

LB/TH/01/2026
TH6100

**TRAVEL DEMAND MODELING WITH MOBILE
NETWORK BIG DATA.**

(STUDY AREA: WESTERN PROVINCE OF SRI LANKA)

N. K. Bhagya Jeewanthi

(188062D)

Doctor of Philosophy

Department of Transport Management and Logistics Management

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

**TRAVEL DEMAND MODELING WITH MOBILE
NETWORK BIG DATA.
(STUDY AREA: WESTERN PROVINCE OF SRI LANKA)**

N. K. Bhagya Jeewanthi
(188062D)

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Transport Management and Logistics Management

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

DECLARATION OF ORIGINALITY

I declare that this is my work. This thesis does not incorporate without acknowledgment 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 acknowledgment is made in the text.

Signature:

Date: 05/08/2025.

COPY RIGHT STATEMENT

I hereby grant the 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 in part in future works (such as articles or books).

Signature:

Date: 05/08/2025

STATEMENT OF THE SUPERVISORS

The above candidate has carried out the research for the PhD under my supervision.

Name of the supervisor: Prof. Amal S Kumarage

Signature of the supervisor:

Date:07/08/2025

Co- Supervisor: Dr. Shehan Perera

Signature of the supervisor:

Date:08/08/2025

ACKNOWLEDGEMENT

Prima Facea, I sincerely thank my research supervisor, Senior Prof. Amal S. Kumarage. I have been fortunate to have a supervisor like him who guided me from start to end, continuously followed up on my progress, and gave me advice at each step of the research. I am very grateful for your encouragement and continuous attention throughout the research. Further, I extend my gratitude to Dr. Shehan Perera for the continuous guidance he gave me throughout the period.

Special thanks should go to the postgraduate research coordinator, Prof. A.I.T. Gamage, and former coordinators, Dr. T. Sivakumar and Dr. Varuna Adhikariwattage, for conducting research progress reviews and supporting the process. I would like to express my profound gratitude to Prof. Asoka Perera, Head of the Department of Transport and Logistics Management of the University of Moratuwa.

A very special thank you goes to Sriganesh Lokanathan (Former Team leader, LirneAsia) for lending me the data. I also thank Mr. Buddhi Weerasinghe and Mr. Anuradha Munasinghe, who contributed immensely to the primary data processing and analysis stages. I thank you all for trusting me in providing data and for the guidance given from start to finish. I gratefully acknowledge the funding received for my PhD from the AHEAD project.

I am sincerely thankful to all the staff members and 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. Your encouragement, advice, and opinions are really appreciated.

ABSTRACT

The increasing travel demand necessitates dynamic traffic control mechanisms, which rely heavily on human mobility data for effective implementation. Traditional gathering of mobility information, such as household surveys, is costly and time-consuming. However, the advent of big data offers an unprecedented opportunity to analyze human mobility on a large scale in a cost-effective and real-time manner. Specifically, Call Detail Record (CDR) data, a subset of Mobile Network Big Data (MNBD), provides valuable insights into human mobility patterns. Yet, this data requires extensive processing to convert into structured travel information suitable for integration into travel demand models.

In response to this challenge, this study introduces a novel methodology that leverages CDR data to inform the four-step modeling process, encompassing Trip Generation/Attraction, Origin-Destination (OD) information, and Network Assignment. Utilizing a stay-based approach with pseudo-anonymized CDR data, the methodology identifies trip nodes by tracking the presence of mobile phone users within specific temporal and spatial cells. This approach facilitates extracting individual mobility characteristics, including activity sequences and trip generations/attractions. Moreover, it allows for determining trip purposes based on the origins and destinations and generating OD matrices from CDR data using mapped spatio-temporal data. Individual commuting routes are then defined based on the frequency of appearances and weighted caller activities along each potential route.

The methodology was applied to a CDR dataset comprising randomly selected 400,000 users over one month within the Western Province of Sri Lanka. The findings were validated against data from traditionally collected household visit surveys. The validation results demonstrated a significant correlation at each stage of the four-step process, underscoring the potential of repurposed, privacy-preserved mobile phone data to significantly enhance dynamic travel demand analysis.

Keywords – Mobile Network Big Data, Call Detail Record, OD Matrix, Network Assignment

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY	i
COPY RIGHT STATEMENT	ii
STATEMENT OF THE SUPERVISOR.....	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
LIST OF FIGURES.....	x
LIST OF TABLES	xiii
LIST OF ACRONYMS.....	xiv
LIST OF APPENDICES	xv
1 INTRODUCTION.....	1
1.1 Background.....	1
1.2 Mobile Network Big Data	3
1.3 Purpose of the Research	7
1.4 Research Gap	9
1.5 Research Objectives	12
1.6 Scope of the Research.....	15
1.7 Overview of the Thesis.....	16
2 LITERATURE REVIEW.....	17
2.1 Traditional Travel Demand Modelling.....	17
2.1.1 Trip Generation and Attraction.....	18
2.1.2 Trip Distribution	19
2.1.3 Network Assignment	19
2.2 Use of Digital Data in Transport Analysis	20
2.3 CDR and Human Mobility	22

2.3.1	Origin Destination Matrices.....	22
2.3.2	Mobility Pattern Recognition and Significant Location Identification.....	24
2.3.3	Route Estimation from CDR data.....	25
2.4	Use of CDR in the Sri Lankan Context	25
2.5	Drawbacks in Call Detail Record Data.....	27
2.6	Techniques Used in Overcoming the Limitations of CDR.....	28
2.7	Research Gap Identification	29
3	METHODOLOGY	31
3.1	Study Area	32
3.2	Data Description – CDR Data	33
3.2.1	Data Privacy and Security Measures	33
3.2.2	Attributes in CDR Data.....	33
3.3	Household Visit Survey Data	34
3.4	Sampling Technique	36
3.4.1	Identifying the Sample Required from the Lowest Populated DSD.....	39
3.5	Overview of the Research Methodology	42
4	OD ESTIMATION FROM CDR DATA.....	46
4.1	Initial Processing	46
4.1.1	Removing NA	46
4.1.2	Splitting Time	47
4.1.3	Load Share Effect	47
4.1.4	Spatial Mapping.....	51
4.2	Descriptive Statistics	53
4.2.1	Temporal Measures.....	53
4.2.2	Spatial Measurements	61
4.3	Identifying Movements from CDR Data	69
4.4	Behavioral Representation of the Subscribers.....	72

4.5	Home-based Trip Identification and validation.....	80
4.5.1	Stage 1: Constructing a trip from CDR Data	80
4.5.2	Stage 2: Identifying Significant Locations	83
4.5.3	Stage 3: Trip Extraction and Classification	88
4.5.4	Stage 4: Estimation of Origin-Destination Matrices.....	90
4.5.5	Results and Validation	94
4.6	Conclusion	103
5	NETWORK ASSIGNMENT	104
5.1	Methodology	105
5.2	Stage 1 – Initial Data Processing	106
5.2.1	Defining the Sample	106
5.2.2	Enroute Cell Identification.....	107
5.2.3	Route-Cell Matching.....	109
5.2.4	User-Case Example.....	110
5.3	Stage 2 – Defining Trip Statistics.....	111
5.4	Stage 3- Determining the Route Choice	115
5.5	Stage 4 – Validation using STRADA.....	116
5.5.1	Introduction to STRADA.....	116
5.5.2	Results and Validation	121
5.6	Summary and conclusion.....	126
6	SUMMARY AND CONCLUSION.....	127
6.1	Utilizing CDR for the Four-Step Demand Modeling Process.....	127
6.2	Validation with HVS Data.....	128
6.3	Application of CDR-based Findings	129
6.4	Summary of Key Contributions of the Study	130
6.5	Limitations of the Study	132
6.6	Future Research	133

7	REFERENCES	135
8	APPENDIX A - SAMPLE REPRESENTATION FROM EACH DSD.....	143
9	APPENDIX B - HOUSEHOLD SURVEY QUESTIONNAIRE	144
10	APPENDIX C - SUMMARY OF THE LITERATURE REVIEW	145

LIST OF FIGURES

Figure 1-1 Data Transmission Flow in Mobile Telecommunication Systems	4
Figure 1-2 Standard Voronoi Cell Format.....	5
Figure 1-3 Attributes within a CDR Entry.....	7
Figure 1-4 Relationship among urban computing and big data.....	9
Figure 2-1 Illustration of the Traditional Travel Demand Model.....	18
Figure 3-1 District and DSD segregation in the Western Province of Sri Lanka,.....	32
Figure 3-2 Overview of the sampling population -Comparison with HVS population.....	37
Figure 3-3 Graphical Representation of Multi-Staged Random Sampling technique	39
Figure 3-4 Conceptual Framework of the Study	42
Figure 3-5 Methodology of the study at Macro level	44
Figure 3-6 Summary of expected outputs from the study	45
Figure 4-1 Removing NA from the CDR data.....	46
Figure 4-2 Splitting Time component within CDR Data.....	47
Figure 4-3 Graphical Representation of the Load Sharing Effect	48
Figure 4-4 Methodolog used within the study for Load Share Identification	49
Figure 4-5 Discrepancy Analysis Between CDR and HVS Distributions Across Speed Thresholds	50
Figure 4-6 User Connectivity among Cell Towers.....	52
Figure 4-7 Summary on Spatio-temporal measures derived from CDR Data.....	53
Figure 4-8 Caller Activity Distribution of CDR Sample.....	55
Figure 4-9 CDF of Caller Activities	55
Figure 4-10 Hourly Distribution of Caller Activities of the CDR Sample.....	56
Figure 4-11 Average Inter-Event Time of the CDR Sample	57
Figure 4-12 Cumulative Distribution of AIT.....	57
Figure 4-13 Maximum Inter Event Time Distribution of the CDR Sample.....	58
Figure 4-14 CDF of MIT	59
Figure 4-15 Day-wise Caller Activity Distribution of the CDR Sample.....	60
Figure 4-16 CDF of Day wise caller activity distribution	60
Figure 4-17 Cell Appearance Distribution of the CDR Sample	62
Figure 4-18 Cumulative Distribution of the cells visited	62
Figure 4-19 Saptial Activity Space of user 1 as derived from CDR data.....	64
Figure 4-20 Spatial Activity space of user 2 as derived from CDR data.....	64

Figure 4-21 Activity Space Distribution derived from CDR data	65
Figure 4-22 CFD of Activity Space	66
Figure 4-23 Entropy Values distribution derived from CDR data.....	68
Figure 4-24 CDF of Entropy Value	68
Figure 4-25 Generation of Time-Stamped Matrices from CDR data	71
Figure 4-26 User Behavior Model 1 Based on CDR Data	73
Figure 4-27 User Behavior Model 2 Based on CDR Data	74
Figure 4-28 User Behavior Model 3 Based on CDR Data	75
Figure 4-29 User Behavior Model 4 Based on CDR Data	75
Figure 4-30 User Behavior Model 5 Based on CDR Data	76
Figure 4-31 User Behavior Model 6 Based on CDR Data	77
Figure 4-32 Socio-Economic Classification Using Spatio-Temporal Features of CDR Data.....	79
Figure 4-33 Approach for OD Estimation Based on CDR Data.....	81
Figure 4-34 Theoretical Trip Structure.....	81
Figure 4-35 Personalized Activity Trajectories	82
Figure 4-36 User Mobility Mapping Across Time Periods Using CDR Data.....	83
Figure 4-37 Significant location categorization based on User Behaviour	84
Figure 4-38 Integration of Appearance and Activity Layers.....	84
Figure 4-39 CDR-Based Core Time Window Identification.....	85
Figure 4-40 Trip Classification Based on Origin-Destination Combinations	89
Figure 4-41 Spatial Mapping of Segmented User Locations to Administrative Zones.....	91
Figure 4-42 Cell Tower Allocation to District and DSD Zones via GIS.....	91
Figure 4-43 Trip Type Classification Across DSD and District Levels	94
Figure 4-44 Validation of Trip Distribution derived from CDR data at Intra District Level	96
Figure 4-45 Validation of HBW Trip Distribution derived from CDR data at at DSD level.....	96
Figure 4-46 Validation of HBO Trip Distribution derived from CDR data at at DSD level.....	97
Figure 4-47 Validation of NHB Trip Distribution derived from CDR data at at DSD level.....	97
Figure 4-48 Intensity map of Intra District Trips by Trip Type	98
Figure 4-49 Trip Distance Distribution Across All Trip Types.....	101
Figure 4-50 Hourly Distribution of HBW trips	102
Figure 4-51 Hourly Distribution of HBO trips	102
Figure 4-52 Hourly distribution of NHB trips.....	103
Figure 5-1 Methodology for Inferring Travel Routes Using CDR Data	105
Figure 5-2 Identification of Enroute Locations from Call Detail Records	107

Figure 5-3 Filtered User Set for Route Estimation Using CDR Data.....	108
Figure 5-4 Location-Based Mapping of Daily Commute Patterns	110
Figure 5-5 Mapping CDR derived locations onto the Most Probable Travel Path.....	111
Figure 5-6 Steps in deriving the trip statistics(a) and (b)	113
Figure 5-7 Route Choice Representation of a Single User.....	114
Figure 5-8 Steps in deriving the trip statistic (c)	114
Figure 5-9 Route Choice generation function via CDR Data.....	116
Figure 5-10 Initial Interface of STRADA.....	119
Figure 5-11 Sample OD format in STRADA	119
Figure 5-12 Parameter interface of STRADA	120
Figure 5-13 Network Assignment Methodology with CDR Data and STRADA,	121
Figure 5-14 Distribution of the Route Choices.....	122
Figure 5-15 Sample Route Assignment comparison	124
Figure 5-16 Route Assignment comparison with CDR Vs STRADA.....	125
Figure 5-17 Route Assignment comparison of CDR Vs STRADA after outlier detection.....	125
Figure 5-18 Route assignment from CDR Data (A) and Traditional Data (B).....	126

LIST OF TABLES

Table 1-1 Scope and Direction of the Study	12
Table 3-1 Attributes in a CDR Entry	34
Table 3-2 Estimation of the target sample size from the HVS	35
Table 3-3 Total Sample Calculation according to the Multistage Random Sampling Method ...	41
Table 3-4 Sample Comparison-HVS and CDR Sample	41
Table 4-1 Sample Representation of Load Shared Records	50
Table 4-2 Cell tower coverage at district level of Sri Lanka	52
Table 4-3 Sample Entropy calculation.....	67
Table 4-4 Socio-Economic Segmentation of Users Using CDR data.....	77
Table 4-5 Time-of-Day Categorization for Regularity Analysis	87
Table 4-6 Regularity Measurement of Cell Tower Visits by Time Category.....	87
Table 4-7 Significant Location Extraction via Regularity and Ideal-Typical Matching	88
Table 4-8 Trip Type Classification Using Sample CDR Entries	90
Table 4-9 Classification of Intra- and Inter-Zone Trips at District and DSD Levels	93
Table 4-10 Comparison of Trip Type Proportions: CDR vs. Household Travel Survey Data	95
Table 4-11 Expansion of HBW Trips to Actual Trip Counts Using Population Scaling	98
Table 4-12 Comparison of Expanded HBW Trip Counts (CDR vs. HVS) at District Level	99
Table 5-1 Example Calculation of Determining Route Choice for a Single User.....	115
Table 5-2 Steps in Traditional Network Assignment.....	118

LIST OF ACRONYMS

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

LIST OF APPENDICES

Appendix A	Sample Representation from each DSD	134
Appendix B	Household Survey Questionnaire	144
Appendix C	Summary of the Literature Review	145

CHAPTER 1

INTRODUCTION

1.1 Background

A secure and well-operated transportation system plays a pivotal role in a nation's economic strength and the overall quality of life for its communities. Urban areas witness a significant volume of travel. While water-based urban transport is limited to only a few cities worldwide, air transport is generally impractical for urban travel. Consequently, the primary means of transportation available for urban passenger travel are land-based, encompassing private modes such as walking and private motor vehicles and various public transportation services.

The increasing need for efficiently moving large quantities of goods and people over substantial distances within reasonable timeframes has highlighted the importance of reliable traffic and transportation planning. As a result, there has been a significant focus on planning and implementing transportation initiatives in recent history. Today, transportation investment and effective planning have become an integral part of economic and development strategies. These plans often include developing transit systems, pedestrian and cyclist facilities, and measures to manage transportation demand (Meyer, 2016).

Transportation planning, as the initial step in the implementation of transportation systems, constitutes a significant technical process that heavily relies on computer models and advanced tools to simulate the intricate interactions inherent in transportation system performance. Effective transport planning is vital for enhancing the livability and efficiency of cities, thereby ensuring their preparedness for the challenges of future eras (Batty et al., 2012). This intricate transportation planning process often necessitates engagement with a broad spectrum of stakeholders, primarily focusing on the infrastructure users. The fundamental action in planning transport initiatives is the measurement and analysis of people's movements. Comprehensively scrutinizing these movements is critical to developing transportation infrastructure and services that operate efficiently while minimizing the burden on end-users.

Typically, information corresponding to human travel activity or movements may be acquired using different approaches, such as survey-based, passive, activity-based, and device-based (Doyle, 2014). Survey-based approaches follow a method of gathering data from people engaged in a survey. Survey methodology targets instruments or procedures that ask one or more questions that may or may not be answered. Researchers conduct statistical surveys to make inferences about the population being studied. Ordinary travel surveys assemble human activity data, preferences, and behavioral data through questionnaires (Hwang & Fesenmaier, 2003; Dieleman et al., 2002). The veracity of survey-based approaches is widely influenced by the memory of each individual participating in the survey. Additionally, the cost associated with conducting a survey is expensive and laborious. Due to this, the frequency of surveys is low, often leading to small sample sizes and an increased risk of sampling bias. A lower response rate due to response burden is another problem encountered, but certain complex data types, like preferences or opinions, can be gathered easily. Moreover, the existing studies from CDR generate only the screen line flows without the OD details.

In contrast to survey approaches, passive sensing methods utilize static sensors to infer human activity. Examples of passive sensing approaches include sensing devices, traffic cameras, and flow counters (Cardelino, 1998; Doyle et al., 2010; Morris & Trivedi, 2008; Xu et al., 2005; Zhang et al., 2012). Some of these devices' functionality relies on Bluetooth connectivity or Wi-Fi activity. The coverage area of sensing devices generally depends on the spatial distribution of device placement, and the accuracy of passive approaches is highly contingent on the precision of these devices. However, implementing a sensor-based survey with nationwide coverage is prohibitively expensive from a cost perspective. Consequently, sensor-based data gathering is typically limited to local, small-scale applications.

On the other hand, activity-based approaches involve collecting data from user interactions with various services and applications. This data typically includes when, where, and how specific interactions occurred. Activity-based approaches include data from platforms such as Facebook, Instagram, and Twitter, used to gather public opinions and map social connections (Bollen et al., 2011; Marvell Solutions, 2010). Additionally, tracking consumer purchases, bank card transactions, and even phone calls fall under activity-based approaches (Kononenko & Kukar, 2007; Capizzi & Ferguson, 2005). However, activity-based approaches may

encounter sampling-related issues due to the coverage area of the service network and usage fluctuations. As discussed above, the characteristics of each approach, including data availability, accuracy, scalability, cost, privacy concerns, and information content, vary drastically (Doyle, 2014). Therefore, careful consideration should be given when selecting the data type and acquisition method methodology. In recent decades, researchers have increasingly turned to using time-stamped location data derived from mobile network providers, combining aspects of both passive and active approaches in modeling human behavior. This data type is commonly referred to as Mobile Network Big Data. Technological advancements, including improved data retrieval, processing, and storage facilities, have enhanced the usability of such big data, motivating researchers to incorporate these datasets into their studies.

1.2 Mobile Network Big Data

Big Data encompasses a vast collection of data characterized by its substantial volume, which grows exponentially over time. Examples of Big Data include social media, stock market, consumer, medical, and Global Positioning System (GPS) data, among others. Mobile Network Big Data (MNBD), a subset of Big Data, comprises large-volume, complex, and expanding datasets derived from the usage of communication devices and collected by mobile operators (Iqbal et al., 2014). These datasets share characteristics similar to other types of Big Data, such as high velocity and variety, and carry significant value (Priya, 2017).

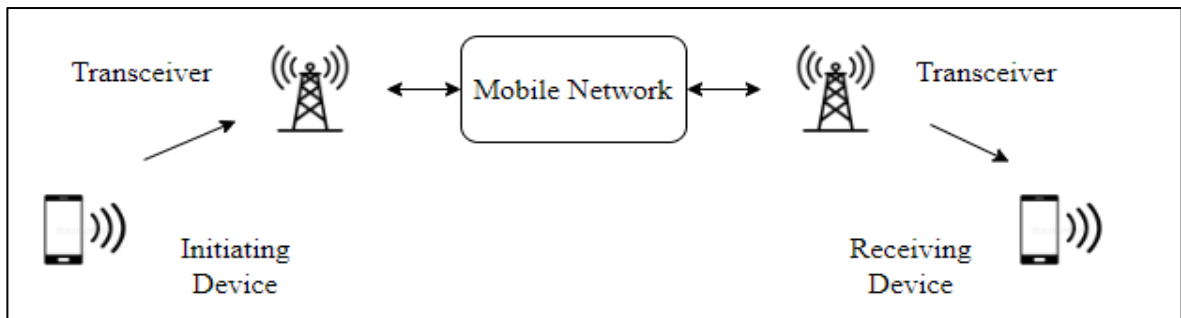
The quantity of data generated from mobile phone networks has witnessed a remarkable increase, driven by the constant escalation in the population of mobile phone users. Mobile subscriber penetration has surpassed 100% in most developed and developing countries. This extensive mobile community provides access to substantial amounts of data, given that the activities and movements of mobile phone users serve as indicators of mobility patterns. Consequently, mobile phone networks can be considered large-scale human sensing platforms.

A mobile phone network can be defined as a geographically distributed radio network that facilitates communication, including voice, text, or data transmission, between two or more

devices (Korhonen, 2003). Each device within the network typically maintains a wireless connection with a fixed location, referred to by various names such as a cell tower or base transceiver station.

Figure 1-1

Data Transmission Flow in Mobile Telecommunication Systems



Note: Source: (Doyle, 2014)

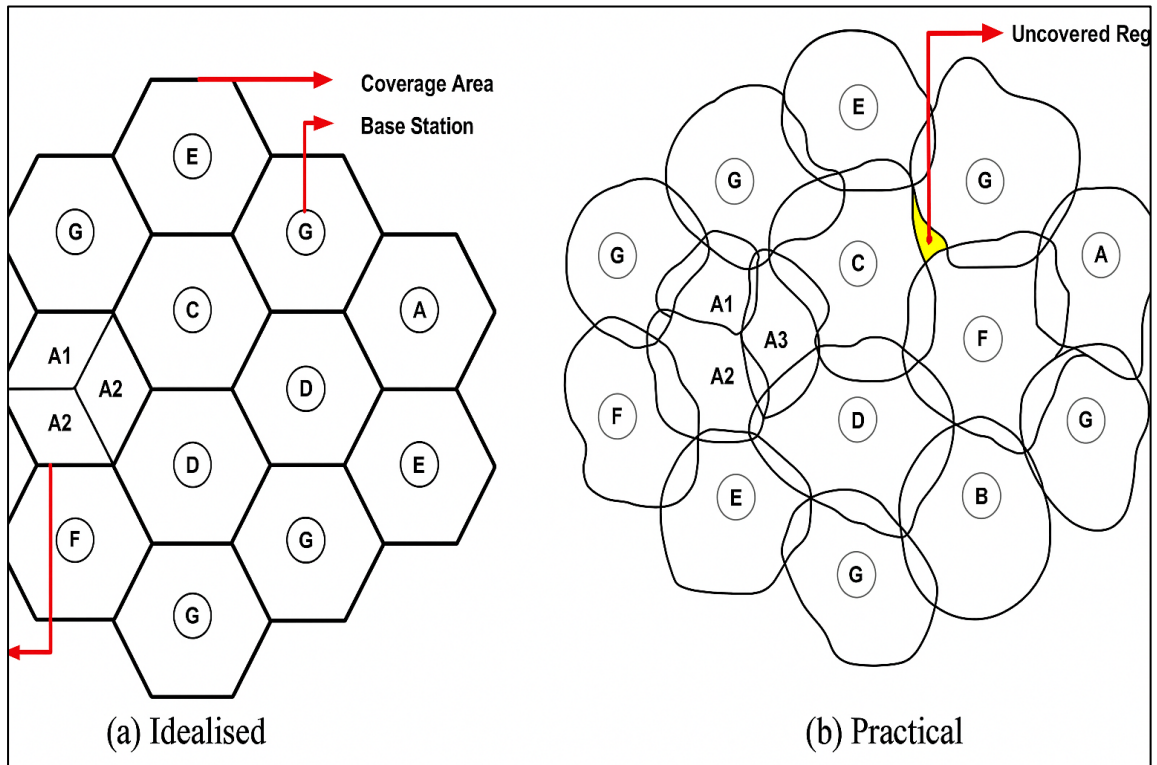
Each cell tower possesses a distinct coverage area, a scope typically determined by capacity requirements. This coverage may extend in rural settings across several hundred square kilometers, whereas, in urban environments, it may be limited to only a few square kilometers. Figure 1.1 illustrates the connection process, depicting the flow of communication from the initiating device's connected transceiver through hierarchical network elements to the destination cell, ultimately reaching the receiving device (Doyle, 2014). This communication flow model applies to both intra-cell and inter-cell communications. Figure 1.2 provides a visual representation of the typical layout of cell coverage areas, often referred to as Voronoi cells (Xie et al., 2016). Although these cells are commonly depicted in hexagonal shapes, it is essential to note that this representation is an idealized illustration and may not precisely reflect the actual boundaries of the cells.

