UNDERSTANDING TRAVEL TO WORK ATTRIBUTES USING MOBILE NETWORK BIG DATA

(STUDY AREA: WESTERN PROVINCE)

N. K. Bhagya Jeewanthi

(168046N)

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Department of Transport and Logistics Management

University of Moratuwa

Sri Lanka

February 2018

DECLARATION OF ORIGINALITY

I declare that this is my own work and this thesis 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.

Signature: .... Date: ....

COPY RIGHT STATEMENT

I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/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: …………………….

STATEMENT OF THE SUPERVISOR

The above candidate has carried out the research for the Degree of Master of Science under my supervision.

Name of the supervisor: …………………………….

Signature of the supervisor: .... Date: ....

ABSTRACT

As people become more mobile, urban traffic patterns become more complex, creating a need for more continuous transportation planning processes. Currently, manual and online surveys are the primary source for such analysis. However, such data collection while being expensive, takes time and is often outdated by the time it is made available for analysis. Mobile Network Big Data (MNBD) which concerns large data sets has the potential to supplement such traditional data sampling programs. Call Detail Records (CDR) which is a subset of MNBD is readily available as most of the telecommunication service providers maintain such data. Thus, analyzing CDR leads to an efficient identification of human behavior and location.

This research uses the CDRs of nearly 10,000 mobile phone users in the Western Province (WP) of Sri Lanka for a period of three months for the analysis of their caller locations in order to determine their mobility patterns. In analyzing the CDRs, the frequency of making calls from a specific location is identified, classified them into potential home and non-home locations based on the regularity and time of day and week these calls were generated from each such location. Users are thereby categorized hierarchical levels based on the regularity of presence within the study area, identification of province of residence and by typical employment categories across the sample of users.

An estimate of home-based work trips made within the Western Province was identified using the CDR and validated by comparing with the origin-destination matrix for work trips calculated from an extensive survey of 35,000 households under the CoMTrans Study (JICA, 2014) obtaining a fit of 76%. The successful validation and the identification of sources of errors in CDR data provides direction for further research in using CDRs for travel estimation and the identification of the appropriate comprehensive data mining techniques.

Key words – Mobile Network Big Data, Call Detail Record, Home-based Work trips

ACKNOWLEDGEMENT

Prima Facie, I would like to offer my heartfelt gratitude to my research supervisor, Senior Prof. Amal S. Kumarage. I have been fortunate to have a supervisor like him who guided me from start to the end; in the meantime, followed up my progress continuously and guided me with a bunch of advices in each step of the research. I am very grateful for your encouragement and continuous attention throughout the research.

Special thanks should go to the post-graduate research coordinator Dr. Varuna Adhikariwattage and former coordinator Dr. Indika Sigera, for conducting research progress reviews and the support given through the process.

I would like to express my profound gratitude to Dr. T. Sivakumar, former Head of the Department of Transport and Logistics Management of University of Moratuwa and the current Head of the department. Mr. S. N. Bentotage for the opportunity created.

A very special gratitude goes to Sriganesh Lokanathan (Team leader, LirneAsia), Chairman of the review panel for lending me the data. I also thank Isuru Jayasooriya, Danaja Maldeniya (Researchers - LirneAsia) including Sriganesh Lokanathan who immensely contributed in the stages of primary data processing and analyzing. Specially thank you all for keeping trust on me, in providing a data and the guidance given from start to the end. I gratefully acknowledge the funding received towards my M.Sc. from the University of Moratuwa by the SRC grant (SRC/ST/2016/14) for "Mobility studies from mobile big data for Colombo".

I am sincerely thankful to all the staff members and the non-academic staff of the Department of Transport and Logistics Management. I am also thankful to my family members, colleagues, senior colleagues, and friends who supported me in many ways throughout the research. Encouragements, advices, and opinions of yours are really appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MNBD	Mobile Network Big Data
CDR	Call Detail Records
VLR	Visitor Location Register
CA	Caller Activities
HVS	Household Visit Survey
WP	Western Province
DSD	Divisional Secretariat Divisions
GND	Grama Niladhari Divisions
BTS	Base Transceiver Station
O- D MATRIX	Origin – Destination Matrix

CHAPTER 1 - INTRODUCTION

1.1 Background

The major movements within urban areas are related to the travel undertaken by people. Water-based urban transport is found only within few cities of the world and air transport is unsuited 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. In the Sri Lankan context almost 100% of the urban travel are land-based, thus a significant attention should be given in planning and implementing proper transport initiatives.

Human mobility patterns are critical sources of information for designing, analyzing and enhancing transport planning activities. These transport initiatives require mobility rich data to become a continuous process. Currently, manually carried out roadside surveys or household surveys done once every few years provide input data for such initiatives. Such data collection while being expensive, takes time and is often outdated by the time it is available for analysis. Those drawbacks of the existing methodologies generate the necessity for more advanced data extraction techniques to analyze human motion.

The recent adoption of ubiquitous computing technologies, including mobile devices by very large portions of the world population has enabled the capturing of large-scale spatiotemporal data about human motion (Marcos, et al., 2010). Large-scale penetration of mobile phones serves as a dynamic source to derive insights on human mobility, due to the generation of big data. Big data plays an important role in today's world in understanding the more and more social processes.

Mobile Network Big Data (MNBD) concern large-volume, complex, growing data sets derived from the way people use communication devices (Iqbal, Charisma, Choudhury, Wangb, & Gonzalez, 2013). These Mobile Network Big Data have the ability to locate mobile devices in time and space with a certain accuracy using the mobile network infrastructure that includes location information of the mobile device (Tiru, 2014). Such data collected through phone networks, GPS or Bluetooth had been used in a wide range of mobile phone applications including location services, navigation, tracking and etc. In the present scenario Call Detail Records (CDR) data, which is collected by mobile phone carriers for billing purposes is the most common type of MNBD used in a variety of transport studies.

The current study focuses on understanding the behavioral patterns of mobile users within the study area, i.e. the Western Province of Sri Lanka. CDR data were used by transforming its attributes to behavioral attributes and home visit survey data which is an important basis for model development is used to conduct the validation of CDR findings. Validation ensures that the CDR data emphasize a significant relationship with the actual human mobility insights while adhering to the basic concepts in transportation.

Mobility analysis provides outputs that align with some of the intermediate and final results of the traditional forecasting framework. The section below draft the basic concepts of transportation along with the traditional framework, which acts as the foundation for deriving the CDR outputs.

1.2 Introduction to basic transport concepts

Travel has always been viewed as derived from the demand for activity participation (McNally & Mickael, 2007). Users of transport or the trip makers consume the specific service, not because of its direct benefit, but to fulfill other different purposes like work, school, recreation and etc. All these movements have a rational behavior which makes it predictable and takes place in a finite space or a closed system.

1.2.1 Identifying trips

A trip, in general terms defined as a single journey made by an individual between two points by a specific mode of travel and for a defined purpose (Planning Tank, 2017) and in most cases the basic unit of travel behavior. An individual’s general movements can be depicted as in Figure 1.1. Accordingly, the individual travel for the purpose of work with different modes. Also, there are multiple stops in between the travel to work. Still, this is considered as one trip for a defined purpose.

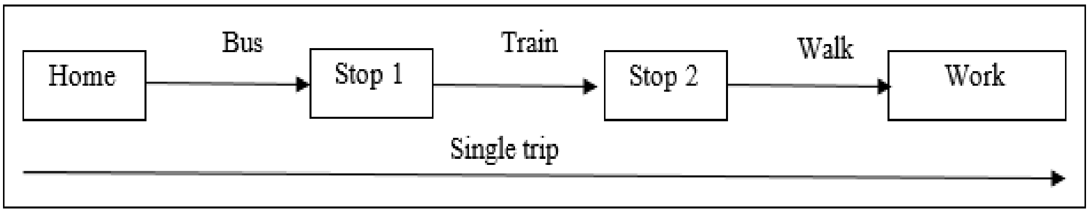


Figure 1.1: Concept of a trip

As stated earlier, each trip is undertaken to satisfy different purposes. People move for a grate variety of reasons, they got to work, they go for shopping, take children to school, pop out for lunch and etc. In simple terms, people move for a multitude of reasons which results in extra-ordinary crisscross patterns in their journey of travel (Buchanan, 1976) also these purposes have a specific mobility context in time of travel. Though individual's make multiple trips corresponding to multiple purposes, it is difficult to classify trips to the above degree due to its randomness and the data sparsity involved, unless for large studies like household surveys. A general study can be modeled for the following three purposes.

- Home – Based Work (HBW)
- Home – Based Other (HBO)
- Non – Home based (NHB)

Home-based trips are those having one end of the trip, either origin or the destination at the home of the person making the trip while those that neither begin nor end at the trip maker's place of residence are termed as non-home based trips. In case of home-based other trips, either origin or the destination will be the home location and other end would be for a different purpose excluding the work end.

When a person makes two trips during the day (to work and back to home) the home is the origin and the destination as is the place of work. In terms of productions and attractions, the place of residence generate productions and the place of work generates attractions (Wright & Ashford, 1989). Figure 1.2 below indicates the general illustration of home-based and non-home based trips.

All these movements are studied by aggregating the origins and destinations known as transport zones or traffic analysis zones (TAZ). This is the geographical unit which is commonly used in most of the transport planning models. The spatial extent of the zones may vary based on the purpose of the study.

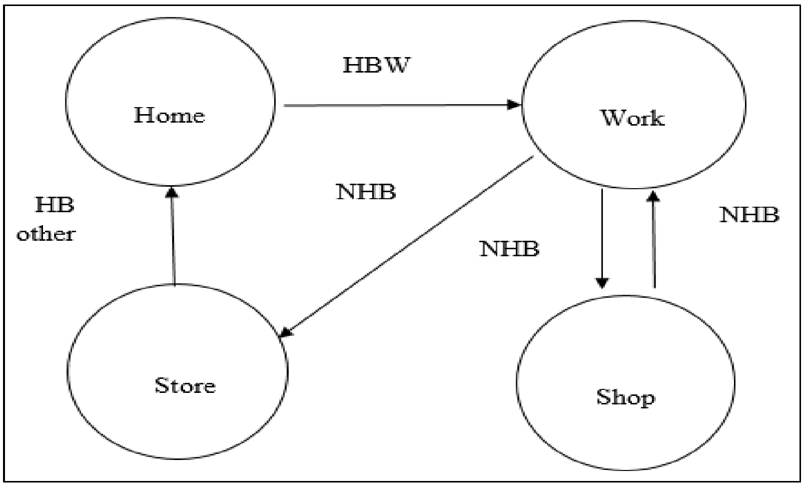


Figure 1.2: Home based work trips

1.2.2 Transportation forecasting

Transportation forecasting is the process of estimating the total no. of people or vehicles that will use the given facility in the future (Meyer & Miller, Urban transport planning, 2000). Four step model is the conventional transport forecasting technique used as the center of most travel demand models. The process initiates with trip generation followed by destination choice, mode choice, and route choice as shown in Figure 1.3.

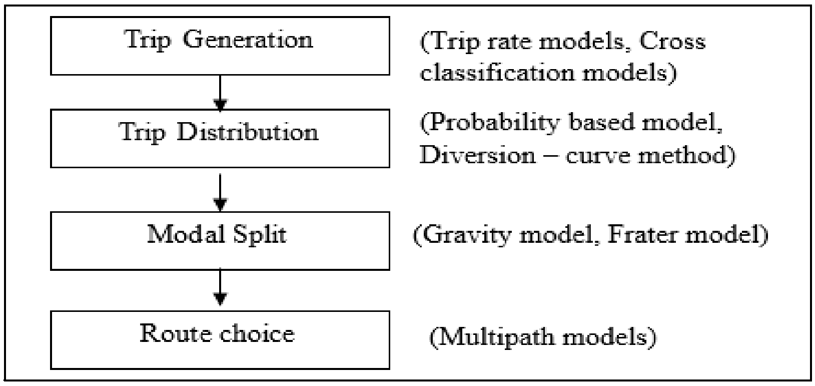

Trip Generation
(Trip rate models, Cross classification models)
Trip Distribution
(Probability based model, Diversion – curve method)
Modal Split
(Gravity model, Frater model)
Route choice
(Multipath models)

Figure 1.3 : Four step model
Source: 2015 Regional Travel Demand Model

Trip generation is the analytical process that provides the relationship between the urban activities and travel (Wright & Ashford, 1989). Trip generations depend on the land use and socio-economic characteristics of the considered area while conventional trips are defined as those that are the productions of a particular zone and attracted to other specific land uses. Trips that originate or terminate within each zone are known as trip ends from origins or destinations. The first step of the model forecast the number of trips that will begin and terminate from each travel analysis zone within the region for a typical day of the year. Total trip generations are calibrated by using the observations taken from the variety of surveys mentioned above.

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 (Papacostas, 1987). As an example, the trip productions of each zone I obtained from the trip generation phase are distributed among the trip-attracting zones J. The trip volume that a zone J would attract depends on its relative attractiveness. Generally, non - residential areas are the attractors and the residential areas are the trip generation points. Figure 1.4 below demonstrates the trip distribution of residential and non-residential zones. During the peak hours of the morning, people travel from home to workplaces. Which make the residential areas as the production zone and the non-residential areas into attraction zone. The scenario happens vice-versa during the peak hours of the mornings.

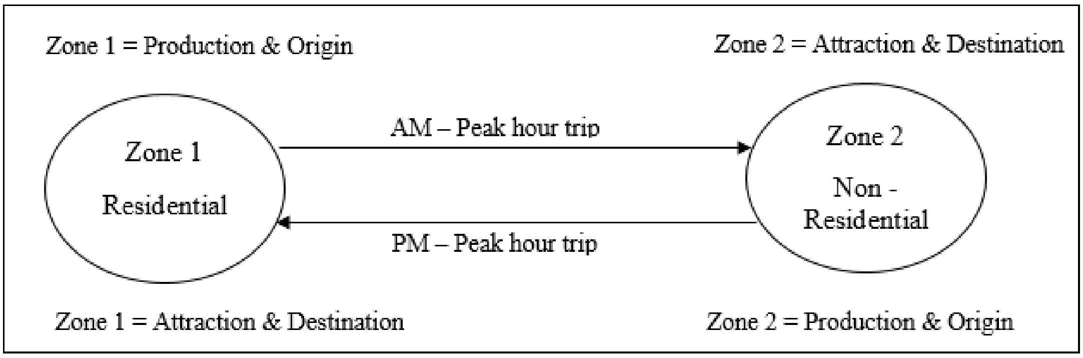
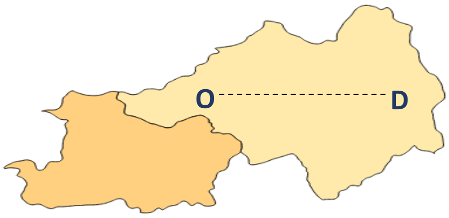
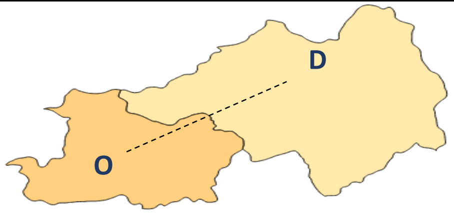
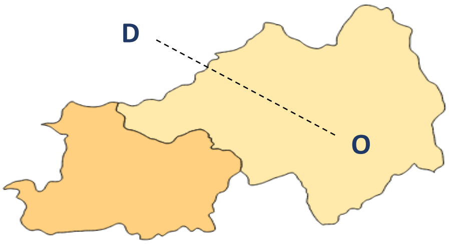
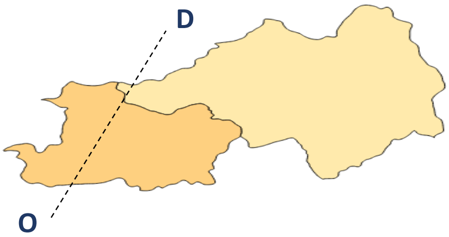

Zone 1 = Production & Origin
Zone 2 = Attraction & Destination
Zone 1
Residential
AM – Peak hour trip
PM – Peak hour trip
Zone 2
Non - Residential
Zone 1 = Attraction & Destination
Zone 2 = Production & Origin

Figure 1.4: Characteristics of residential and non - residential zones

Further, movements can be divided into four main categories based on the origins and the destinations of the trips as in Table 1.1. Movements were segmented by considering the origins and destinations in context with study area and the traffic analysis zones. The overall behavior of the users can easily be identified by categorizing the trips under the following categories. Whether the users are moving within the same zone or they move to other areas, whether the zone is a transit zone and etc.

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.	

The modal split comes as the third step of the model. This step determines which vehicle will be utilized when going from one zone to the other or in between the same zone. Mainly auto, transit, bicycle, walking and etc. The complexity of the step depends on the study area, based on the availability of the transit modes. Travelers can account for a wide variety of choices including, whether people drive alone or carpool and what specific type of transit they take like a bus, subway, commuter rail or ferry and etc. Final step is the route choice. This step will take all of the trips from mode choice and assign them to a transportation network. This will determine what route or path trips will take in going from zone to zone. In this step, the travel demand of the traffic zones interacts with the supply provided by the road and transit systems. There are several methods of trip assignment. As an example, in one such method, each individual’s path is determined by factors such as minimum travel time and the congestion that can arise from too many vehicles using the same particular route.

Information about travel is required to model the trips using above techniques. That data collection is considered as the most expensive and time-consuming part of the transportation model development process and currently, surveys are the source for such data aggregation. Transportation surveys investigate when, where and how people travel within areas. Within this context, a number of transport surveys were conducted to obtain various data on human and vehicular movements. Home Visit Surveys (HVS), Cordon Line Surveys (CLS), Screen Line Survey (SLS), Trip Generation Survey (TGS), Travel Speed Survey (TSS) are few examples of such data collection surveys. The main purpose of the current study is to take the initial steps in supplementing these surveys with CDR data by establishing a relationship between CDR and traditional data.

1.3 Introduction to Mobile Network Big Data

Big data is an information technology term which refers to high volume, velocity, and variety of data (Beyer & Lanley, 2012). High volume with rising quantities is the biggest feature of big data. The diversity of the different data types and the generation methods create the structural differences in big data. Administrative data (digitized medical records, insurance records, tax records), stock market fluctuation data, online trading data, credit card usage data and mobile positioning data are few examples of diversified big data sources.

Mobile positioning is the technical ability to locate devices in time and space using available network infrastructure with a certain accuracy. Mobile positioning data or Mobile Network Big Data is one of the main categories of big data generated by the use of mobile phones. Such mobile devices generate data through Wi-Fi networks, GPS, Bluetooth and such data had been involved in a wide range of applications including location service, navigation and etc.

These mobile positioning data are introduced in different names based on their purposes and collection techniques. The current study uses the CDR data which is one of the most commonly used categories of big data. Call detail records are collected by every mobile network carrier primarily for their billing purposes. This event-driven data is normally generated during incoming and outgoing voice calls and text messages. This unique identifier allows linking each phone owner with their Caller Activities (CA). CDR generate the time-stamped locations of the users. All the phone numbers are replaced by a computer generated unique identifier by the operator to preserve the anonymity and the privacy of the users. A single CDR contains the following data,

- A random ID number of the phone,
- Device and phone number
- Exact time and date
- Call duration
- Location in latitude and longitude of the cell tower that provided the network signal.

When a person makes a call, all this information gets recorded as an array of data. But these CDR are recordings of human mobile phone usage and in transportation, the interest is on travel patterns. Hence the processing of CDR leads to the accumulation of mobility rich information on human insights.

1.4 Purpose of the research

As humans become more mobile, travel becomes more complex and it has become a necessity to come up with a continuous process to plan transport initiatives. The planning of these initiatives needs analysis in mobility aspects, which leads to the requirement of frequently updated data sets. The capacity of the existing mechanisms is beyond the requirement and the current study take the initiatives in supplementing these traditional mechanisms with mobile network big data.

The study uses the rational behavior of trips as the basis of analysis. Rational behavior makes the travel regular and predictable. When the trip making patterns are regular it overlaps with the corresponding call making patterns up to a certain extent. This results in regular call records at specific points in the same time period of the day. Study employee this concept to derive trip making patterns of mobile users.

The ultimate purpose of the study is to explore the potential for MNBD to generate high-value transport insights by presenting a significant relationship between MNBD and traditional outputs while aligning with traditional transport forecasting methods including trip generation and trip distribution steps.

