the exact parameters involved in other countries will decrease the accuracy of the results as the socio-economic backgrounds of the countries are different to each other.

Data sparsity is one of the main limitations embedded in CDR data, Due to this the data points are not continuous which might result in incorrect behavioral interpretations.

Another main limitation is the limited sample size of the data. Since the consent for the app installation is low, the total number of data generated are low. Due to this limited availability of data the model accuracy becomes low.

5.2 Future Research

There are many opportunities to extend this research further. Labeling the rich sample of secondary activities are initial points of human mobility recognition. Research can be extended further to different directions in data fusion. Other than Call Detail Records there are other types of big data sources like GPS, Point of interest (POI) data and etc. More elaborate and unique human mobility insights can be identified by employing these data with more sophisticated data mining techniques.

References

[1] Constantinos S Papacostas. Fundamentals of transportation engineering. 1987.

[2] Marcos R Vieira, Enrique Frías-Martínez, Petko Bakalov, Vanessa Frías-Martínez, and Vassilis J Tsotras. Querying spatio-temporal patterns in mobile phone-call databases. In *2010 Eleventh International Conference on Mobile Data Management*, pages 239–248. IEEE, 2010.

[3] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[4] Margus Tiru. Overview of the sources and challenges of mobile positioning data for statistics. In *International Conference on Big Data for Official Statistics. United Nations Statistics Division (UNSD) and National Bureau of Statistics of China*, pages 28–30, 2014.

[5] Michael Meyer and Eric J Miller. Urban transport plan, 2000.

[6] Paul H Wright and Norman J Ashford. *Transportation engineering: planning and design*. 1989.

[7] Mark A Beyer and Douglas Laney. The importance of 'big data': a definition. *Stamford, CT: Gartner*, pages 2014–2018, 2012.

[8] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58:240–250, 2015.

[9] Tapas Saini, Kruti Barot, Amritanshu Sinha, Rajesh Gogineni, Rajesh Krishnan, Venkata Srikanth, Shikha Sinha, and Rakesh Behera. Estimating origin-destination matrix using telecom network data.

[10] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology*, 17(1):3–27, 2010.

[11] Ying Zhang. User mobility from the view of cellular data networks. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 1348–1356. IEEE, 2014.

[12] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[13] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–219, 2017.

[14] Cory M Krause and Lei Zhang. Short-term travel behavior prediction with gps, land use, and point of interest data. *Transportation Research Part B: Methodological*, 123:349–361, 2019.

[15] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5):275–286, 2003.

[16] Lin Liao, Donald J Patterson, Dieter Fox, and Henry Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311–331, 2007.

[17] Marcela A Munizaga and Carolina Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18, 2012.

[18] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings*

of the 19th international conference on World wide web, pages 1029–1038, 2010.

[19] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*, 33(5):727–748, 2006.

[20] Nina Glick Schiller, Linda Basch, and Cristina Szanton Blanc. From immigrant to transmigrant: Theorizing transnational migration. *Anthropological quarterly*, pages 48–63, 1995.

[21] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences*, pages 995–1004. IEEE, 2013.

[22] MKDT Maldeniya, Sriganesh Lokanathan, and Amal S Kumarage. An assessment of mobile network big data-based insights for transport planning in sri lanka. *Colombo, Sri Lanka 3 rd and 4 th June, 2016*, 2016.

[23] Manoranjan Dash, Kee Kiat Koo, James Decraene, Ghim-Eng Yap, Wei Wu, Joao Bartolo Gomes, Amy Shi-Nash, and Xiaoli Li. Cdr-to-movis: Developing a mobility visualization system from cdr data. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1452–1455. IEEE, 2015.

[24] Ziliang Zhao, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Ling Yin. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9):1738–1762, 2016.

[25] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2:1001, 2012.

[26] Erik Mellegard, Simon Moritz, and Mohamed Zahoor. Origin/destination-estimation using cellular network data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 891–896. IEEE, 2011.

[27] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6), 2014.

[28] Yan Leng et al. *Urban computing using call detail records: mobility pattern mining, next-location prediction and location recommendation*. PhD thesis, Massachusetts Institute of Technology, 2016.

[29] Olle Järv, Rein Ahas, and Frank Witlox. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135, 2014.

[30] Victor Powell and L Lehe. Principal component analysis. *URL http://setosa.io/ev/principalcomponent-analysis*, 2015.

[31] Samiul Hasan, Christian M Schneider, Satish V Ukkusuri, and Marta C González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2013.

[32] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, (4):36–44, 2011.

[33] Sri lanka telecom - annual report. 2016.

[34] Gabriel Kreindler and Yuhei Miyauchi. Commuting and productivity: Quantifying urban economic activity using cell phone data, 2015.

[35] Kaushalya Madhawa, Sriganesh Lokanathan, Rohan Samarajiva, and Danaja Maldeniya. Understanding communities using mobile network big data cprsouth 2015. 2015.

[36] Kaushalya Madhawa, Sriganesh Lokanathan, Danaja Maldeniya, and Rohan Samarajiva. Using mobile network big data for land use classification. In *Communication Policy Research South Conference*, 2015.

[37] Rohan Samarajiva. Policy commentary: mobilizing information and communications technologies for effective disaster warning: lessons from the 2004 tsunami. *New Media & Society*, 7(6):731–747, 2005.

[38] Danaja Maldeniya, Amal Kumarage, Sriganesh Lokanathan, Gabriel Kreindler, and Kaushalya Madhawa. Where did you come from?: where did you go?; robust policy relevant evidence from mobile network big data. 2015.

[39] Danaja Maldeniya, Sriganesh Lokanathan, and Amal Kumarage. Origin-destination matrix estimation for sri lanka using mobile network big data. In *Proceedings of the 13th International Conference on Social Implica tions of Computers in Developing Countries, Negombo, Sri Lanka*, 2015.

[40] Zhan Zhao, Jinhua Zhao, and Haris N Koutsopoulos. Individual-level trip detection using sparse call detail record data based on supervised statistical learning. In *Proc. Transp. Res. Board 95th Annu. Meeting*, pages 1–18, 2016.

[41] Balázs Cs Csáji, Arnaud Browet, Vincent A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications*, 392(6):1459–1473, 2013.

[42] Ming-Heng Wang, Steven D Schrock, Nate Vander Broek, and Thomas Mulinazzi. Estimating dynamic origin-destination data and travel demand using cell phone network data. *International Journal of Intelligent Transportation Systems Research*, 11(2):76–86, 2013.

[43] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying

important places in people's lives from cellular network data. In *International Conference on Pervasive Computing*, pages 133–151. Springer, 2011.

[44] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, pages 1–9, 2013.

[45] Kentaro Toyama and Ramaswamy Hariharan. Modeling location histories, May 18 2010. US Patent 7,720,652.

[46] Urban transport system development project for colombo metropolitan region and suburbs. 2014.

[47] Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: parsing and modeling location histories. In *International Conference on Geographic Information Science*, pages 106–124. Springer, 2004.

[48] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. In *Advances in neural information processing systems*, pages 427–434, 1995.

[49] Yoshua Bengio and Paolo Frasconi. Input-output hmms for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249, 1996.

[50] Yi Zhang, Xiao Qin, Shen Dong, and Bin Ran. Daily od matrix estimation using cellular probe data. In *89th Annual Meeting Transportation Research Board*, volume 9, 2010.