Mobile operators generate two primary types of data, namely "Static data" and "Collected data" (International Telecommunication Union [ITU], 2022). "Static data" comprises Home Location Register (HLR) data, containing information about all subscribers and their service contract preferences. In contrast, "Collected data" encompasses two main types: Visitor Location Register (VLR) data and Call Detail Record (CDR) data. These "Collected data" can locate mobile devices in both time and space with significant accuracy, utilizing the mobile

network infrastructure that includes location information for the mobile device (Dong et al., 2014).

Figure 1-2

Standard Voronoi Cell Format



Note: Source: (Xie et al., 2016)

VLR is a database that contains information about subscribers roaming within a specific Base Transceiver Station (Pourmoradnasseri et al., 2019). It maintains a comprehensive list of all devices being handled by the associated BTS. Beyond data generated from calls, VLR also includes information on phone presence, rendering it suitable for real-time applications. Typically, VLR datasets are larger, given that locational data for each user is collected every second.

Call Detail Record Data, or CDR is generated during voice calls, mobile internet usage, and other active interactions. Mobile operators primarily collect call detail records for billing purposes. This event-driven data is typically generated during incoming and outgoing voice calls and text messages. For example, a corresponding CDR record is created when a user

initiates a call or sends an SMS. No information is generated when the mobile phone is not in use. A unique identifier lets each phone owner link with their Caller Activities (CA).

Compared to VLR, CDR data provides a more accurate reflection of users' mobility patterns, given that the transitionally updated VLR data may potentially result in more location-based data, leading to a higher sampling frequency. Unlike CDR data, VLR data does not introduce location uncertainty from the perspective of human mobility sensing. For instance, CDR records are created during a long-distance journey only when an activity is initiated. In contrast, VLR records are created whenever the user changes the Base Transceiver Station (BTS) along the path. Although VLR data is rich in information, CDR data is the most commonly used dataset for research due to mobile operators' reluctance to store VLR data long-term. CDR data are typically required for legal activities and billing purposes, making them highly available. Furthermore, the storage and analysis feasibility of CDR data are also high.

CDR data generates time-stamped locations for the users. To preserve their anonymity and privacy, the operator replaces all phone numbers with computer-generated unique identifiers. A single CDR contains a wide array of information, as shown in Figure 1.3, which will be explained later.

Here are some examples of the most essential attributes.

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

Figure 1-3

Attributes within a CDR Entry

SUBSCRIBER_ID	CALL_TIMESTAMP	LOCATION_ID	LONGITUDE	LATITUDE	DSD_CODE4	PROVINCE_NAME	DISTRICT_NAME	DS_NAME	BTS_AREA
1001998411	2013-07-09T11:03:37.000+05:30	29425	80.18484	6.93261	1115	Western	Colombo	Hanwella	12932753
1001998411	2013-07-20T18:29:42.000+05:30	29171	80.15710	6.92025	1115	Western	Colombo	Hanwella	14601418
1001998411	2013-07-24T10:35:26.000+05:30	28658	80.05414	6.90609	1112	Western	Colombo	Homagama	5551219
1001998411	2013-07-01T13:36:56.000+05:30	29425	80.18484	6.93261	1115	Western	Colombo	Hanwella	12932753
1001998411	2013-07-12T18:08:59.000+05:30	29425	80.18484	6.93261	1115	Western	Colombo	Hanwella	12932753

In addition to the spatial and temporal attributes mentioned earlier, CDR data also capture other information crucial for extracting socio-demographic parameters. For instance, the phone type/model can serve as a measure to infer the user's income level and is available in CDR data. Furthermore, details such as the country of registration and state information can be utilized to infer the user's home city (Leng, 2016).

Chapter 4 will elucidate the process of feature extraction from CDR data in detail. When a person initiates an activity, all the information above is recorded as an array of data. However, it's important to note that while CDR records capture human mobile phone usage, the focus in transportation studies is on travel patterns. Therefore, processing CDR data accumulates rich mobility information, providing valuable insights into human mobility patterns.

1.3 Purpose of the Research

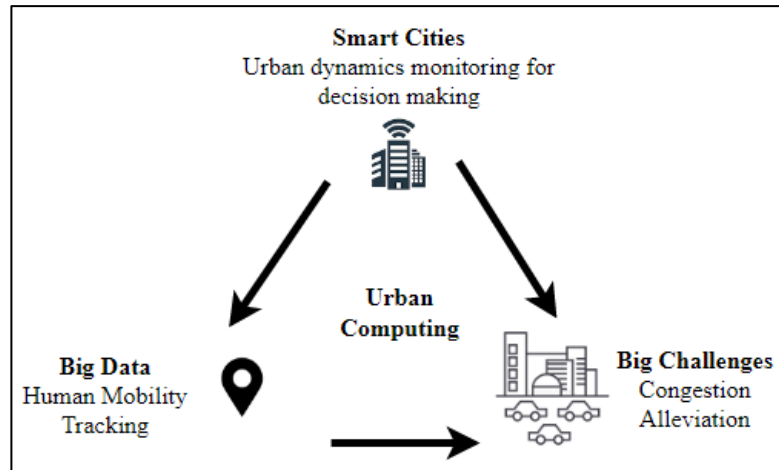
As highlighted in the background explanation, the travel landscape has become more intricate with the growing population and increasing mobility. Establishing a continuous process for measuring human activity has become essential to address the complexities of modern travel. The ability to monitor movement plays a crucial role in implementing effective transport policies, planning adequate infrastructure, and ensuring timely operations. Regularly updated datasets, rich in mobility information, are vital for accurately measuring human mobility, thus facilitating the smooth functioning of these critical aspects. Traditional mechanisms currently in place often exceed capacity requirements, and the associated costs of transportation surveys prompt the exploration of low-cost and scalable alternatives. Therefore, this study complements traditional processes with Mobile Network Big Data.

The utilization of Mobile Network Big Data is driven by the opportunities it offers in the development of smart cities. The overarching goal of building a smart city is to enhance the quality of life through the strategic use of technology (Batty et al., 2012). As illustrated in Figure 1.4, tracking human mobility facilitated by the usage of Mobile Network Big Data contributes to dynamic decision-making processes aimed at alleviating significant challenges, such as traffic congestion. The primary data source for this study comprises anonymized Call Detail Records (CDR) obtained from a mobile operator in Sri Lanka. As mentioned earlier, CDR is a subset of Mobile Network Big Data, encompassing voice, SMS, and data records of communication. It includes information such as the activity's start time, serving cell tower, and duration of the activity, among other details. Since individuals generally carry their mobile phones with them for a significant portion of the time, CDR becomes a mobility-rich, frequently updated, and cost-effective data source. While CDR-derived information is applied in various applications, including dynamic population estimation and land use pattern recognition, the primary objective of this study is to explore the potential of Mobile Network Big Data in generating valuable transport insights. This is achieved by establishing significant relationships between Mobile Network Big Data and traditional outputs, aligning with traditional transport forecasting models.

Traditional travel demand modeling encompasses four steps: trip generation and attraction, origin-destination estimation, route assignment, and mode estimation, which will be detailed in Chapter 2. This study primarily focuses on the first three steps and excludes the mode estimation step. The exclusion is due to spatial uncertainty and temporal sampling issues, making it challenging to relate CDR positional estimates to transportation-related features, particularly speed, which is crucial for mode estimation.

Figure 1-4

Relationship among urban computing and big data



Note: Source (Batty et al., 2012)

Alongside the CDR information, this study incorporates the concept of rational behavior of trips as the basis of analysis. Rational behavior refers to the regular and predictable nature of travel. When trip-making patterns are regular, they often coincide with corresponding call-making patterns to some extent. This leads to the occurrence of regular call records at specific points in time during the day. Even when no records are available, the behavior can still be predicted based on previous data. The study employs this concept to derive trip-making patterns of mobile users.

1.4 Research Gap

Despite the global rise in using mobile phone data for transport analysis, substantial methodological and empirical gaps persist—especially in the Sri Lankan context—with regard to the effective integration of Call Detail Records (CDR) into transport modeling frameworks. While CDR datasets offer broad spatial and temporal coverage at relatively low cost, their utility remains underexploited due to several critical limitations.

Gap 1: Inadequate trip detection methodologies for sparse CDR data (Linked to Objective #1)

The first major limitation is the lack of robust frameworks for detecting trips and inferring individual mobility from CDR data. Unlike continuous GPS logs, CDRs are event-driven

and temporally sparse, with records triggered only during active phone use. Furthermore, technical noise, such as the Load Sharing Effect (LSE)—which results in apparent but false tower transitions due to load balancing—can create misleading interpretations of user movement. Current studies often bypass or insufficiently address these issues, leading to erroneous trip counts and spatial misallocations. There is thus a pressing need for methodologies that can effectively distinguish genuine travel behavior from artifacts, enabling accurate mapping of individual presence patterns over time.

Contribution 1: This study develops a comprehensive methodological framework for processing raw CDRs into meaningful individual mobility trajectories. It introduces a speed-based filtering approach to eliminate the Load Sharing Effect and a presence-pattern-based location characterization method to improve the detection of stay points and trips—transforming sparse and noisy data into structured mobility timelines suitable for travel analysis.

Gap 2: Limited application of CDR data in travel demand estimation and network assignment (Linked to Objective #2)

While several studies have explored the use of mobile data for aggregate OD matrix estimation, very few have leveraged CDR data for more advanced transport modeling stages such as trip purpose classification, trip type categorization, and especially network assignment. Route choice modeling remains underexplored due to the spatial granularity of tower-level data and its asynchronous nature. However, considering the vast availability of CDRs, especially in data-scarce environments, these limitations can be mitigated through tailored approaches. What is missing in the literature is a unified framework that transforms CDR-derived mobility patterns into inputs for traditional transport modeling components, such as four-stage or activity-based models.

Contribution 2: This research introduces a novel travel demand modeling framework based on CDR data. It includes methods for OD matrix construction, trip type identification (home-based work, home-based other, non-home-based), and network assignment by inferring route choices based on trip frequency and temporal regularity.

The model aligns digital data processing with conventional transport analysis workflows, offering a bridge between big data and traditional modeling paradigms.

Gap 3: Lack of empirical validation against conventional survey data (Linked to Objective #3)

Another critical gap is the absence of rigorous validation studies that compare CDR-derived travel patterns with those obtained from Household Travel Surveys (HTS). While mobile data offers advantages in scale and granularity, its use in official transport planning remains limited without empirical benchmarks. Particularly in Sri Lanka, there is no established precedent for validating CDR-based models against household surveys at both intra- and inter-district levels.

Contribution 3: This study conducts a comprehensive validation of trips extracted from Mobile Network Big Data (MNBD) using Household Visit Survey (HVS) data across three districts in Sri Lanka's Western Province. It offers comparative insights at both district and Divisional Secretariat Division (DSD) levels, highlighting where and how CDR-derived trip estimates align—or deviate—from conventional data sources.

Summary of Contributions

This thesis addresses three core gaps in the literature through a set of tightly integrated methodological contributions:

1. Trip Detection Framework: Accurately infers individual mobility from raw CDRs, addressing noise and sparsity.
2. CDR-Based Travel Demand Modeling: Constructs OD matrices, classifies trip types, and performs route assignment within a traditional modeling context.
3. Empirical Validation: Validates findings through direct comparison with household travel survey data in a Sri Lankan case study.

By addressing these gaps, this research enhances the methodological foundation for using digital mobile data in transport modeling, particularly in resource-constrained or data-scarce environments where traditional data collection is limited.

1.5 Research Objectives

Aligned with the aims and the identified research gap, this study delineates three primary research directions, as outlined in Table 1.1. These directions include the feature extraction of Call Detail Record (CDR) Data, the estimation of travel demand, and the validation of the Model of Network-Based Data (MNBD). The research objectives are established to proceed within these specified paths. The specific objectives are detailed as follows.

Table 1-1

Scope and Direction of the Study

Categories	Research direction
Feature extraction from CDR Data	Processing the sparse data to a readable format to estimate travel attributes.
Travel Demand Estimation	Implementing the four-step - travel demand estimation model.
Validation of MNBD results	Validation of the derived distributions with traditional survey data

Objective #1

Develop a methodological framework to address the trip detection issues of digital data by inferring individual mobility.

As previously noted, Call Detail Record (CDR) data are inherently sparse, with records generated solely upon initiating an activity. This characteristic necessitates substantial processing to render the data appropriate for travel-related analyses or estimations. The research questions targeted by this objective are outlined below:

- Addressing the Load Sharing Effect (LSE)

After extracting a sample dataset, it is crucial to accurately deduce activities and trips from the CDR data. This requires filtering out the noise associated with the mobile service provider's tower-to-tower call balancing. This process can create the illusion of false movements, known as the Load Sharing or Load Balancing Effect. This phenomenon

negatively impacts transport predictions by potentially inflating the counts of trips between origins and destinations beyond their actual numbers, leading to misidentification of activity or travel behavior. The study addresses the Load Sharing Effect by implementing a speed-based filtering technique.

- Characterize user locations by presence patterns based on the longitudinal observations from the Call Detail Record.

Following the noise filtration, the analysis can proceed to the next phase. Each call generates a Call Detail Record (CDR) encompassing several critical attributes, as previously mentioned. These attributes facilitate the timestamping of the caller's mobile phone ID in conjunction with the tower, facilitating the connection to the phone network and thereby pinpointing a mobile phone user's location when placing or receiving a call. This process subsequently enables the determination of the tower zone most frequently accessed by each mobile phone user. By tracking the time from the initial call made from a specific location to the time of the last call from the exact location, it becomes possible to ascertain the minimum duration a user spent within a particular tower zone. Compiling all CDRs attributed to a single user aid in gathering diverse details and categorizing user locations. Once these locations are labeled and timestamped, the CDR data emerges as a valuable resource for transportation modeling and estimation.

Objective #2

Development of a novel methodology for handling the steps in travel demand estimation and network assignment to supplement traditional models using digital data.

The primary aim of this study is to develop a travel demand estimation model that can be applied to real-world scenarios. The second objective addresses this goal using processed Call Detail Record (CDR) data. In pursuit of this target, the following research questions will be explored within the study:

- OD Matrix Estimation from CDR data

Human activity-travel behavior is characterized by intricate patterns of stays and movements across space and time. An individual may either be traveling or stationary at a particular

location. This distinction is evident in the rate at which the connected cell tower changes; it is high during travel and remains constant when the user is stationary. Such differences can be effectively represented through a spatiotemporal matrix of users. This representation aids in estimating the Origin-Destination (OD) matrix, which is a crucial step towards generating trips and identifying points of attraction.

- Defining trip types

By mapping all Call Detail Records (CDRs) of an individual to a spatiotemporal matrix, it becomes feasible to gather detailed information such as:

Home Location: This is identified as the most frequently visited location by users, typically during nighttime and weekends. These periods are characterized by prolonged stays, suggesting the location's significance in the user's daily life.

Work Locations: Users regularly visit these places on weekdays, with visits noticeably diminishing on weekends. Extended duration at these locations implies substantial time spent by users, indicative of work or similar activities. Although labeled "work," this category also includes activities like education that exhibit a similar regularity and pattern, often called commutes. These represent routine and repetitive travel behaviors.

Other Locations: In addition to the primary locations of home and work, users visit other places that are less significant, referred to as non-dominant or non-significant locations.

Identifying the endpoints of trips facilitates the determination of home locations and the classification of trips as home-based, including those that originate from home. Within this categorization, it is possible to distinguish work-based trips. This approach comprehensively explains home-based work trips and other travel behaviors.

- Network Assignment from CDR Data

This section focuses on identifying Home-Based Trips by analyzing the regularity of individuals' travel patterns. Trips made by individuals are aggregated over time to infer probable route choices. The estimation of trips, categorized by origin, destination, and mode, is then applied to the transportation network. This analysis assumes that users aim to minimize

their travel time, individually or system wide. The ultimate objective is to ascertain the total number of trips undertaken via a selected route.

Objective #3

Validate the trips extracted from MNBD with trips estimated from the Household survey.

To ensure the methodology's accuracy and applicability to future predictions, it is crucial to validate the findings derived from Call Detail Record (CDR) data against existing, actual data. Therefore, this study uses household visit surveys to corroborate the findings.

1.6 Scope of the Research

The Western Province, which includes the districts of Colombo, Gampaha, and Kalutara, was selected as the study area for validation purposes. This region is subdivided into 47 Divisional Secretariat Divisions (DSDs).

This study categorized trips into three types: Home-based Work, Home-based Other, and Non-Home-based. These trips were validated at the district level, encompassing intra-district (within the same district) and inter-district (between districts) trips. However, validation at the DSD level was limited to inter-district trips, excluding intra-district trips due to the relatively small geographic size of DSDs. Typically, a DSD encompasses only a few cell towers, often no more than three. Since trip calculations are based on changes between cell towers and movements within the confines of a single cell tower do not register as significant changes, identifying trips within a single DSD could lead to a higher margin of error. This limitation arises from the challenge of detecting movements within the coverage area of a single cell tower.

Home-based work trips emerged as the most common and accurately identifiable category among the various possible trip purposes, such as shopping, visiting friends and relatives, and school travel. These trips were specifically extracted for the network assignment task. Subsequent chapters will present detailed discussions on trip identification across different spatial levels.

1.7 Overview of the Thesis

The thesis is organized into seven chapters as follows.

Chapter 1 introduces the background, motivations, research objectives, and the specific data, namely Call Detail Records (CDR). Aligned with the motivations and research objectives, the subsequent chapters cover three main areas: trip generation and attraction using CDR data, OD Estimation, and Network assignment using CDR data.

Chapter 2 The literature review introduces CDR and human mobility and discusses three main areas of CDR usage in transportation: OD Estimation, Mobility identification, and Network assignment. This chapter specifically focuses on CDR usage in Sri Lanka and highlights the research gap through a thorough review of existing literature.

Chapter 3 provides an overview of the research methodology, along with a brief introduction to the study area and the data used. The section on data provides a comprehensive explanation of the CDR data and the ground truth data used for validation purposes. Additionally, the section critically explains the sampling technique used for the study.

Chapter 4 primarily discusses the OD estimation process using CDR data. It begins with the initial data processing activity and provides a brief introduction to the spatiotemporal statistics that can be derived from CDR data. The chapter then delves into a detailed explanation of the user behavior extraction process from the CDR record, along with interpretations of the extracted information.

Chapter 5 focuses on the network assignment step using commuting trips. This section introduces a novel methodology for deriving trip statistics from CDR data at the individual or disaggregate level and later aggregating them for the final data generation. The findings obtained through this methodology are validated using existing ground truth data.

Chapter 6 The final chapter of the study includes a summary of the findings, conclusions drawn from the research, and recommendations based on the results. It also discusses the limitations of the study and provides suggestions for future research in the field.

CHAPTER 2

LITERATURE REVIEW

The study of human mobility patterns and the acquisition of empirical data have been areas of live research on numerous occasions. Many state-of-the-art approaches aim to work with data generated through wireless tracking devices. Wireless communication methods have significantly transformed how people communicate, work, and engage in day-to-day activities worldwide. As a result, mobile phone data, one of the most prevalent forms of wireless communication, can be utilized to derive spatio-temporal information about the whereabouts of anonymous phone users, enabling the analysis of their social mobility patterns. Mobile phone data has been employed in various transportation-related investigations, including urban planning and economic activity analysis. This chapter provides an overview of the theory of transport and studies in the transportation field that utilize mobile data. It initially discusses the traditional travel demand modeling process, then introduces the use of digital data in transport analysis, discusses different aspects of transport studies, and discusses the methodologies employed. The latter part of the chapter describes the limitations of existing techniques using Call Detail Records (CDR) and discusses the research gap in detail.

2.1 Traditional Travel Demand Modelling

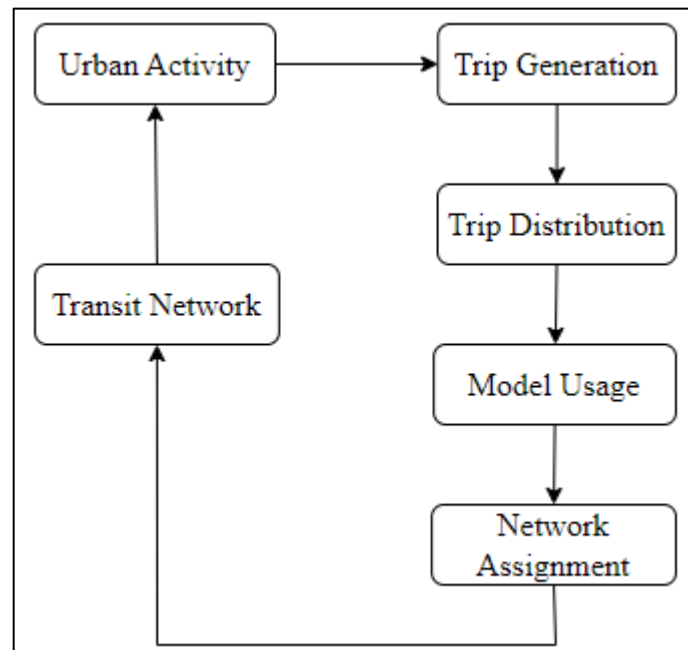
Travel demand modeling primarily aims to quantify travel within a transportation system, which arises from the geographical separation of various urban activities. This necessitates transportation, whose availability is depicted through the features of highways and transit networks.

At the core of travel demand modeling lies the traditional four-step modeling process, which has been the fundamental approach in this domain for over 25 years. This process has undergone significant refinement to represent urban travel patterns accurately. The traditional travel-demand modeling process is divided into four essential stages, as depicted in Figure 2.1 of Ortúzar and Willumsen (2011). This section focuses on the first, second, and fourth stages, which are instrumental in the analysis process (Ortúzar & Willumsen, 2011).

- Trip Generation/ Attraction identifies the number of trips that will be made.
- Trip Distribution determines where the trips will go.
- Mode usage predicts how the trips will be divided among the available modes of transport.
- Trip assignments predict the routes the trips will take, resulting in traffic forecasts for the network.

Figure 2-1

Illustration of the Traditional Travel Demand Model



Note: Source: (Ortúzar & Willumsen, 2011)

2.1.1 Trip Generation and Attraction

Trip Generation is when different moments due to human activities are converted into numbers of trips. In this aspect, the area or space taken might be similar, and the number of trips generated may differ. For example, a school and a shopping complex might occupy the same space, but the trips generated differ. Modelers try to quantify the liaison between urban activity and travel in the trip generation or attraction stage.

In the present scenario, survey data provide the inputs for this trip generation analysis. Travel survey data within the study area provide the types of trips made and the count of

each category. Once they relate to the land use patterns, trip generation can be modeled. The output of the trip generation analysis is the number of trips generated from a zone and the number of trips attracted to a zone. In addition, at this stage, the purpose of trips is also recognized. That can also vary based on the data availability and project purpose. The Comtrans study defines trip purposes in two methods ((Japan International Cooperation Agency [JICA] & Oriental Consultants Co., 2014). As of the first one, there are seven categories: to a home, work, shopping, business, private matters, school, and other, while the second grouping considers three clusters: home-based work, home-based other, and non-home based. All these trip categorizations are very useful since they provide an understanding of the person's behavior.

2.1.2 Trip Distribution

At the end of the trip-generation stage, a study will identify the number of trip productions and trip attractions each zone will have. But some other questions need to be answered,

- Where do the attractions in Zone 1 come from?
- Where do the productions go?
- What is the zone-to-zone travel volumes?

Trip distribution procedures arbitrate where the trips produced in each zone will go and how they will be divided among all other zones in the study area. There are several methods of trip distribution analysis, such as the Gravity model, the Fratar method, and the intervening model.

2.1.3 Network Assignment

Trip Assignment is the process by which planners predict the trips' paths. During this step, vehicles or people are assigned to the road and public transport network. A rational traveler's basic assumption when defining the route choice is that everyone tends to select the route that offers the least individual cost.

Although the above assumption is made, there are several factors that influence the route choice between a particular origin and the destination. These include journey time, distance, monetary cost (fuel and other), type of road (motorway, trunk road, secondary road),

congestion and queues, reliability of travel time, and habit. Due to the difficulty of modeling all these in a practical scenario, approximations are inevitable.

Trips are assigned to the existing road network or transit using a mathematical algorithm that arbitrates the traffic as a function of time, volume, capacity, or impedance factor. The following are different trip assignment mechanisms developed based on the above criteria. The most commonly used of these is the User Equilibrium Assignment (Chang & Dresser, 1990).

- User Equilibrium Assignment
- Incremental Assignment
- Capacity Restrain Assignment
- Transit Assignment
- Linear Programming Assignment

User Equilibrium Assignment

User equilibrium methods provide a more behaviorally realistic representation of traffic patterns than other techniques. An equilibrium assignment is constricted to each link travel impedance, link capacity, and, additionally, the total travel impedance of the network. The methodology generates an assignment where the total travel impedance is minimized. The concept behind the user equilibrium method is based on Wardrobe's first principles, and the results generated from the first iteration are the same as those generated from the all-or-nothing assignment method.

2.2 Use of Digital Data in Transport Analysis

The current generation of transport models takes advantage of the growing availability of data from mobile phones, GPS tracks, transactions, and online activity patterns to assess and forecast people's behavior when moving around a city and their interactions with different locations (Krause & Zhang, 2019). With the prevalence of GPS-enabled smartphones and in-car technologies, the collection of moving vehicle data has become more common and accessible than ever before. Studies have employed various techniques to forecast destinations in real time using trajectory data and past GPS coordinates. Markov models have been used to determine the most probable and probabilistic

following locations based on a subset of prior locations, and probabilistic models have been employed to forecast future vehicle destinations (Ashbrook & Starner, 2003). Some studies have focused on creating multi-modal destination prediction models considering modes other than personal vehicles (Liao et al., 2007).

In addition to GPS data, transaction data has been incorporated for journey forecasts. Smart-card payment systems, such as "Smart Trip" in Washington and "Sica" in Tokyo, provide valuable travel data and are integral to current public transportation fare collection systems. Researchers have proposed and examined methods to obtain Origin-Destination matrices from smart-card data for multi-modal public transportation systems based on anticipated alighting stops for journey segments (Munizaga & Palma, 2012).