1.5 Research gap

A large no. of transport studies have been carried out in different countries, including Sri Lanka using CDR. Different methodologies have been adopted for a range of transport studies, including O-D matrix estimation (Alexander, Jiang, Murga, & González, 2015), (Saini et al., 2015) identifying meaningful places (Ahas, Silm, Ja, Saluveer, & Tiru, 2010) to the characterization of human mobility (Zhang, 2014) which will be discussed in detail within literature review. Though the existing studies present comprehensive mechanisms, machine learning based algorithms act as the foundation of the study. Studies show a deficiency in adhering to fundamental transport concepts which leads to several conceptual issues as discussed below.

Initially, several parameters should be established with respect to sample selection, identifying stays, identifying movements and etc. It would be inaccurate to use the exact parameters used in other countries, in the same manner to define measurements, due to the difference in socio-economic factors which have a direct influence in both trip making and mobile usage patterns. As an example, a human mobility characterization study in Estonia, in the sample selection they have used only the users who had recorded more than 26 different days per month (Jarv et al., 2013). Estonia being a more developed country compared to Sri Lanka, the socio-economic factors of the two countries are different to each other and using the exact parameters leads to inaccurate outputs. There are also issues with the trip identification based on CDR. The subsection below indicates a simple example of such movement state identification.

Trip detection based on CDR

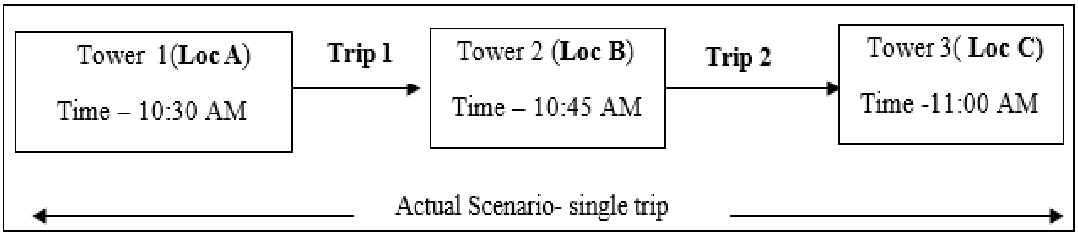


Figure 1.5: Trip Estimation with CDR

Figure 1.5 above illustrates the issues with the movement state identification. Assume that a person is making a number of calls during his travel from origin to destination, in location A, B, and C. Then the majority of the existing techniques identify those as a collection of trips. First trip from A to B and second trip from B to C. But in the actual scenario, the individual is making a single trip. Ultimately the total trips exceed the traditional survey data during trip validation.

Above mentioned facts are a few examples of the issues with the existing mechanisms. The current study focuses to fill these gaps by analyzing CDR aligning to fundamental transport concepts. Acceptability of the techniques used were proved by the validation of CDR outputs with the traditional survey data.

1.6 Research Objectives

In line with the background described above, this study identifies two research directions as shown in Table 1.2. Mobility pattern mining and the MNBD validation were the two analysis categories identified to support the two research directions. The research objectives were set to continue in these directions and the detailed objectives are depicted in the following session.

Table 1.2: Research direction

Categories	**Research direction**
Mobility Pattern Mining	Segmentation and profiling users from cell phone data
Validation of MNBD results	Validating the identified distributions with the traditional survey data

Objective 1: Develop a methodology to estimate the basic travel to work attributes of users within the Colombo Metropolitan Region using MNBD.

The first step of the data analysis is to examine the travel behavior of the users. Any location visited by an individual could be categorized as a home or a non-home location. Understanding the home and non-home locations help to dynamically monitor the commuting trip pairs on regular basis. All these home and non-home locations are interconnected to each other with different trips. Moreover, high frequency of appearing in a location ensures relatively a high accuracy in interpretations compared to other random user points. Home location is one point an individual spent most of their time and generally work location is the other point out of all the non-home locations. This is mainly due to the high stability of home and workplaces compared to other locations. Also, home and work points are the origins and destinations of commuting trips during peak hours. Typically that results in a variation of the population distribution during different time periods. The study targets to identify these home-based work trips within the Colombo Metropolitan Region from the existing CDR excluding other trips to random locations as the main outcome.

Secondly, by taking home and work points as the basis it is possible to segment users into different profiles based on their type of employment or how they move within these two locations with time. Basically, by analyzing the presence patterns of users on particular days of the week and time of the day, there can be different behavior patterns which create different user profiles. The study extracts these profiles as to in line with direction mobility pattern mining.

Objective 2: Validate the trips extracted from MNBD with trips estimated from Household survey

It is important to validate the CDR findings with the existing actual data to ensure the accuracy of future predictions. Therefore the study to validate its findings with trip estimates using Household Visit Survey data in the WP.

By achieving the above objectives, the main research questions this thesis aims to answer include:

1. Characterize user locations by their presence on different time periods of the day based on the observations from CDRs.
2. Develop a method to categorize user locations into home, workplaces and other based on the presences patterns.
3. Identify the underlying common behavioral structures at user locations across the sample and segment them into named clusters.
4. Validate the extracted information with the Household Visit Survey data.

1.7 Data sources

- The study uses 3 months (May, June, and July) of Call Detail Records for nearly 10,281 randomly selected SIMs from a mobile operator in Sri Lanka were provided for this research by LirneAsia- A regional ICT policy and regulation think tank. The data are completely pseudonymized by the operator, the phone numbers have been replaced by a unique computer generated identifier.
- The main data source used for the purpose of validation is obtained from the Household Visit Survey data collected from surveys within the Western Province conducted by CoMTrans study. In understanding travel demand, the household surveys provide transportation information, including socio-demographic information, travel time, trip purposes, travel modes etc.

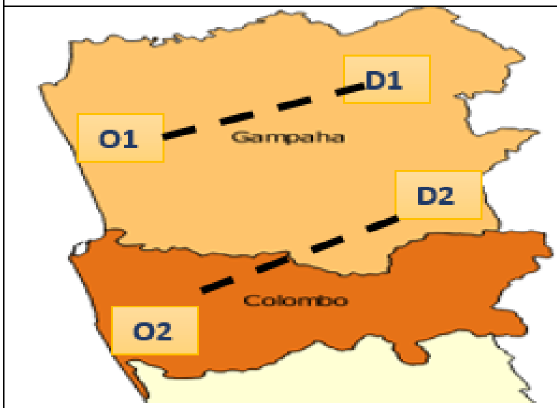
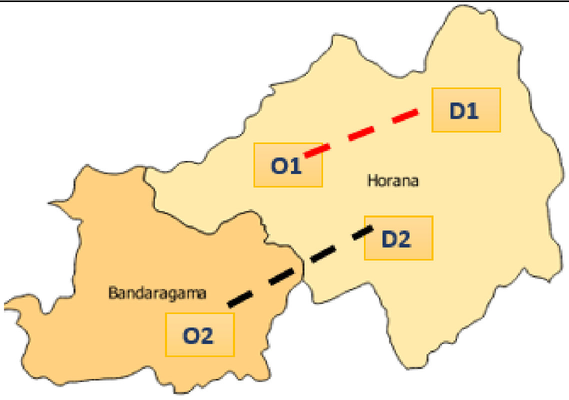
1.8 Scope of the research

Considering the possibility of validation, the Western Province was taken as the study area covering three districts; namely, Colombo, Gampaha, and Kalutara having a total of 47 Divisional Secretariat Divisions (DSD).

Both intra and inter trips were validated at the district level. But only the inter trips were validated by eliminating the intra trips at DSD level (Table 1.3). The main reason is, DSDs have comparatively a smaller geographical area. Generally, few cells (Maximum 3 cells, less than 3 cells in most of the cases) fit into a single DSD. Trips are calculated with the changing of the cell towers. But when the geographic area is limited, the majority of the movements takes place within the coverage of a single cell, which does not indicate a significant change in the cell tower. As the movements within one cell area cannot be identified, calculating trips at intra DSD level leads to a larger proportion of errors. Furthermore, only the work trips were identified at district and inter DSD levels.

Out of all the trip purposes such as shopping, visiting friends and relations, traveling to school and etc., the most common and the accurately identifiable home-based work trips were extracted and used for the validation purpose. A detailed description of work trip identification at these spatial levels is explained in the following chapters.

Table 1.3: Identified trip types

Type of trip	Description
	Both Intra trips (O1-D1) and inter trips (O2-D2) were validated at the district level.
	Only the inter DSD level trips (O2-D2) were validated by eliminating the intra DSD level trips (O1-D1).

1.9 Overview of the thesis

This thesis consists of five main chapters consisting of an introduction, Literature review, Introduction to HVS data, Data analysis, Data validation and Conclusion to achieve the ultimate objectives of the research. The section below gives a brief introduction to the content of each chapter.

First chapter gives an overview, including the purpose of the research, its importance, its contribution towards filling the knowledge gap that exists in the area of using mobile positioning data in transport aspects. This chapter also describes the scope of the research.

Second chapter explains the literature around the research area which has also been used in understanding travel attributed from Call Detail Record Data. A Literature review was carried out under five subsections related to the use of CDR which will give the reader a comprehensive understanding of the current findings in the scope of this research.

Third chapter focuses on the methodology and the findings from the Household Visit Survey. This will explain the methodology adopted for the survey and the findings in relation to the current study.

This chapter also gives a clear idea about the different purposes of travel and how they are distributed within areas. Main findings related to work, home attributes from these traditional methods will be used for the validation purpose of the study.

Fourth chapter aims to segment users into different profiles based on their behavior identified through the mobile activity. It also explains the basic methodology followed in understanding the human behavior and the statistical techniques used to process the data, eliminate the noise from the sample of users, and methodology followed in identifying these significant locations using CDRs.

Fifth chapter contains the validation of the work home distribution derived from the CDR data with the Household Visit Survey data obtained from the urban transport system development project for the Western Province, conducted by the Japan International Cooperation Agency.

Last chapter includes the summary of the findings, conclusions, and recommendations. It also provides information on the limitations of the study, and suggestions and recommendations for future research.

CHAPTER 2 - LITERATURE REVIEW

There is an increasing amount of literature arising from the study of travel patterns from electronic data, as the spatiotemporal data collected from any mobile device is a very convenient tool to understand the behavioral patterns of people (Khan, Ali, & Dev, 2015). This chapter gives an overview of the studies in the transportation field using CDR data, followed by the introduction to mobile positioning data and a discussion of the methodologies implemented in the behavioral identification, issues with the existing techniques of using CDR and the identified research gap in a detailed manner.

2.1 Mobile positioning data

Primarily there are two types of mobile positioning data categorized based on their collection method as “active positioning data” and “passive positioning data” (Esko, 2013). 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 (Kaisler, Armour, Espinosa, & Money, 2013). Visitor Location Register (VLR) and Call Detail Records (CDR) are two main inclusions under passive positioning data. VLR data generate 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 the 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. Considering all these the most common, and one that is universally captured by all mobile operators are passive positioning data in the form of CDR (Maldeniya, Lokanathan, & Kumarage, Origin - destination matrix estimation for Sri Lanka using Mobile Network Big Data, 2015). Compared to other network related data like GPS, CDR is easily available as most of the telecommunication service providers maintain such data (Dash et al., 2014). CDR had been used in a wide range of studies related to transportation and other aspects including economic activities, urban planning an etc.

2.2 Use of Call detail records in transportation

Studies had been focusing on different mobility aspects within their research areas. Trip extraction using CDR had been ranging from O-D matrix (Origin – Destination matrix) estimation (Alexander et al., 2015), (Saini et al., 2015), (Caceres, Wideberg, & Benitez, 2007) (Mellegard, Moritz, & Zahoor, 2011), (Wang, Schrock, Broek, & Mulinazzi, 2013), identifying meaningful places (Ahas et al., 2010), (Isaacman et al.,2010) to the characterization of human mobility (Zhang, 2014), (Schneider, Belik, Couronne, Smoreda, & Gonzalez, 2013).

O-D matrices had been developed with different techniques, As an example, in a study of Mumbai CDR data are 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 (Saini et al., 2015). Apart from that majority of the studies have developed algorithms to assign mobile phone towers extracted from CDR to traffic nodes (Mellegard et al., 2011). But most of these studies have focused more on computational issues and the relationship between the mobile phone O-D and the traffic O-D have not been explored in detail.

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 (Ahas et al., 2010). Other than that, most of the studies focus on identifying major home and work anchor points (Saini et al., 2015), (Kevin, Kung, Greco, Sobolevsky, & Ratti, 2014). 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. (Leng, 2013). In case of human mobility characterization, statistical models have been used in identifying the variance in the number of individual’s activity locations (Jarv et al., 2013)) 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 meaning full locations to users (Jarv et al., 2013).

Table 4 given below indicates studies and different methodologies followed by different researchers within their corresponding scope of analysis. Studies can be classified into three main categories as (i) identifying significant locations, (ii) profiling users and (iii) O-D matrix estimations. In identifying significant locations majority have followed the concept of finding the stay locations and assigning them as a home or work location based on the behavioral structures. Principal Component Analysis (PCA) is one such technique used in identifying the common structures. Generally, PCA is a technique used to emphasize variation and bring out strong patterns in a dataset (Powell & Victor, 2014). Some have considered the most frequently used cells by month (monthly activity locations) or day (daily activity locations) and assigned significant points based on them. In profiling users, one study had segmented the users based on the total number of work locations identified. Finally, in deriving O-D matrices, the general technique followed is the identification of stays and generating tower to tower or geographical location based matrices. Table 2.1 outlines the methodologies followed by various studies based on the discussed techniques.

Table 2.1 Methodologies of 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 (Leng, 2013)	• Capture when, how often and how long each user appears at each user location. • Identify the common behavioral patterns and daily routines from the identified appearances with Principal Component Analysis (PCA) • Identify home and workplaces based on the similarity in the presence patterns.

<table>
<tr><th>Attribute</th><th>Name of the research</th><th>Methodology followed</th></tr>
<tr><td rowspan="2"><b>Identifying significant locations</b></td><td>Understanding monthly variability in human activity spaces:
A twelve-month study using mobile phone call detail records (Jarv et al.,2013)</td><td><ul><li>Assign meaningful monthly and daily activity locations using mathematical linkage analysis technique.</li><li>Total no. of caller activities made within locations on considered days or months is used as the main influencing criteria for identifying locations.</li></ul></td></tr>
<tr><td>A Hierarchical Approach for Identifying User Activity Patterns from Mobile Phone Call Detail Records (Hasan et al., 2015).</td><td><ul><li>Finding the usual stay locations of the users from one-month CDR data.</li><li>Work time considered as 9 am to 5 pm while other period as off hours. Distinguished two main groups as regular and irregular workers.</li><li>Identified the distance of travel to work and some other random locations within these two groups.</li></ul></td></tr>
<tr><td><b>Profiling Users</b></td><td>Identifying Significant Places Using Multi-Day Call Detail Records (Yang, Zhu, Wan, & Wang, 2014)</td><td><ul><li>Define the time between nightfall and dawn as Home time (7 pm -8 am) and the daytime as Work time (8 am-7 pm)</li><li>Identify the stop with the longest journey time as home and work respectively</li><li>Captured 4 types of users based on their multiday significant places of the visit as follows.</li></ul>
<table>
<tr><th><b>Group</b> 1</th><th>Home</th><th>Work</th></tr>
<tr><td>1</td><td>1</td><td>0</td></tr>
<tr><td>2</td><td>1</td><td>N<=3</td></tr>
<tr><td>3</td><td>1</td><td>N>3</td></tr>
<tr><td>4</td><td>0</td><td>N</td></tr>
</table></td></tr>
</table>

Attribute	Name of the research	Methodology followed
O-D matrix estimations	Estimating Origin-Destination Flows Using Mobile Phone Location Data (Calabrese, Lorenzo, Liu, & Ratti, 2011)	• Trip determination -All consecutive points for which the radius < 1 km fused together such that the centroid becomes a virtual location • Once the virtual locations are detected, the stops and trips were evaluated as paths between users' positions at consecutive virtual locations. • The geographical area under analysis is divided into regions to derive the O-D matrices
	Development of origin–destination matrices using mobile phone call data (Iqbal et al., 2013).	• Generate tower-to-tower transient O-D matrices using and trips occurring within different time periods. • Converted the corresponding nodes of the traffic network to node-to-node transient O-D matrices. • Derive the actual 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 (Alexander et al.,2015)	• The records were first converted into clustered locations at which users engage in activities for an observed duration, to be home, work, or other depending on observation frequency based on triangulated mobile data. • Probabilistically infer departure time using survey data on trips in major US cities. Trips are then constructed for each user between two consecutive observations in a day.

2.3 Use of CDR in Sri Lankan context

In Sri Lanka, the mobile penetration has increased from 87% - 123% in between 2011 to 2016 (Sri Lanka - Telecoms, 2016). Which suggests that almost every person is using a mobile phone for their daily activities. On the other hand, there can be people who are using more than one SIM and some do not use any at all. But anyway, as the majority of the population use these handheld mobile devices, it would be reasonable to assume that the travel patterns of a random sample of users would represent the population of mobile users.

CDR had been used in a wider range of studies, including economic, urban planning, transportation and etc. in Sri Lankan context. CDR data can be used to get insights on human mobility when combined with other geographical and various demographic data.

The research on quantifying urban economic activity using cell phone data (Miyauchi, Gabriel, & Yuhei, 2015), study the link between commuting flows and urban economic activity using fine spatial and temporal variations. CDR has also been used to understand communities using communication patterns by applying several community detection algorithms (Madhawa, Lokanathan, Samarajiva, & Maldeniya, 2015). Land use classification being an important concept in urban planning is one other aspect involved with CDR in Sri Lankan context. (Samarajeewa, Madhawa, Lokanathan, & Maldeniya, 2015). This explores the potential of leveraging massive amounts of human mobile usage to derive conclusions at spatiotemporal activities of the masses and provide a useful measure for activity based classification of land use. Apart from the above, CDR had also used in disaster management and disaster resilient development aspects (Samarajeewa, 2005)

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 (Maldeniya et al., 2015) which uses different techniques like stay based approach, frequency-based approach and transient-based approach in deriving OD-matrices using CDR (Maldeniya et al., 2015) In the stay-based approach, a stay is defined by a continuous series of records such that,

- Any two records of the series are not more than one km distance apart – 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, such that, T<1hr

The movement between two stays was introduced as a trip by the stay-based approach. In the transient-based approach, if the cell tower utilized for each call is different and records are separated by a time interval of 10 min <T<1 hour the movement is taken as a trip. Table 2.2 below, illustrates an example of a trip identification using stay based approach.

Table 2.2: Trip identification in stay based approach

Time	**Time difference (T)**	**Tower connected**	**Distance (D)**	
10:20 AM	10 min <T<1 hr.	23432	D < 1KM	**Stay 1**
10.32 AM		24321		
				Trip 1
10: 50 AM	10 min <T<1 hr.	23653	D < 1KM	**Stay 2**
11: 10 AM		24321		

Apart from that in the frequent- trip based approach, frequent locations are identified from the daily sequences of an individual over a period of 1 year. The movement between the identified most frequently visited locations were calculated as trips to generate O-D pairs.

However, there are variations in the results of the traditional transport surveys and MNBD estimations. Such differences have to be understood before MNBD can be used to replace traditional transport forecasts.

2.4 Limitations of using CDR in Travel predictions

In general, use of big data has different issues due to its high volume and the data complexity. Considering the CDR data obtained from the telecommunication service providers, though CDR can be obtained more frequently and economically there are limitations which need to be considered when using them in mobility estimations.

A critical limitation of CDR data for mobility analysis is data sparsity (Zhao, Jinhua, & Koutsopoulos, 2014). A data entry is generated only when a communication activity is initiated (Hasan et al., 2015). A user may be observed to move from zone B to zone E, but his/her 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 unobserved in the CDR (Iqbal et al., 2013). Most of the user records are sparsely and irregularly distributed and trip estimation based on such data is incomplete. This data sparsity also leads to the problem of movement state identification, where it is difficult 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 cell tower used by the network (Hasan et al., 2015). 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 spatial granularity of the location data depends on the density of a cell tower and the density of cell towers vary with the area. The coverage area of the cell towers varies depending on the socio-economic features including population density (Csaji et al., 2013)

One other limitation of the CDR data is often there are changes in the tower in the data with no actual displacement which is introduced as the load sharing effect (Iqbal et al., 2013). This is mainly because the operator often balances call traffic among adjacent towers by allocating a new call to the tower that is handling lower call volumes at that movement. Due to this process, the towers providing the signal varies without actual displacement of the user.