Mobile phone data, which includes location and timestamp information for activity destinations, is less structured than traditional travel survey data or as accurate as GPS data regarding frequency and precision (Zheng et al., 2010). However, out of all the available data sources, mobile phone data, collected as a byproduct of mobile phone billing processes, is the most affordable and easily accessible information that can be used for trip predictions. Due to its massive volume and prolonged observation duration, mobile phone data can provide insights into human mobility on an unprecedented scale. Although such data may be sparse in space and time (Ratti et al., 2006), they can still capture spatiotemporal information about mobile phone users' travels at the cellular tower or even finer-grained levels, depending on the positioning technology used by service carriers (Jiang et al., 2016).

Mobile positioning data is primarily divided into two categories: "active positioning data" and "passive positioning data," based on how they are collected (Bwambale, 2018). Active positioning data involves requesting a device's location, while passive positioning data is gathered historically without specific requests. Both types can be historically monitored in real time (Kaisler et al., 2013). Passive positioning data mainly includes Call Detail Records (CDR) and Visitor Location Register (VLR). VLR data generation is logged every time a subscriber switches to a different serving base station, providing a higher-resolution representation of human motion than CDR. However, analyzing VLR data can

be challenging due to the high volume of logs, storage requirements, and processing expenses. As a result, CDR in the form of passive location data is the most collected by all mobile service providers (Lokanathan & Kumarage, 2016). CDR is more readily available than other network-related data, such as GPS, because most telecommunication service providers store this data (Dash et al., 2015).

2.3 CDR and Human Mobility

The conventional method of estimating is based on a small survey sample of daily travel; the travel demand depends on creating analytical techniques and drawing conclusions that allow a small sample to represent daily travel accurately. On the other hand, mobile network data doesn't record trips even if it offers better geographic coverage of travelers and a longer duration for travel behavior observation. So, developing new processes, techniques, and algorithms for deriving trips and activities is necessary.

Call Detail Records were heavily utilized in prior research to examine individual human mobility by using new techniques and algorithms. Human daily trips are planned according to daily activities and anchor places, which may also be referred to as each person's activity space (Jiang et al., 2016). Individual trajectories, which indicate a person's movement in location over time, can be used to define activity space (Zhao et al., 2016). A substantial amount of literature uses this specific activity space to investigate various aspects of travel. The analysis of travel or activity space has been mainly studied under three main sections below.

- Origin Destination Matrix Estimation
- Mobility Pattern Recognition and Significant Location Identification
- Route Choice Estimation

2.3.1 Origin Destination Matrices

In order to use mobile network data to derive OD matrices, a set of procedures need to be considered. At first, a trajectory must be extracted from mobile network connectivity using

timestamp and connected location information. Later, a technique must be devised. Trajectory segments are divided by stop locations to convert an extracted trajectory into trips. Setting the strategic temporal and spatial thresholds is an example of a trajectory transformation technique. This can be considered an effort to find the changes in connectivity to nearby locations and thereby identify the traveler's stop and stay locations (T. Luo et al., 2017; Valerio et al., 2018). The detected stop locations may be used as either the starting point for the subsequent trip or the ending point for the previous trip. The modeling of trip distribution and trip generation uses these origin and destination pairings (Demissie et al., 2019). Ghosh, Sarkar, and Nagesh (2023) identify users' home and work zones by analyzing the most frequented cell tower locations during nighttime and daytime, respectively, and distinguish between stationery and pass-by locations to improve OD flow estimation. Developed origins and destinations can be grouped into a conventional fixed Traffic Analysis Zones, such as official boundaries, grids, and zones based on Voronoi diagrams, or an adaptive zoning scheme where O-D regions change depending on the density of the spatial demand (Demissie & Antunes, 2016; Phithakkitnukoon et al., 2010; Gundlegård et al., 2016; Mungthanya et al., 2019).

Previous studies follow different methods for the OD matrix estimation as summarized in Appendix C. Fekih et al. develop a method by which trips are extracted from the spatiotemporal traces of users based on a minimum stationary time assumption that enables determining activity (stop) zones for each user (Fekih et al., 2020). Santi et al. also used a similar approach in OD estimation, where a minimum stationary time is applied to detect stays and pass-by locations (Phithakkitnukoon, 2022). Following an identical approach, Mamei et al. identified individual trips based on movements between cells (sequence of CDR) or of inferred routine trips known to be taking place, and then, individual trips are aggregated to create the OD matrix (Mamei et al., 2019). While these methods provide a way to convert CDR data to ODs, they showcase a trip-based approach that mimics trip segments of travel based on the appearance of people in space and time. Bwambale et al. predict individuals' demographic group membership relativeness based on their phone usage characteristics and then use this relativeness as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups (Bwambale et al., 2017). This study uses voluntarily collected CDR

data and individuals' demographic information. One other developed a method to generate tower-based transient origin-destination (OD) matrices for different periods and convert them to node-to-node transient ODs in the road network for Boston and San Francisco, and similar studies have been conducted in Dhaka, Bangladesh as well (Iqbal et al., 2014; M. Wang et al., 2013).

However, it is important to resemble the individual's spatiotemporal behavior before aggregating them into traffic zones for validation purposes. Since individuals are the smallest unit of trip makers, it is important to understand and capture their behavior. Previous studies do not systematically map and analyze users' movement states and visualize their behavior in matrix layers, especially how trip aggregation takes place over time.

2.3.2 Mobility Pattern Recognition and Significant Location Identification

Human daily travels are organized based on activities and anchor locations that are important in daily life. Studies used different methods to capture these anchor points and estimate the trips between these anchor points using CDR data. Most of all, the studies use frequency and regularity of the appearance in particular cells to label them as anchor locations (work, home, or other). Luo et al. combine CDR data with road network data to identify trajectories to decrease base station positioning errors. Next, the trajectory regularity is measured, and home and work locations are estimated using the regularity (Luo et al., 2020; Zagatti et al., 2018). Zagatti et al. have also used a similar technique to identify these significant locations. Some studies use clustering techniques, where cell towers that are spatially close are clustered together, which could also be represented by clusters containing towers that frequently channel call activities (F. Wang & Chen, 2018). Mamei et al. cluster CDRs together based on the spatial parameter and weigh clusters based on time to determine those associated with frequented places (Mamei et al., 2016). Rodrigues, Veloso, Alves, and Bento (2023) further highlight the effectiveness of using CDRs to analyze population mobility changes during pandemic contexts by comparing mobility data from two time periods—just after lifting COVID-19 measures and the same period the following year—revealing significant shifts in movement patterns. Similarly,

Li, Ran, Tsai, and Williams (2023) combined radius of gyration, movement regularity, and motif types from CDR data, applying K-means clustering to identify mobility behavior types and correlating them with socio-economic indicators to understand disparities in pandemic responses.

Though many researchers have addressed this trip detection based on anchor points, their methods do not capture individualistic behaviors, where the same regularity and frequency criteria are used for all the individuals in the sample or clusters. It is more challenging to combine these data with person-level attributes, and previous studies have not developed data fusion frameworks to specifically address this challenge.

2.3.3 Route Estimation from CDR data

The target of the network assignment is to assign OD pairs derived from CDR data to the road network. This can be conducted in two steps. The first step is route set generation, in which a set of alternative routes for each OD pair is determined. Most studies use OpenStreetMap for this purpose (Lwin et al., 2018; Sakamane et al., 2020). The second step is the flow distribution, which distributes the total flow in the OD pair among the routes in the route set, generally completed using defined cell paths (Breyer et al., 2018). However, no previous study has used CDR data which cooperates with individual caller attributes when deciding the route choice. The OD pairs are typically assigned using a simple process based on appearance frequencies, yet this can be more systematically completed by considering individual caller frequencies, appearance frequencies, the number of actual trips made, and interpolated trips. Petrovic and Crivellari (2023) demonstrate the feasibility of using a dual-Transformer framework for CDR-based trajectory reconstruction, where the first Transformer performs map matching on GPS traces and the second reconstructs road segments from sparse cell-level data, showing acceptable performance despite limitations from data sparsity.

2.4 Use of CDR in the Sri Lankan Context

Between 2011 and 2021, Sri Lanka's cellphone penetration increased from 87% to 141% (Mathematics, 2016), indicating widespread usage of mobile devices for daily activities. It is worth noting that some individuals may use multiple SIM cards, while others may not use mobile devices at all. Given the high prevalence of mobile device usage, it is

reasonable to assume that the travel habits of a random sample of mobile users accurately represent the population of mobile users.

CDR has been employed in various studies in Sri Lanka, including those focused on economics, urban planning, transportation, and other domains. By combining CDR data with geographic and demographic information, it is possible to gain insights into human movement patterns. Research utilizing acceptable spatial and temporal variations of mobile phone data has examined the relationship between commuter flows and urban economic activity, quantifying urban economic activity (Kreindler & Miyauchi, 2015). Community detection algorithms have also been applied to CDR data to understand communication patterns and identify communities (Madhawa et al., 2015). Furthermore, CDR has been used in land use classification, a critical aspect of urban planning (Madhawa et al., 2015). These studies explore the potential of using large volumes of mobile usage data to infer spatio-temporal activities of the population and provide practical indicators for activity-based land use classification. Additionally, CDR has been utilized in certain areas of disaster management and resilient development.

One noteworthy study in Sri Lanka's transportation field involves the estimation of origin-destination matrices using CDR. Various methods, such as the stay-based approach, frequency-based approach, and transient-based approach, have been employed to derive OD matrices using CDR data (Maldeniya, Kumarage, et al., 2015; Maldeniya, Lokanathan, et al., 2015).

In the stay-based approach, a stay is described as a continuous set of recordings so that;

- Connecting cell towers should be separated by less than 1 km.
- The total duration of the records should be greater than 10 minutes.
- Two recordings are separated by a period of less than 1 hour.

The shift between two stays signaled the introduction of a trip using the stay-based methodology.

If the cell tower used for each call is different and the records are separated by a time gap of 10 minutes to 1 hour, the movement is regarded as a trip under the transient-based approach. Additionally, a person's daily sequences over a year can determine their frequent locales by utilizing a frequent trip-based technique. To create O-D pairs, the travel between the most frequently visited places was estimated as trips. However, before MNBD may be utilized to replace conventional transport forecasts, there are differences between the results of the traditional transport surveys and MNBD estimations that must be understood.

2.5 Drawbacks in Call Detail Record Data

Despite the rich mobility information encapsulated within Call Detail Record (CDR) datasets, several challenges emerge, affecting their utility in real-world applications, including transportation initiatives. These challenges complicate the analysis process and limit the applicability of CDR data.

The primary concern is the spatial resolution of CDR data. The resolution is relatively low, with user locations approximated based on the coordinates of the cell tower to which their mobile device is connected. This approximation makes it challenging to pinpoint the precise location of users. The coverage area of cell towers varies with land use; urban areas typically have higher cell density than rural areas. For instance, the average coverage area per cell tower in the Colombo district is approximately 2.3 square kilometers.

In contrast, in more sparsely populated areas like the Kalutara district, it can extend to about 18.3 square kilometers. Consequently, location estimation accuracy decreases, especially in areas with dense cell coverage where frequent tower switching further complicates location determination. Some mobile operators, primarily in developed countries, offer triangulated CDR data, significantly reducing location estimation errors by capturing the regularity of user behaviors.

Another significant drawback of CDR data is its event-driven nature; records are generated only upon specific events such as making/receiving phone calls, sending/receiving messages, or web browsing. This characteristic renders the dataset sparse, creating a disparity between observed and actual user behaviors in the dataset. The gap is notably wider among infrequent

mobile users than frequent ones. To mitigate this issue, analysis often focuses on data from frequent mobile users to enhance the accuracy of travel estimations.

The "load sharing effect" presents an additional challenge, necessitating substantial processing to address the errors it introduces. This effect arises when operators manage call traffic by diverting calls to adjacent cell towers based on each tower's current load. This practice can lead to incorrectly identifying user locations and misclassifying stationary periods as a movement.

Together, these limitations introduce various inaccuracies in determining user locations, durations of stays, and identification of pass-by points, highlighting the complexities involved in leveraging CDR data for transportation analyses.

2.6 Techniques Used in Overcoming the Limitations of CDR

Most researchers have employed various techniques to mitigate errors in CDR data and ensure accurate interpretation of the results. One crucial factor many studies consider is location identification, which involves determining the places users visit. Different methodologies have been used to obtain these results. A common approach is trajectory smoothing, where a series of visited places is created by filtering the sequence of CDR based on time, speed, or proximity to designated center positions for nearby locations (Csáji et al., 2013; M. H. Wang et al., 2013; Isaacman et al., 2011). In some cases, locations have been grouped based on spatial distribution without considering the time factor, even if the points of interest were visited on different days (Jiang et al., 2013a; Hariharan & Toyama, 2004). This methodology helps understand the places individuals visit most frequently as part of their routine.

Another vital aspect is movement state identification, which determines whether a user has stayed at or passed by a location. Many studies establish a time threshold to classify movement states (Calabrese et al., 2011; Maldeniya, Kumarage, et al., 2015). For example, a lower threshold of 10 minutes and an upper threshold of 1 hour are commonly used. Additionally, the records must be at least 1 km apart. If a record sequence meets these criteria, it is classified as a stay site, and travel between two stays is considered a trip. However, this technique is more effective for high-frequency data, marking many

data points as pass-by. To address this, locations are connected to previously visited places, and the 10-minute lower threshold helps reduce incorrect trip identification caused by stationary individuals connecting to nearby towers (Maldeniya, Kumarage, et al., 2015).

Data sparsity can lead to discrepancies between observed and actual trip ends, distinguishing between "observed end locations" and "hidden end locations." One study specifically addressed this issue by introducing a 1-hour upper constraint to reduce the likelihood of concealed visits during observed displacements (Zhan et al., 2016). The choice of this threshold is often empirical, considering that sparse data tends to overestimate the occurrence of hidden visits. Despite its significant implications for trip identification and the extraction of origin-destination pairs, many studies pay less attention to this problem.

2.7 Research Gap Identification

The literature review reveals that CDR has been applied across various transportation-related tasks, including estimating origin-destination (O-D) matrices, identifying significant locations, and understanding human mobility patterns. Despite the substantial body of research on leveraging CDR for transportation issues, it is observed that the depth of theoretical application often does not meet expectations. Notably, discussions on fundamental trip concepts, such as trip purposes and differentiation between intra-city and inter-city journeys, appear to be underrepresented in existing studies. Additionally, analyzing data at an individual level before aggregating findings is seldom adopted, marking a significant gap that this study aims to address.

A further limitation within the current scope of research is the absence of labeled CDR data, which restricts the potential for deriving precise travel estimations. This study seeks to bridge this gap by focusing on the labeling of CDR data to facilitate meaningful travel analysis. Moreover, the issue of load sharing, an inherent complication within CDR data, has not received comprehensive examination in prior studies. This research proposes a model to estimate travel demand based on CDR data by addressing these challenges and incorporating all processed data.

An essential aspect of travel that emerges from the literature is its purposeful nature, underlining the importance of incorporating trip objectives into the analysis. Previous research has often delineated geographic and temporal boundaries for study based on the socioeconomic status of users, employing heuristic methods and arbitrary criteria. Recognizing the significance of socioeconomic factors, this study intends to integrate and adapt these criteria to the context of Sri Lanka, enhancing the relevance and applicability of the findings.

In summary, this research not only aims to fill the gaps identified in previous studies by emphasizing individual-level analysis and addressing the challenges associated with labeled CDR data and the load-sharing effect but also seeks to refine the application of socioeconomic criteria within the framework of transportation analysis. By validating the proposed methodologies against existing transportation data, this study advances our understanding of travel demand estimation using CDR data. The following section will detail the methodology adopted in this research, outlining the approaches for data labeling, analysis, and model development.

CHAPTER 3

METHODOLOGY

This section examines the methodology employed to extract and visualize information from data provided by mobile network operators that is either directly or indirectly related to human activity. The primary objective is to outline the process used to analyze human behavior to estimate travel demand. The methodology chapter is organized into four main parts, each dedicated to a specific aspect of the methodological approach. This structured format offers a clear and concise understanding of the procedures involved in leveraging mobile network data for travel demand estimation analysis.

- **Exploration of the Study Area:** This initial step involves a comprehensive examination of the geographical and socio-economic characteristics of the area under study. Understanding the context within which the mobile network data operates is crucial for accurate analysis and interpretation.
- **Data Understanding and Feature Extraction:** This phase analyzes Call Detail Record (CDR) data to identify key features that can inform human behavior. Additionally, it introduces the concept of ground truth data, which serves as a benchmark for validating the findings derived from the CDR data.
- **Population Sampling Techniques:** This section outlines the strategies for selecting a representative population sample. The objective is to ensure that the sample accurately reflects the broader population's behavior patterns, enabling reliable travel demand estimations.
- **Description of the Research Methodology:** The final section provides a detailed account of the methodology adopted. It encompasses the methods used for data collection and analysis and the approach for integrating various data sources to estimate travel demand effectively.

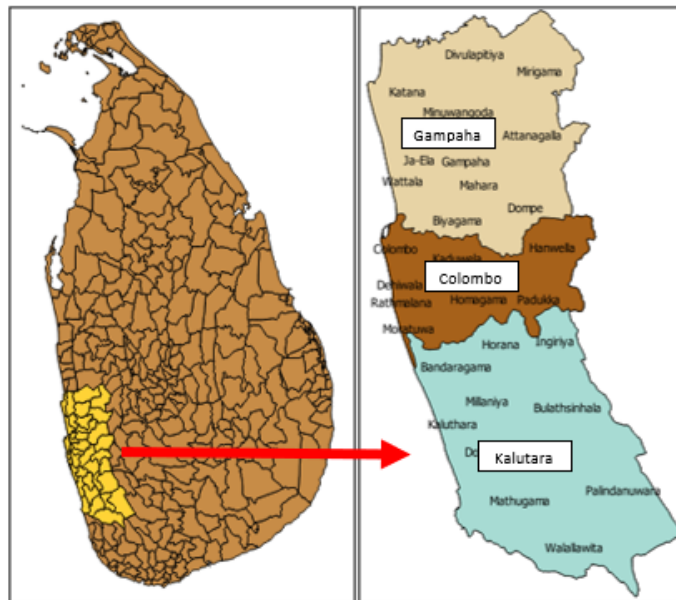
3.1 Study Area

The Western Province (WP) of Sri Lanka has been selected as the focal area for this study, motivated by the pressing need to address the escalating transport demand and the availability of comprehensive data. As the foremost administrative division within the nation, the Western Province is home to approximately 5.8 million residents, as reported by the 2012 census. This province encompasses three districts, Colombo, Gampaha, and Kalutara, which comprise 48 Divisional Secretariat (DS) divisions, as illustrated in Figure 3.1.

Both road and rail systems are utilized in public transportation within the WP. The road network within the confines of this study spans approximately 3,684 square kilometers. It has been documented that the average trip production rate per individual stands at 2.9, with motorized transport modes accounting for 78% of this (DSRSL Ministry of Transport et al., 2014). Traffic congestion becomes significantly pronounced during peak travel hours, reducing the average vehicular speed on roads across the Western Province to a mere 15 kilometers per hour (Shaadya Ismail, 2018). Moreover, mobile phone penetration in the WP is estimated to exceed 100%, indicating a high level of mobile network usage among the inhabitants.

Figure 3-1

District and DSD segregation in the Western Province of Sri Lanka,



3.2 Data Description – CDR Data

This research employs a dataset comprising Call Detail Records (CDR) furnished by a mobile operator in Sri Lanka, provided explicitly for this study by LirneAsia, a regional think tank specializing in ICT policy and regulation. The CDR, collected at the operator's Mobile Switching Center, encompasses detailed records of voice calls, Short Message Service (SMS), and data transfers. This section elaborates on the diverse features encapsulated within the CDR, outlines the processes involved in data cleansing, and discusses their application in analyzing human travel behavior. Additionally, it presents insights into the methodology for aligning cell coverage areas, which is critical for accurately interpreting the spatial dimensions of the data about human movement.

3.2.1 Data Privacy and Security Measures

Ensuring data privacy and secure handling of sensitive information was a key priority throughout the research process. As mentioned in previous chapters, the dataset used in this study was fully pseudonymized prior to its transfer to the study. Pseudonymization involved replacing personal identifiers with non-identifiable codes, ensuring that individual users could not be directly or indirectly identified.

In addition to pseudonymization, different data protection protocols were carried out during the data handing process. Access to the data was restricted to authorized personnel only and it was required to sign data confidentiality agreement before using the data. The data was stored on secure university servers with restricted network access. No data was stored on personal devices or external storage media. In addition, the data will be retained only for the duration necessary to complete the research objectives and will be securely deleted.

3.2.2 Attributes in CDR Data

Sri Lanka's telecommunications landscape comprises five main network operators, whose market shares range from 11% to 44% (Oxford Business Group, 2023). The dataset utilized in this study originates from a single operator, which boasts a subscriber base of just over 13 million individuals. This figure represents approximately half the country's total population

of 22.16 million, with the network providing services via 2G and 3G telecommunications technologies. The data used in this study corresponds to the month of June 2013.

Turning to the dataset attributes, as delineated in Table 3.1, the voice calls and SMS records are stored in their originating files. These logs capture a wealth of information, including the timing of each call, the anonymized IDs of both the calling and the called subscribers, the duration of each call, and the cell towers servicing the initiating and receiving ends of the calls at their commencement. Similar details about SMS activities are encapsulated within the SMS-originating logs. As a foundation for subsequent analytical endeavors, the attributes are deemed crucial for identifying subscriber behavior.

Table 3-1

Attributes in a CDR Entry

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

3.3 Household Visit Survey Data

The Western Province has witnessed a significant surge in transport demand in recent years. Numerous transport development projects and plans have been implemented to address the anticipated challenges in meeting this demand and solving related problems, accompanied by various surveys to gather reliable transport data. Among these, the Household Visit Survey

(HVS), conducted by the Ministry of Transport with technical assistance from the Japan International Cooperation Agency (JICA), stands out for its validation role in the current study due to its scope and comprehensiveness. This survey, one of the most extensive transport surveys conducted in Sri Lanka, encompasses the entirety of the Western Province, including the Colombo, Kalutara, and Gampaha districts. The survey area is subdivided into 2,496 Grama Niladhari Divisions and 47 Divisional Secretariat Divisions. The reluctance of individuals to participate in such surveys necessitated an increase in the original sampling size from the planned 3% to 4% to ensure representativeness ((JICA) & ORIENTAL CONSULTANTS CO., 2014). Table 3.2 presents the number of households interviewed in each district, reflecting the scale and distribution of the survey effort across the Western Province

Table 3-2

Estimation of the target sample size from the HVS

Criteria	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

Note: Source: Urban Transport Development project for CMR and Suburbs

Four distinct types of survey forms were utilized to facilitate a comprehensive data collection for analysis, each targeting different household and individual mobility characteristics. The initial form, distributed to each participating household, aimed to capture socio-economic data. This included information on the household's address, income level, vehicle ownership, expenditure patterns, and preferences regarding urban transportation options.

Subsequently, a second questionnaire was provided to every household member. It was designed to collect detailed information on individual transportation costs, the addresses of workplaces or educational institutions, the nature of their workplace or school, and specifics regarding their habitual commuting routes. A specialized form was introduced for households

that had experienced a change in residence within the past five years. This instrument was intended to gather insights into the motivations behind their relocation and particulars about their former residences. The final form, directed at individual household members, concentrated on trip information. This encompassed data on the origins and destinations of trips, their purposes, chosen modes of transportation, any transfer points encountered, and the timing of departures and arrivals. The collection of trip data was focused predominantly on a typical weekday.

Key among the data points for travel details were the purposes of the trips and the zones where trips ended. These elements were of particular importance for validating the analysis conducted. The survey findings indicated that an estimated 10 million trips occurred on a typical weekday within the Western Province. Of these, approximately 2 million were identified as Home-based Work (HBW) trips, with the remainder categorized as either Home-based Other (HBO) or Non-Home-based (NHB) trips.

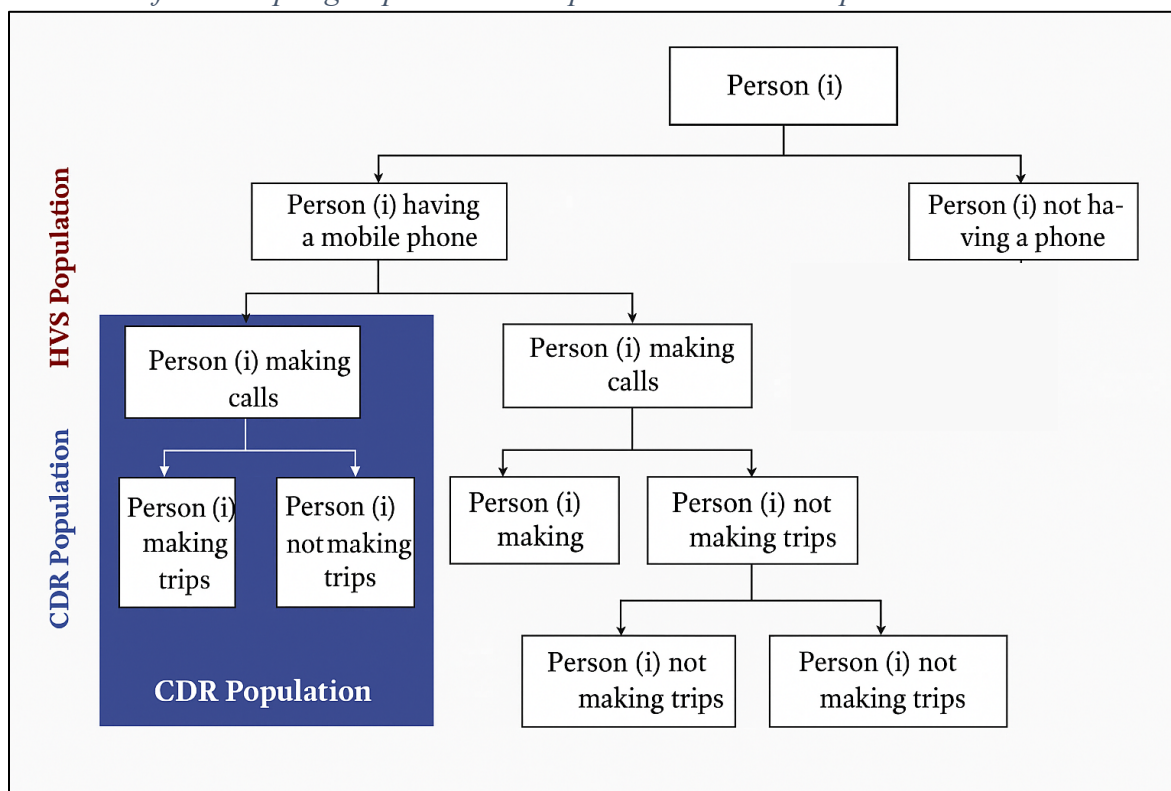
In summary, the Household Visit Survey provides a rich and detailed account of travel behavior across the Western Province, capturing socio-economic, mobility, and trip-related data at both the household and individual levels. Notably, the HVS data was collected during the same period as the Call Detail Records (CDR) dataset in year 2013, ensuring temporal alignment between the two data sources. This synchronicity enhances the reliability of comparative and integrative analyses, allowing for a robust validation of mobility patterns derived from the CDR data using independently collected survey-based information

3.4 Sampling Technique

Traditional travel data collection methods typically aim to represent the entire population, capturing details of all individuals regardless of their connectivity to cellular networks or engagement in trip-making activities, as illustrated in Figure 3.2. In contrast, the population captured through Call Detail Records (CDR) encompasses only a subset of the broader population, including individuals who own mobile phones and engage in call-making activities.

Figure 3-2

Overview of the Sampling Population -Comparison with HVS Population



Although the CDR population is theoretically a subset, the proportion of non-included categories is significantly small, attributed to the high mobile phone usage penetration rates. Furthermore, the general population engaged in trip-making activities, particularly those aged between 14 and 65, can be practically considered mobile phone users. This assumption is based on the widespread adoption of mobile phones within this demographic, thereby allowing for a reasonable approximation of the trip-making population through CDR data (DSRSL Ministry of Transport et al., 2014).

While mobile phone penetration is nearly universal in urban Sri Lanka, certain groups—such as very young individuals, the elderly, or low-income individuals without mobile devices—may be underrepresented in CDR datasets. This study assumes that trip-making is minimal among those not captured in the dataset, and thus their exclusion has limited impact on mobility analyses. Nevertheless, this assumption introduces a potential bias that is noted and acknowledged. Additionally, individuals with multiple SIM cards or shared

devices may contribute disproportionate or irregular activity, possibly influencing trip count metrics. This was mitigated by including only users with consistent activity over the selected one-month period.

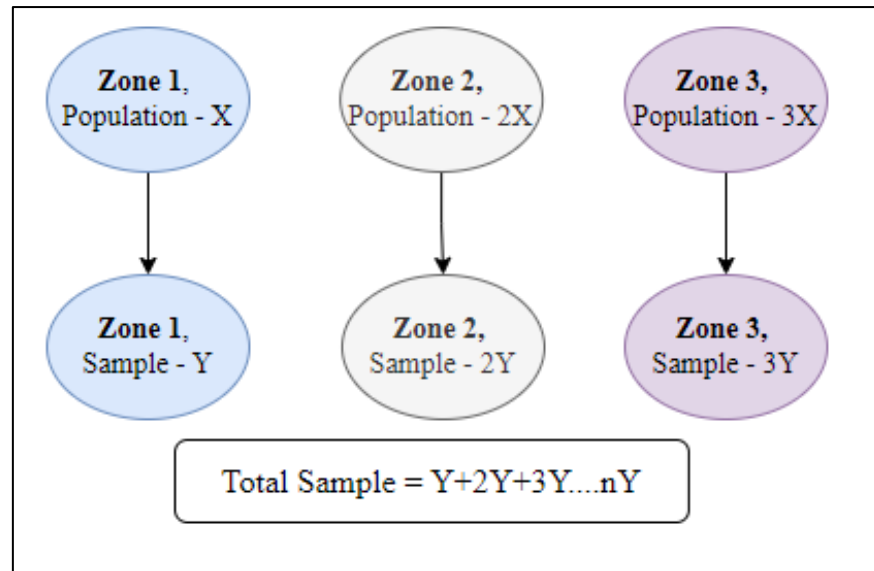
To achieve a representative sample that mirrors the general population, this study adopted a multi-stage random sampling technique to select the analysis sample from the original Call Detail Record (CDR) population, comprising 13 million unique users. The sheer volume of the CDR population rendered a comprehensive analysis impractical due to the substantial computational resources and time such an endeavor would demand.

The multi-stage sampling methodology involves sequential sampling, each utilizing progressively smaller units. This technique is comparable to cluster sampling, as it initially segregates the population into distinct groups. This practice is particularly prevalent in national surveys that span geographically widespread populations. In this research context, the sample extraction stages were aligned with the minimum resolution required for DSDs within the analysis framework.

The process commenced with identifying the minimum sample size necessary to represent the demographic within the least populated DSD accurately. Subsequently, this figure served as a benchmark to calculate the requisite sample sizes from the remaining DSDs, employing the population ratios of the lowest populated DSD as a reference point. For a comprehensive understanding of the multi-stage random sampling method utilized in this study, Figure 3.3 offers a visual depiction.

Figure 3-3

Graphical Representation of Multi-Staged Random Sampling technique



3.4.1 Identifying the Sample Required from the Lowest Populated DSD

The following criteria are generally considered when deciding on a representative sample (Naing, 2003)(US Census Bureau, n.d.). Among different sets of specific strategies, the level of precision, confidence level, and degree of variability seemed to be given considerable weight when ensuring representativeness.

- The level of precision - Range in which the actual value of the population is estimated to be Calculated. This is also sometimes called sampling error. Generally, the range is expressed in percentage points (e.g., ± 5 percent)
- The Confidence Level—The confidence is based on the Central Limit Theorem. The key concept in the central limit theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the actual population value, and the values obtained by these samples are normally distributed. If a 95% confidence level is selected, 95 out of 100 samples will have the actual population value within the range of precision specified earlier.

- *Degree Of Variability* - Refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The less variable (more homogeneous) a population, the smaller the sample size. 0.5 indicates the maximum variability in a population

In his book *Statistics-An Introductory Analysis*, Yamane provides a simplified formula to calculate sample sizes, as shown below. A 95% confidence level and a maximum variability of 0.5 is assumed in developing that (Naing, 2003).

$$n = \frac{N}{1+N(e^2)} \text{-----}(1)$$

N= Population size

n = Sample Size

e = Error Rate / Level of Precision

Madurawala has the lowest population out of all the 47 DSDs, and the above formula is used to calculate the sample required for the least populated Madurawala DSD. An error rate (e) of 2%¹. Is assumed

$$N = 34,381, e = 0.02$$

$$n_{\text{madurawala}} = 34,381 / (1 + 34,381 * 0.02^2) = 2330$$

Consequently, to maintain a 95% confidence level, a 2% margin of error, and a variability of 0.5, a minimum of 2,330 users from Madurawala is required. The sample sizes required from other Distance-Size Distributions (DSDs) were calculated by applying the population ratio of Madurawala to the respective DSDs and multiplying this ratio by 2,330. The specific calculations and resultant sample sizes for each DSD are detailed in Table 3.4 below.

¹ Calculated the total values by changing the error rate from 1%-5%, 2% generated the optimum solution by considering the effort required and acceptable sample size.

Table 3-3*Total Sample Calculation according to the Multistage Random Sampling Method*

	DSD_name	Population	Ratio of Population	Calculation	Sample required from each DSD
Base DSD	Madurawala	34,381	1	1*2330	2330.00
	Colombo	323,257	9	9*2330	21907.12
	Kolonnawa	191,687	6	6*2330	12990.63
	Kaduwela	252,041	7	7*2330	17080.82
	Homagama	237,905	7	7*2330	16122.82
	Hanwella	113,807	3	3*2330	7712.70
	Padukka	65,267	2	2*2330	4423.14
	Maharagama	196,423	6	6*2330	13311.58
	Sri Jayawardhanapura kotte	107,925	3	3*2330	7314.08
	Thimbirigasyaya	238,057	7	7*2330	16133.12
	Dehiwala	88,962	3	3*2330	6028.95
	Ratmalana	95,506	3	3*2330	6472.44
	Moratuwa	168,280	5	5*2330	11404.33
	Kesbewa	245,232	7	7*2330	16619.37
					↓
Total					398861.0171

Note: ** A Detailed calculation table is attached in Appendix A.

Table 3-4*Sample Comparison-HVS and CDR Sample*

HVS Survey	CDR_ Suggested Sample
Sampling Rate = 4%	Sampling Rate = 11%
Error Rate = Between 4% to 5%	Error Rate = 2%
No. of Individuals = 140,000	No. of Individuals = 396,531 $\approx$ 400,000

Based on these calculations, it is established that 396,531 users are necessary to uphold the stipulated criteria of a 95% confidence level, a 2% margin of error, and a variability of 0.5. This figure represents approximately 11% of the total sample, with the trip-making population within the Western Province estimated at 3.6 million individuals. Subsequently, a random sample of 400,000 users was extracted for one month (July 2013) to facilitate the analysis. As delineated in Table 3.4, the sample size thus obtained is statistically robust, offering a significant improvement in representativeness over the data traditionally collected through Household Visit Surveys (HVS).

In conclusion, by combining a multi-stage probabilistic sampling strategy with an understanding of mobile phone penetration and usage behavior, the study ensures that the CDR sample meaningfully reflects the active trip-making population. At the same time, the study openly acknowledges potential *biases and exclusions*, ensuring that conclusions drawn from the analysis are framed within appropriate bounds of generalizability.

3.5 Overview of the Research Methodology

Figure 3-4

Conceptual Framework of the Study

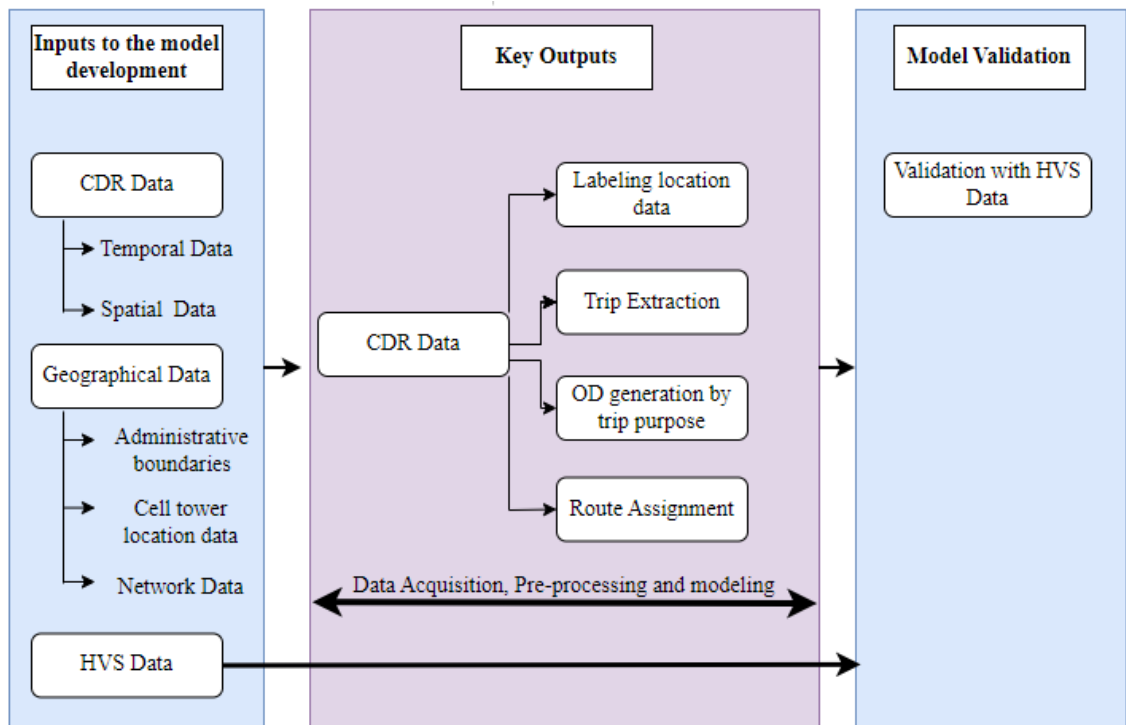


Figure 3.4 presents the study's conceptual framework, while Figure 3.5 outlines the methodology at a macro level. This section briefly overviews each methodological layer, offering initial insights into the research approach. Figure 3.6 depicts the expected outputs graphically.

Input Layer: This initial layer introduces the preliminary data into the model for subsequent detailed processing. The analysis commences with a 400,000 Call Detail Record (CDR) sample data collected over one month. Additionally, geospatial data concerning cell tower locations and administrative boundaries are utilized as foundational inputs for modeling. The Household Visit Survey (HVS) dataset, crucial for the final validation stage, is also integrated as primary input data.

Preprocessing Layer: At this juncture, data is either excluded or adjusted to ensure the use of accurate observations in the modeling phase. The data's structure is refined for clarity and interpretability, focusing on addressing the load-sharing effect, a prevalent error in CDR data, during this phase.

Moment State Layer: This layer identifies sequences of coordinates with timestamps for each user, revealing the digital movements of subscribers within the observational period. Records are associated with specific activity types based on location, though this layer of activity remains obscured in passive-positioning CDR data and necessitates significant processing to uncover. Based on user location patterns, the moment state layer categorizes locations into home/work and other significant points. It utilizes changes in location appearance to delineate trips and stays from the processed dataset.

Travel Demand Estimation Layer: This pivotal layer is constructed upon the traditional four-step modeling process. This layer's data analytics and interpretation component is detailed in two thesis chapters: "OD Estimation by Purpose" and "Route Choice Estimation for Commuter Trips." Data from the moment state layer undergoes comprehensive processing to align with the requirements of the conventional Travel Demand Estimation (TDE) process.

Output Layer: This final layer yields the study's outputs. The results are validated against traditionally collected data by situating them within administrative zonal levels. Ultimately, this stage showcases the refined CDR-based Travel Demand Estimation model, marking the culmination of the research endeavor.

Figure 3-5

Methodology of the study at Macro level

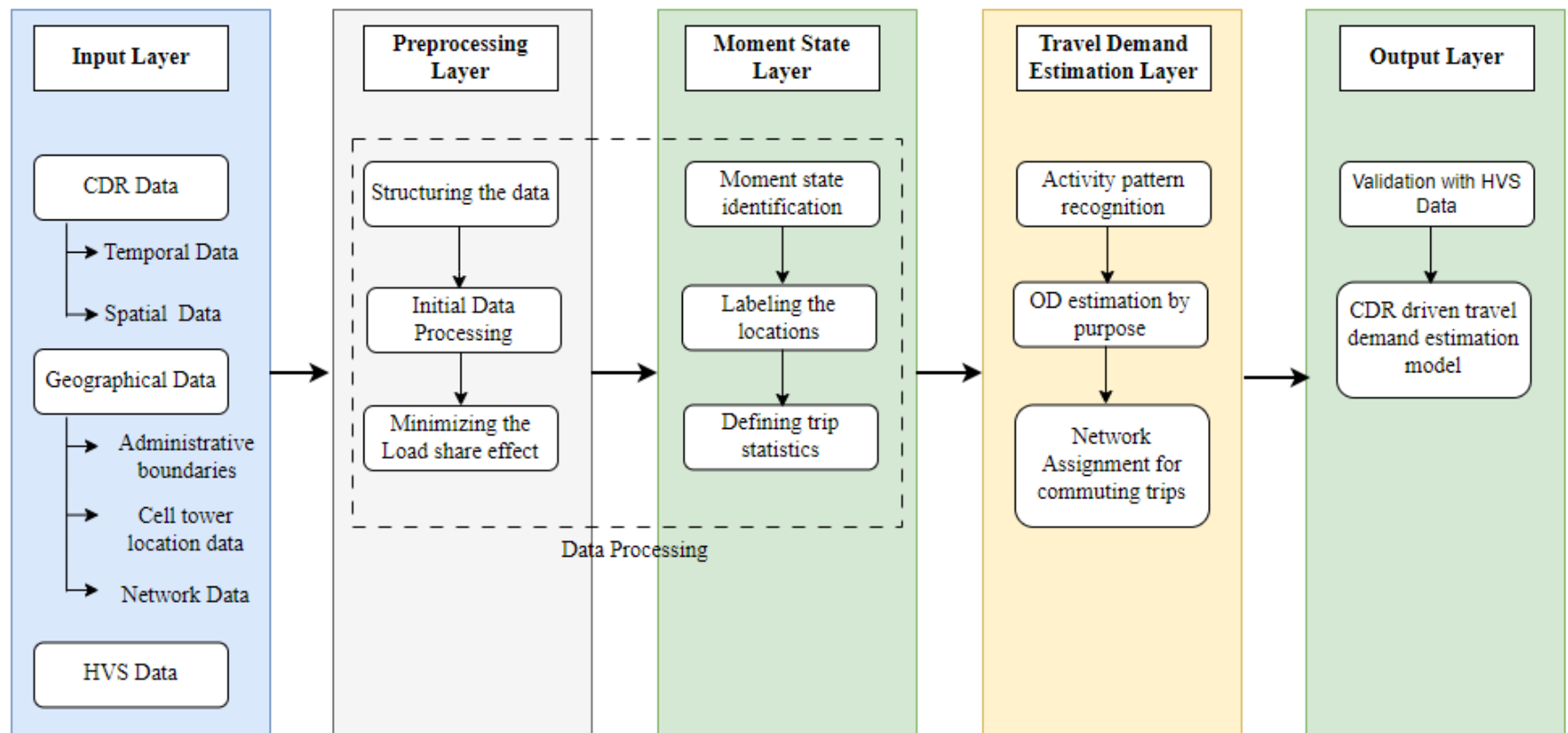
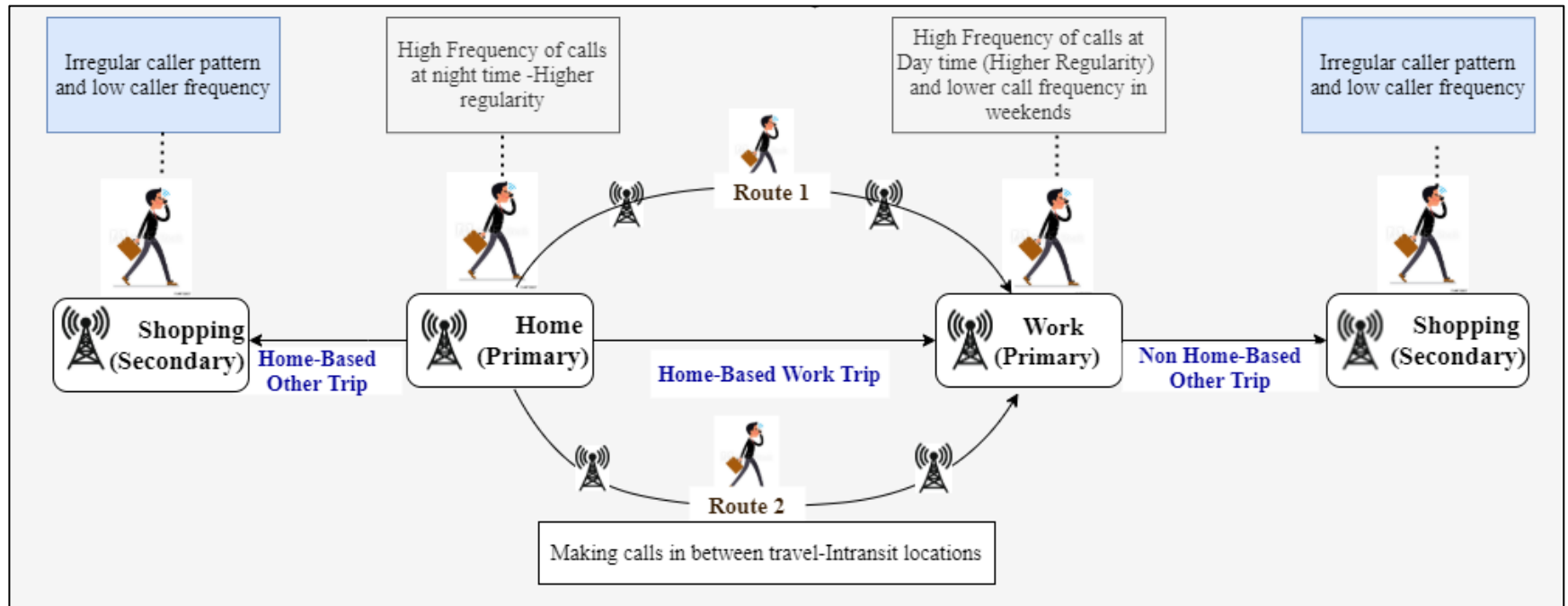


Figure 3-6

Summary of expected outputs from the study



CHAPTER 4

OD ESTIMATION FROM CDR DATA

This chapter commences with a succinct overview of the initial processing undertaken on the Call Detail Record (CDR) data. Subsequently, the descriptive statistics extracted from the CDR data are elucidated in Section 4.2, while Section 4.3 delineates the process of moment identification within the CDR data. Section 4.4 explores the behavioral representation inferred from the observed movements. The penultimate section provides a comprehensive overview of the Origin-Destination (OD) matrix estimation methodology derived from the processed CDR data. The chapter concludes in Section 4.6 with a discussion of the themes addressed throughout the chapter. It reflects on the implications of utilizing CDR data for travel demand estimation and considers how these findings might influence future research directions.

4.1 Initial Processing

The initial step in accurately inferring human behavior from Call Detail Record (CDR) data involves the elimination of noise present within the generated data. Subsequently, the data must be structured appropriately to facilitate smoother analysis. The following tasks are undertaken to render the CDR data in a format conducive to analysis:

4.1.1 Removing NA

Figure 4-1

Removing NA from the CDR data

Device Name	ID	Time Stamp	Call Durat	Cell_ID
2 MOTOROLAW156	1016940145	20130702204240	185	32867
2 MOTOROLAW156	1016940145	20130505200613	19	32867
2 MOTOROLAW156	1016940145	20130611211413	NA	32867
2 MOTOROLAW156	1016940145	20130505202015	129	32867
2 MOTOROLAW156	1016940145	20130509082321	29	32867
↓				
Device Name	ID	Time Stamp	Call Durat	Cell_ID
2 MOTOROLAW156	1016940145	20130702204240	185	32867
2 MOTOROLAW156	1016940145	20130505200613	19	32867
2 MOTOROLAW156	1016940145	20130505202015	129	32867
2 MOTOROLAW156	1016940145	20130509082321	29	32867

Non-availability, or the presence of "NA" values, is a prevalent issue in raw data inherent to both survey data and large datasets. The presence of "NA" values poses challenges in data analysis and in the interpretation of results. Consequently, an initial step in data preparation involves formatting the data to exclude rows containing "NA" values, as illustrated in Figure 4.1.

4.1.2 Splitting Time

Time is a crucial attribute that reveals subscribers' temporal behavior. While most attributes in the dataset are systematically organized, the time feature requires transformation into a more interpretable format. This process involves dividing the time attribute into manageable components, as depicted in Figure 4.2

Figure 4-2

Splitting Time component within CDR Data

ID	Time Stamp	Call Duration(Seconds)	Cell_ID	
1016940145	20130702204240	185	32867	
1016940145	20130505200613	19	32867	
1016940145	20130505202015	129	32867	
1016940145	20130509082321	29	32867	
↓				
ID	Date	Time (Hr:Min: Sc)	Call Duration(Seconds)	Cell_ID
1016940145	7/2/2013	8:42:40 PM	185	32867
1016940145	5/5/2013	8:06:13 PM	19	32867
1016940145	5/5/2013	8:20:15 PM	129	32867
1016940145	5/9/2013	8:23:21 AM	29	32867

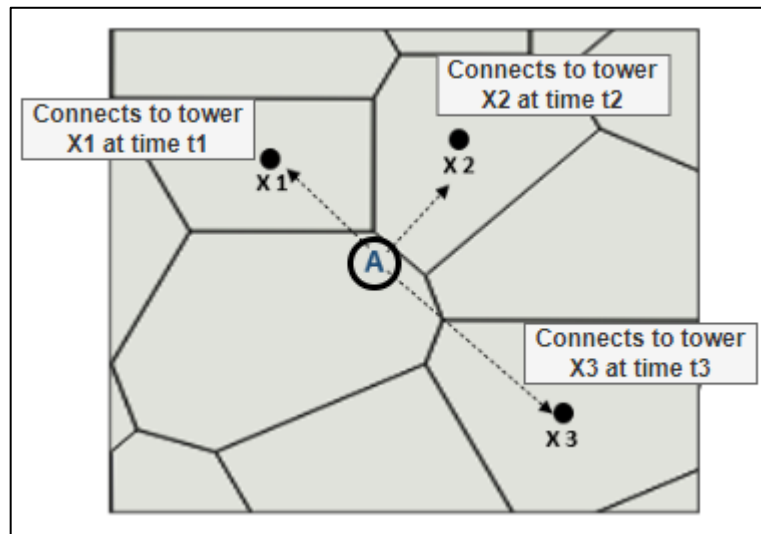
4.1.3 Load Share Effect

The phenomenon known as the load share effect represents a significant challenge in analyzing Call Detail Record (CDR) data, contributing substantially to data noise. This effect arises when mobile network operators distribute call traffic across adjacent towers to manage call volume efficiently, assigning new calls to the tower currently under less load. Consequently, the tower providing signal coverage for a user may change without the user's physical movement, creating the illusion of movement where none exists. This scenario, graphically depicted in Figure 4.3, can significantly skew transport predictions. It may lead to inflated trip counts between origins and destinations and misidentifying activity or travel behaviors.