Apart from these, there are identifiable issues with the three approaches described in the above section. As examples in stay-based approach where a trip is defined as a movement between two stays, the transient locations or random locations cannot be captured. Also, the long-distance trips cannot be identified and those trips are taken as a collection of small trips. In, frequent based approach the location sequence is considered frequent if it occurs on at least 10% of the daily sequences of the individual which disregard the Ad-hoc movements. All these limitations together create different errors in identifying the location of users, identifying the stays of users, identifying pass by points,

2.5 Techniques used in minimizing the limitations of CDR

Studies have used different methods to reduce the errors of CDR and interpret the most precise results. Location identification is one such main factor focused by most of the studies. The main idea is to identify the sequence of locations visited by users and the results were obtained through different techniques. The majority had produced a sequence of visited locations using different algorithms (Wang et al., 2013), (Isaacman et al., 2010), (Csaji et al., 2013) which is referred by the term trajectory smoothing. As an example, the sequence of CDR was taken within a certain time threshold and filtered using time, speed or by assigning a center location for close by points. Such algorithms facilitate in identifying the common or the significant location sequences of users. Some studies had identified the locations as clusters based on their spatial distribution without considering the time factor. Finally, the points of locations were consolidated even though they were visited in different days (Jiang et al., 2013), (Hariharan & Toyama, 2010). 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 (Calabrese et al., 2011), (Mellegard et al., 2011), (Maldeniya et al., 2015). As an example, a threshold of 10 min for the lower boundary and 1-hour upper boundary is used and also the records should be more than 1km apart (Saini et al., 2015). 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 (Maldeniya et al., 2015).

There exists a distinction between “Observed end locations” and “hidden end-location” which occurs due to differences in observed trip ends and the actual trip ends that had created with the data sparsity as discussed earlier. One study had explicitly addressed this issue. 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 reduce the chances of hidden visits occurring during observed displacements (Zhao et al., 2014). The selection of the threshold is heuristic based and they tend to overestimate the existence of hidden visits with sparse data. The majority of the works pay less attention to this issue, which has serious implications on trip detection resulting wrong extraction of O-D pairs.

2.6 Research gap identification

In conclusion, the literature review shows 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. Although there is a large body of literature in the current practices of using CDR in transportation aspects, the theoretical involvement of the methods is still lacking.

Transportation is a derived demand created with the human needs. Human behavior leading to satisfying these needs is the foundation of transportation. Therefore, the predictions can be done more accurately by starting from the transport concepts which had been evaluated over the years based on human behavior. The majority of the studies focuses on identifying the locations using CDR, which had been a fundamental problem of CDR data that is relevant not only to transport applications but also for other areas like geography, urban planning and etc.

Apart from that the attention on the basic trip concepts like trip purposes, intra-inter trips are comparatively low with regards to previous studies. Travel always takes place for a defined purpose and it is entwined with these concepts. Therefore, these ideas should be prioritized in the analysis work. Also, the existing studies use heuristic approaches and arbitrary criteria in setting up different location and time boundaries which depends on the socio-economic background of the users. These socio-economic criteria should be adjusted to the Sri Lankan context and included in the analysis.

This research attempts to fill this gap by exploring a methodology to derive travel estimations using CDR by focusing on the purpose of home-based work. The study uses the time windows and other socio-economic criteria by aligning with the Sri Lankan context. The validity of the methodology is proven by the comparison of the CDR data with the existing transport data.

CHAPTER 3 – INTRODUCTION TO HVS DATA USED FOR RESULT VALIDATION

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. The current study uses the household visit survey (HVS) data (JICA, 2014) to validate its findings. Understanding the process of conduct in these surveys is an initial requirement. This chapter gives the introduction to the household visit survey data, the methodology followed in the data collection and attributes used from the survey for the purpose of validation.

3.1 Introduction to Household Visit Survey

Transport demand in Western Province has increased remarkably over the past few years (JICA, 2014). To cope with the anticipating transport demand and problems, several transport development projects and plans were carried out. Different surveys were conducted as a part of this to obtain reliable transport data. Out of different types of surveys, Household Visit Survey (HVS) is used for the validation of the current study. HVS conducted by Ministry of Transport with the technical support of Japan International Cooperation Agency (JICA) was one of the largest and comprehensive transport surveys carried out in Sri Lanka.

3.2 Methodology of the household visit survey

The study area of the HVS covers the boundaries within the Western Province (Appendix A) which includes Colombo, Kalutara, and Gampaha districts consisting of 2496 Grama Niladhari Divisions and 47 Divisional Secretariat Divisions. Sample households were randomly selected from the available lists of addresses obtained from the electoral registration data and the study ensured the fair spatial distribution of the sample across the survey zone. The original sampling size was estimated to be 3%, but within the sampling work, it was extended to 4% due to the fact that people hesitate to participate in these kinds of surveys. Table 3.1 indicates the no. of households interviewed within each district.

Table 3.1: Estimation of the target sample size

	Colombo	**Kalutara**	**Gampaha**	**Total**
Population	2,309,809	2,294,641	1,217,260	5,821,710
Sampling rate	3%	3%	3%	3%
Average household size	4	4	4	4
no. of households surveyed	17,500	17,300	9,200	44,000

Source: Urban Transport Development project for CMR and Suburbs

Four different types of survey forms were given to each household. The first form was given to each household and collect data on household information such as the socio-economic background that includes address, level of income, vehicle ownership, expenditure, preference regarding urban transportation and etc. In addition to that, a second questionnaire was given to each member of the household to gather information like cost of transportation, the address of workplace/school, type of workplace and information regarding ordinary commuting trips to workplace/school. One other form was given to households who have changed their residence within the last five years. That collects data like the reason for moving and information on previous housing. The last form gathers the trip information (Appendix B) and it was given to each member of the household. Information about the trips of each member of the household, including origin and destination, trip purpose, mode of transport, transfer point, departure and arrival time were collected. Mainly the trip information was gathered on an average weekday. Out of the several attributes collected on travel details, the trip purpose and trip end zones were primarily used for the validation work of the current analysis. The section below describes the different trip purposes employed within the HVS data.

3.3 Trip purpose composition

The questionnaire of the household survey is encrypted with seven main purposes. Total number of trips made by each household under these seven purposes were recorded for the purpose of analysis. Figure 3.1 below indicates the seven primary trip purposes and their compositions in the survey result and the Table 3.2 briefly explain on each of the considered purposes.

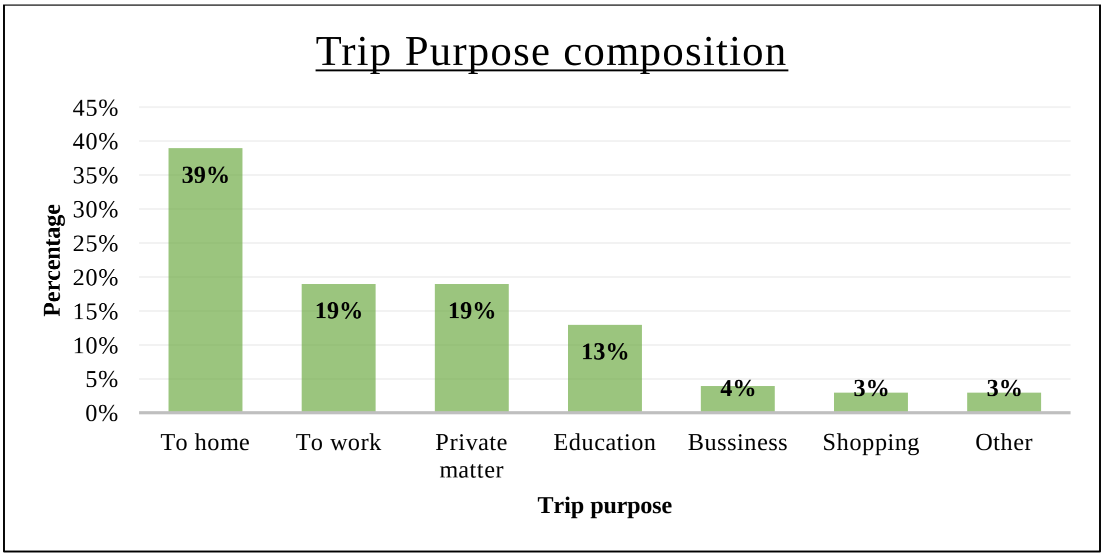


Figure 3.1: Composition of trips by trip purpose
Source: Urban Transport Development project for CMR and Suburbs

Table 3.2: Description of the 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	Commutes 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.

The rate of trip making primarily depend on the residential and the working population of the considered area (Wright & Ashford, 1989) . Although different urban settings have different trip compositions, most of the trips undertaken are home-based work trips. Generally, the work trip is the longest trip of the day and an average person spend most of the time either at work or home. Characteristics of the worker generally depend on the supply and the location of jobs and housing, availability, cost and convenience of the various modes of commuting (McGucking & Srinivasan, 2011). Commute trips bear an importance to transport planning beyond its share of travel. Commuting is still considered to be predominantly a weekday activity tied to morning and evening hours, which are defined to be peak hours. The no. of work trips had risen with the no. of workers. Nearly 87% percent of the people aged 20 and older are in the workforce, which is 45% of the population in the Western Province (Source: Sri Lanka Labor force survey: 2012). For the majority of the community, commuting is regular in frequency, time of departure and trip ends. In addition, they are highly concentrated in time and space.

From the survey, it is estimated that around 10 million trips were made on a weekday within the Western Province. Out of these, around 2 million trips are counted as home-based work 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.

The behavior or the daily routine of an individual may change with the type of employment. Other than the identification of work-home trips, the current study also expects to segment the users into different profiles based on their behavioral structure. Supporting that it would be important to identify the different employment types recognized by the HVS data. Such that the overall picture of the working population can be viewed. Table 3.3 shows the different occupation categories and the percentage of users belong to each category. The study has identified 17 working categories along with different behavioral structures.

Table 3.3: Occupation types and the corresponding percentages

No.	**Type of occupation**	**Percentage of users**
1	Managers, Senior officials, Legislators	6%
2	Professors, Lectures, Teachers	4%
3	Researchers, Engineers, Medical doctors, etc.	2%
4	Government employees, Technicians, etc.	9%

No.	Type of occupation	Percentage of users
5	Brokers, Sale/ Business service agents	4%
6	Office clerks, customer service clerks	8%
7	Travel attendants, guides, chefs, etc.	1%
8	Sale persons at shop, stall, and market	10%
9	Policemen, firefighters, security guards	3%
10	Farmers, fishermen, forestry workers	3%
11	Workers in construction, textile, etc.	10%
12	Drivers, operators of vehicle & machinery	11%
13	Housemaids, cleaners, helpers	3%
14	Labors	15%
15	Street vendors, elementary occupations	3%
16	Armed forces	1%
17	Other	8%

Although there are 17 categories, most of them are having similar behavioral structures. First, four categories which include managers, lecturers, government employees, doctors generally have an immobile work time behavior, where they do not travel within their working time. That accounts for a total of 21% of the working population. The rest of the 79% generally has a mobile behavior where they have to travel to places during their working time. The study uses these behavioral changes in different time windows to capture user profiles from CDR data.

3.3.1 Distribution of HBW trips

As stated earlier, for the HVS data it had estimated that nearly are made 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 origins and destinations as in Table 3.4. Appendix C indicates the total no. of trips and their percentages at inter DSD level.

Table 3.4: Distribution of Home-based work trips

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

Appendix E shows the density of the trips attracted and originated per zone within the WP along with the district boundaries. The higher density of trip attractions can be observed inside the Colombo Metropolitan Area (CMA), which shows that the majority of the workers have their workplaces inside CMA and its surroundings. Also, the majority of the trips seems to be originated within the Colombo district as per O-D matrix calculations, approximately 17% Trips per day are from outside the CMA ending inside the city.

3.3.2 Hourly fluctuations of home-based work trips

Figure 3.2 depicts the hourly fluctuations of the home-based work trips. The trips starting from the workplace to the home locations and trips starting from the home locations to work place were graphed against the time of the day. Trip starting times were considered for the current plot.

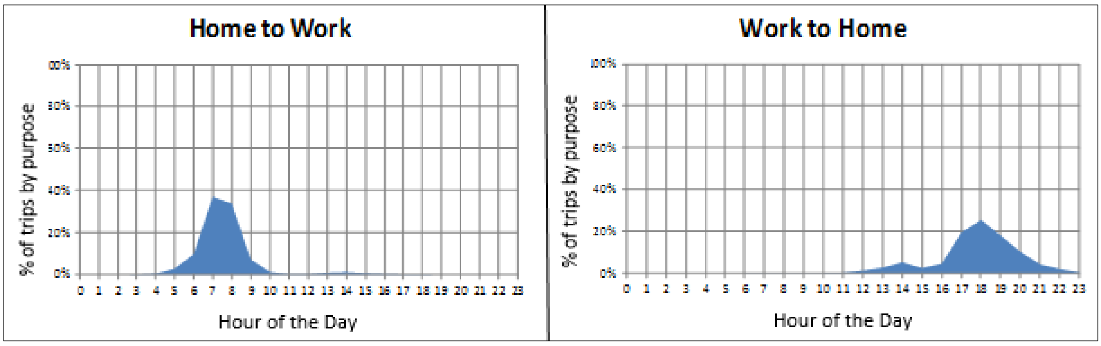


Figure 3.2: Hourly distribution of home-based work trip
Source: Urban Transport System Development Project for CMR and suburbs

Two peaks can be seen from 5 PM to 8 PM and considering the inverse trip, the majority seems to be moving between 6 AM to 9 AM. This proves that the majority travel to their workplaces between 6 AM -9 AM and come back home between 5 PM to 8 PM. These time windows help for future analysis in allocating separate time windows for work and home-stay periods. The slope of the graph is mainly due to the different starting and ending times of the workers and their distance to work.

3.4 Discussion

This chapter summarized the methodology used for the data collection within the household visit survey and the attributes extracted for the purpose of validation. Trip purposes and the distribution at the district and DSD level were two such points which will be elaborated in future chapters.

Although HVS contains precise data compared to the existing data, there were records of the few issues within the primary stages. Within the initial analysis of the HVS data, it had been discovered that a sample bias of households exists with income levels. The data seems to be over-represented by the households with lower income levels (below LKR 40,000 per month). That issue had been adjusted as the household income directly affect the trip making patterns. Apart from that, the commuting trips were well recorded by the participants, but the other trip purposes were not fully captured due to the perception of interviewers as those were of less importance.

CHAPTER 4 – ANALYSIS OF CALL DETAIL RECORDS

CDR from cell phone data, which is considered as a large-scale data source to track individual movements and aggregate mobility patterns at different spatial scales. But the data require extensive processing to derive meaningful results. Chapter 4 explores a methodology to create value from these CDR data by analyzing them logical manner from data extraction to the end objective of data validation.

4.1 Introduction

Behavioral patterns of users are different from user to user and that leads to a difference in their pattern of making calls. Generally, the behavior is influenced by socio-economic and other different demographic features like income level, education and etc. The work-home locations and the type of employment are two main aspects which influence in behavioral changes.

In general, home and work locations are the two most influential locations in an individual's daily routine. Therefore, home and work locations are inferred to be the building blocks of different aspects in the transportation field like generating O-D matrices, understanding trip purposes including the characterizing of human mobility. Many of the past studies have also used this home - work concept in their work along with different techniques (Kevin et al., 2014), (Ahas et al., 2010)

The other main influencing socio-economic factor is the employment category in which the users are engaged in. Therefore, in order to analyze the human behavior, it is important to be aware of the specific characteristics of the working nature. As an example, a government worker reaches the office at 8.30 a.m. and leave the office by 4.30 p.m. on most days. Private workers work till late in the evening and some private workers are employed even during weekend mornings. In case of drivers or salesmen, they are not having a specific place of work. At the same time, there are housewives who have not employed shows relatively an immobile nature. This chapter explores a methodology to cluster these users into different profiles based on their similarity within the mobility behavior. Mainly the following three objectives will be achieved by the end of this chapter.

- Extracting the presence of users within significant locations.
- Identifying the distribution of work and home locations within the administrative boundaries (District and DSD level).
- Identifying common presence structures within the sample.

4.2 Methodology

This section explores the basic methodology of segmentation and profiling the users. Main concepts used are the user location and presence or the appearance of the users within the locations. The methodology will be ordered with the following three folds.

1. Narrowing down the sample of users by processing raw CDR.
2. Extracting the presence of users from CDR data.
3. Proposing a method to infer home and work locations based on CDR.
   - Extract the behavioral models which could be captured via CDR by using the timely presence of users.
   - Apply the methodology to the sample of users to extract the shared behavioral patterns across the sample.

4.2.1 Data description

The study uses the records of randomly selected 10,281 users from the main sample of 10 million users for a period of three months which includes May, June, and July of the year 2013. Data were provided by a single mobile operator along with both the incoming and outgoing caller activities. Mainly a single entry delivers the user Id, time of the call, cell tower connected and the caller duration. Cell tower connected helps in mapping the user location with the geographical location of the user during the time of the call. Household visit survey data which had been used for the validation purpose had also been collected and processed in the year 2013. This eliminates the need for use of expansion factors in the process of data validation.

4.2.2 Methodological framework

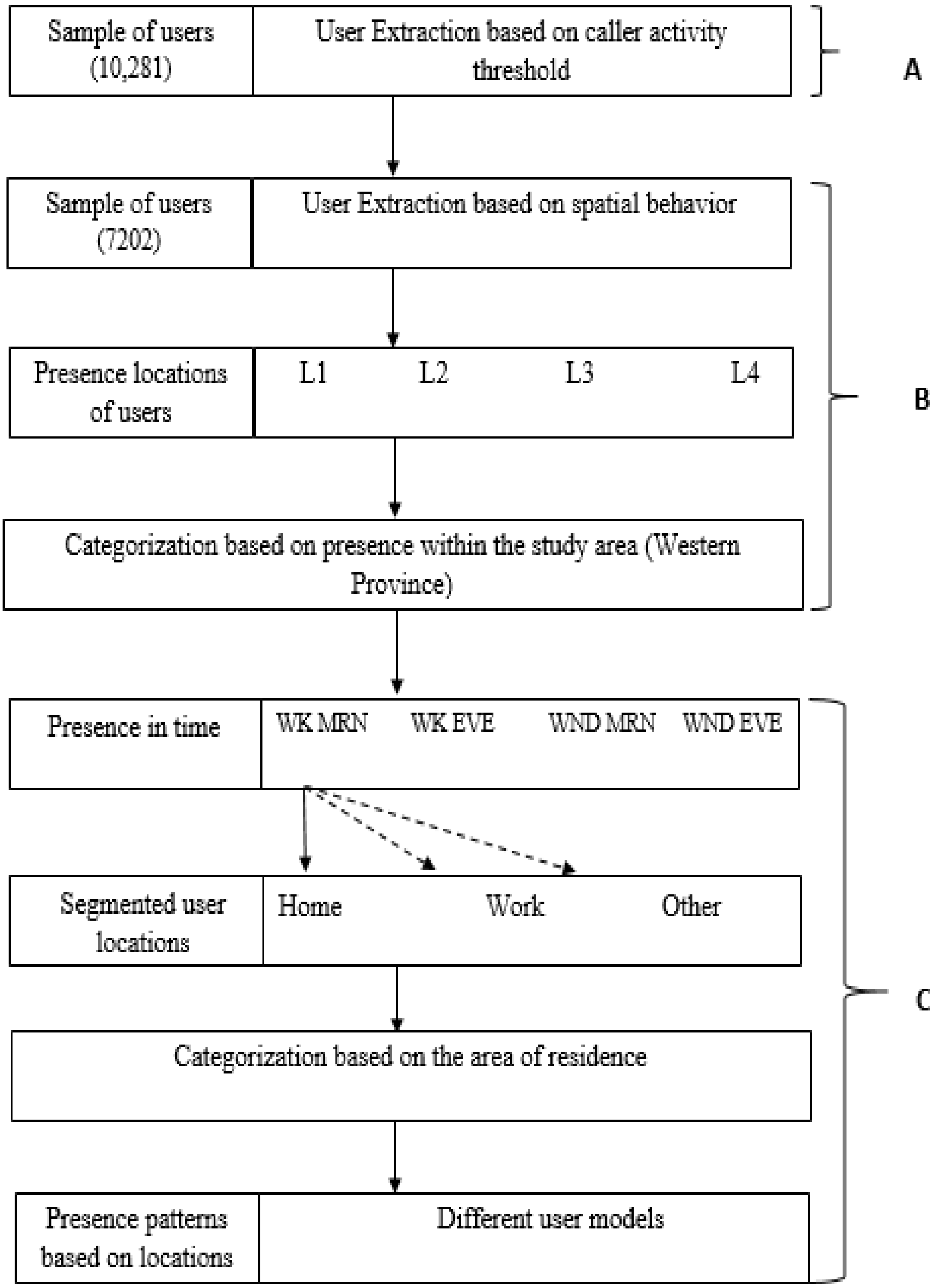

Sample of users (10,281)
User Extraction based on caller activity threshold
A
Sample of users (7202)
User Extraction based on spatial behavior
Presence locations of users
L1
L2
L3
L4
Categorization based on presence within the study area (Western Province)
B
Presence in time
WK MRN
WK EVE
WND MRN
WND EVE
Segmented user locations
Home
Work
Other
Categorization based on the area of residence
C
Presence patterns based on locations
Different user models

Figure 4.1: Methodological framework of the analysis

Research analysis was conducted on three main topics. The summary of the content under each topic is drafted below.