Several studies have proposed methodologies to mitigate the impact of load sharing, employing techniques such as trajectory smoothing and spatial clustering (Csáji et al., 2013), (Jiang et al., 2013), (P. Wang et al., 2012). The approach taken in this work to address the load-sharing effect involves a speed-based filtering technique, a subset of trajectory smoothing methods (M. H. Wang et al., 2013). This method calculates the speed between two consecutive records, identifying any record exhibiting an implausible travel speed due to load sharing.

Figure 4-3

Graphical Representation of the Load Sharing Effect

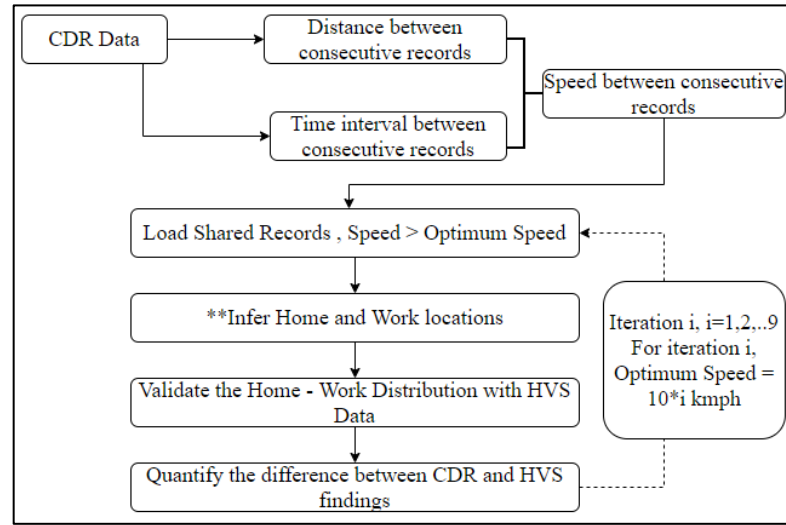


This study use the methodology depicted in figure 4.4 where, Call Detail Records (CDRs) are sequentially arranged for each subscriber, facilitating the computation of the travel speed between two consecutive logs. Speed calculation necessitates knowledge of the distance covered and the time elapsed between these records. The analysis assumes that the geographic midpoint of the cell tower's coverage area, represented by the centroid coordinates, approximates the user's location. Thus, the latitude and longitude data for tower centroids determine the distance between two successive CDR locations. After the speed calculation, records exhibiting anomalously high travel speeds are earmarked as potentially resulting from load sharing.

A subset of users with distinguishable home and workplace locations is examined to establish a benchmark for discerning load-shared records. Their movements between these points are analyzed to identify the Distance-Size Distributions (DSDs) for home and work settings. The analysis explores a range of speed thresholds, incrementing from 10 km/h to 90 km/h in steps of 10 km/h, to isolate records potentially affected by load sharing. At each threshold, the method identifies the users' work and home locations anew, comparing these locations with those documented in the Household Travel Survey (HVS) data. This comparative analysis, conducted at the DSD level, evaluates the concordance between the location data derived from CDRs and the empirical data obtained from the HVS, thereby facilitating a refined approach to mitigating the impact of load sharing on the analysis of CDR-based mobility patterns.

Figure 4-4

Methodology used within the study for Load Share Identification



In evaluating the findings, both percentages and the Kullback–Leibler divergence, a metric quantifying the deviation of one probability distribution from another (equation 2) are employed to ascertain the disparities between distributions derived from CDR and HVS data across various speed thresholds. The optimal speed threshold for mitigating the impact of load sharing in CDR data is the speed that minimizes the discrepancy percentage. Utilizing the Kullback–Leibler divergence method, for two discrete probability distributions, P and Q , defined over the same probability space, R , the divergence from Q to P is calculated to quantify the relative difference between these distributions.

$$(P||Q) = \sum_{x \in R} P(x) \log(p(x)/Q(x)) \text{ ----- (2)}$$

Figure 4-5

Discrepancy Analysis Between CDR and HVS Distributions Across Speed Thresholds

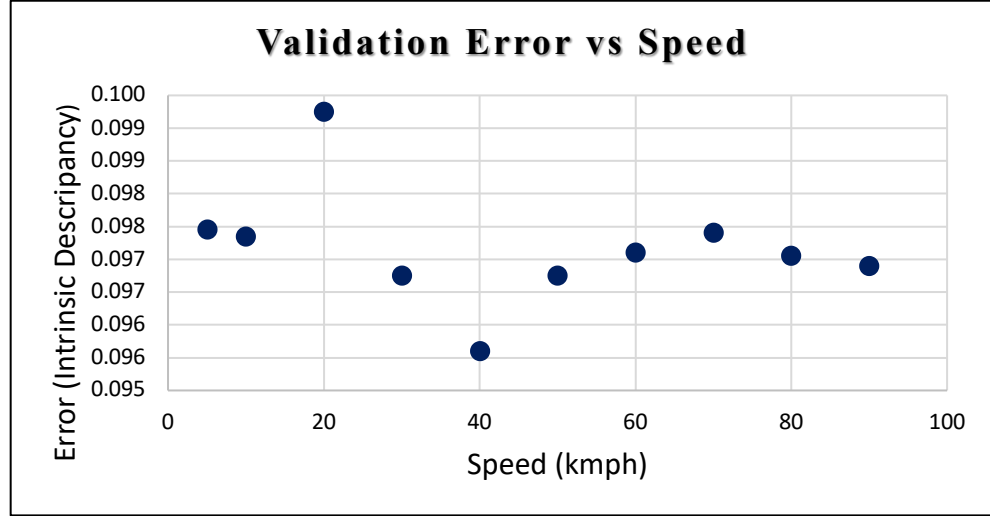


Table 4-1

Sample Representation of Load Shared Records

User ID	Cell ID	Time of Caller Activity	Speed between consecutive records (Kmph)	No. of days appeared in the Cell ID	Replaced Cell ID
5055875	21373	5/1/13 17:32	1.32	65	21373
5055875	23379	5/1/13 17:34	245.75 > 40Kmph	28	21373

The discrepancy was minimal at a 40 km/h threshold for the sample under consideration, as depicted in Figure 4.5. Consequently, records exhibiting consecutive speeds exceeding 40 km/h were classified as load-shared, as illustrated in table 4.1. These accounted for approximately 11% of the total records analyzed.

4.1.4 Spatial Mapping

Spatial mapping significantly contributes to understanding travel demand by shedding light on users' spatial behaviors. When a user places a call, the nearest cell tower receives the signal and logs the location data. This data encompasses geospatial coordinates within the WGS 84 / UTM zone 44N coordinate reference system, notably latitudes, and longitudes. This system is uniformly applied throughout this thesis to ensure consistency in spatial data analysis.

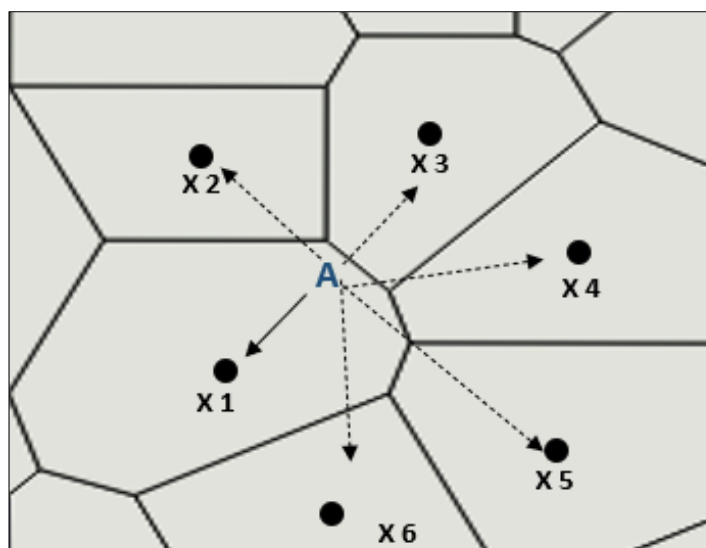
The technique of Voronoi tessellation is utilized to estimate the coverage areas of cell sites, drawing on the geospatial coordinates and network types of cell towers (Doyle et al., 2014).). It is crucial, however, to recognize that several factors can influence the accuracy of this tessellation method in reflecting accurate cell coverage. These include the characteristics of the communication channel, the area's topography, and the cell tower's physical parameters, such as the frequency, tilt, height, and transmission power (Doyle et al., 2014). The level of spatial accuracy in pinpointing the location of mobile device users is directly related to the coverage area of a network's antenna (Khan et al., 2015).

For instance, figure 4.6 illustrates the case of User A, who has the potential to connect to any nearby tower. If User A initiates a call at 9:30 pm on May 23, 2013, from the location shown, they will connect to a specific tower during the call. In this scenario, towers labeled X1 through X5 represent different cell towers, with their coverage areas delineated accordingly, positioning each tower at the center of its respective area. If User A connects to Tower X1, we can deduce that User A was within the coverage area of Tower X1 at the time of the call.

Cell towers or user locations can be associated with existing administrative zones for a deeper analysis. This association is essential for interpreting and analyzing the data, facilitating a comprehensive understanding of spatial patterns and user behavior.

Figure 4-6

User Connectivity among Cell Towers



Within Sri Lanka, specifically in the Western Province, there are 763 cell towers. By overlaying the spatial layer of these cell towers with the Divisional Secretariat Division (DSD) and District spatial layers, it is possible to accurately identify the corresponding DSD and District for each cell tower. Table 4.1 provides a detailed enumeration of cell towers across each administrative zone. Furthermore, the analysis reveals that the average coverage area of these towers varies significantly, ranging from 2.37 square kilometers to 18.37 square kilometers, depending on the district. This variation underscores the diversity in cell tower distribution and coverage across different geographic areas within the province.

Table 4-2

Cell tower coverage at district level of Sri Lanka

District	No. of cell towers	Average coverage area (sq km)
Colombo	295	2.37
Gampaha	253	5.48
Kalutara	87	18.37

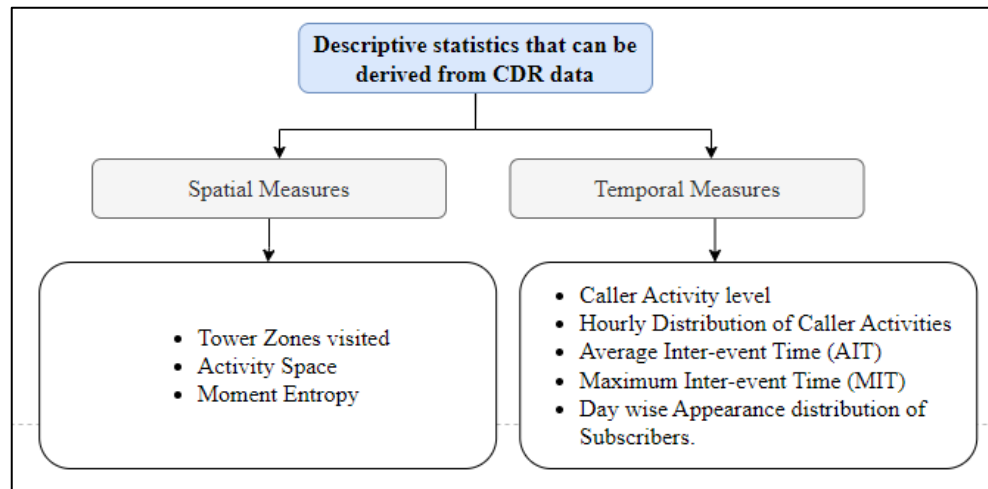
Following the initial data processing phase, the subsequent step involves identifying potential, measurable spatiotemporal attributes within the processed data. The forthcoming section of this chapter will concentrate on this specific aspect.

4.2 Descriptive Statistics

As discussed in earlier chapters, Call Detail Records (CDR) encompass a broad spectrum of data about subscriber behavior. This data can primarily be classified into two categories: temporal measures and spatial measures, as illustrated in Figure 4.7

Figure 4-7

Summary on Spatio-temporal measures derived from CDR Data



4.2.1 Temporal Measures

Temporal measurements quantify the durations of and intervals between events, providing a discrete resolution of measurements about time. Time is crucial in evaluating human behavior, particularly in understanding travel patterns. The frequency and timing of travel, be it daily, weekly, hourly, or at specific times of the day, are valuable insights derived from the Call Detail Record (CDR) data analysis. Key temporal outputs generated from this analysis include caller activity levels, the hourly distribution of caller activities, average and maximum inter-event times, and day-wise subscriber appearances. These metrics offer a detailed perspective on the temporal dynamics of subscriber behavior, facilitating a deeper understanding of movement patterns and preferences.

Caller Activity level

The caller activity level quantifies the number of events (phone call records) observed for each user over one month. Figure 4.8 depicts the frequency distribution of these events, indicating the number of events recorded per user. Subsequently, Figure 4.9 demonstrates the cumulative distribution function (CDF) of observations per user.

From the analyzed sample, it is evident that approximately 70% of users have fewer than 30 events recorded within the period, and around 10% have fewer than five records. Analyzing the average daily caller activities per user (as shown in Figure 4.9), it emerges that about 95% of users experience less than one event per day. In contrast, a smaller fraction, 10%, has fewer than 20 daily records. The infrequent activity observed for certain users could be attributed to various factors, such as changing their SIM card during the observation period, using another SIM card from a different mobile operator, or leaving the study area. Conversely, a minimal proportion of subscribers (0.07%) exhibit an exceptionally high number of observations (exceeding 100), potentially due to device anomalies or the use of a communication device in a specific location, such as a shop.

To produce reliable results and mitigate the impact of outliers, certain measures were implemented. Specifically, the analysis excluded the top 5% of users with the highest daily observation rates and the bottom 5% with the lowest rates. Moreover, users with a record count ranging between 2 and 845 during the observation period were included for further analysis. This approach ensures a balanced representation of user activities while minimizing the influence of extreme values on the overall analysis.

Figure 4-8

Caller Activity Distribution of CDR Sample

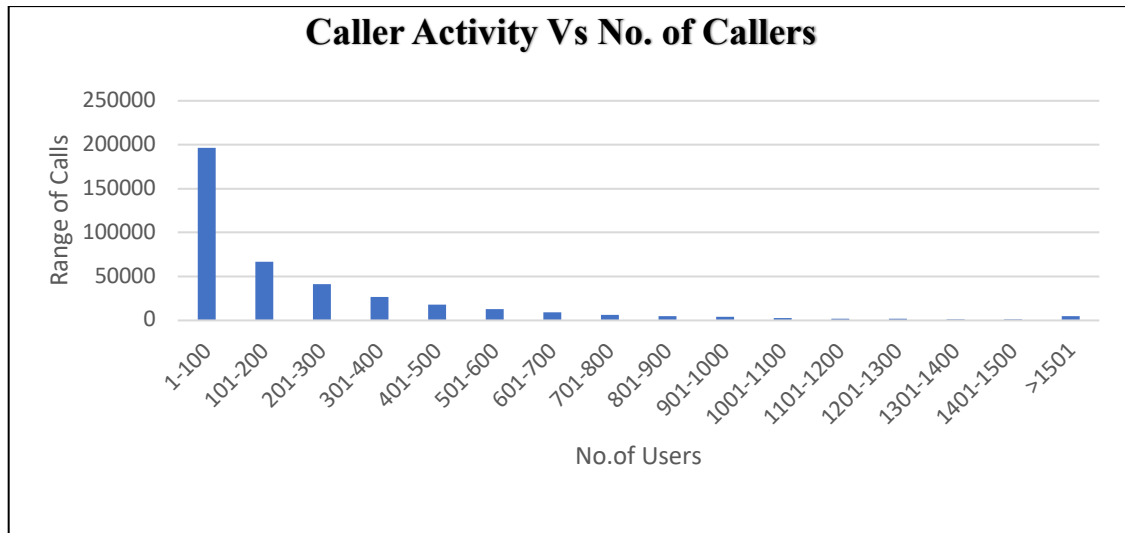
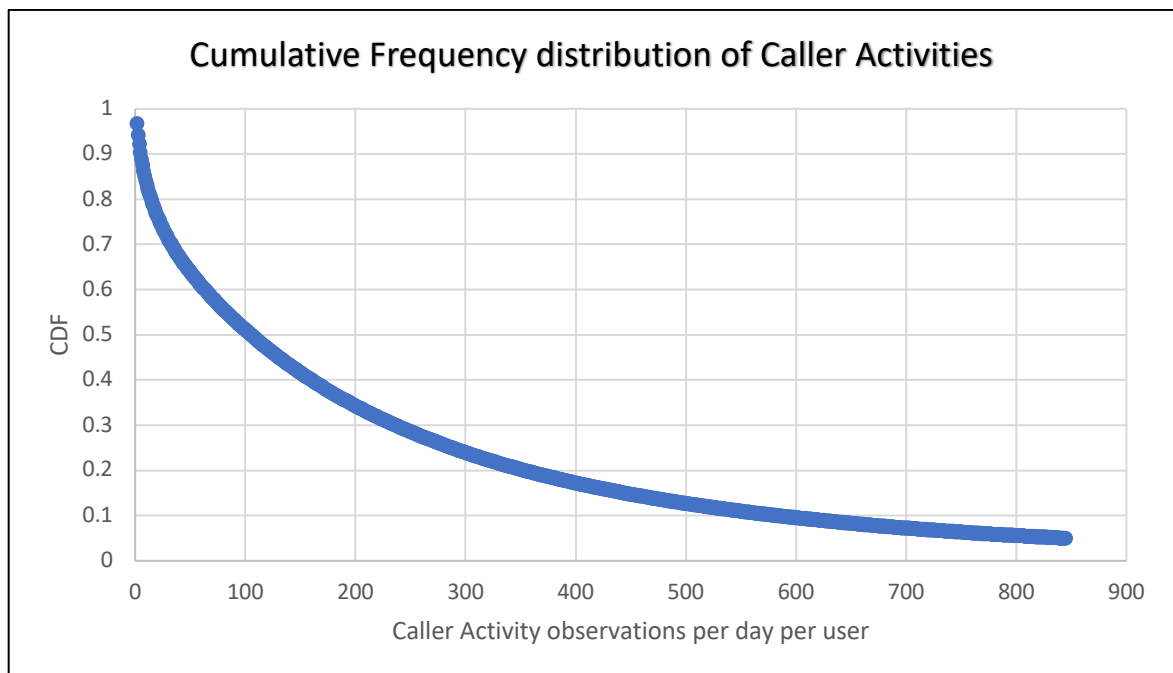


Figure 4-9

CDF of Caller Activities



Hourly Distribution of Caller Activities

Figure 4-10

Hourly Distribution of Caller Activities of the CDR Sample

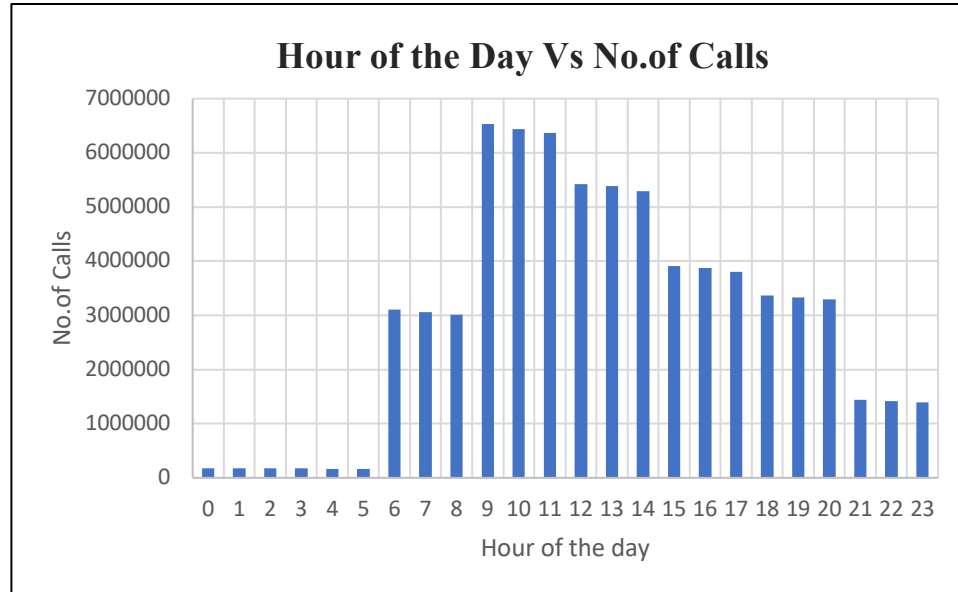


Figure 4.10 above illustrates the hourly distribution of caller activities throughout the day. It is observed that caller activities are significantly higher during the 6 am-8 pm time window compared to other periods. This pattern can be attributed to the fact that the 9 pm-5 am window encompasses sleeping or late-night hours, while the 6 am-8 pm interval aligns with working or highly active hours. Consequently, the contributions to caller activities during the 9 pm-5 am period are primarily from individuals who work night shifts or are employed in nighttime roles. This distinction highlights the relationship between human activity patterns and the temporal distribution of caller activities.

Average Inter-event Time

This measure is crucial when analyzing individual temporal data, offering an overview of the average time between a subscriber's consecutive records. As depicted in Figure 4.11, the Average Inter-event Time (AIT) spans from 3 seconds to 711 hours, with the mean being approximately 26 hours. Figure 4.12 presents AIT's cumulative distribution function (CDF), indicating that 93% of subscribers have an AIT of less than 1 hour. Meanwhile, the 1% of devices classified as hyperactive exhibit a significantly lower AIT, around 1 minute. This

analysis provides valuable insights into the frequency and timing of subscriber activities, highlighting the variability in user engagement with the network.

Figure 4-11

Average Inter-Event Time of the CDR Sample

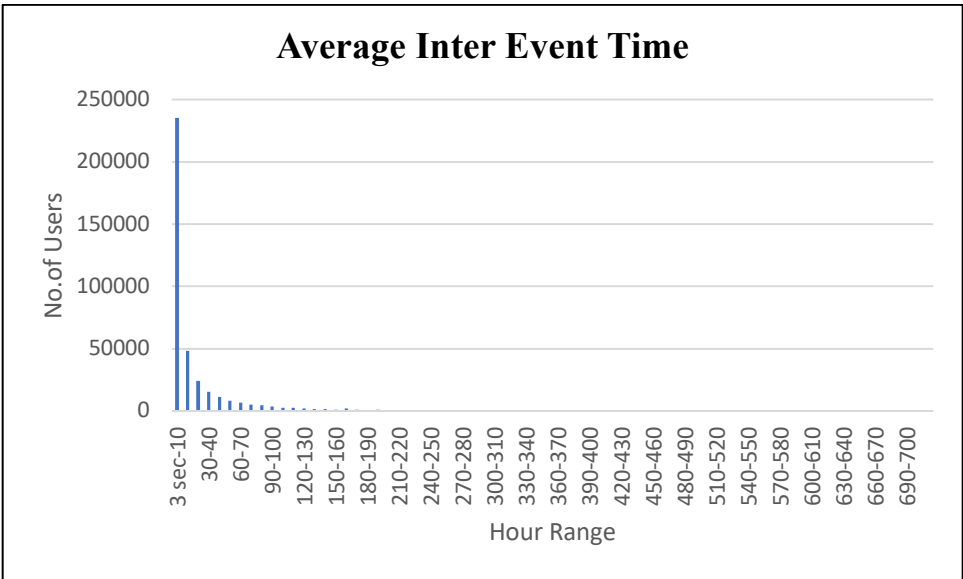
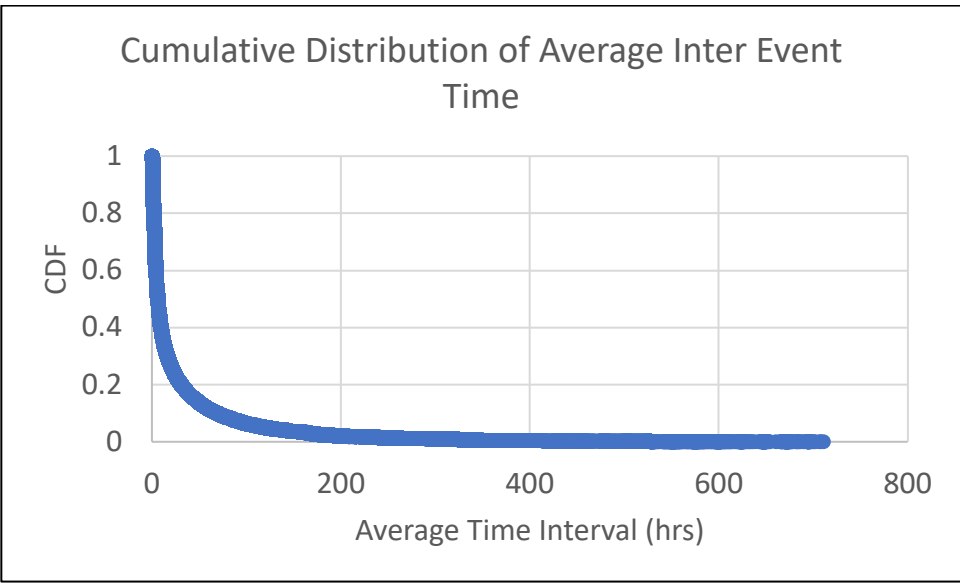


Figure 4-12

Cumulative Distribution of AIT



Maximum inter-event time (MIT)

Interpreting user behavior solely based on Average Inter-event Time (AIT) may lead to inaccuracies, particularly because device activity tends to be lower during certain times of the day. To mitigate this issue, it is recommended to focus on a highly active time window, such as 6 am to 9 pm, as identified in Figure 4.13. This approach ensures that periods of minimal activity, like late night to early morning, are excluded from the analysis. Consequently, the Maximum Inter-event Time (MIT) between consecutive records for each subscriber was analyzed to construct the subsequent graphs. The analysis reveals that approximately 70% of users have an MIT of less than three days. This adjusted methodology provides a more accurate reflection of user engagement by concentrating on peak activity periods.

Figure 4-13

Maximum Inter Event Time Distribution of the CDR Sample

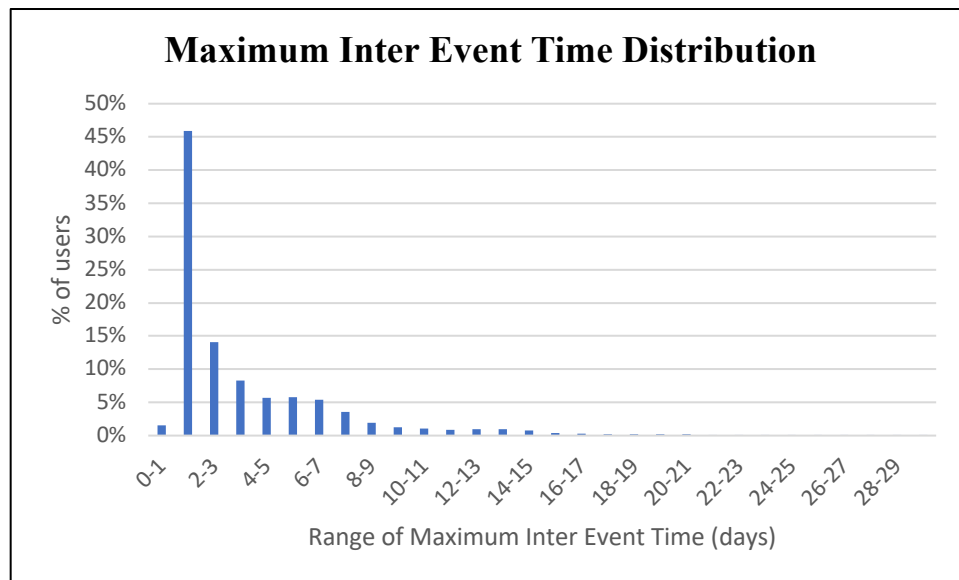
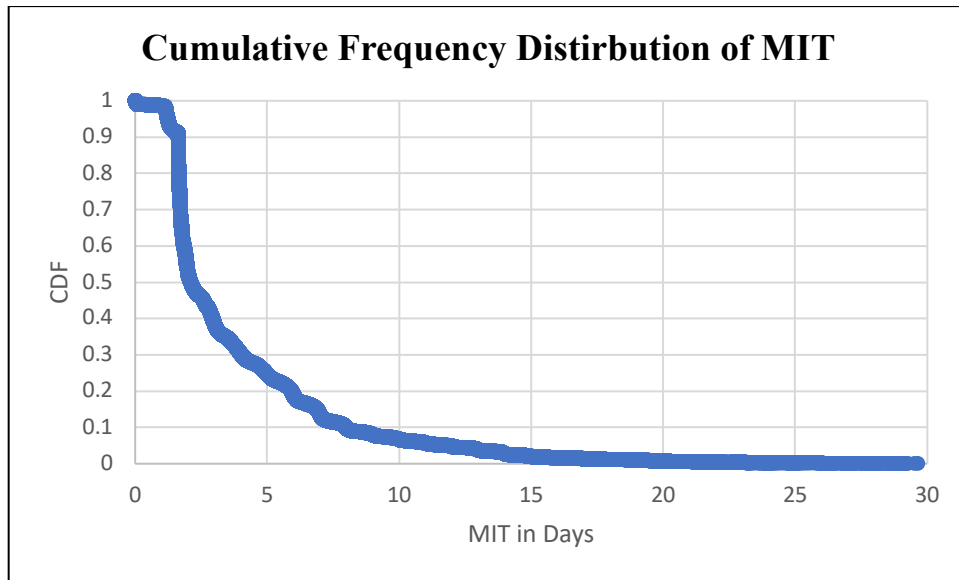


Figure 4-14

CDF of MIT



Day-wise Appearance distribution of Subscribers.

This measure is essential for examining the regularity of subscriber behavior. It offers an overview of the number of days within the observed period during which users have engaged in caller activities. As illustrated in Figure 4.14, 4% of users had records for only a single day, whereas 6% of subscribers made calls on all days under consideration. Figure 4.15 displays the cumulative probability distribution of this measure, providing insights into the distribution of user activity across the observed timeframe. This analysis helps understand the consistency of subscriber engagement with the network.

Figure 4-15

Day-wise Caller Activity Distribution of the CDR Sample

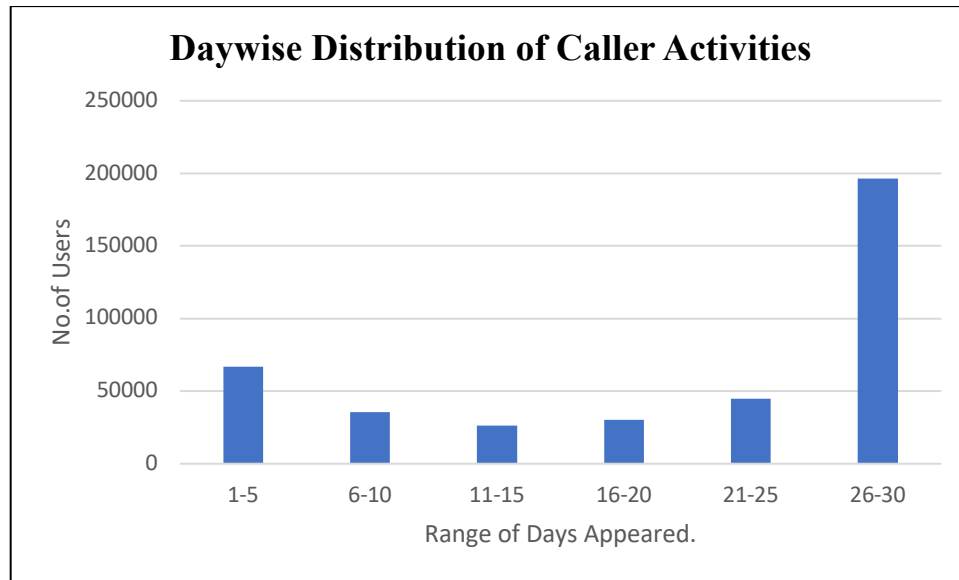
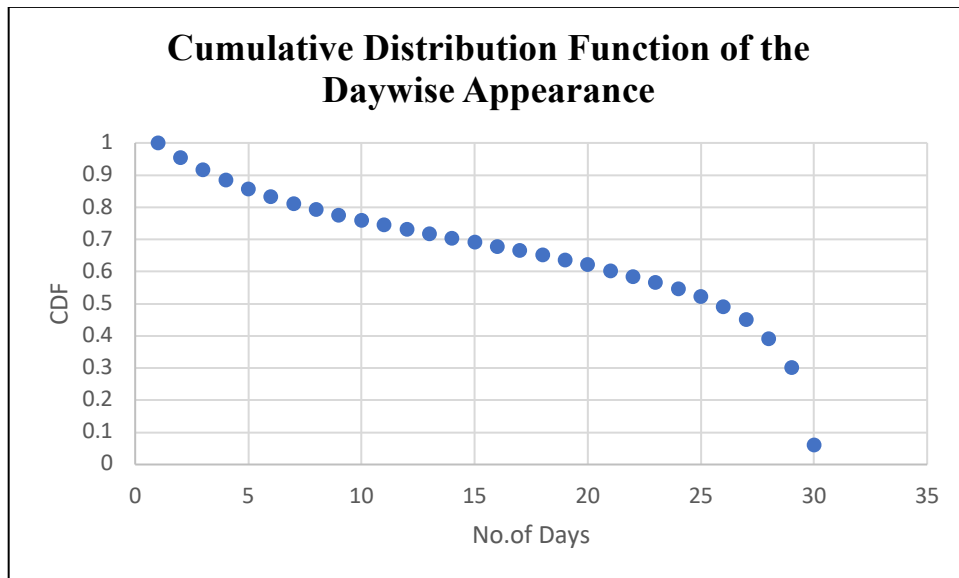


Figure 4-16

CDF of Day wise caller activity distribution



4.2.2 Spatial Measurements

Spatial measurements play a pivotal role in quantifying the extent and distances between objects, offering insights into the geographical positioning of subscribers from a human behavioral perspective. Beyond time, location serves as a fundamental pillar in modeling human behavior. From the Call Detail Records (CDR), various spatial attributes have been identified, which are crucial for understanding patterns of movement and interaction. Two key spatial concepts, Activity Space and Entropy, emerge as significant in this context and will be discussed in detail. These concepts help capture the complexity and diversity of subscriber movements, providing a deeper understanding of spatial behavior within the framework of human activity.

Distribution of the Tower Zones Visited.

This metric is pivotal for assessing the mobility of subscribers. As previously mentioned, the study area encompasses 763 tower zones. The range of tower zones users visit varies from 1 to 142, clearly indicating subscriber mobility. Subscribers who have visited fewer tower zones demonstrate lower mobility or contribute minimally to overall trip-making, whereas those visiting a larger number of zones contribute significantly to travel demand. Figure 4.17 highlights that 8% of users had visited only a single cell tower zone. Figure 4.18 presents the cumulative probability distribution, revealing that 99% of users had visited fewer than 55 cell zones. Furthermore, the concept of tower zones visited is expanded upon through measures such as Entropy and Activity Space, which will be elaborated on in the subsequent section. This analysis underscores the importance of spatial mobility patterns in understanding and modeling human behavior within the context of the study area.

Figure 4-17

Cell Appearance Distribution of the CDR Sample

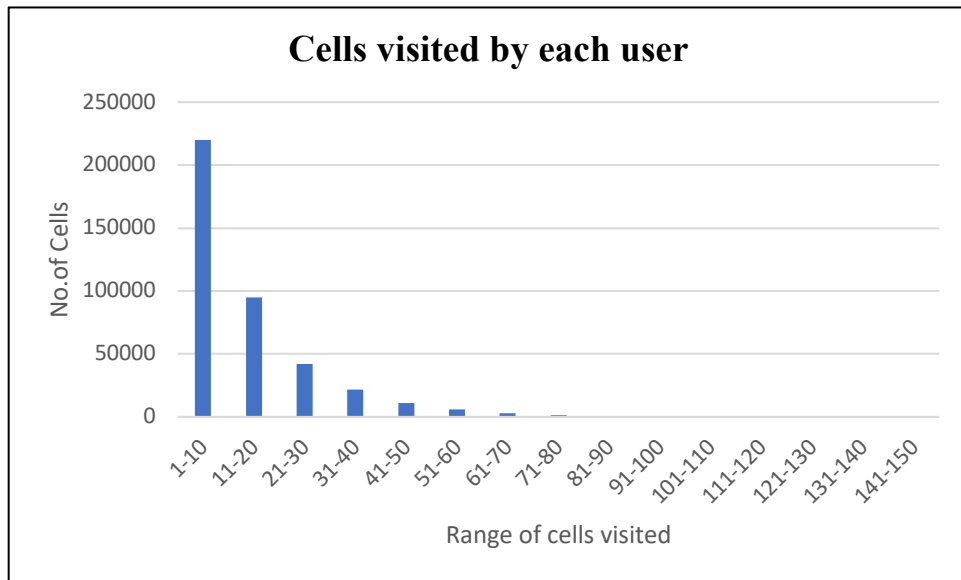
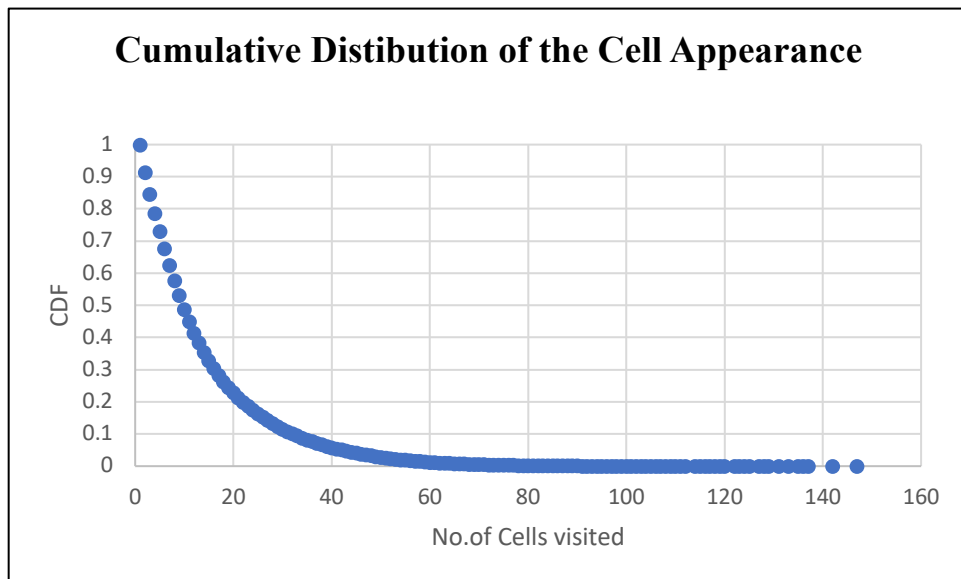


Figure 4-18

Cumulative Distribution of the cells visited



Activity Space

Activity space is defined as the geographical area encompassing the locations where individuals carry out their daily routines, encompassing work, leisure, and other activities. Understanding the concept of activity space is vital in geography and urban studies for analyzing spatial behavior and mobility patterns. A key metric used to quantify activity space is the radius of gyration.

The radius of gyration is a statistical measure that quantifies the spatial dispersion of visited locations around a central point. It is calculated as the root mean square of the distances from each visited location to the center of mass of these locations. Typically, the center of mass is considered a significant anchor point in an individual's daily life, such as the home or workplace, due to its central role in structuring daily activities.

In this study, the center of mass for each user is determined based on the tower zones they have visited. The radius of gyration is subsequently calculated by assessing the distances between these tower zones and the identified center of mass. This calculation offers valuable insights into the scale and distribution of individuals' activity spaces, highlighting the diversity in movement patterns and the range of locations that constitute their daily spatial interactions. The methodology for calculating the Radius of Gyration is detailed in Equation 10, providing a mathematical framework to quantify these spatial dynamics.

Where,

R_i = Position of the user at time i

R_{cm} = Center of mass for all the temporally located locations for the user

N = Total number of distinct locations

$$R_G = \sqrt{\frac{1}{N} \sum_{i=1}^N (R_i - R_{cm})^2} \text{ -----(3)}$$

Figures 4.20 and 4.21 provide illustrative examples of two subscribers, highlighting differences in their activity spaces. Despite visiting a similar number of locations, the

subscriber depicted in Figure 4.19 exhibits a smaller activity space when compared to the subscriber in Figure 4.20. This difference is attributed to the geographical spread of the visited locations; the locations visited by the subscriber in Figure 4.20 are more dispersed, indicating a larger activity space. This comparison underscores how the radius of gyration considers the number of visited locations and their spatial distribution, offering a nuanced understanding of individual mobility and spatial behavior.

Figure 4-19

Spatial Activity Space of user 1 as derived from CDR data

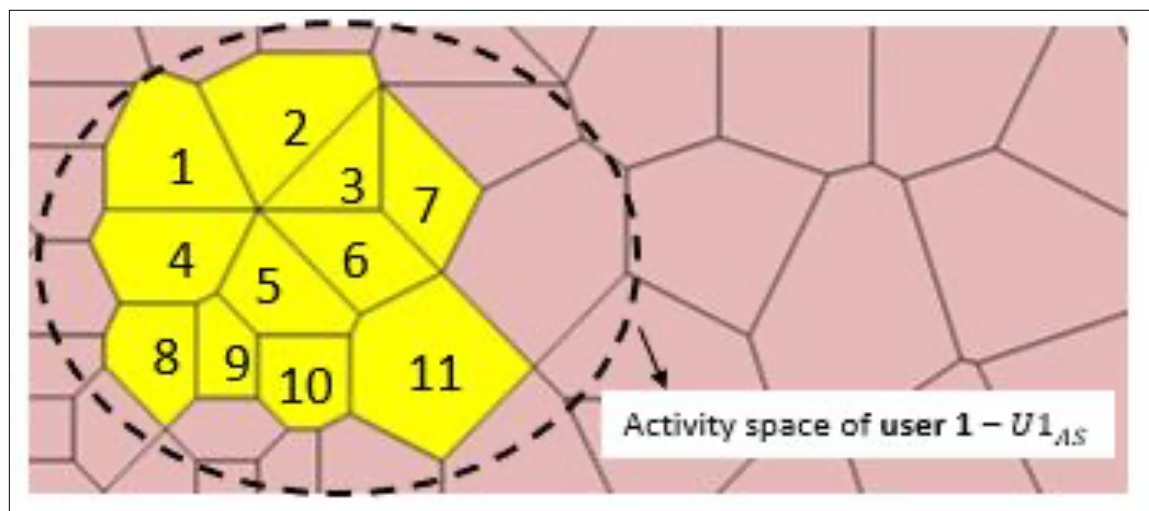
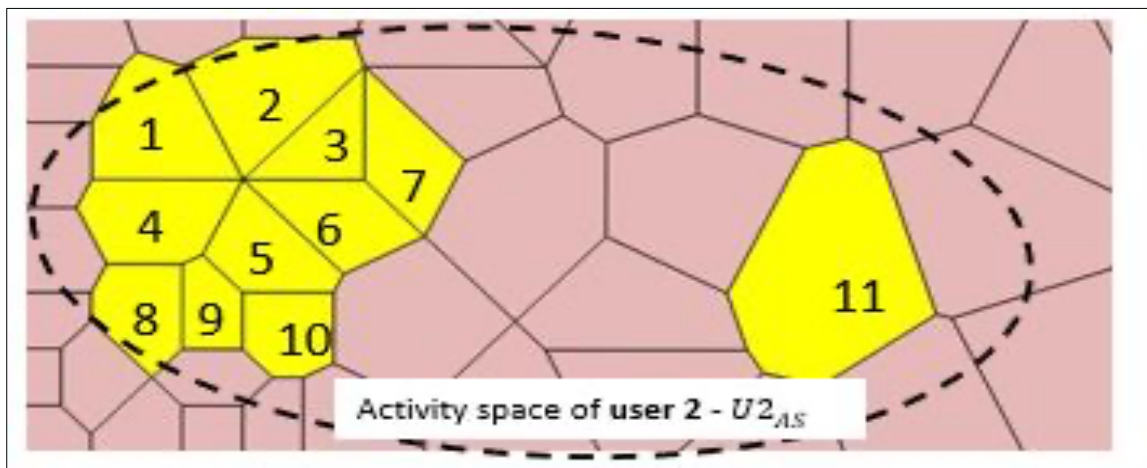


Figure 4-20

Spatial Activity space of user 2 as derived from CDR data



As demonstrated in the graph presented in Figure 4.21 the Activity Space of the studied subscribers ranges from 0 to 55 kilometers. Furthermore, the cumulative probability distribution depicted in Figure 4.22 reveals that approximately 99% of users have a radius of gyration measuring less than 23 kilometers. Conversely, a marginal 1% of users exhibit a notably smaller radius of gyration, approximately 0.02 kilometers. This data provides insight into the spatial mobility patterns of the population, indicating a significant concentration of activities within relatively confined areas for the vast majority. In contrast, a small fraction of the population exhibits extremely localized movement. patterns.

Figure 4-21

Activity Space Distribution derived from CDR data

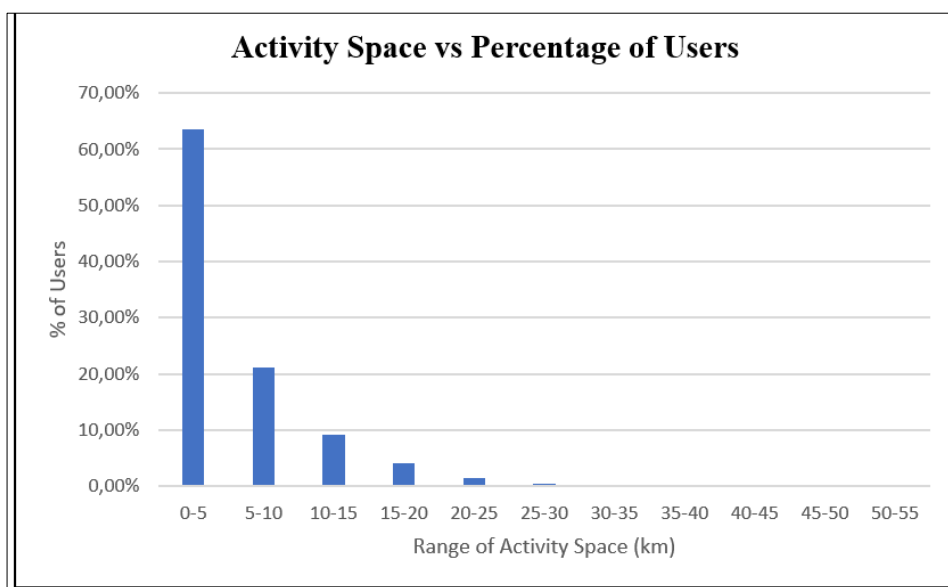
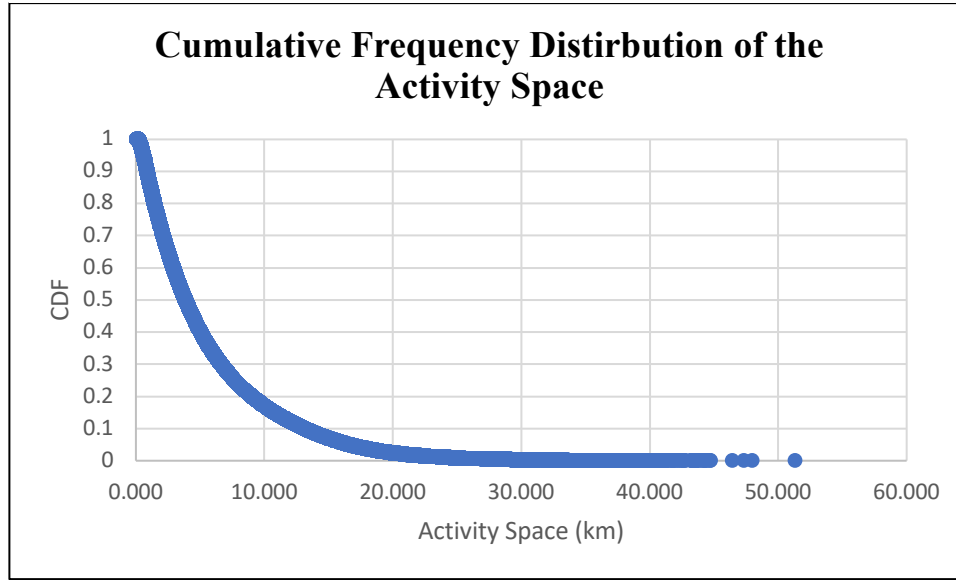


Figure 4-22

CFD of Activity Space



Movement Entropy

Movement Entropy (E) is generally considered a measure of the heterogeneity of visitation patterns (Zhao et al., 2016). It can be calculated using the equation below.

$$E = - \sum_{i=1}^n P_i \log_2 P_i \text{----- (4)}$$

Where,

n = Number of distinct locations visited / Number of cell tower locations visited by the subscriber

P_i = Probability that location i is visited by the subscriber

Figures 4.24 and 4.25 display the distribution of entropy values and the cumulative distribution of moment entropy, respectively. Analysis of these figures indicates that 99% of users have an entropy value of less than 0.28, while a small fraction, about 1%, exhibit a much lower entropy measure of approximately 0.002. These findings suggest a predominance of relatively predictable visitation patterns among most users, with a minimal proportion displaying highly predictable patterns.

Table 4.2 further elucidates these findings, demonstrating that entropy values tend to increase with the heterogeneity of visitation patterns. This is visually represented through the gradation of colors in each row, where darker shades correspond to higher levels of heterogeneity. This pattern underscores the relationship between entropy and the diversity of an individual's visitation behavior, reinforcing that greater unpredictability in location visits is associated with higher entropy values.

Table 4-3

Sample Entropy calculation

No. of locations and frequency in each location	Calculation	Result
Location A, probability 1	$(-1 * \log_2 1)$	0
Location A visited one time, and Location B visited four times	$(-0.2 * \log_2 0.2) + (-0.8 * \log_2 0.8)$	0.72
Location A visited five times, and Location B visited five times	$(-0.5 * \log_2 0.5) + (-0.5 * \log_2 0.5)$	1
A, B, C, and D locations were visited two times each	$(-0.25 * \log_2 0.25) + (-0.25 * \log_2 0.25) + (-0.25 * \log_2 0.25) + (-0.25 * \log_2 0.25)$	2

In conclusion, integrating temporal and spatial measurements provides a comprehensive understanding of user behavior. The forthcoming section of this thesis will delve into the representation of user behavior, leveraging the processed data to elucidate patterns and insights. This analysis aims to further our comprehension of how individuals interact with their environment over time and space, contributing significantly to the overarching goals of the study.