A - User extraction based on caller activity threshold.

Initially, users were extracted from the main sample of 10,281 based on the total no. of calls (Which may further referred to as caller activities) made by them within the considered time period of three months. The extraction criteria will be discussed later in detail.

B - Identifying the user presence

Extracted users will further be divided into categories based on their spatial behavior within the Western Province of Sri Lanka. This section basically includes, how to identify user presence based on the day and the time of the day along with the procedure followed in cumulating the total presence. By the end, three user categories were identified as follows,

- Low presence within WP
- Medium presence within WP
- High presence within WP

C - Identifying work and home locations

The locations were further assigned to home and work based on the frequency of the presence in identified locations. Mainly three user categories were identified by the province of the residence,

- Residence within WP
- Residence outside WP
- Inter provincial commuters

Afterwards, the work – home distribution at district and DSD level were identified (Results are validated in chapter 5). Different user models were implemented as the last step, by categorizing users based on their behavior within work and home locations in different time periods of the day. The methodology that categorizes the users into profiles is typically a clustering technique based on a heuristic approach. Clustering is a widely-used tool to discover underlying structures and group similar objects. It helps to find patterns in a collection of unlabeled samples by organizing items that are similar in some way (Leng, 2013). In the current work, 24 hours of the day is divided based on the core activity periods and the top frequent calls of users during each time category were used to derive different user profile models. Different time categories employed are depicted in the chapter along with the features used for clustering application.

4.2.3 Value creation with Call Detail Records

Call Detail Records, being the most common type of data in the interest of the transportation community, contain the time-stamped locations of the user which is the exact requirement in trip estimation. As an example, after processing the CDR data a single entry takes the following format.

Table 4.1: Sample of Data Record- Single user data after pre-processing

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

Table 4.1 can be interpreted, such that user A took a call at cell ID 23452 on 23/07/2013 at 9.30 am. Which means that this user had been to, stayed or passed by this particular location on this date at 9.30 am. User A had also made a call by connecting to the same cell ID at 10:30 AM. There are three possible explanations for this scenario,

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

Considering the last entry, the user had moved to a different location by 6:30 PM. In order to create value from these entries within the transport studies, it is important to analyze whether these are random points visited by the user or a pass by points or whether this is a significant location in the particular user's daily life. But there are various methodological challenges which need to be addressed when working with CDR in order to get the targeted interpretation preciously.

Following sections discuss the methodological framework of the study in detail. The analysis is continued in three sections as discussed in the methodological framework.

4.3 A - User extraction based on caller activity threshold.

Initially, individual records of CDR were processed such that data can be interpreted in an accurate manner. Initial records consist of the time of the call, the day of the week and cell tower locations. A single user record takes the format in Table 4.2. The total no. of calls or the caller activities made within the time period are different from user to user. Due to this dynamic nature of the records, it is difficult to define the exact threshold (intensity of caller activities that a certain user should exceed to be qualified for further analysis) of caller activities to extract users for further analysis. Even the users who are with fewer records are a different category where the frequency of the mobile usage is less in number. Accordingly, users who have records in all the three months were considered for the analysis. As an example, Table 4.2 indicates the dispersion of caller activities of a single user. This particular user does not have at least a single record in the 3rd month of the data set, such users were initially eliminated from the main sample of users and left with 7202 users for further analysis.

Table 4.2: User behavior illustration – Absence of records in the 3rd month

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

Figure 4.2 illustrated below, summarize the results of the first user extraction. 3079 users were disqualified for further analysis as the CDR were not consecutive in

monthly context. The remaining 7202 users were carried forward for the analysis in section B.

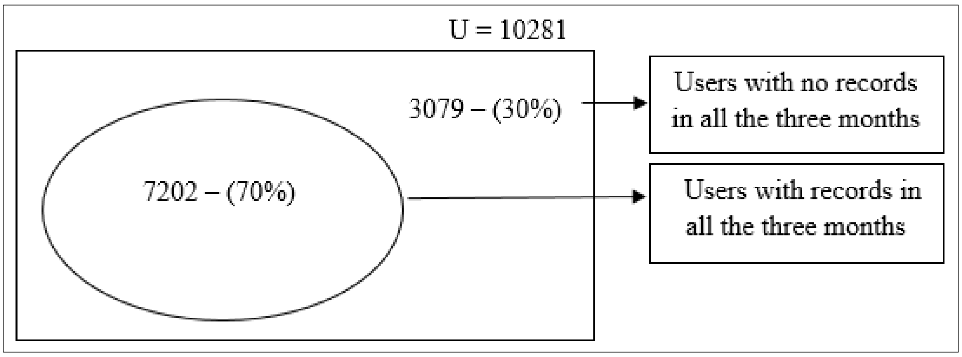

U = 10281
3079 – (30%)
7202 – (70%)
Users with no records in all the three months
Users with records in all the three months

Figure 4.2: Initial user identification based on the nature of the records

4.4 B - Identification of user Presence

In general definition, presence is remaining or visiting a certain place. Turning this into the CDR context, in order to identify the presence of users, there are two main parameters which need to be addressed. That includes presence location (where) and the time of the day (At which date/ mornings or evenings) as shown in Figure 4.3. This section aims to identify these two parameters and categorize users accordingly. Western Province being the study area considered for the analysis, users who present within the Western Province required to be prioritized.

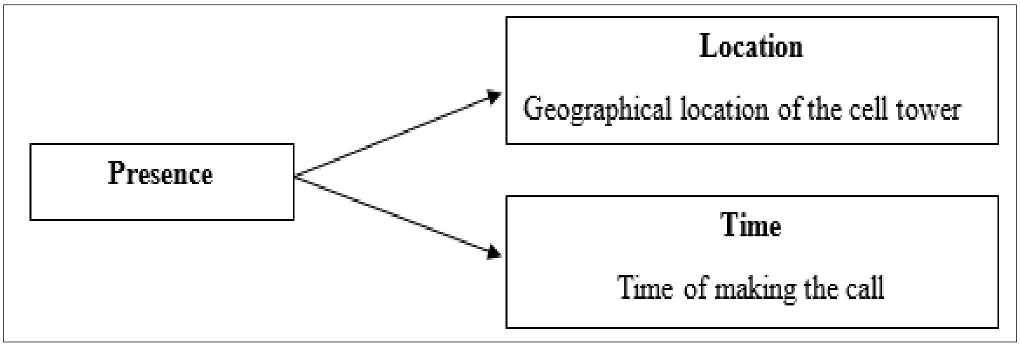


Figure 4.3: Defining a presence

4.4.1 Presence location identification

Geographical information was obtained from the geographical coordinates (latitude and longitude) of network antennas. The precision of the spatial accuracy of the mobile device activity corresponds to the coverage area of a network antenna (Hasan et al., 2015). The coverage area is not spatially fixed and varies according to population density. A single CDR contains important information including the location to which the user connects when making the call. That data is the critical factor in deriving the location parameter. Figure 4.4 is a graphical illustration of the above fact.

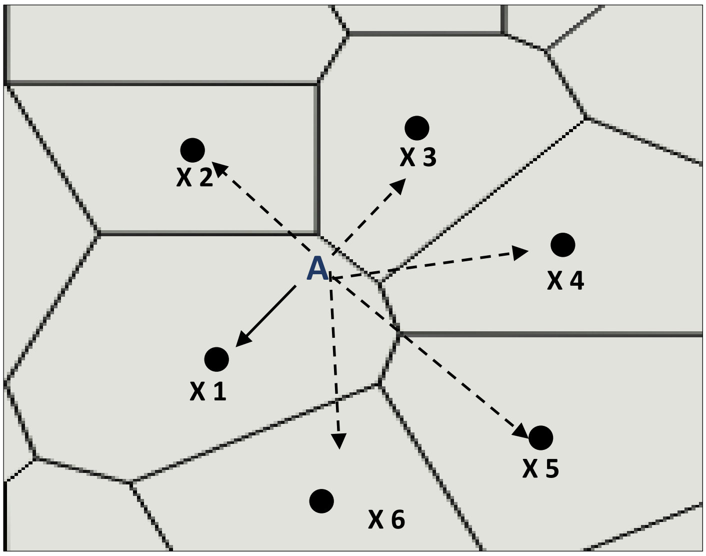
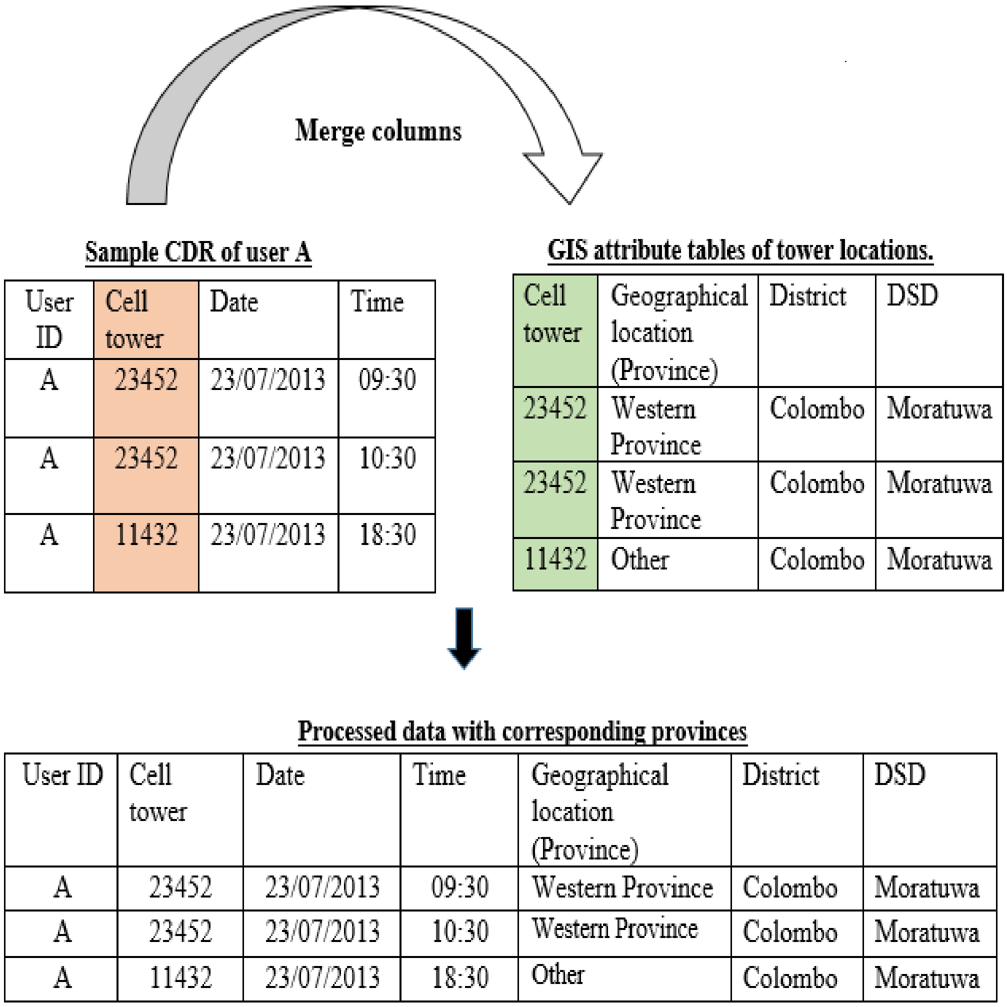


Figure 4.4: Example of user location identification (Gampaha DSD)

User A has the possibility of connecting to any tower as shown in Figure 4.4. Assuming that user A is making a call at 9:30 PM on 23/05/2013 at the demonstrated location and once he dialed the numbers he gets connected to a tower. X1-X6 are the cell towers and their coverage areas are bounded as shown above by keeping the tower as the center point. Assume that user A gets connected to tower X1, it can be concluded that user A had been on that particular location (Anywhere within the coverage area of tower X1) at 9.30 PM on 23/05/2013. But it’s difficult to estimate and compare the locations in this current format of tower name. Therefore, particular geographical locations should be assigned to these towers. The current study assigns geographical locations to provincial, district and as well as in DSD (Divisional Secretariat Divisions) level excluding the GNDs (Grama Niladhari Divisions). In order to perform that, the geographical area to which this cell tower belongs should be identified and it was completed through the GIS software. Finally, CDR entries and the GIS outputs were merged together to identify the geographical location of the particular user at the considered time



Sample CDR of user A

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

GIS attribute tables of tower locations.

Cell tower	Geographical location (Province)	District	DSD
23452	Western Province	Colombo	Moratuwa
23452	Western Province	Colombo	Moratuwa
11432	Other	Colombo	Moratuwa

Processed data with corresponding provinces

User ID	Cell tower	Date	Time	Geographical location (Province)	District	DSD
A	23452	23/07/2013	09:30	Western Province	Colombo	Moratuwa
A	23452	23/07/2013	10:30	Western Province	Colombo	Moratuwa
A	11432	23/07/2013	18:30	Other	Colombo	Moratuwa

Figure 4.5: Identifying administrative areas from tower locations

After processing, an individual's data takes the above format (Figure 4.5). As an example, the user A had located within the WP during the morning and he had also gone outside the WP during the evening of the particular day. District and the DSD can also be assigned to the cell towers in a similar manner. For this initial study, the spatial regularity considered is the provincial boundary, but for further studies, it is important to take this into finer levels including districts and DSDs. The Table 4.3 below shows the total no. of DSDs within the districts of the WP and their coverage areas. The best available validation data was generated by the trip distribution counts at the Divisional Secretariat Division (DSD) level for the Western Province (Maldeniya et al., 2015).

Table 4.3: Cell coverage areas

Province	**District**	**No. of DSDs**	**No. of Cells**	**Area (km^2)**	**Area per cell (km^2)**
Western Province	Colombo	13	307	699	2.2
	Kalutara	13	138	1958	11.5
	Gampaha	14	324	1387	4.2
	Total	**40**	**769**	**3684**	

Assigning DSDs to cell towers

As shown in Figure 4.6, errors considering coverage areas arise for the cells that overlap with the geographical boundaries. Yellow color highlighted cells are overlapping with the boundaries and always an error occurs when deciding to which particular district do we assign the considered cell. As an example, cell A may belong to Mahara as well as Gampaha DSD. For the current study, boundary cells are assigned randomly to the DSDs ("A" as a cell in Gampaha and "B" as a cell in Mahara) such that errors will be eliminated due to cancelling with each other (By assigning A to Gampaha there is an overestimation and it gets cancelled by the underestimation created due to the assigning of cell B to Mahara). The scenario with the GNDs are more complicated and the error percentage that can occur is large compared to DSDs.

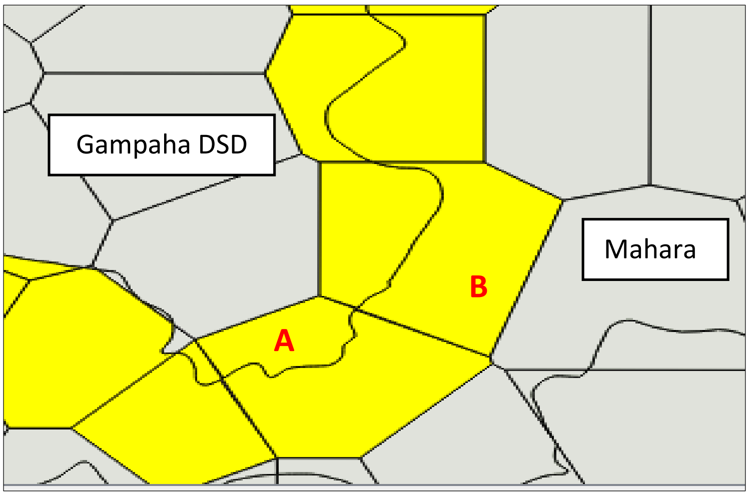


Figure 4.6: Cell towers within the Gampaha DSD

4.4.2 Time parameter

Once the locations are identified, it is important to get an idea about the time parameter of user’s behavior within the Western Province. Accordingly, the total no. of days the users had visited the WP were cumulated as in Table 4.4.

Table 4.4: Example_ cumulating the total no. of days within the Western Province

User ID	Cell tower	Date	Time	Geographical location (Province)	
A	23452	23/07/2013	09:30	Western Province	Day 1
A	23455	23/07/2013	10:30	Western Province	
A	11432	23/07/2013	18:30	Other	
A	23452	24/07/2013	08:45	Western Province	Day 2
A	23456	24/07/2013	11:50	Western Province	
A	44512	25/07/2013	07:45	Other	

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

From the sample of 7202 users (left from the initial extraction), only 56% (5758) of users had at least a single presence within the Western Province for the period of three months and Figure 4.7 indicate the summary of extracted users up to the discussed analysis.

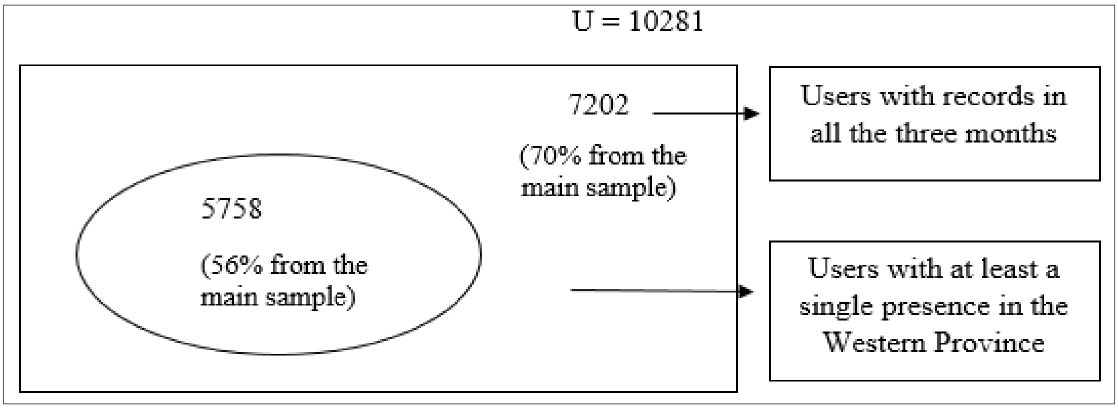

U = 10281
7202
(70% from the main sample)
5758
(56% from the main sample)
Users with records in all the three months
Users with at least a single presence in the Western Province

Figure 4.7: Initial user identification based on the nature of the records

4.4.3 Categorization based on the presence in the Western Province

Further, this 56 % of users were categorized into three segments based on the total no. of days each had visited the Western Province as follows. Appearance percentages to categorize users were heuristically taken in the current analysis.

- **Low presence within WP (258) –** From the considered period of 92 days, this category has appeared within the WP only within 5% of the total of 92 days. (Total presence <=5 days)
- **Medium presence within WP (2256) –** These users are the middle category which is not frequent and at the same time not infrequent. They have appeared in the Western Province from 5% - 75% from the total of 92 days. (5 days < Total presence < = days 69)
- **High presence within WP (3244) –** These users have appeared within the Western Province more than 75% of the days form the 92-day period. (Total presence > 69 days.

4.5 C - Identifying work and home locations

Next step of the analysis is to identify the work and home locations of the extracted users (5758 users). Accordingly, the data requires extensive processing and analysis. Referring to literature and based on a heuristic approach the cumulative presence of each user within particular locations and periods were used as the basis of the analysis in deriving work and home locations.