Figure 4-23

Entropy Values distribution derived from CDR data

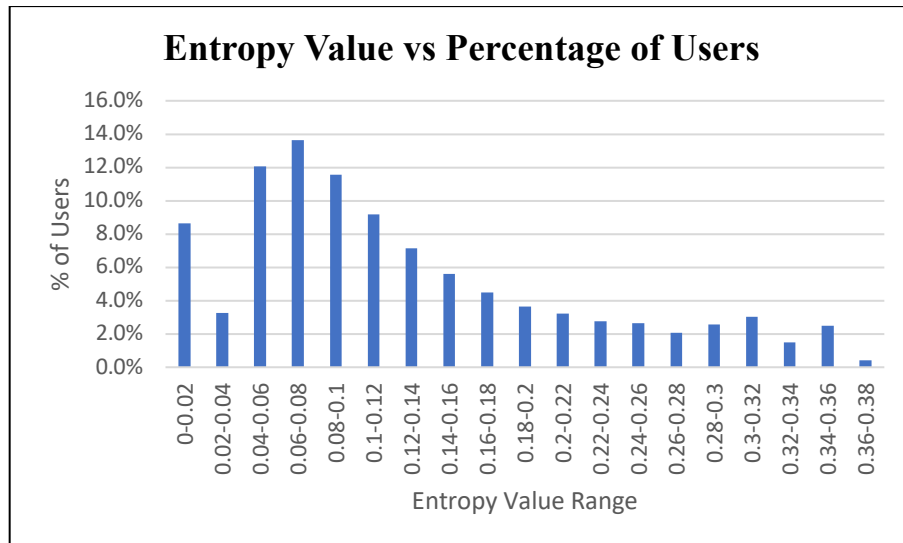
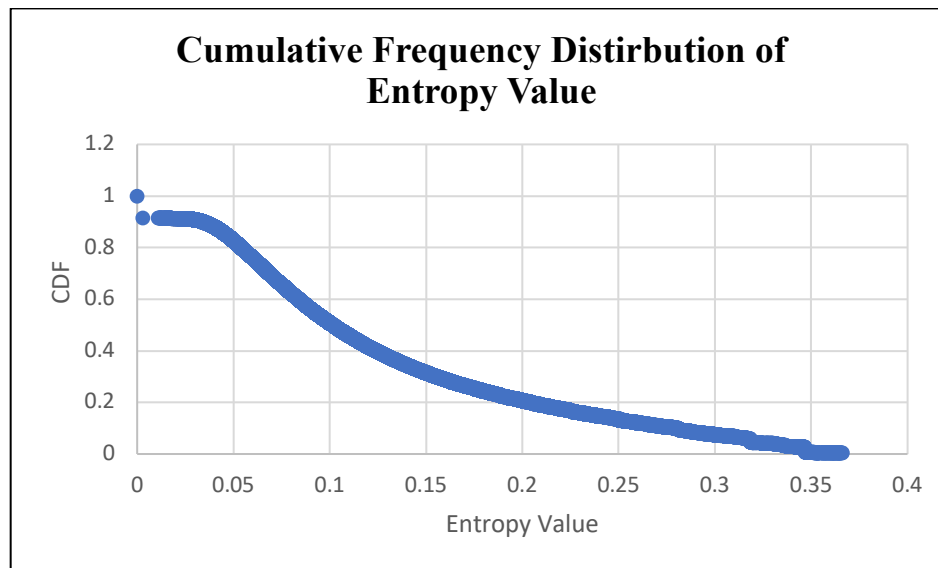


Figure 4-24

CDF of Entropy Value



4.3 Identifying Movements from CDR Data

Human activity and travel behavior encapsulate movement and stationary periods within spatial and temporal contexts. This behavior for an individual may manifest as either transit, marked by frequent changes in the connected cell tower reflecting movement, or as stationary periods, during which the connection to a specific cell tower does not change over time. Call Detail Records (CDR) data, providing spatial location information at designated temporal intervals, enables the mapping of user behavior through spatiotemporal connectivity analysis. By reviewing CDR tables and selecting entries chronologically aligned with specific cell tower connections, it is possible to extract data on individual cell tower activities. These entries, indicative of cell tower connections, are amalgamated with geographical information of the cell towers to construct a spatio-temporal cell activity matrix.

This process involves organizing cell tower activity data in a temporal sequence and integrating it with spatial data to produce a spatio-temporal matrix. This matrix offers an exhaustive depiction of user behavior, highlighting spatial movements and temporal engagements.

The foundation of this analysis is the Caller Activity Matrix, which estimates user locations based on their call activities within a ten-minute frequency. This matrix logs the cumulative number of calls made by users. Subsequently, the Spatial Appearance Matrix is derived from the Caller Activity Matrix. However, it only records users' presence (denoted by 1) or absence (denoted by 0) in the respective spatiotemporal cells based exclusively on on-call activities. Given CDR data's sparse and irregular nature, characterized by significant intervals between user movements, a consistent location is presumed for users until a subsequent call activity is recorded, leading to the formation of the Appearance Interpolated Matrix. This approach facilitates a more continuous and representative depiction of user behavior across spatial and temporal dimensions, notwithstanding the inherent limitations of the data. For example, assume that user i makes a call by connecting to cell tower C_z at time T_j on day D_k . Then the Appearance (A) can be depicted as,

$$F(A) \sim U_i C_z T_j D_k \text{ ----- (5)}$$

- i = No. of users (n)
- z = Total no. of cell tower areas (m)
- j = Time windows with a frequency of 10 mins (1, 2, ...144)
- x = Total no. of days of the selected sample (p)

Assume that the user i makes the next call at time T_{j+5} by connecting to cell tower C_{z+1} . Then, the cell tower at T_{j+1} , T_{j+2} , T_{j+3} , and T_{j+4} equals C_z .

If the first appearance is A_1 , and the second appearance is A_2

$$A_1 \sim U_i C_z T_j D_x$$

$$A_2 \sim U_i C_{z+1} T_{j+1} D_x$$

If $z \neq z+1$ when $j \approx j+1$, then $A_1 \neq A_2$ (TR=Trip)

If $A_1 \neq A_2 \approx \text{TR}=1$

$$A_2 \neq A_3 \approx \text{TR}=2$$

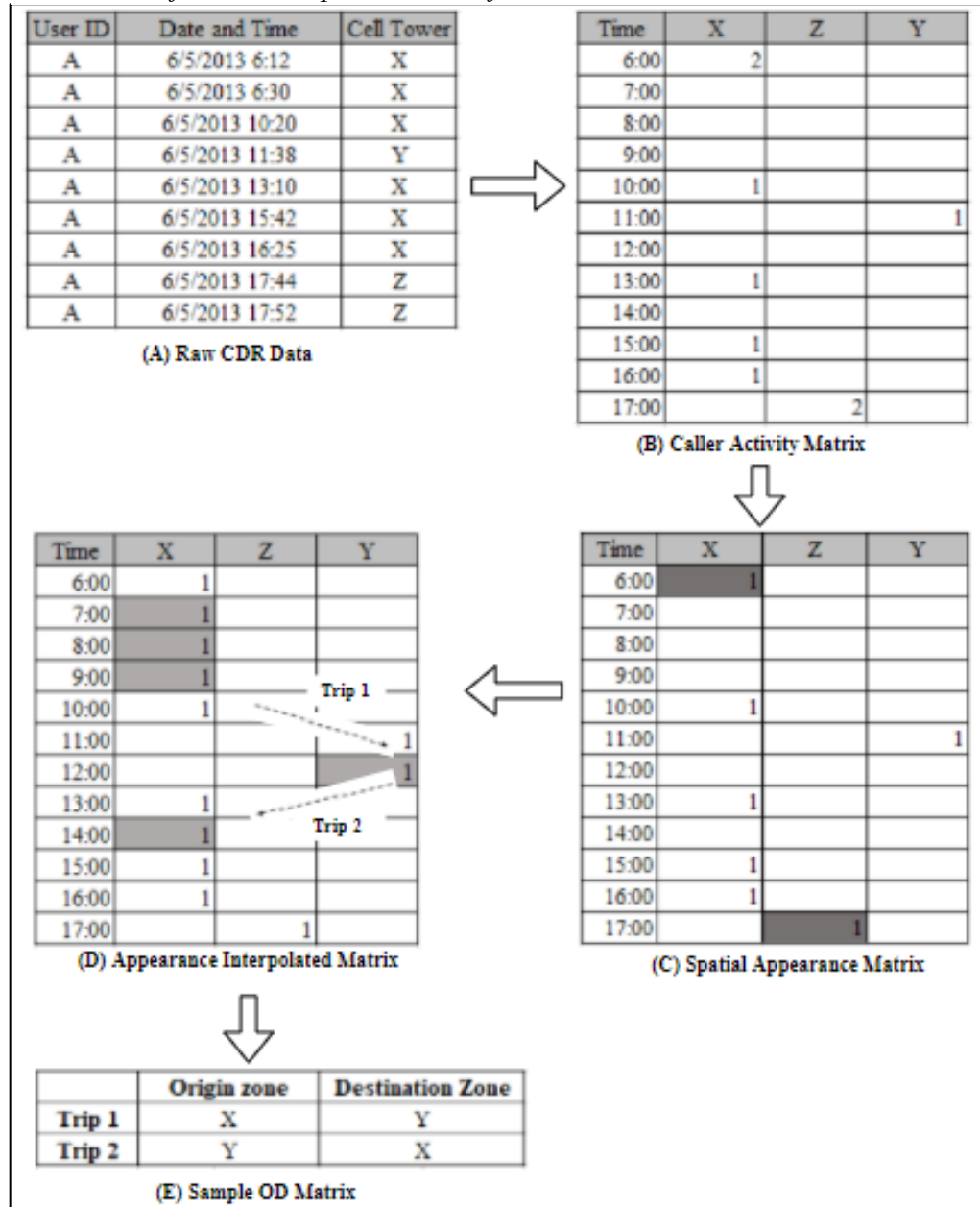
$$A_3 \neq A_4 \approx \text{TR}=3$$

$\vdots$ $\vdots$
 $\vdots$ $\vdots$

$A_{n-1} \neq A_n \approx \text{TR} = \sum_{i=1}^n \text{TR}i$ (n = Appearance of the final user in the sample at the last time window of the last day of the considered period)

Figure 4-25

Generation of Time-Stamped Matrices from CDR data



In the analysis of user appearances, a trip is discerned when there is a change in location from one appearance (A_i) to the subsequent one (A_{i+1}). This shift indicates a relocation and thus signifies a trip. For instance, Figure 4.25 displays a portion of an appearance interpolated matrix, depicting a scenario where a user remains at location X for a duration of 3 hours, subsequently transitions through various locations between X and Y, and ultimately settles for an extended period at location Y. Although the Spatio-Temporal matrix for each day per user offers a detailed portrayal of individual behavior, a more sophisticated approach is required to accurately identify trips on an aggregate level, considering the frequency of calls and regularity over time. Since users do not initiate calls at every juncture of their travel and not all changes in cell towers necessarily equate to distinct trips, merely consolidating Spatio-Temporal matrices for each day per user could result in significant inaccuracies in trip estimations.

To mitigate this, it is imperative to aggregate Spatio-Temporal matrices for each user across days, generating a consolidated Spatio-Temporal matrix per user. This aggregation process incorporates the concept of regularity, which is elaborately discussed in Section 5.5 of the study. It elaborates on the methodology employed to approximate travel demand by leveraging regularity in replicating the travel demand modeling process. The subsequent section of the study is dedicated to presenting various behavioral models that draw upon the Spatio-Temporal behaviors of users, utilizing these meticulously aggregated matrices for a nuanced understanding of travel patterns and demand.

4.4 Behavioral Representation of the Subscribers.

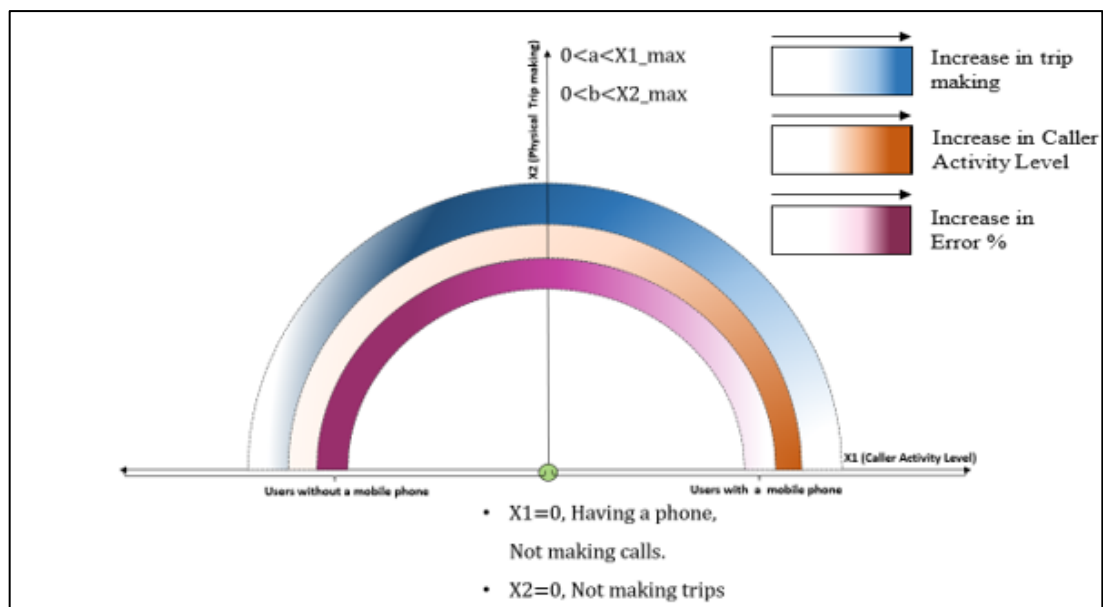
This section examines the impact of caller activity levels on trip-making, how they influence behavior, and the estimation of overall travel demand. Six distinct behavioral models were identified (figure 4.26-4.31), each represented on a two-dimensional plane. This plane uses one axis to depict trip-making and another to represent the caller's activity level. To illustrate variations in caller activity levels, trip making, and their respective contributions to errors in travel demand modeling, a color-coding scheme is employed, as detailed in the table key. The intensity of the color in each category increases progressively to indicate an escalation in each of the criteria above. Samples with a negative value on the X-axis are categorized as

individuals without mobile phones, highlighting a distinct behavioral model within travel demand estimation. Further users can be divided into different employment or socio-economic baskets considering the behavior as in table 4.3 and figure 4.32

Category 1

Figure 4-26

User Behavior Model 1 Based on CDR Data



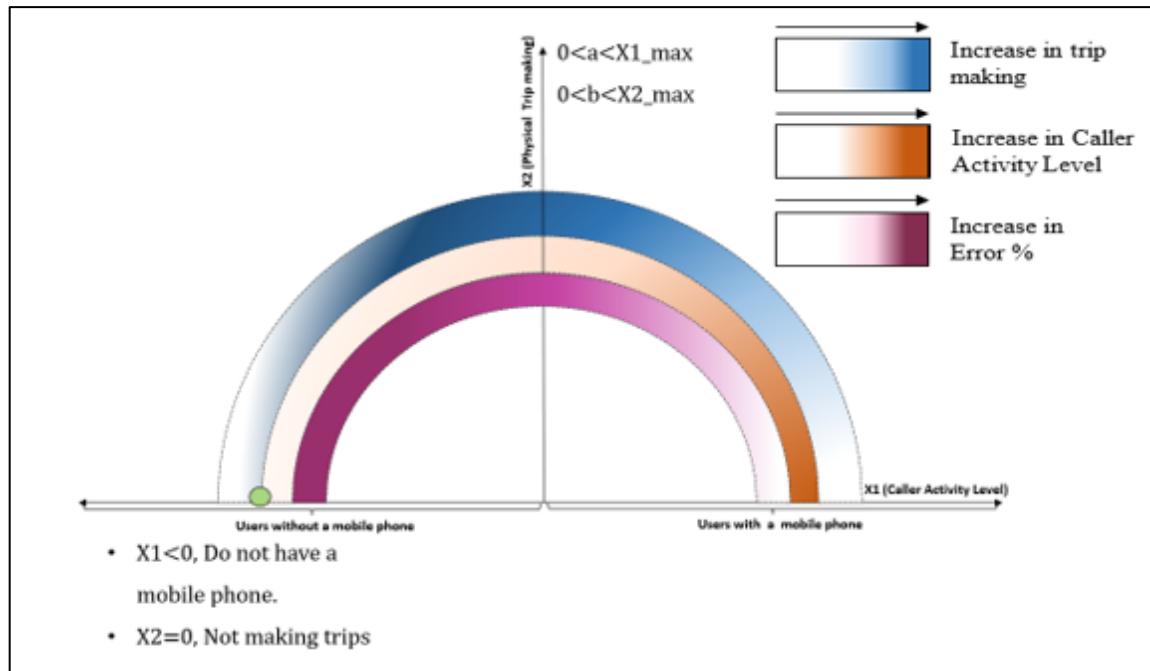
A sample positioned at the specified point on the graph indicates an individual without a mobile phone, hence not engaging in call activities or making trips. This category does not influence the travel demand modeling process, as it involves no physical movements that would utilize transportation resources.

Category 2

In this category, the sample represents individuals without a mobile phone (indicated by a negative value on the X-axis) who do not undertake trips. Like Category 1, this group has no impact on Travel Demand Estimation, as their lack of mobility does not contribute to transportation resource consumption.

Figure 4-27

User Behavior Model 2 Based on CDR Data

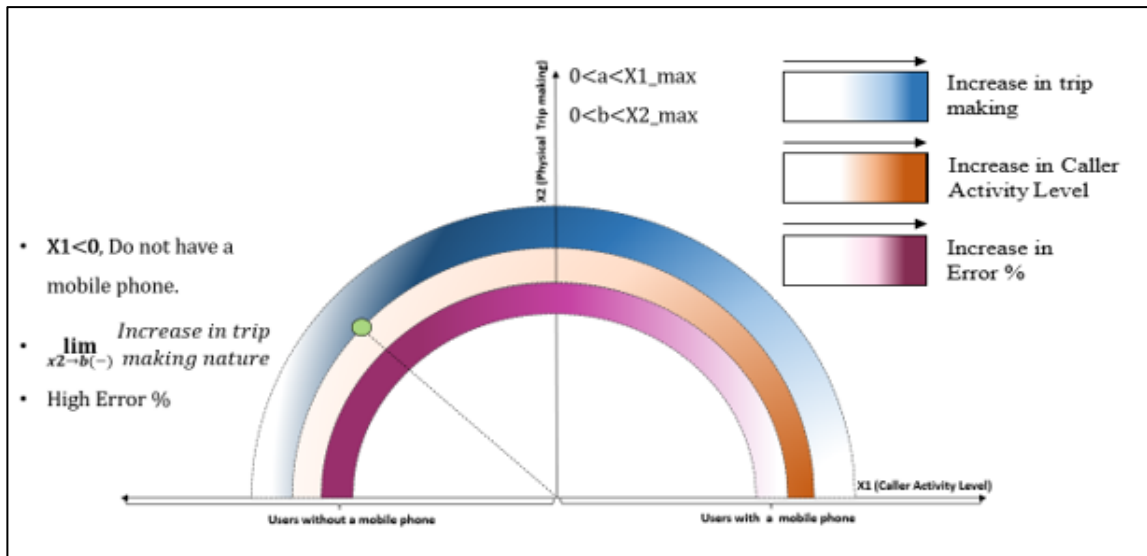


Category 3

This category is pivotal, as it encompasses individuals actively making trips but cannot be tracked due to their lack of mobile phone usage. Such travelers introduce significant errors in the Travel Demand Estimation process, leading to an underestimation of recorded trips. However, in the contemporary era, the number of commuters not utilizing mobile phones is minimal, as the majority view mobile phones as essential items.

Figure 4-28

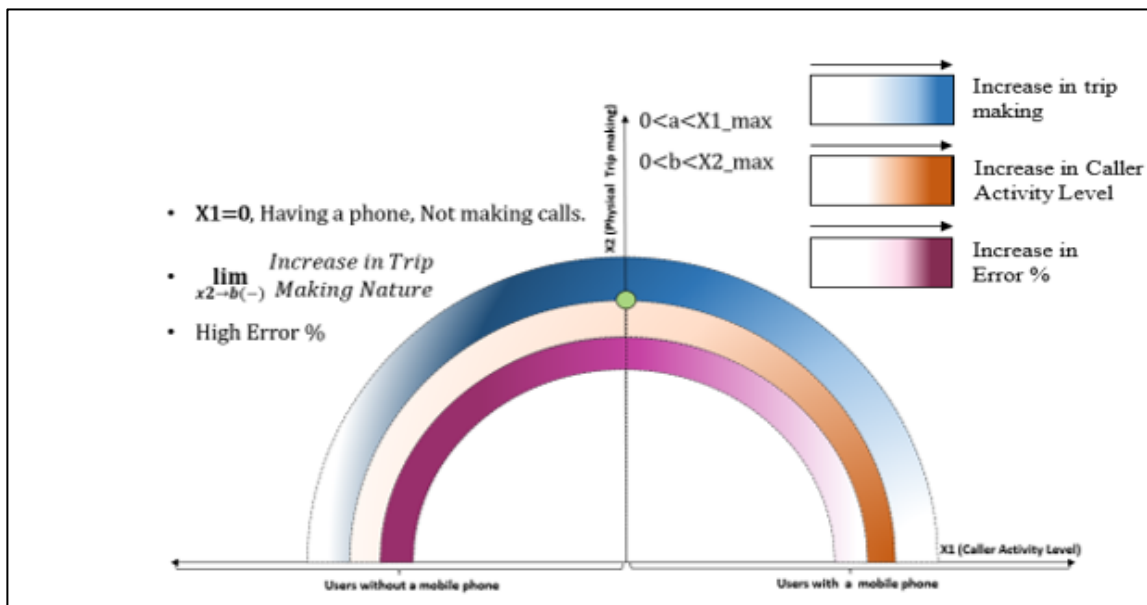
User Behavior Model 3 Based on CDR Data



Category 4

Figure 4-29

User Behavior Model 4 Based on CDR Data



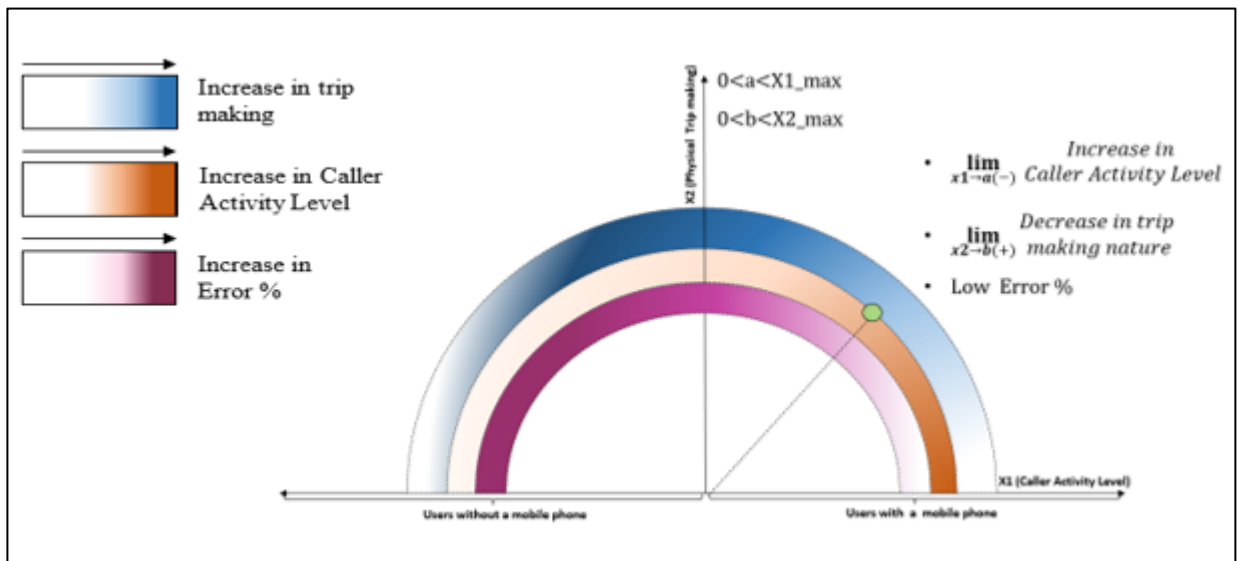
This category closely mirrors Category 3, with the key distinction being a higher frequency of trip-making. Consequently, the error introduced by Category 4 in the Travel Demand

Estimation is more pronounced compared to Category 3. This is attributed to the increased activity level that remains untracked due to the absence of mobile phone usage among these individuals.

Category 5

Figure 4-30

User Behavior Model 5 Based on CDR Data



When samples are located toward the upper right quadrant, caller activity levels increase, enhancing the ability to track movements. Consequently, whether users make a high or low number of trips, these movements are accurately recorded, leading to a significantly lower error impact on the Travel Demand Estimation process than other categories. This improved tracking capability is pivotal in refining the accuracy of travel demand models by capturing a more comprehensive picture of user mobility.

Category 6

As samples are positioned in the upper right quadrant, an increase in caller activity levels is observed, substantially enhancing movement tracking. This results in accurately documenting both high and low frequencies of trips, thereby significantly reducing the error impact on the Travel Demand Estimation process compared with other categories. This enhancement in tracking capabilities is crucial in refining the accuracy of travel demand models, offering a more detailed and comprehensive understanding of user mobility patterns.