Considering the working population, their work starting and ending times are different to each other and respectively the period of staying at homes times are almost different to each other. This fact needs to be considered when deriving the work and home locations. In order to avoid the time overlapping of general working hours and home-stay hours, the presence of users at particular locations within the core working, home-stay period was taken into consideration. This will eliminate the difference due to commuting times and work starting times. Aggregated caller activity patterns were observed in order to identify these core periods.

Figure 4.8 indicates the aggregated caller activity pattern of the day. The pattern indicates two peaks as bonded by boxes, where one during the mornings (8 AM – 1 PM) and comparatively a smaller peak during evenings between 7 PM – 10 PM. Which suggest that people are more active with mobile devices within these two periods. The accuracy of the results seems to be high when using time windows with high caller activity levels and this concept was proven by the validation of the sensitivity analysis conducted on identifying home-stay period, which will be discussed in chapter 5.

Previous studies have used different time periods to extract work and home locations. A Study in Dhaka have used 9 AM-5PM to distinguish work presence (Hasan et al., 2015) core hours for work, which are between 8 AM to 5 PM on a weekday, and core hours at home, which are between 7 PM and 7 AM, are taken into account to identify home and work locations by a study in Tokyo (Arai & Shibasaki, 2013). In the study of using mobile positioning data to model locations by Rein Ahas, 17:00 hours is taken as the demarcation point such that calls before 17:00 is taken as work points and calls after as home points (Ahas et al., 2010)

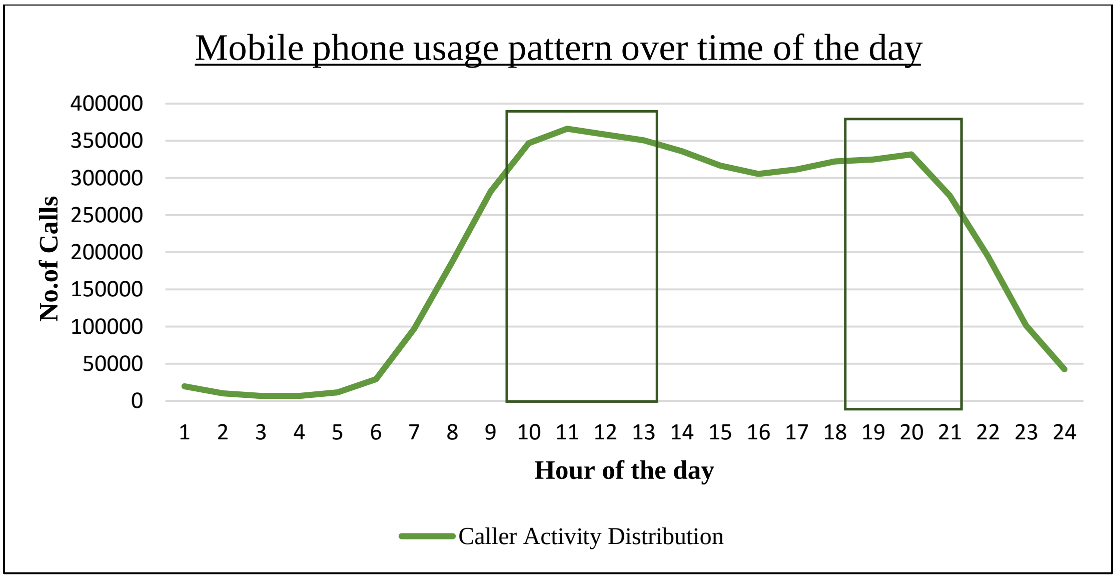


Figure 4.8: Aggregated caller activity pattern of users across the sample of 5758 users

Considering all these facts, the study takes the core period as 10 AM-4PM during mornings to identify work and 7 PM - 4 AM to identify home locations. Figure 4.9 shows the general illustration of the core working period and home period respectively.

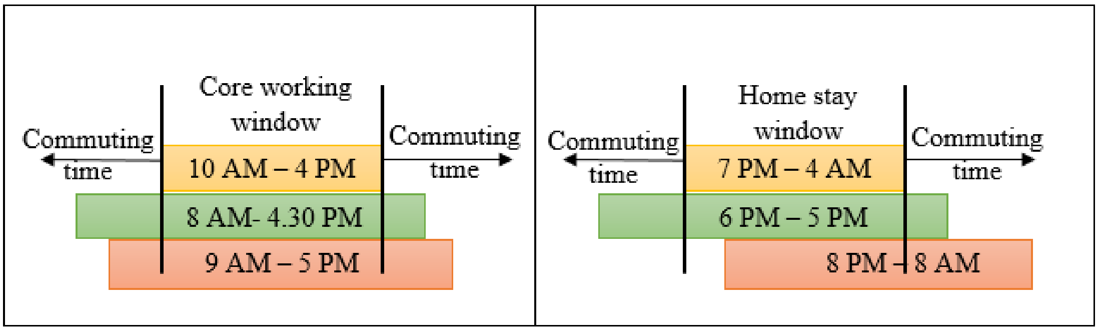


Figure 4.9: Illustration of the core work/ home-stay period

Legend

🟨	User Type 1
🟩	User Type 2
🟧	User Type 3

By adopting this methodology, users who have their working time from 8 AM – 5PM or from 9 AM – 6 PM, all can be captured. In the other end, both the users who come home early and late can be analyzed by employing these core time periods. (Acceptability of these time periods were proven by the validation of the work home distribution in the next chapter)

4.5.1 Aggregating presence of users

As the next step, we sum the total presence in each morning and evening across weekdays and weekends. These time categories are aggregated separately to identify different presence patterns. Time categorization employed is illustrated in Table 4.5.

Table 4.5: Division of time categories

Weekday mornings (**10 AM - 4 PM**)	Weekend morning (**10 AM - 4 PM**)
Weekday evenings (**7 PM - 4 AM**)	Weekend evening (**7 PM - 4 AM**)

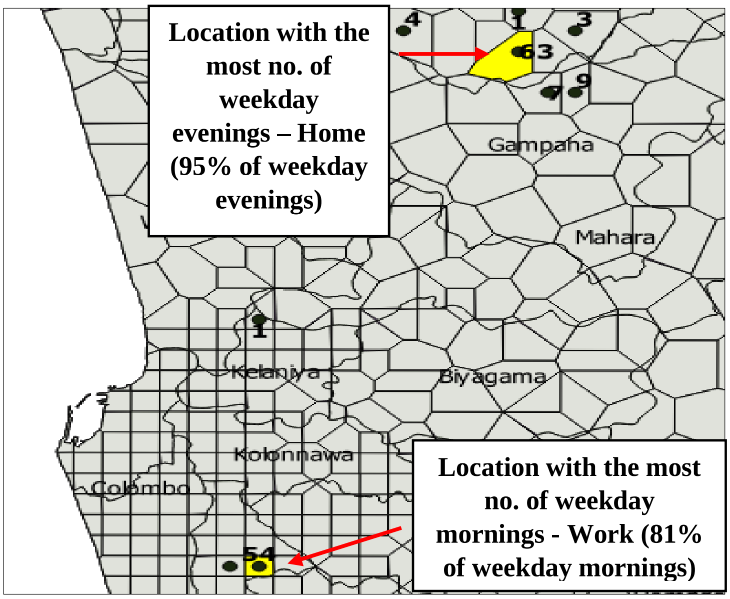
When calculating the total presence, the study considered the totals in terms of days but not the total no. of calls made. For an example assume that a user makes X (X>1) no. of calls on a weekday morning at cell Y. Then that is counted as one but not as X records. Such that the processed data can be interpreted as the total no. of weekday mornings a user had appeared in a particular cell, Total of weekday evenings and etc. Table 4.6 shows an example of a processed user. The user had appeared on 60 weekday mornings in cell number 25634 and 34 weekday mornings in cell number 25636 and etc.

Table 4.6: Presence of a single user

User	**Cell**	**Time**	**Frequency of Presence**
X	25634	Weekday morning	60 **Work**
X	25636	Weekday morning	34
X	25633	Weekday evening	63 **Home**
X	37419	Weekday evening	10

Work - Considering user X, there are three two main locations with 60 and 34 presence respectively in weekday mornings. Out of these two cell towers, 25634 on which this user X had present most frequently (60 weekday mornings) was taken as his work location.

Home – Similarly user X, had presence in 63 days on 25633 and 10 days on 37419 in weekday evenings. Out of these two, cell tower 25633 which had the highest no. of nights was considered as his home location.



Location with the most no. of weekday evenings – Home (95% of weekday evenings)

Location with the most no. of weekday mornings - Work (81% of weekday mornings)

Figure 4.10: Illustration of the spatial distribution of a single user

Accordingly, the cell with the top frequent presence during night times was taken as the home location while the cell with the top frequent presence during weekday mornings was considered as the work location. The validity of these identified work and home locations will be analyzed within the next chapter. A random user’s work and home locations along with other visited points are illustrated in Figure 4.10.

4.5.2 Minimizing the errors of CDR

There are two main issues with CDR that requires further attention. Two issues recorded are the load sharing effect and trip truncation. The current study focuses on minimizing these issues with different approaches, but a significant attention requires to be given to these aspects in further researches.

Load sharing effect

The load sharing effect occurs due to the cell tower switching without actual moments of the user. This leads to the overestimations of the total moments of the user. As an example, assume that user makes a call by connecting to cell tower A and a few minutes later he makes a call by waiting at the same location. But the second call was made by connecting to cell tower B. This will give an incorrect interpretation as the user had been present in both cell coverage areas of A and B. But in actual scenario he had been only within the coverage area of A. In the current analysis this error had been minimized up to a certain extent by using the cells with the top frequent appearance of location identification. In further studies, it is recommended to use different data mining techniques to minimize the existing issue.

Trip truncation

Figure 1.5 in chapter 1 introduces a basic error in trip identification, which leads to trip truncation. Basically, a single trip is identified as a no. of trips with the various parameters followed. In the current study, this error is minimized up to a great extent by prioritizing on the trip purposes (home-work trips), rather than focusing on different moments in between.

4.5.3 Overview of the sample

Considering the night time location there are users with same top frequent location during weekday nights and weekend nights and some users have different top frequent locations for those two time periods. If the locations are the same the accuracy of claiming that the particular user is a residence within that cell area is comparatively high. Therefore, the study considered those two behavior patterns as two categories and analyzed separately in future sections. Based on the home and work locations identified by the above methodology and the cell towers of the particular significant locations, the users were basically categorized into three segments as low mobile users, highly mobile residents and expatriate residents (Table 4.7).

Table 4.7: Basic user categorization according to user appearance

User type	**No. of users**	**Description**	
Low mobile users	3121 – (54%)	Most frequent cell during Weekday mornings, weekend evenings and weekday evenings are same.	Does not indicate a significant mobility, these might be workers resided closer to their home locations or un-employed users.
Highly mobile residents	1193 – (21%)	Most frequent cells during Weekday evenings and weekend evenings are same while weekday morning location is different.	Users who have clearly identifiable wok and home locations
Expatriates	1444 – (25%)	Most frequent cells during Weekday evenings, weekend evenings are different to each other.	Users who might be the residence in one location but travel to work by staying at a different location.

Above is a simple categorization of the users to understand the overall picture of the selected sample. In general, these immobile users might be working closer to their home locations or they may not be working at all and that accounts for the highest percentage of the users. At glance, the second type consists of the ones with clearly identifiable wok and home locations. This category of IDs was further used for validation purpose. In case of expatriate residents, they have high mobility compared to others and those seem to be moving to different locations during weekend evenings.

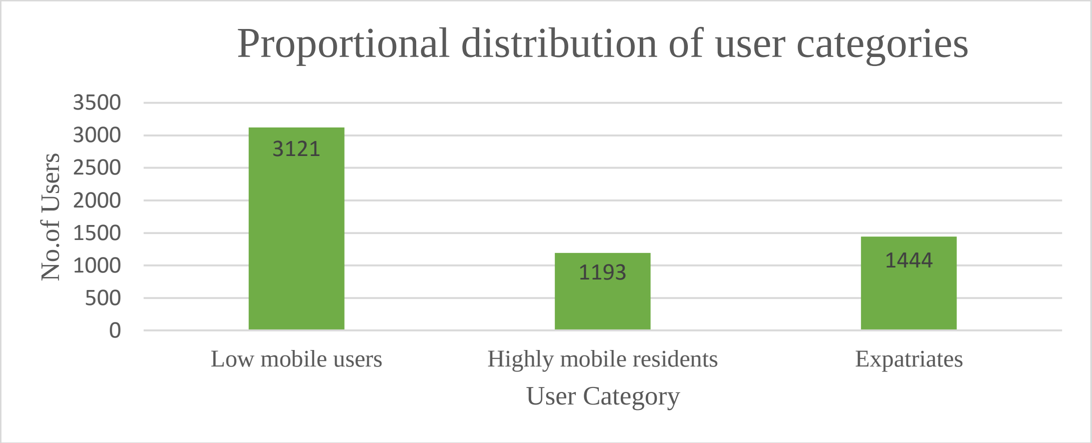
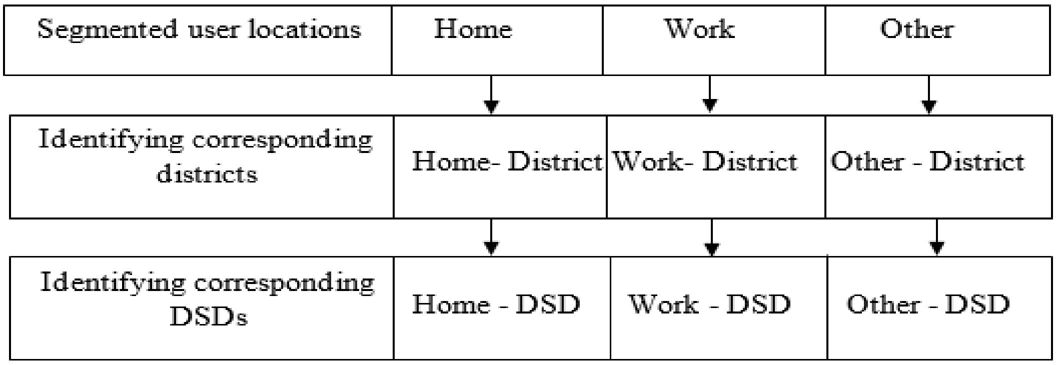


Figure 4.11: Overview of the sample

4.5.4 Framework for identifying work – home distribution

This section analyses in the distribution of home and work locations in a geographical context. The hierarchical approach followed is illustrated in Figure 4.12.



Segmented user locations	Home	Work	Other
Identifying corresponding districts	Home- District	Work- District	Other - District
Identifying corresponding DSDs	Home - DSD	Work - DSD	Other - DSD

Figure 4.12: Identifying work - home distribution

Once top frequent locations of the users were identified, the corresponding district and the DSDs was assigned to the cells by merging spatial data with cell tower records.

Only the highly mobile residents who have different top frequent locations for morning and night times were considered for comparison. In addition both the locations of the considered users were within the WP. Identified users were cross-classified as an O-D matrix based on their work home locations at district level. Percentage of home - work distribution obtained from CDR and the HVS sample are shown in table 4.8 and 4.9.

<u>CDR data</u>

Table 4.8: Home - work distribution _mobile data

Home/Work		Trip Attractors			**Total**
		Colombo	**Gampaha**	**Kalutara**	
Trip Generators	Colombo	52%	2%	2%	**4%**
	Gampaha	12%	21%	0%	
	Kalutara	5%	0%	7%	
Total		**17%**			

(→ = Aggregation of percentages)

Example – 52% of highly mobile residents have their work and home location within the Colombo District, 2% of users have their home location in Colombo district and the work location within the Gampaha district.

<u>HVS data</u>

Table 4.9: Home -work distribution_ HVS data

Home/Work		Trip Attractors			**Total**
		Colombo	**Gampaha**	**Kalutara**	
Trip Generators	Colombo	45%	2%	1%	**3%**
	Gampaha	10%	28%	0%	
	Kalutara	4%	0%	10%	
Total		**14%**			

(→ = Aggregation of percentages)

As the arrow indicates, the Cumulated percentages of trips attracted to Colombo districts from other two districts and cumulated trips generated from Colombo to other two districts are closer to each other as shown below. The necessary deviations seems to be well explained by the CDR data.

4.5.5 User categorization based on the residential area.

Once the home and work locations were identified as above, the sample of 5758 users were further categorized into three sectors as follows. The provincial location of the residence was used as the criteria for categorization.

- **Residence within Western Province** – These users have their top frequent location during night times within the Western Province. From these users, some had the same top frequent location during weekday and weekend night times and some had different top frequent locations for weekday nights and weekend nights. Within this category regardless of the location similarity, one or both the locations were within the Western Province.
- **Residence outside the Western Province** – Most frequent nighttime locations were outside the Western Province.
- **Inter provincial commuters** – This group of users had different top frequent locations during weekday nights and weekend nights and for these users, either location were outside the Western Province

4.5.6 User Model implementation based on the presence pattern of the day.

Next step of the analysis was the identification of different user models. The work and home locations identified from the above methodology were used as the basis of the model development. A single day is divided into different time categories and the cells with the most frequent appearance during each of these time categories were identified. Finally, different user models were predicted based on the similarity of the identified cell to the home and work location. Predicted models and the criteria used for the user selection are shown in Table 4.10.

Table 4.10: User Model Description

Model name	**Assumption of Working Conditions**
Public sector worker	General working hours are between 8.30 AM to 4.30 PM and do not work during the weekends. Office workers are included in this category. Government shift workers like police, security and etc., or workers with an exceptional time of working were not captured by this category.
Private sector worker	Working hours generally change from company to company. Starting time may vary between 8 AM -10 AM and ending time normally vary from 5.00 PM- 7.00 PM. Some companies work on weekends too. The core working period somehow disperses between 10 AM - 4 PM and added to that they may work after 4 PM.
School worker	School employees generally work from 7.30 AM- 1.30 PM and that period remain constant compared to other categories. Students and teachers are included within this category but school administration staff who normally work till 4.30 PM are exclusions to the current category
Highly mobile users	Users within this category don't have a fixed place of work like the above categories. They travel from place to place for their work. Salespeople, drivers are inclusions of this category. They generate the highest no. of trips compared to other categories with low regularity for any location.
Highly immobile users	This category of users are not employed or in general, they do not change their home location considerably. This category may include housewives or even the shop owners who have their boutiques closer to their home locations may fall within this category. Most of the self - employed people likely to be elements of this set.

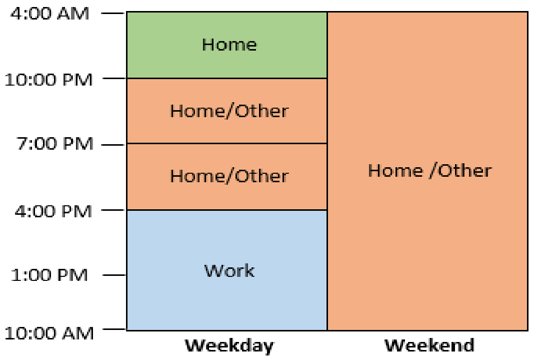
Top frequent cells of the users were analyzed within different time categories as follows. Also, the selection of the time windows was done by aiming the core working periods of the considered employment categories as discussed in Table 4.11.

Table 4.11: Description of time categories

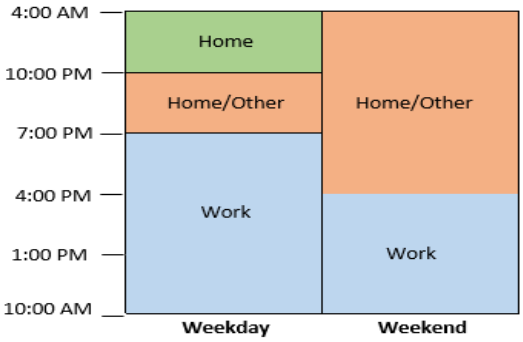
Time Category	Description
10 AM -1 PM	Core working period of school, public, and private workers.
1 PM - 4 PM	Included within the core working period of public and private workers. Support to distinguish school workers from other sectors.
4 PM - 7 PM	Included within the working period of a private-sector worker and support to identify public and private sector workers separately.
7 PM - 10 PM	Use to identify home location more accurately and distinguish users with high mobile activity from other time periods
10 PM - 4 AM	Core periods correspond to home-stay

Furthermore, Table 4.12 demonstrate the different models identified along with the criteria used to filter each model. Users who fulfill the criteria in a similar manner were clustered together to derive the totals of each category. Different colors are used in each time segments to indicate the similarity of the cell usage within time windows.

Table 4.12: Demonstration of the models

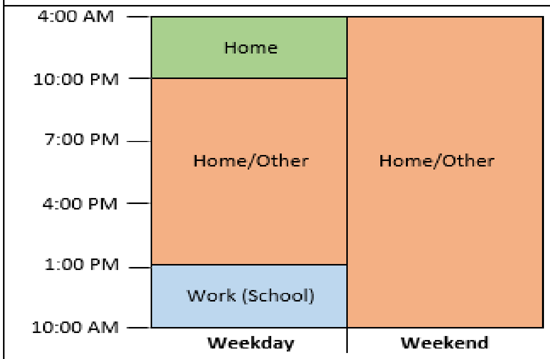
Model 1 – Public sector (No. of users - 202)	
 4:00 AM 10:00 PM 7:00 PM 4:00 PM 1:00 PM 10:00 AM Home Home/Other Home/Other Work Home /Other Weekday Weekend	• From 10 AM- 4 PM the cell with the most frequent appearance is the same as work and the work cell does not appear on weekends.

Model 2 – Private sector (No. of users - 383)



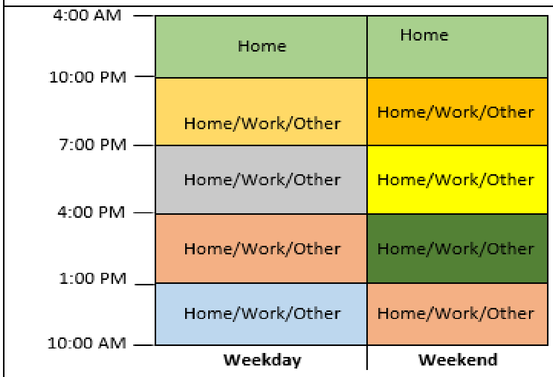
- From 10 AM- 7 PM the top frequent cell is similar to work for some users and some have the records within the work location during weekend mornings (Saturday morning)

Model 3 – School (No. of users - 35)



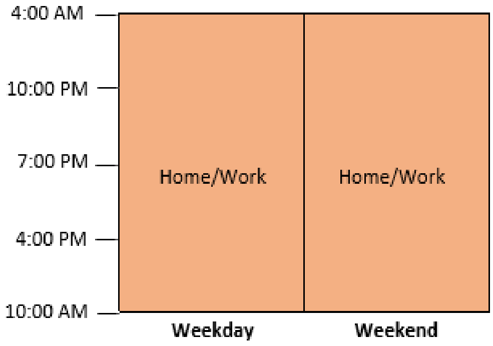
- Top frequent location from 10 AM – 1PM equals to work cell on weekday mornings. From 1PM the cells are different with no appearance during weekend mornings.

Model 4 – Highly mobile users (No. of users - 1024)



- Top frequent cells during the considered time categories are different to each other. These users do not possess a specific place of work

Model 5 – Highly immobile users (No. of users - 1941)



- Top frequent cells during the considered time categories are almost equal to each other.

4.6 Summary of user categorization

Dynamic characteristics of travel behavior were analyzed within the earlier sections based on the total presence or appearances of users within particular locations in considered time windows. Figure 4.13 illustrates the summary of all the identified categories. But a significant percentage of users (34% from the sample of 5758) could not be classified to a specific employment category. The major reason for this effect is the irregularity of the records. An analysis of activity engagement should present a regularity in daily or weekly presence of the identified locations. But these undefined users did not possess this regularity in their caller activities.

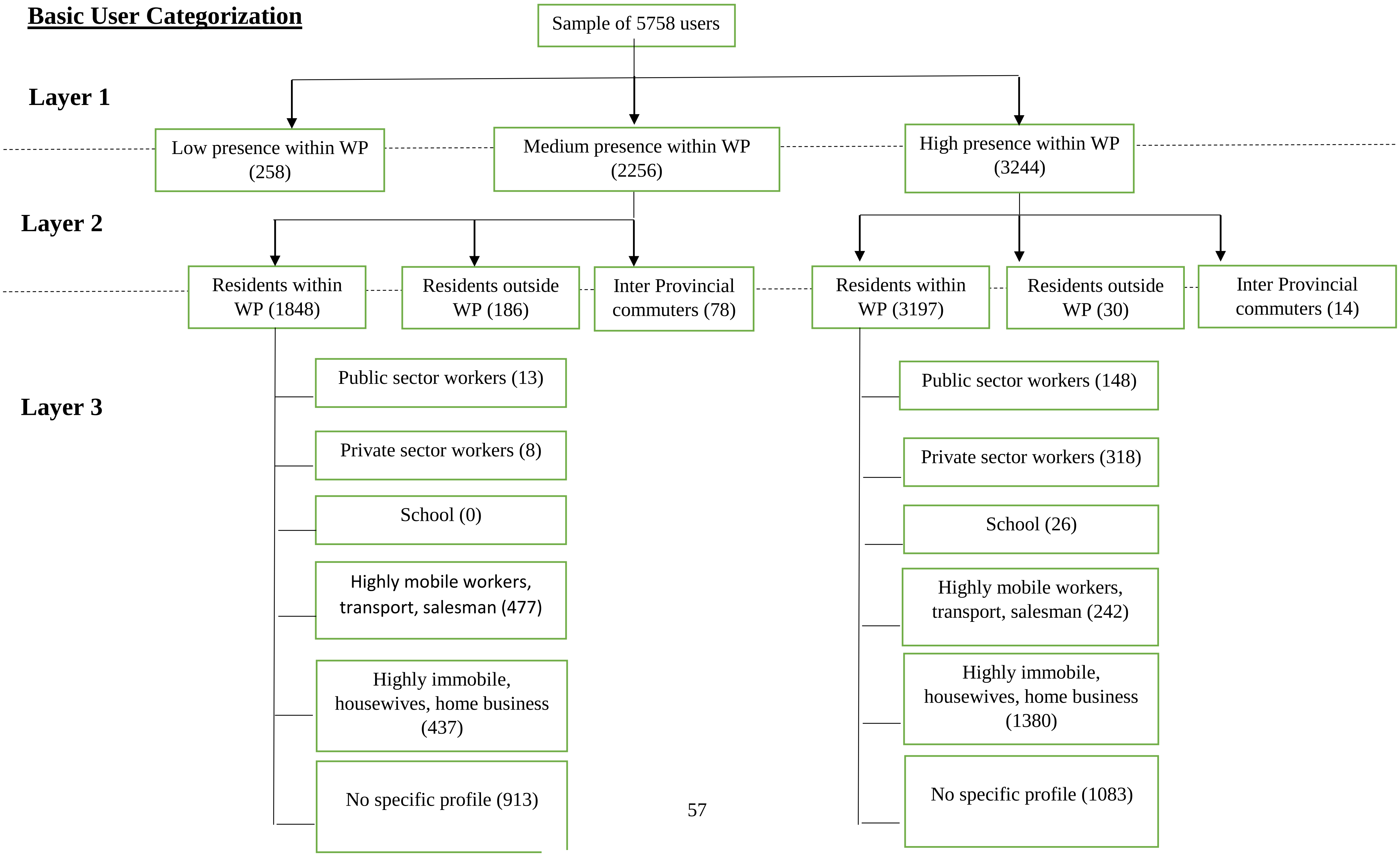

Basic User Categorization
Sample of 5758 users
Layer 1
Low presence within WP (258)
Medium presence within WP (2256)
High presence within WP (3244)
Layer 2
Residents within WP (1848)
Residents outside WP (186)
Inter Provincial commuters (78)
Residents within WP (3197)
Residents outside WP (30)
Inter Provincial commuters (14)
Layer 3
Public sector workers (13)
Private sector workers (8)
School (0)
Highly mobile workers, transport, salesman (477)
Highly immobile, housewives, home business (437)
No specific profile (913)
Public sector workers (148)
Private sector workers (318)
School (26)
Highly mobile workers, transport, salesman (242)
Highly immobile, housewives, home business (1380)
No specific profile (1083)

Figure 4.13: Summary_ User categorization

CHAPTER 5 – VALIDATION OF WORK-HOME DISTRIBUTION

5.1 Introduction

Accurate estimation of human behavior is crucial for transport estimations. Due to the regularity of human behavior (Hasan, Schneider, Ukkusuri, & Gonzalez, 2013) we were able to infer home and work stay locations for users from CDR data. But it's important to ensure the accuracy and reliability of the data to use them in future transport estimations and researches. The current chapter discusses the validation of the findings and explores that there is a possibility to use frequently updated CDR instead of time-consuming travel surveys to estimate total trip generations and attractions.

5.2 Result validation

Home and work distribution identified from the CDR at DSD level were statistically validated through regression and F-test analysis with the household survey data within the following sections. F test analysis is used for the validation, as the requirement is to compare the variations between two samples of data. If the variations are high the F value will be higher than 0.05 and low variations provide a F value less than 0.05

The analysis was done for 40 DSDs within the Western Province and the 40 DSDs were finalized by balancing the difference in two data types with the boundaries. In the HVS data, separate statistics were collected for Fort/Pettah, Maradana, Thimbirigasyaya, Wellawatta, Borella Kotahena, Kollupitiya, Kurunduwatta, and Bambalapitiya DSDs. But in mobile data, this group of DSDs was processed as Colombo. Therefore the current analysis also continued by taking each of these DSDs under Colombo DSD.

5.3 Statistical analysis

Following the previous steps, only the highly mobile residents who have clearly identifiable work and home locations within the Western Province were used for the validation purpose. Validation is completed in three sections, such that work distribution for one time window which is the core working period and home distribution for two time windows as the core home presence period and the night period with highest no. of caller activities. Work distribution validation needed only one analysis as the peak period of mobile usage during mornings was already embedded within the core working period. In case of home location distribution, time window with the highest match out of the two was used for further analysis as mentioned in chapter 4.

5.3.1 Validation of workplace distribution

Results obtained from the work location were validated with the HVS data by considering the time window as 10 AM - 4 PM and following steps describe the process of the analysis.

Step 1 – Calculate the total no. of weekday mornings (10 AM - 4 PM) that each user had appeared on each of the cells. Select the cell with highest no. of appearances as the work location and record the corresponding DSD for each user's work location. Next, the percentages of work locations were calculated for each of the DSD as in Appendix C and the corresponding HVS results were obtained to start the statistical tests.

Step 2 - F-test

The P- values of the two data sets were analyzed to determine whether mobile and HVS data are statistically significant. Data were tested with 95% significant level using Minitab software. A significant level of 0.05 indicates a risk of concluding that a relationship exists where there is none in actual effect.

> **Hypothesis - H_1**: There is a significant relationship between the mobile data results and the HVS data.
>
> **Null Hypothesis - H_0**: There is no significant relationship between the mobile data results and the HVS data.

If the p-value is greater than the significance level (0.05) the effect is not statistically significant and if the p-value is less than or equal to the significance level, then the effect for the term is statistically significant which reject the null hypothesis.

Step 3 – Interpretation of the test results

F-test produces a p-value of 0.000 which is smaller than 0.05 and that essentially reject the null hypothesis. The two data sets, mobile and HVS data seems to have a significant relationship between each other and the Pearson correlation of mobile and HVS have significantly a high value of 0.89.

Once the two data sets are fitted to a linear model the R-squared value lies as 80%, which says that the 80% of the variation in the HVS or the traditional survey data can be explained by the model comprised of the mobile data (Appendix D). A high R^2 value does not necessarily indicate that the model meets the model assumptions (Behavior of mobile data). It should be checked with the residual plots to verify the influence of mobile data. The probability plot assesses how closely the data points follow the fitted distribution line. If the specified theoretical distribution is a good fit, the points fall closely along the straight line. As indicated in Appendix D the points seem to be falling in line with the straight line which suggests the linear model of mobile data is a good fit to predict travel data.

In conclusion, the methodology of work location identification with mobile data is applicable and statistically proven at DSD level. Refer to Appendix G in order to have a more clear understanding of the geographical demonstration of work location distribution for both HVS and CDR datasets.

5.3.2 Validation of home distribution

Home locations are validated at two time windows as in table 5.1 and the most statistically validated time window was used for the other analysis work.

Step 1 - Calculate the total no. of weekday evenings (10 AM - 4 PM) that each user had appeared on each of the cells for both A and B time windows separately. Select the cell with highest no. of appearances in A and B categories as the home location and record the corresponding DSD for each user’s home location. Next, the percentages of home locations were calculated for each of the DSD as in Annex C for both windows and the corresponding HVS results were obtained to perform the statistical tests

Table 5.1: Time window selection in validation of home locations

	Time window	**Reason**
A	10 PM – 4 AM	Considering a general individual, the core home-stay period tends to be within the proposed time window
B	7 PM – 4 AM	According to the aggregated caller activity graph, the majority of the people tends to make calls between 7 PM- 10 PM, which increases the accuracy of the results. Including that period to identify home locations can improve the accuracy of the results

Step 2 - F-test

Results were tested with 95% significant where a significant level of 0.05 indicates a risk of concluding that a relationship exists where there is no actual effect.

Table 5.2: Statistical results of two time categories

	Time window	**Result**	**Interpretation**
A	10 PM – 4 AM	P-values – 0.000 Pearson Correlation – 0.6 R-squared value - 35.9 %	Two data sets possess significant relationship among each other with a medium correlation. But the R-squared value is significantly low which suggest that the total amount of variation in HVS explained by the mobile data accounts only for 35% (Appendix D)

	Time window	Result	Interpretation
B	7 PM – 4 AM	P-values – 0.000 Pearson Correlation – 0.81 R-squared value - 71%	Two data sets possess significant relationship with each other with a high correlation of 0.81. The R-squared value is comparatively high and a 70% variation is explained by the model. (Appendix D)

As the R-squared value is at an acceptable level, a residual plot is carried out for the data of time window B and appendix D shows the normal probability plot graph. Accordingly, the points seem to be in line with the straight line which supports the acceptability of the model. In conclusion, the time window B (7 PM – 4 AM) produce more validated results than the time window A (10 PM – 4 AM). Therefore, the home locations were calculated by considering the time window B.

Furthermore, this proves that the more accurate results can be obtained when the total no. of caller activities are high in frequency. Geographical demonstration of home location distribution with time window B is attached with Appendix F.

5.4 Conclusion

This section discussed the applicability and validity of findings built upon the CDR data which was extracted by excluding the noise embedded with them. The observations suggest that the CDR data reasonably represent the spatial distribution of work and home locations. This chapter provides two key contributions in identifying significant locations from CDR. First, the analysis proposes that the work location can be forecasted accurately by considering the cell with the highest no. of presence during the time window of (10 AM- 4 PM) and secondly home location can be identified by taking the cell with the highest presence during 7 PM-4 AM. The total fit of the data appears to be 76% and the high accuracies indicate that the method is ready for the application. Moreover, the accuracy of the predictions tend to be precious when the total no. of records possess by users are higher.

CHAPTER 6 –SUMMARY & CONCLUSION

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 and etc. In this thesis, the CDR phenomenon is analyzed to create value in the transportation sector. Current chapter summarize the findings of the study in three directions including the methodological framework of using CDR for transport initiatives, applications of CDR based findings and the validation of the CDR findings. Final section of the chapter describes the several ways to expand the current study in future work

6.1 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, characterizing the common behavior patterns using MNBD and validating the extracted information with traditional transport survey data

6.1.1 Methodological framework of using CDR

The first part of the study develops a methodology to extract behavioral patterns, despite the noise embedded in the CDR data. Although CDR has been widely used in human mobility and transportation research, no previous study has touched on generating CDR based results starting from the fundamentals of transportation.

Dataset consists of the CDRs of 10281 users for a period of 3 months. The initial task was to process the raw data by removing the noise embedded within the CDR s. From the main data, the users who had at least a single record in each of the considered months were selected, with that criteria only 70% (7202) of the users were left for further analysis. Out of this 70 % of users, IDs who have at least a single record within the Western Province (as the study area is the Western Province) were filtered for future work. From the main dataset, only 57% (5758) of users had visited the Western Province at least once. Next task was to characterize the locations of these filtered users by their presence during different time periods of the day based on the observations from Call Detail Records. The study employs four time categories as weekday mornings, weekday evenings, weekend mornings and weekend evenings for the user behavioral analysis. Core time period to differentiate mornings and evenings were taken as 10 AM- 4 PM and 7 PM-4 AM respectively. If a user made 30 calls by connecting to a tower on a particular morning of a day, it was identified as the user had been present within the particular tower area and count as a presence in a single weekday morning. Similarly, the individual presence patterns of users during each of these time intervals were interpreted such that total weekday mornings on a particular location, total weekday evenings on a cell and etc., for the considered period of May, June, and July (2013). User locations were identified as cell towers and the towers were merged with files of administrative boundaries to locate towers with corresponding districts or DSDs.

6.1.2 Application of CDR based findings

Considering the application context of the CDR-based findings, initially home and work locations were identified based on the highest frequency of appearance of users in each of the considered time categories. Accordingly, the cell with the top frequent presence during night times was taken as the home location while the cell with the top frequent presence during weekday mornings was considered as the work location. Identified cells of the work and home locations were matched with the corresponding administrative boundaries including districts and DSDs to figure out the work home distribution of the users.

Different user models had been identified as the next step. A single day divided into time segments to match with the work starting and ending periods and the users who shared same behavioral patterns across the segmented time categories were clustered based on a heuristic approach to identify employee categories. In summary three layers were identified by different categorizations. In the first layer, users were clustered into three categories based on their presence in the study area: Western Province as users with low presence within WP, users with medium presence within WP and users with high presence within the WP. The second layer further categorize them based on their residing Province and the third layer identified the shared behaviors based on the employment type and ultimately five categories were identified as public workers, private workers, highly immobile workers, mobile workers and school workers.

6.1.3 Validation with HVS data

This section tests the method on the CDR data with the traditional survey data. In order to carry out further research activities and use CDR in other planning activities, it is important to verify the accuracy of using CDR. As described in the above chapter people move for different purposes but home-based work trips are the most commonly and accurately identifiable category. Therefor the study validate the findings by comparing the distribution of the work and home locations of the sample of users based on two geographical levels as district and DSDs. After the validation at DSD level, results suggest an acceptable fit of 76% with HVS data.

6.2 Limitations of the study

Use of CDR in transport context is an innovative method in Sri Lankan context, therefore very limited no. 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.

Considering the sample size of 10,281 it is much smaller compared to the 13 million mobile population. After the processing of data, the sample size which can be used for further research drops down nearly to 50%. With a limited sample size, the accuracy of validating with exact numbers reduced hence the percentages were used in the current study.

Moreover, the limitations associated with the properties of CDR reduce the accuracy of the findings. Mainly the load sharing effect, which occurs due to cell tower switching reduce the accuracy of the location predictions , but the study minimizes the effect by giving the priority to the most frequently used cell.

Furthermore, with the implemented method, the users who are shift workers can be misclassified. Their work and home-stay periods are different from regular users. If a particular user is on a night shift and the total no. of night shifts was high, he should be spending the majority of the night locations at work. This will lead to an incorrect classification of the home and work locations.

6.3 Future Research

There are many opportunities to extend this research further. Analyzing presence patterns and identifying work and home locations is the initial point of exploring the understanding of human mobility. Different directions on which the research can further be extended are,

- The models discovered for the work and home identification is just to reason out that the methodology is statistically valid. This can further be improved to a higher level by using more socio-economic features and other demographic factors.
- Other than work and home location, there are significant locations that users tend to travel during their daily routine. Other different locations of travel can be identified by employing various spatial data mining techniques and more behavioral patterns can be identified. Based on that the users can be segmented further into finer clusters.
- Moreover, estimated commuting characteristics can further be used to identify the commuting distance, departure and arrival times and etc. Such that the traditional four-step model can be reached to further steps using CDR data. This can also be enhanced by combining with other geographical data sources like land use data.
- Aside from the transportation field, the research can further be extended to social science aspects: Other than Call Detail Records there are other type’s big data sources like GPS and even online transaction, social network data and etc. More clear and different user profiles can be identified by employing these data with more sophisticated data mining techniques.

- The current study does not take the load sharing effect of CDR into consideration. Future works can also be extended to find a proper methodology to minimize the effect from this cell tower switching process and increase the accuracy of the results.

REFERENCES

Ahas, R., Silm, S., Ja, O., Saluveer, E., & Tiru, M. (2010). *Using mobile positioning data to model locations meaningful to users of mobile phones.*

Alexander, L., Jiang, S., Murga, M., & González, M. (2015). *Origin–destination trips by purpose and time of day inferred from mobile phone data.* ScienceDirect.

Arai, A., & Shibasaki, R. (2013). *Estimation of Human Mobility Patterns and Attributes Analyzing Anonymized Mobile Phone CDR.*

Beyer, M., & Lanley, D. (2012). *Gartner*. Retrieved from http://www.gartner.com

Buchanan, C. (1976). Transport for society. *Institution of Civil Engineers.*

Caceres, N., Wideberg, J., & Benitez, F. (2007). *Deriving origin destination data from a mobile phone network.* IET Intelligent Transport Systems.

Calabrese, F., Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE.*

Castillo, E., Menendez, J., & Jimenez, P. (2008). *Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations.*

Csaji, Browet, B., Traag, V., Delvenne, J., Huens, E., Dooren, P., . . . Blondel, V. (2013). Exploring the mobility of mobile phone users. *Physica A.Statistical Mechanics and its Applications,.*

Dash, M., Koo, K., Decraene, J., & Yap, G. (2014). *CDR-To-MoVis: Developing A Mobility Visualization System From CDR Data.*

Esko, S. (2013). *Mobile positioning datasets in mobility studies.*

Farkavcova, Guenther, E., & Greschner, V. (2010). Decision making for transportation systems as a support for sustainable stewardship. *Emerald Insight*.

Gamage, M. (2016). Is Google Loon right technology for Sri Lanka.

Groves, R. (2006). *Nonresponse rates and nonresponse bias in household survey.*

Hariharan, R., & Toyama, K. (2010). Parsing and Modeling Location histories. *Geographic Information Science*.

Hasan, S., Schneider, C., Ukkusuri, S., & Gonzalez, M. (2013). Spatiotemporal patterns of urban human mobility. *Journal of statistical physica*.

Iqbal, M., Charisma, F., Choudhury, Wangb, P., & Gonzalez, M. (2013). *Development of origin–destination matrices using mobile phone call data.*

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2010). *Identifying Important Places in People's Lives from Cellular Network Data.* Springer Berlin Heidelberg.

Jarv, O., Ahas, R., & Witlox, F. (2013). *Understanding monthly variability in human activity spaces : A twelve-month study using mobile phone call detail records.*

Jiang, F., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and opportunities. New York.

JICA. (2014). *Urban Transport system development project for Colombo MetropolitonRegion and Suburbs.*

Kaisler, S., Armour, F., Espinosa, J., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *Hawaii International Conference on System Sciences*.

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

Khan, F., Ali, M., & Dev, H. (2015). A Hierarchical Approach for Identifying User Activity Patterns from Mobile Phone Call Detail Record. *IEEE.*

Kulpa, T., & Szarata, A. (2016). *Analysis of household survey sample size in trip modelling process.* ScienceDirect.

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

Madhawa, K., Lokanathan, S., Samarajiva, R., & Maldeniya, D. (2015). *Understanding communities using Mobile Network Big Data.*

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

Maldeniya, D., Lokanathan, S., & Kumarage, A. (2015). *Origin - destination matrix estimation for Sri Lanka using Mobile Network Big Data.*

Marcos, R., Vieira, Martınez, E. F., Bakalov, P., Martınez, V., Vassilis, J., & Tsotras. (2010). Querying Spatio -Temporal Patterns in Mobile Phone-Call Data bases.

McGucking, N., & Srinivasan, N. (2011). The journey to work in the context of daily travel. *Census Data for Transportation Planning Conference.*

McNally, G., & Mickael. (2007). *The four step model.*

Mellegard, E., Moritz, S., & Zahoor, M. (2011). *Origin/Destination-estimation Using Celluar Network Data.* IEEE.

Meyer, M., & Miller, E. (2000). *Urban Transport Planning.* McGraw-Hill.

Miyauchi, Gabriel, K., & Yuhei. (2015). *Commuting and Productivity: Quantifying Urban Economic Activity using Cell Phone Data.*

Papacostas, C. (1987). *Fundamentals of Transportation Engineering.*

Parry, K., & Hazelton, M. (2012). *Estimation of origin–destination matrices from link counts and sporadic routing data.*

Planning Tank. (2017). Retrieved from http://planningtank.com

Powell, & Victor. (2014). *Principal Component Analysis*. Retrieved from http://setosa.io/ev/principal-component-analysis/

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

Samarajeewa, R. (2005). *Mobilizing information and communications technologies for effective disaster warning.*

Samarajeewa, R., Madhawa, K., Lokanathan, S., & Maldeniya, D. (2015). *Using mobile network big data for land use classifications.*

Schneider, Belik, C., Couronne, T., Smoreda, Z., & Gonzalez, M. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface,.*

Sri Lanka - Telecoms, M. a.-S. (2016). Retrieved from http://www.budde.com.au

Tiru, M. (2014). *Overview of the sources and challenges of mobile positioning data for statistics.*

Wang, M., Schrock, S., Broek, N., & Mulinazzi, T. (2013). *Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data.* International Journal of Intelligent Transportation Sy 1 stems.

Wright, P., & Ashford, N. (1989). *Transportation Engineering.*

Yang, P., Zhu, T., Wan, X., & Wang, X. (2014). *Identifying Significant Places Using Multi-day Call Detail Records.* IEEE.

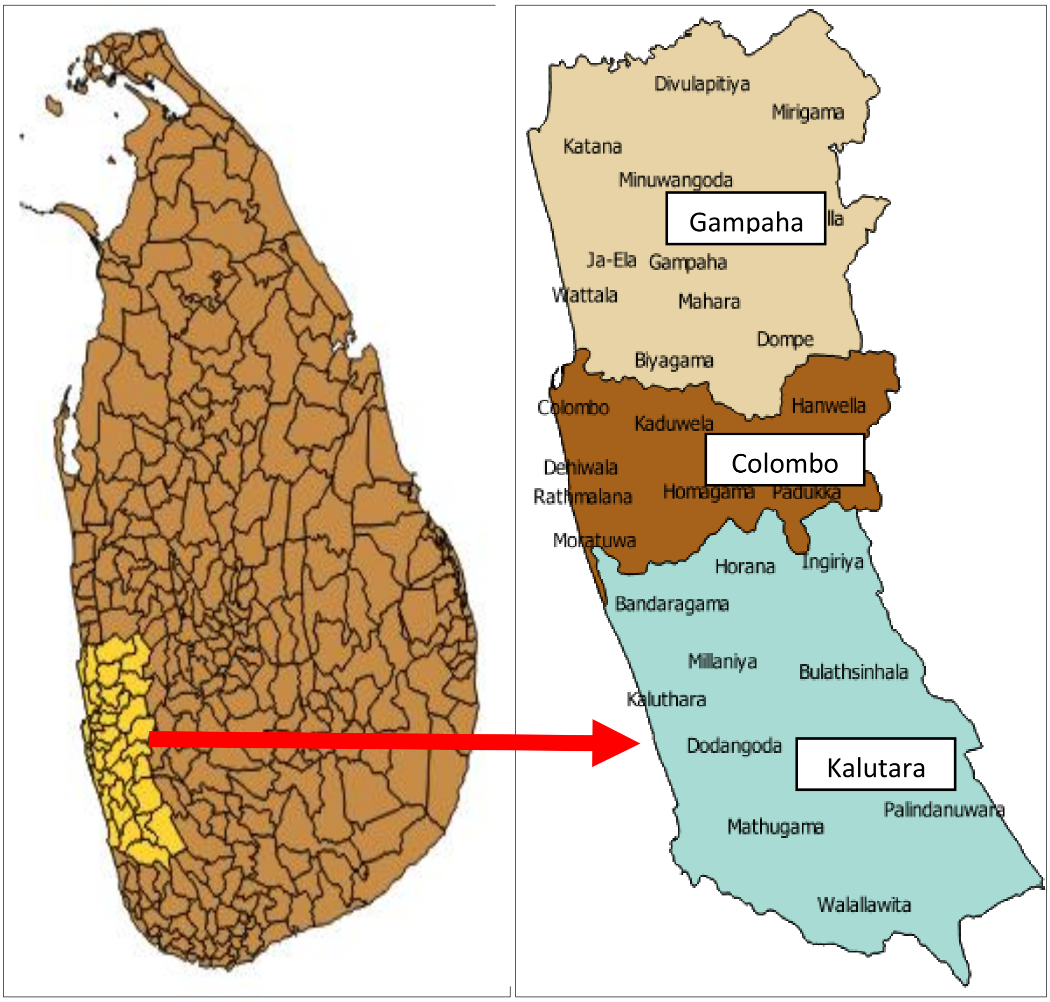
Zhang, Y. (2014). *User Mobility from the View of Cellular Data Networks.*

Zhao, Z., Jinhua, Z., & Koutsopoulos, H. (2014). *Individual-Level Trip Detection using Sparse Call Detail Record Data based on Supervised Statistical Learning.*

Appendices

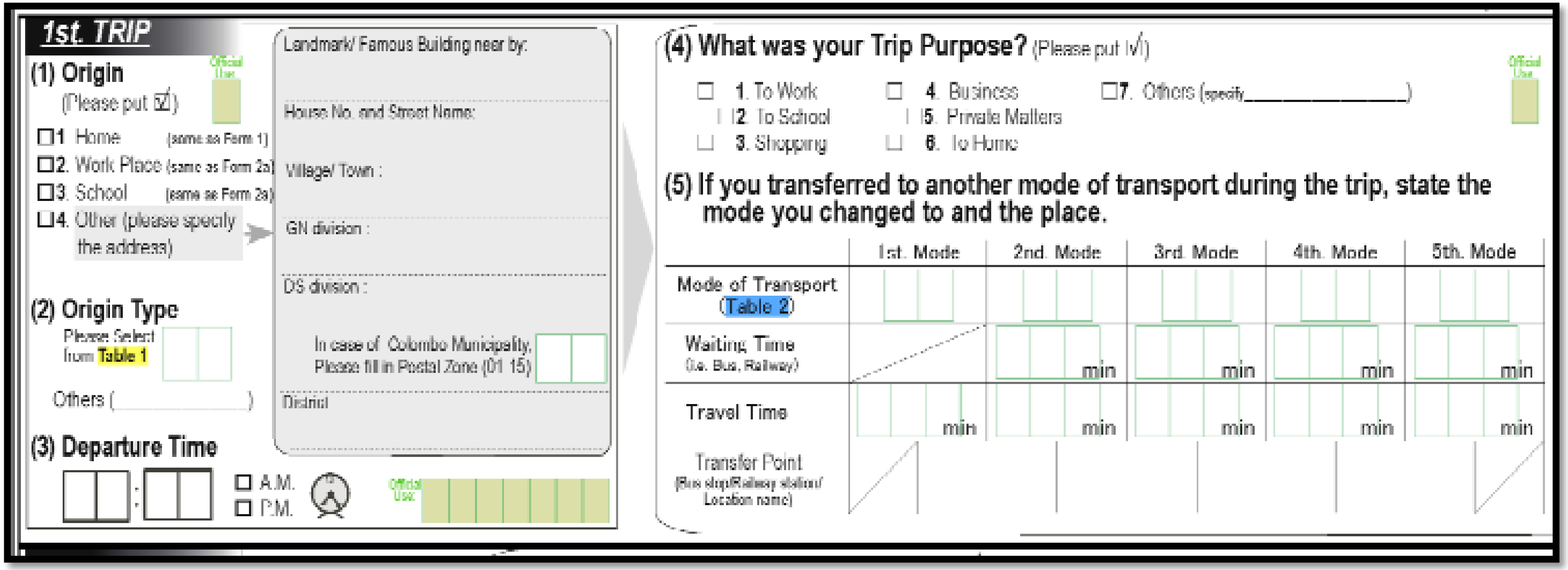
APPENDIX A

Coverage area of the Household Visit Survey



APPENDIX B

Questionnaire- Household survey –gathering trip information (*Source: Urban transport system development project for Colombo Metropolitan Region and Suburbs, Person trip survey*)



1st. TRIP

(1) Origin
(Please put ☑)

Official Use

□1. Home (same as Form 1)
□2. Work Place (same as Form 2a)
□3. School (same as Form 2a)
□4. Other (please specify the address)

Landmark/ Famous Building near by:

House No. and Street Name:

Village/ Town :

GN division :

DS division :

In case of Colombo Municipality, Please fill in Postal Zone (01 15)

District

(2) Origin Type
Please Select from Table 1

Others (____________)

(3) Departure Time

□□ : □□ □ A.M. □ P.M.

Official Use

(4) What was your Trip Purpose? (Please put √)

Official Use

□ 1. To Work
□ 2. To School
□ 3. Shopping
□ 4. Business
□ 5. Private Matters
□ 6. To Home
□7. Others (specify_______________)

(5) If you transferred to another mode of transport during the trip, state the mode you changed to and the place.

	1st. Mode	2nd. Mode	3rd. Mode	4th. Mode	5th. Mode
Mode of Transport (Table 2)					
Waiting Time (i.e. Bus, Railway)		min	min	min	min
Travel Time	min	min	min	min	min
Transfer Point (Bus stop/Railway station/ Location name)					

APPENDIX C

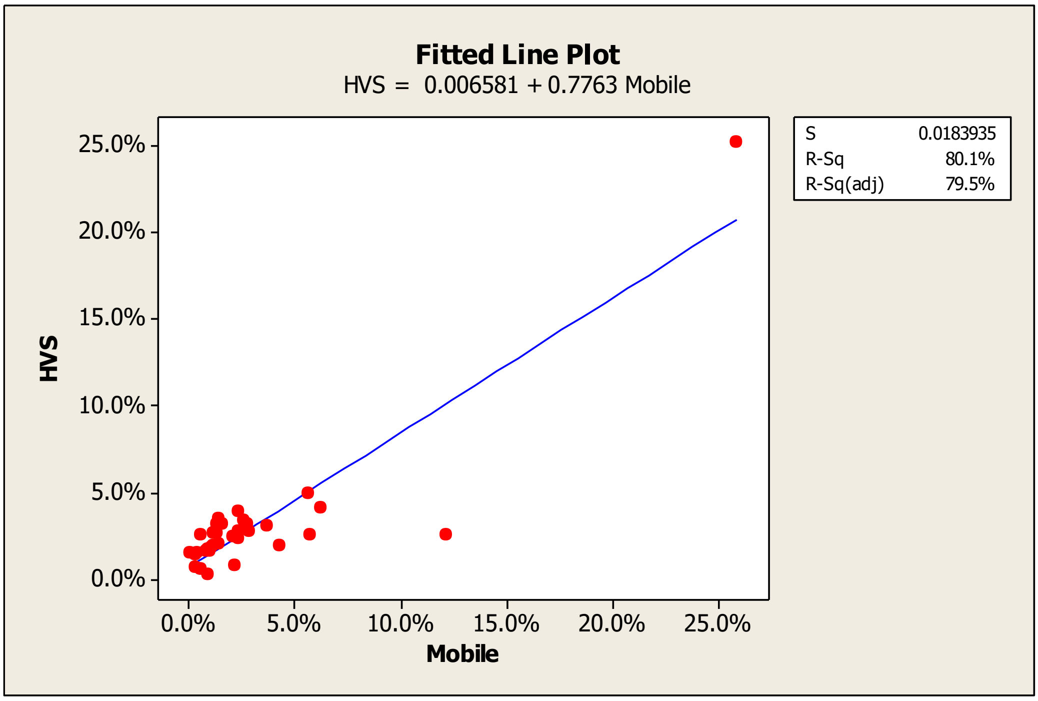
Work - home distribution at DSD level

Home				Work	
DSD	**HVS**	**Mobile (10 PM-4AM)**	**Mobile (7PM-4AM)**	**HVS**	**Mobile**
Agalawatta	0.30%	0.90%	1.10%	0.30%	0.90%
Attanagalla	3.20%	4.30%	4.20%	2.40%	2.30%
Bandaragama	1.70%	3.00%	2.80%	0.80%	2.10%
Beruwala	2.30%	1.60%	1.60%	1.90%	1.20%
Biyagama	3.60%	3.70%	3.60%	3.40%	2.60%
Bulathsinhala	0.80%	1.20%	1.10%	0.60%	0.50%
Colombo	7.80%	7.50%	7.20%	25.1%	25.80%
Dehiwala Mt.Lavinia	2.00%	4.00%	2.50%	1.90%	4.30%
Divulapitiya	2.40%	2.10%	1.90%	1.60%	1.00%
Dodangoda	0.90%	0.70%	0.70%	0.60%	0.50%
Dompe	2.80%	1.00%	0.90%	1.70%	0.90%
Gampaha	3.60%	6.40%	3.90%	2.70%	2.90%
Hanwella	1.90%	1.30%	1.80%	1.60%	0.80%
Homagama	4.60%	5.60%	5.20%	3.10%	3.70%
Horana	2.00%	1.40%	1.40%	1.80%	1.10%
JaEla	3.70%	5.10%	4.70%	3.20%	2.80%
Kaduwela	5.20%	8.60%	6.10%	4.10%	6.30%
Kalutara	2.50%	1.30%	1.40%	2.00%	1.40%
Katana	4.30%	3.70%	3.50%	5.00%	5.60%
Kelaniya	2.30%	2.50%	2.40%	3.10%	1.60%
Kesbewa	5.20%	7.30%	7.00%	2.50%	5.70%
Kolonnawa	3.50%	3.70%	3.40%	2.40%	2.10%
Mahara	3.90%	0.10%	3.50%	1.40%	0.40%
Maharagama	4.40%	0.20%	3.70%	3.50%	1.40%
Mathugama	0.60%	3.90%	0.10%	0.00%	0.00%
Minuwangoda	2.80%	0.10%	1.30%	1.50%	0.40%
Mirigama	2.50%	1.40%	0.80%	1.50%	0.10%
Moratuwa	3.30%	0.70%	3.30%	2.70%	2.30%
Negombo	2.80%	3.50%	1.30%	3.10%	1.30%
Padukka	1.30%	1.30%	0.70%	0.70%	0.40%
Panadura	3.30%	0.60%	3.10%	2.50%	0.50%
Rathmalana	1.80%	0.90%	2.00%	2.70%	1.30%

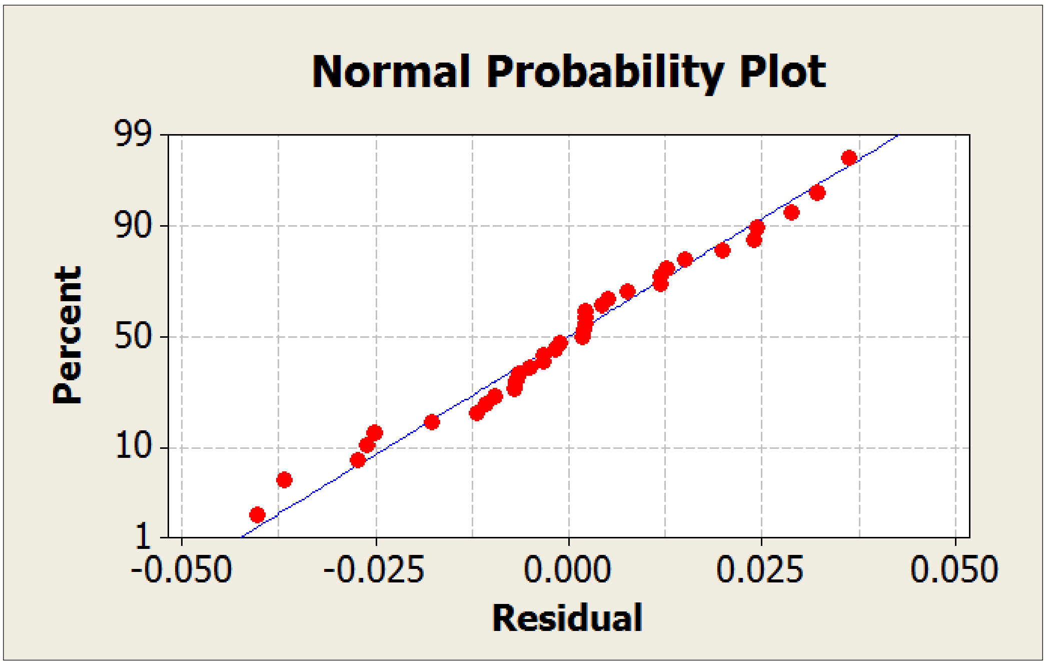
	Home			Work	
DSD	**HVS**	**Mobile (10 PM-4AM)**	**Mobile (7PM-4AM)**	**HVS**	**Mobile**
Sri Jayawardanapura Kotte	2.30%	2.10%	1.80%	3.80%	2.30%
Thimbirigasyaya	1.50%	2.10%	0%	2.60%	12.20%
Wattala	2.90%	5.30%	1.00%	2.60%	1.20%

APPENDIX D

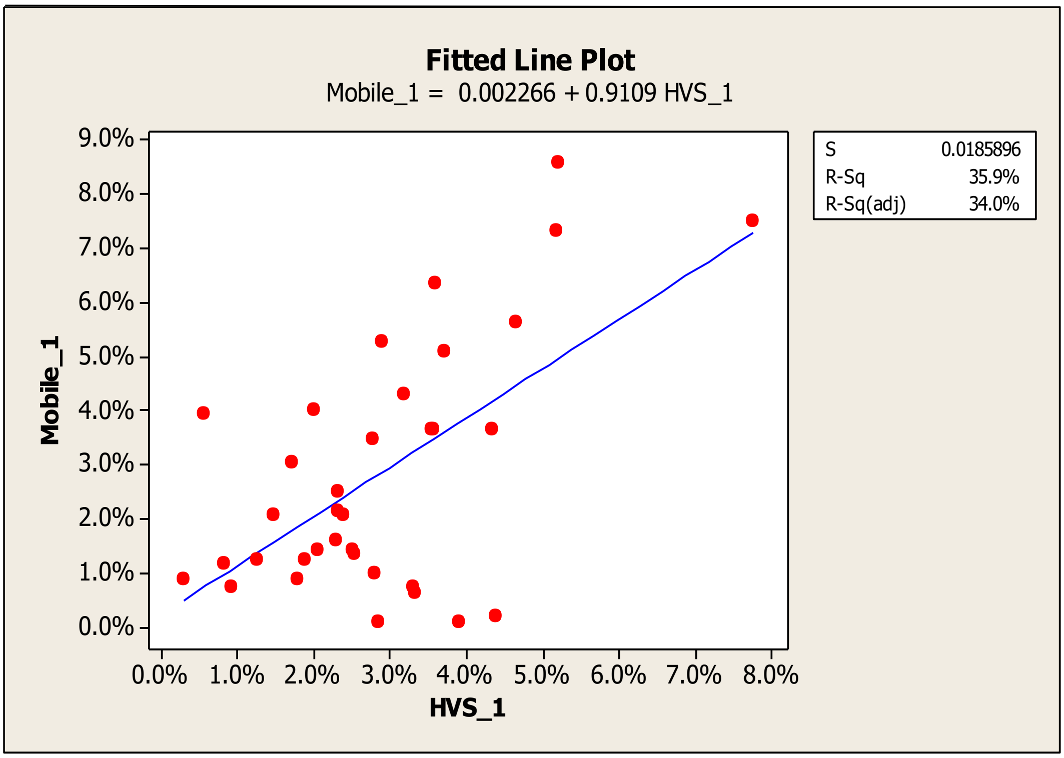
Fitted line plot_ - Work location distribution at DSD level



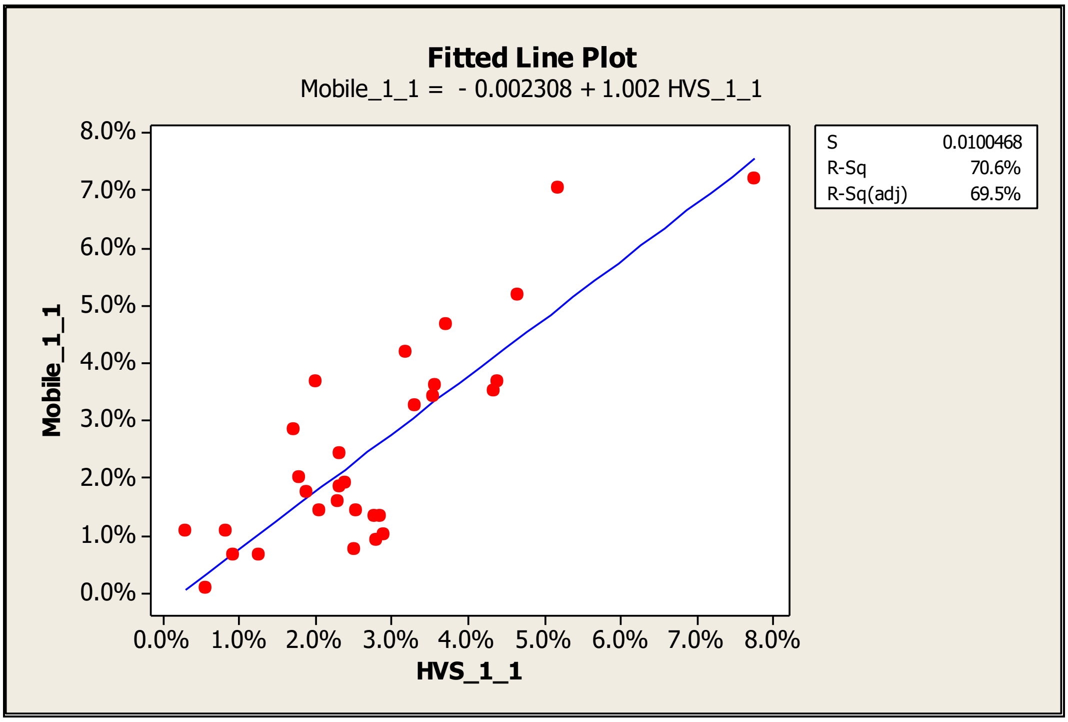
Residual plot of work location distribution



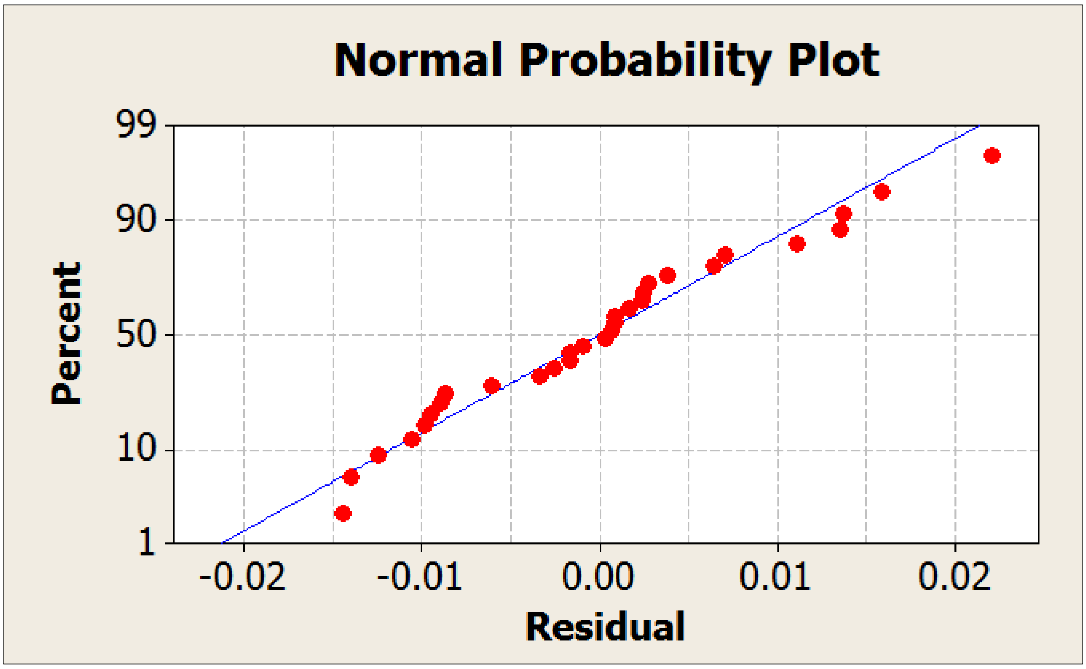
Fitted line plot-Home location distribution at DSD level (Time window-10 PM-4 PM)



Fitted line plot_ Home location distribution at DSD level (Time window-7 PM-4 PM)



Residual plot of Home location distribution (7 PM – 4 PM)



APPENDIX E

CDR Data	HVS Data
Illustration of home location distribution at the district level.	

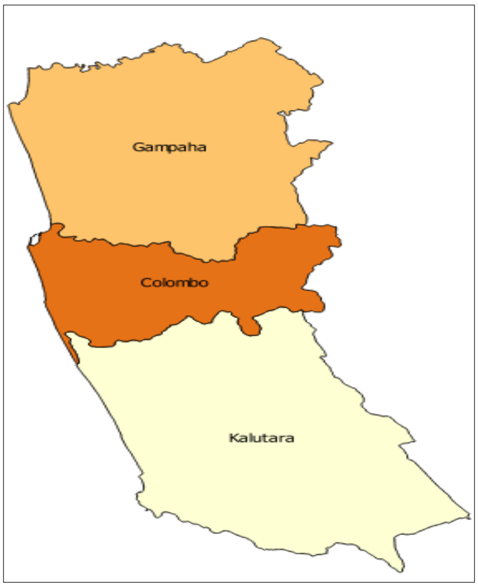
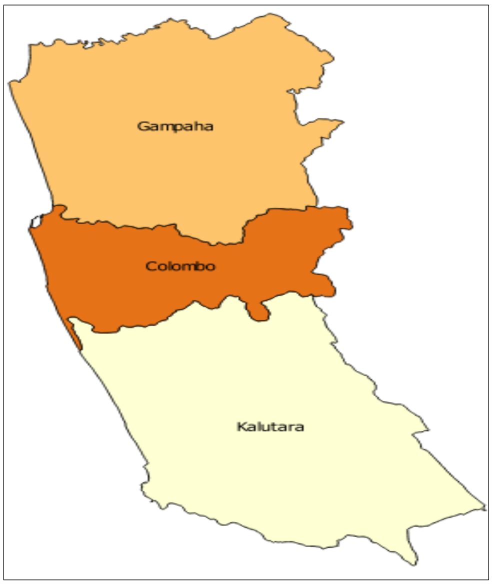
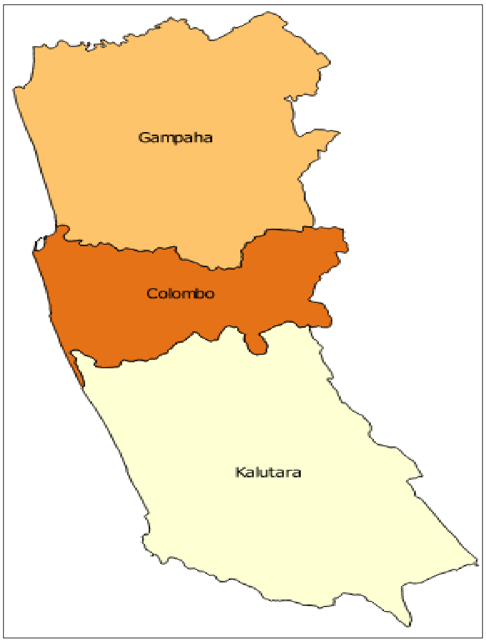
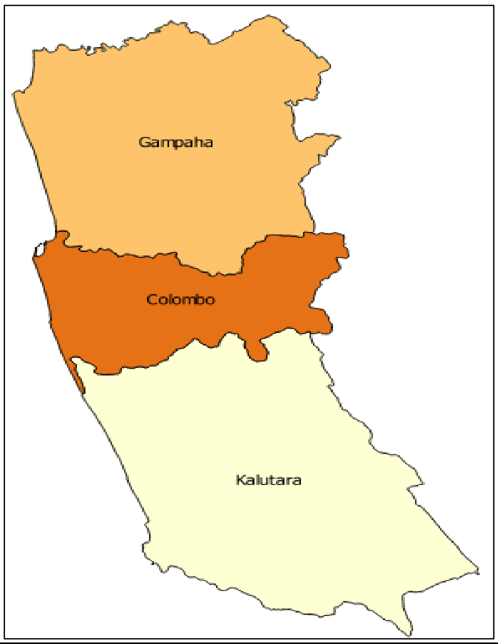

	

Illustration of work location distribution at district level	
	

Legend

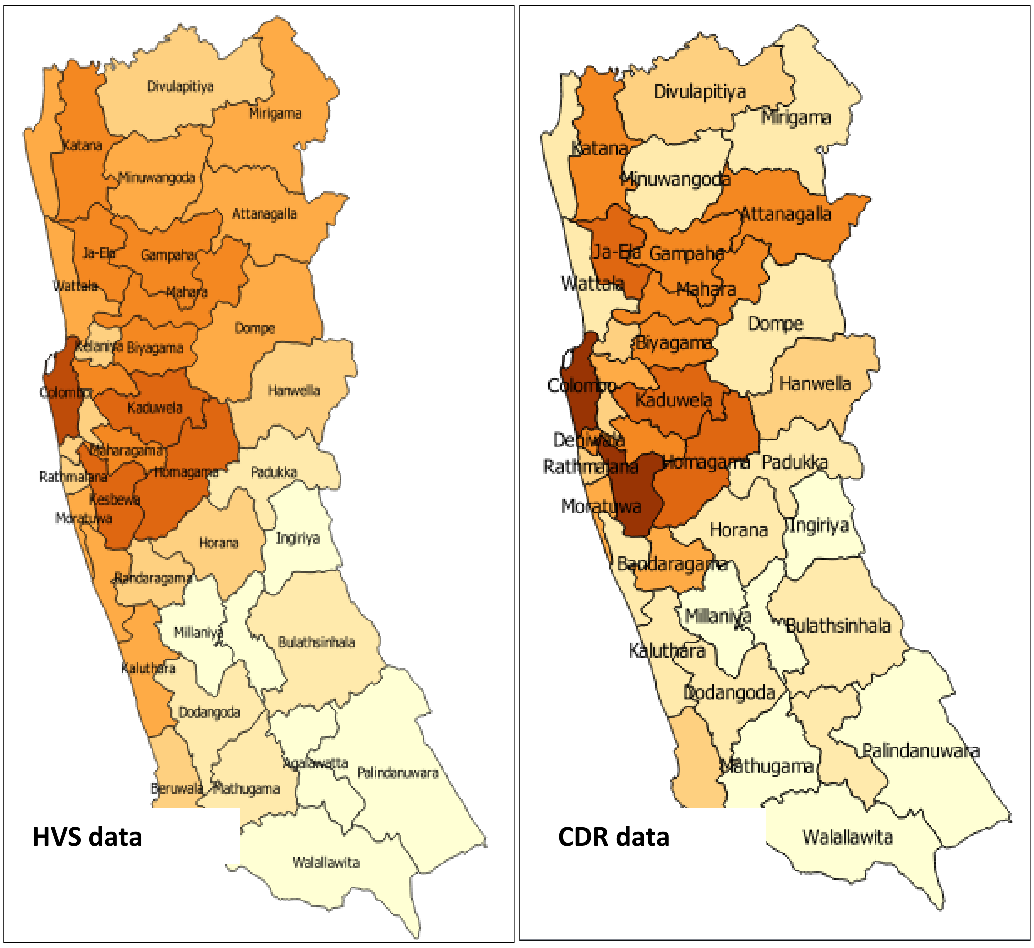
🟨 🟨 🟧 🟧 🟧 🟧 🟫 🟫

→

Increasing home /work population

APPENDIX F

Illustration of home location distribution at DSD level.



Legend

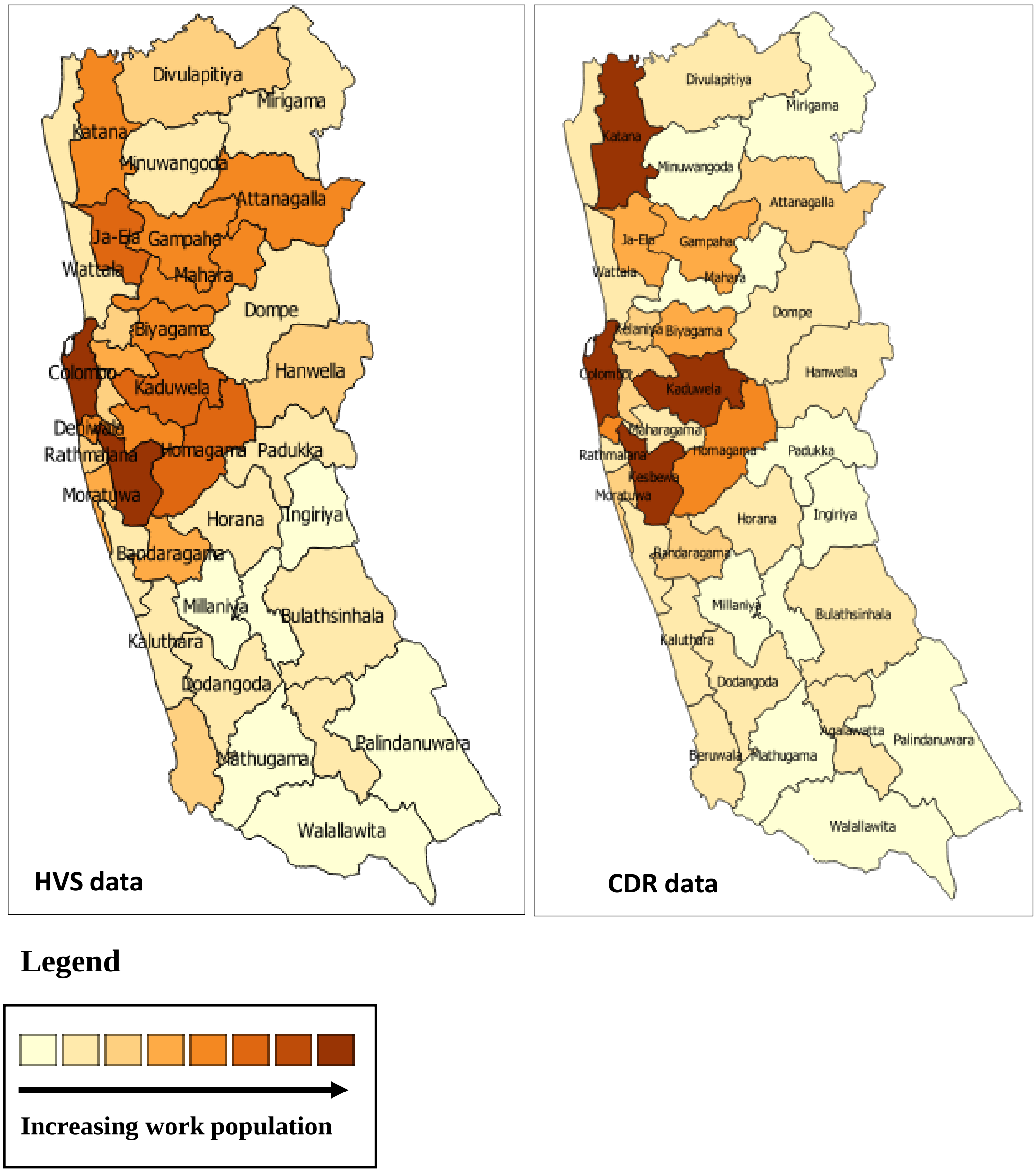
□ □ □ □ □ □ □ □

→

Increasing home population

APPENDIX G

Illustration of Work location distribution at DSD level.


Divulapitiya
Mirigama
Katana
Minuwangoda
Attanagalla
Ja-Ela
Gampaha
Wattala
Mahara
Dompe
Biyagama
Hanwella
Colombo
Kaduwela
Dehiwala
Rathmalana
Homagama
Padukka
Moratuwa
Horana
Ingiriya
Bandaragama
Millaniya
Bulathsinhala
Kaluthara
Dodangoda
Palindanuwara
Mathugama
Walallawita
HVS data
Divulapitiya
Mirigama
Katana
Minuwangoda
Attanagalla
Ja-Ela
Gampaha
Wattala
Mahara
Dompe
Kelaniya
Biyagama
Hanwella
Colombo
Kaduwela
Maharagama
Rathmalana
Homagama
Padukka
Kesbewa
Moratuwa
Horana
Ingiriya
Bandaragama
Millaniya
Bulathsinhala
Kaluthara
Dodangoda
Agalawatta
Palindanuwara
Beruwala
Mathugama
Walallawita
CDR data
Legend
Increasing work population