Figure 4-31

User Behavior Model 6 Based on CDR Data

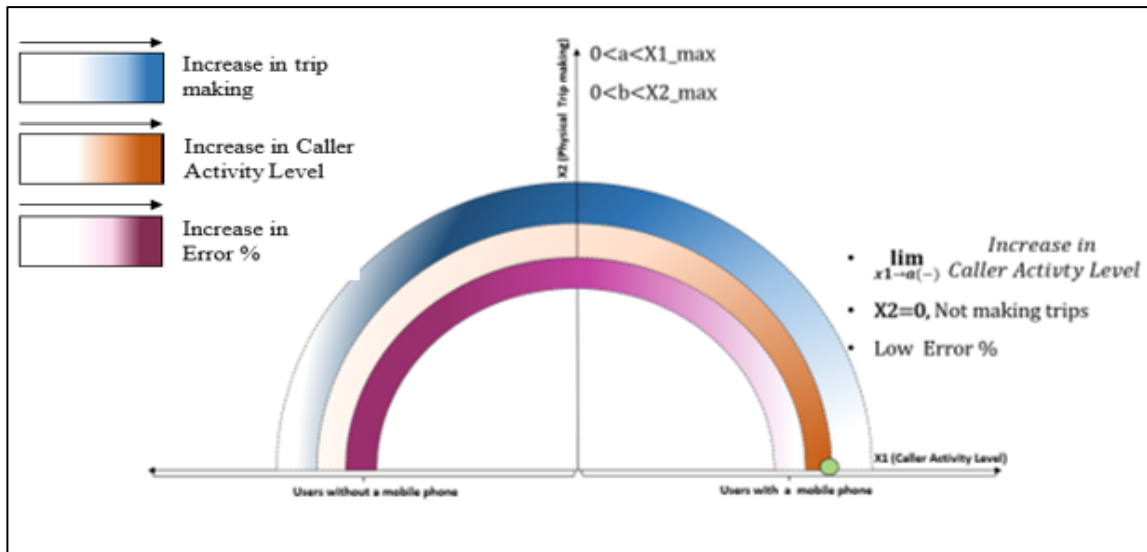


Table 4-4

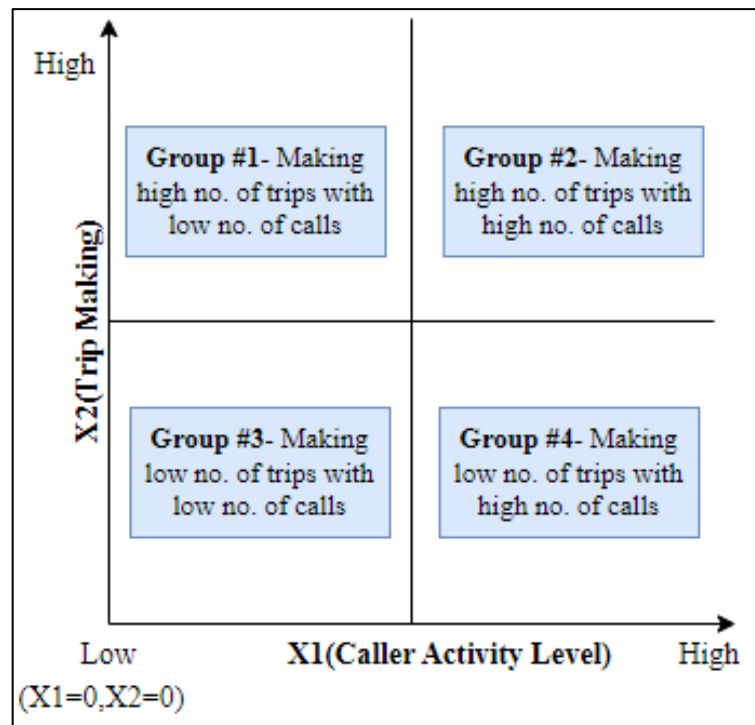
Socio-Economic Segmentation of Users Using CDR data

Group No:	Description
Group 1	This category includes individuals who undertake many trips but make relatively fewer calls. This behavior could be attributed to users who share mobile phones or to specific demographics such as students. Such patterns highlight the diversity in mobile phone usage and its implications for accurately estimating travel demand, underscoring the need to consider various user behaviors in travel demand models.
Group 2	This group encompasses drivers and daily work commuters, primarily the working population. These individuals are characterized by a high frequency of phone calls and a substantial number of trips. This pattern reflects the integral role of mobile communication in the daily routines of working professionals, who rely on it to coordinate travel and work-related

	activities. The behavior of this demographic is crucial for understanding overall travel demand, as they represent a significant segment of mobile phone users and travelers.
Group 3	Adults and young children not engaged in work or education activities represent a distinct category characterized by both low trip-making and call making activities. This demographic segment exhibits unique mobility and communication patterns, reflecting their different daily routines and lifestyle needs compared to the working and schooling populations. Understanding these variations is crucial for a comprehensive analysis of travel demand and mobile phone usage across diverse population groups.
Group 4	Individuals employed near their residences, such as shop owners and domestic workers, constitute a category characterized by a high frequency of calls but minimal utilization of transportation infrastructure. This group's unique mobility pattern, marked by limited reliance on broader transport networks, contrasts with that of other working populations who may commute longer distances. Their communication behavior, evidenced by elevated call volumes, underscores the significance of localized connectivity within their daily operations and social interactions. Understanding these nuanced patterns is essential for tailoring transportation planning and communication infrastructure to meet the population's diverse needs.

Figure 4-32

Socio-Economic Classification Using Spatio-Temporal Features of CDR Data



The figure 44 presents a two-dimensional classification framework that categorizes mobile phone users based on their caller activity level (X1) and trip-making frequency (X2) similar to the previously explained categories. It divides users into four distinct groups to highlight behavioral differences in communication and mobility patterns. Group #1 consists of users who make a high number of trips but have a low level of phone activity, indicating high mobility with minimal digital interaction. Group #2 represents users with both high mobility and high communication activity, suggesting a highly active profile both physically and digitally. Group #3 includes individuals with low levels of both calling and travel, reflecting generally inactive behavior. Group #4, in contrast, shows users with high phone activity but low trip-making, indicating a more stationary lifestyle with frequent digital interaction. This classification aids in understanding the heterogeneity of user behavior in CDR datasets and supports the refinement of mobility analyses by identifying user segments that may influence the accuracy of travel demand estimation.

The next section of the study employs all the above-mentioned features and presents a methodology for purpose-based OD matrix estimation, which is a significant milestone of the research.

4.5 Home-based Trip Identification and validation

This study introduces a multifaceted methodology for processing Call Detail Records (CDRs), as depicted in Figure 4.33. The process for transforming CDRs into purpose-based Origin-Destination (OD) matrices encompasses five distinct stages, as outlined in the diagram. This methodological framework is designed to systematically convert raw CDR data into insightful, purpose-specific OD matrices, facilitating a deeper understanding of mobility patterns and travel behavior.

- Stage 1: Constructing a Trip from CDR data
- Stage 2: Identifying Significant locations
- Stage 3: Trip Extraction and Classification
- Stage 4: OD Estimation by Purpose
- Stage 5: Results and validation

In the following sections, each stage of the methodology will be elaborately discussed, including the validation of the results obtained through comparison with data from the Household Visit Survey. This comprehensive examination aims to ensure the reliability and applicability of the derived Origin-Destination matrices in understanding travel behaviors and patterns.

4.5.1 Stage 1: Constructing a trip from CDR Data

As previously mentioned, a trip is considered the basic unit of travel behavior. It is generally defined as a journey made by an individual between two points and for a specified purpose. Figure 4.34 depicts an example of an individual's trip to work using multiple stops and modes of travel, which are made up of several segments. Work trips are one of many different purposes for which people travel. Still, this whole phenomenon is considered one trip, as all segments of the trip are made for a single defined purpose.

Figure 4-33

Approach for OD Estimation Based on CDR Data

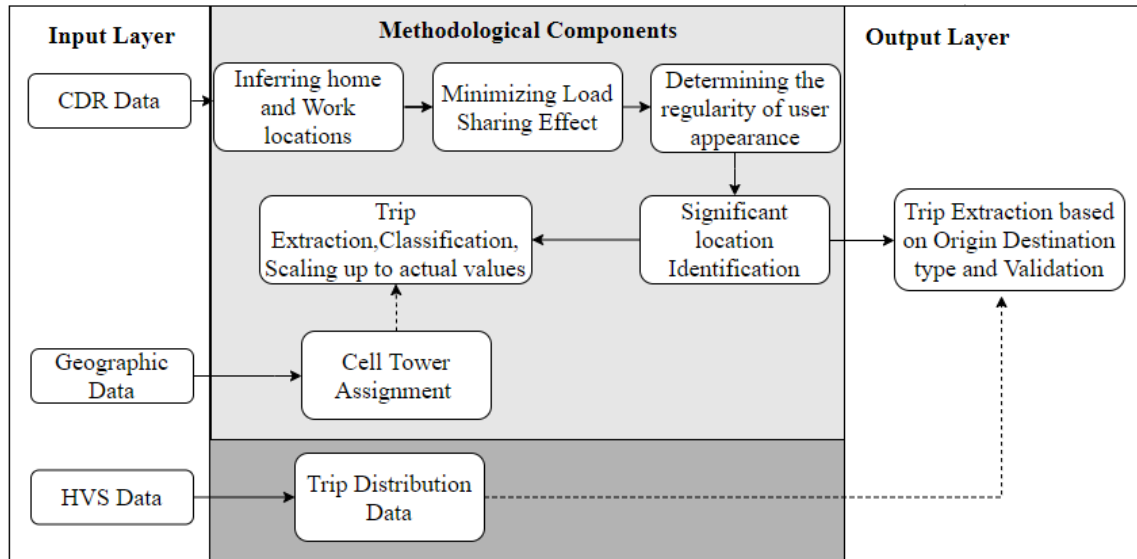


Figure 4-34

Theoretical Trip Structure

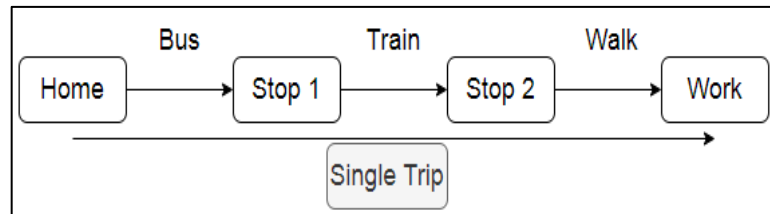
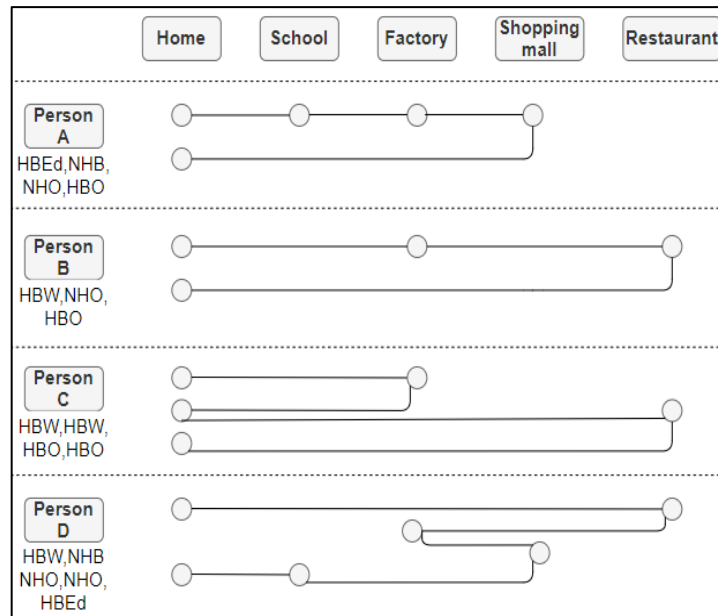


Figure 4.35 presents examples of individual activity patterns articulated through the lens of standard transportation modeling terminology. Individuals frequent various locations based on their specific needs at any given time, with these places of interest shifting as their requirements evolve. This variability is intrinsic to individual mobility patterns. The illustrations in Figures 4.34 and 4.35 elucidate the concept of anchor points, origins, and designated destinations for each journey. While intermediate stops are not the primary focus, the origins and destinations of trips are distinctly highlighted, underscoring the structured nature of travel behavior as individuals move between fixed points of interest.

Figure 4-35

Personalized Activity Trajectories



Identifying origins and destinations is a pivotal step in the travel demand modeling process. When a mobile phone user frequents locations, they will probably also place frequent calls from those sites. Call data analysis facilitates identifying the most visited tower zones for each user, which can be viewed as their inherent origins and destinations for routine trips. Throughout this study, the term "significant locations" will be employed to underscore the importance of these inherent points of interest.

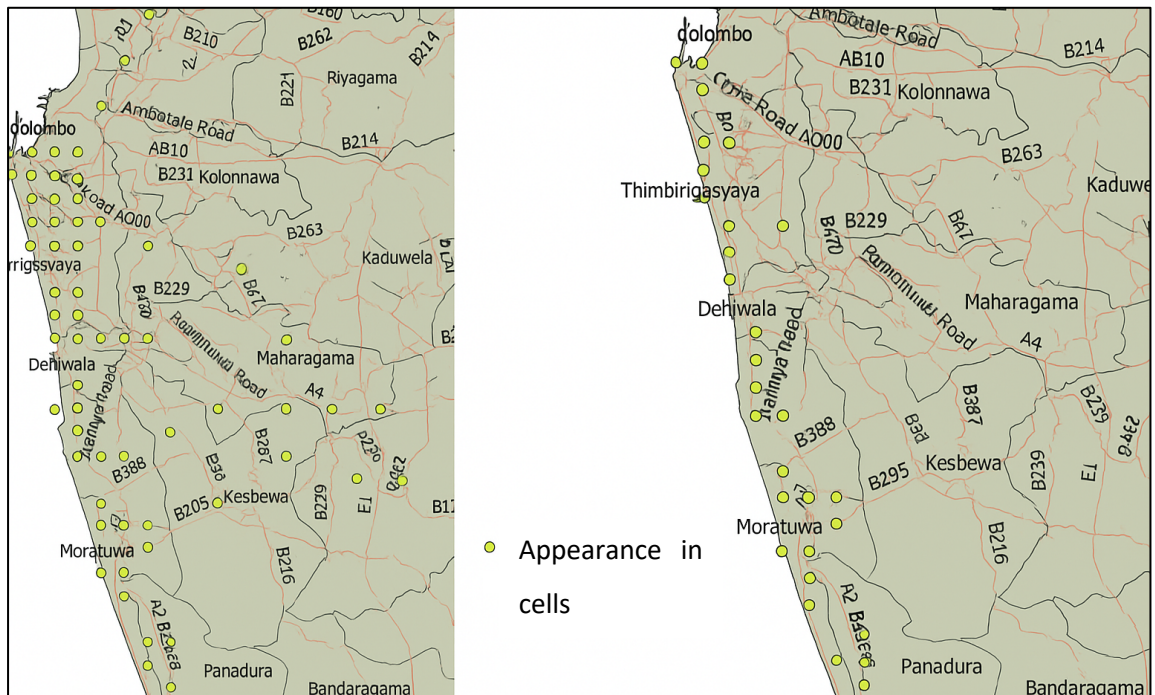
Figure 4.36 provides illustrative examples of user behavior mapped onto a Voronoi cell diagram, with dot size indicating visit frequency. This visualization showcases the appearance frequency of each user on a two-dimensional plane. Although both users exhibit inherent significant locations, establishing a common criterion for their extraction proves challenging. The frequency of appearance and the duration of visits vary considerably between users, rendering a universal criterion, such as appearances exceeding a certain threshold or a specified percentage, inadequate.

Thus, imposing blanket criteria like "locations where the user appeared more than X days" or "locations where appearance percentage exceeds X%" would be erroneous. Instead, each user's significant locations must be individually identified as they serve as the origins and

destinations of their respective trips. Consequently, a tailored methodology is introduced in the subsequent stage to pinpoint these significant locations for each user accurately.

Figure 4-36

User Mobility Mapping Across Time Periods Using CDR Data



4.5.2 Stage 2: Identifying Significant Locations

As previously discussed, the concept of regularity and irregularity in terms of cell visitation is highly subjective and varies from individual to individual, making the measure of cell regularity a distinctly personal metric. Upon identifying the cells that are visited regularly, these can be designated as significant locations within an individual's personal behavioral space. These significant locations are broadly classified into primary and secondary, as depicted in Figure 4.37. The regularity in the spatiotemporal appearances of subscribers forms the foundation for categorizing these locations as either primary or secondary.

Figure 4.38 demonstrates the integration of appearance layers with activity layers to construct a model of individual human mobility patterns. Before identifying significant secondary locations, it is essential to recognize the primary locations, typically home and work. Identifying these primary locations is relatively straightforward compared to identifying

secondary significant locations, which may involve a more complex analysis due to the broader range of activities and purposes they encompass.

Figure 4-37

Significant location categorization based on User Behaviour

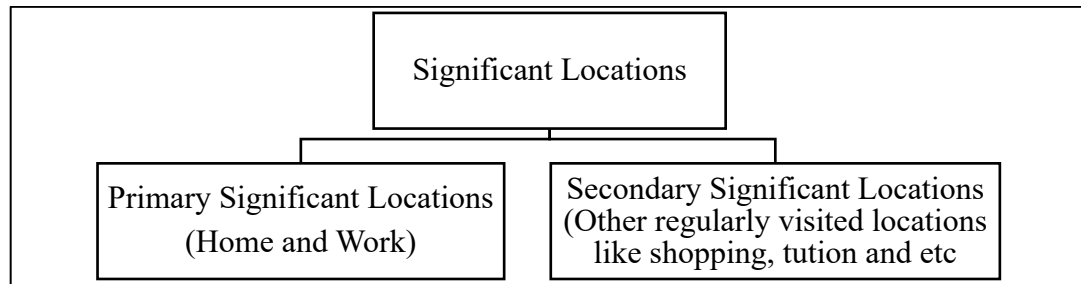
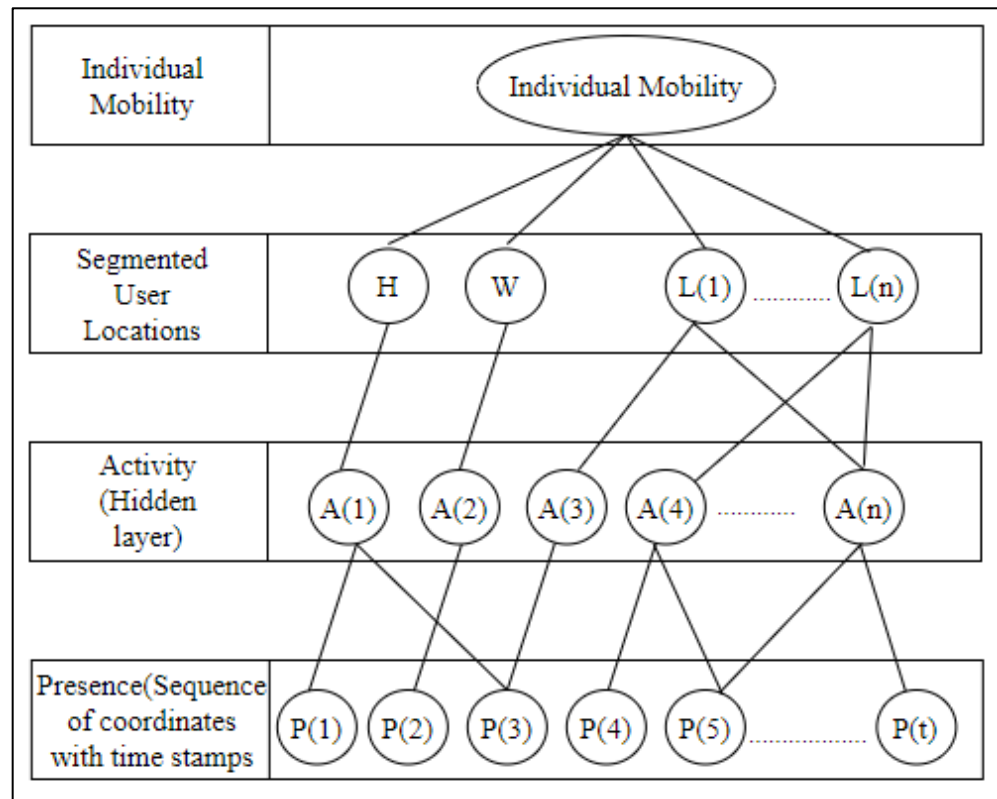


Figure 4-38

Integration of Appearance and Activity Layers

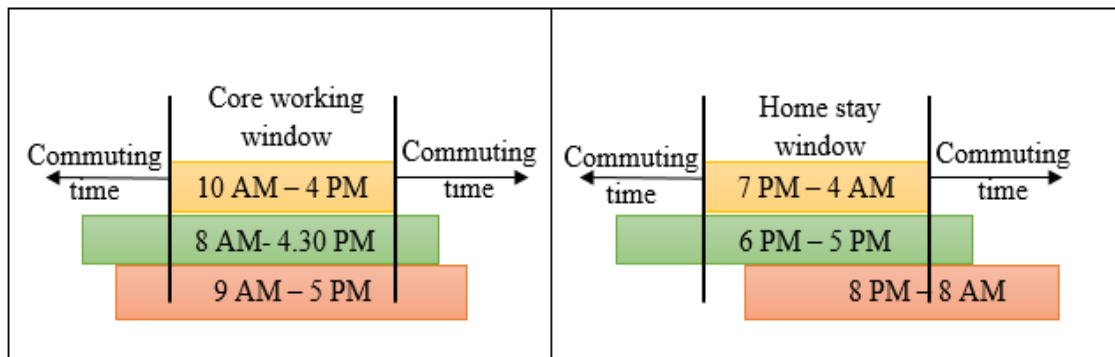


4.5.2.1 Primary – Significant Location Identification

Identifying Home and Work cells for individuals relies on the frequency of their presence during predetermined core time windows. Figure 4.39 illustrates the specific periods utilized for this classification. However, this method presents challenges in accurately identifying separate locations for individuals who work in proximity to their homes but in distinct locations. Additionally, the workplaces of individuals who work night shifts are often misclassified as home locations due to the timing of their presence. This limitation highlights the need to refine the classification criteria to better distinguish between home and work locations, especially for those with non-traditional work hours or locations close to their homes.

Figure 4-39

CDR-Based Core Time Window Identification



The identification of Home and Work locations within the context of mobile phone data analysis involves specific criteria based on user behavior patterns:

- **Home Location:** This is the location where users are most frequently found, characterized by relatively extended stays, predominantly during nighttime hours and often during weekends. The cell tower location that records the highest frequency of user connection from 7:00 pm to 4:00 am is designated as the home location. To enhance the accuracy of identifying a home location, it is verified that the cell tower most frequented during the weekend nighttime aligns with the cell tower used during weekday nighttime, thereby confirming the consistency of the location's significance.

- **Work Location:** This location is identified by the regular presence of a user during weekdays, particularly during standard office hours (10:00 am–4:00 pm), and is typically not visited during weekends. The work location is characterized by extended durations of stay, indicative of the user spending significant blocks of time at this location, which aligns with the concept of commutes or regular, repetitive travel patterns. The cell tower area most frequently utilized during these core office hours on weekdays is classified as the work location.

4.5.2.2 Secondary – Significant Location Identification

Observing subscriber behavior at significant secondary locations reveals that applying the same precise time windows used for identifying primary locations (Home and work) is inappropriate. Consequently, the determination of secondary locations relies on a more flexible approach, measuring regularity through a variety of time parameters. The subsequent section details the methodology employed to label these secondary significant locations, taking into account the nuanced patterns of subscriber mobility and activity.

Measuring the regularity of cells

The study measures the regularity of visits made to each cell in terms of three parameters:

- (a) Regularity by time of the day
- (b) Regularity by day of the week
- (c) regularity of the day of the month.

For this purpose, a single day was divided into different time segments (T_i , $i=1,2\dots10$) as in table 4.5.

The regularity of a particular user (i) being in cell (x) during the time category T_i is directly proportional to the three defined parameters: a , b , and c . Table 4.6 shows an example of a subscriber's regularity measurement.

“ a ” = Total no. of days the user n has visited cell tower x during time category T_i .

“ b ” = Total no. of days the user n has visited cell tower x at time category T_i per week.

“ c ” = Total no. of days the user n has visited cell tower x at time category T_i per month.

The regularity of a particular user (n) at cell (x) in time category $T_i \propto a*b*c$

Table 4-5*Time-of-Day Categorization for Regularity Analysis*

T ₁ = 6am-8am	T ₆ = 4pm-6pm
T ₂ = 8am-10am	T ₇ = 6pm-8pm
T ₃ = 10am-12pm	T ₈ = 8pm-10pm
T ₄ = 12pm-2pm	T ₉ = 10pm-4am
T ₅ = 2pm-4pm	T ₁₀ = 4am-6am

Once the regularity of visiting each cell for each time category is calculated, the regularity measure with the highest score extracts the significant locations from the available list of all possible cells.

Table 4-6*Regularity Measurement of Cell Tower Visits by Time Category*

User ID	Cell_ID	Time category	No. of days (a)	Total no. of days per week (b)	Total no. of days per month (c)	Regularity Measure (a*b*c)
685716430	26889	6 pm-8 pm	59	4	19	4484
685716430	26889	8 pm-10 pm	55	3	18	2970
685716430	26889	10 pm-4 am	48	3	16	2304
685716430	26889	4 pm-6 pm	27	1	9	243
685716430	26889	12 pm-2 pm	25	1	8	200
685716430	26889	2 pm-4 pm	24	1	8	192
685716430	26889	10 am-12 pm	21	1	8	168

Extracting the Significant Locations

As explained earlier, the measurement of cell regularity should be individualistic. Mathematical Linkage Analysis (MLA), derived from graph theory, can address each user's behavior separately. Graph theory models the pairwise relations between objects. In the case

of cell regularity, the cells of each user are arranged in descending order based on their regularity, forming the actual configuration. MLA compares the real configuration to an ideal-typical configuration, where the regularity is the same for the first, second, and third... cells. In other words, in the first ideal-typical condition, only one cell exists; in the second situation, two cells exist with equal regularity. The R-squared value is then calculated between the actual distribution and each ideal-typical condition. The iteration with the largest R-squared value indicates the number of significant locations for the individual. Table 4.6 shows a sample user where the maximum R-squared value is reached in the second iteration. Therefore, the number of significant locations for the selected user can be two. The exact process was carried out for the sample, and the significant locations of each user were identified.

Table 4-7

Significant Location Extraction via Regularity and Ideal-Typical Matching

<i>Actual distribution of Regularity</i>			<i>Ideal-Typical Configuration</i>				
<i>Cell ID</i>	<i>Maximum Regularity measure</i>	<i>% of Regularity Measure</i>	<i>1(%)</i>	<i>2(%)</i>	<i>3(%)</i>	<i>4(%)</i>	<i>5(%)</i>
A	4880	34%	100%	50%	33.33%	25%	20%
B	3444	24%	0%	50%	33.33%	25%	20%
C	574	4%	0%	0%	33.33%	25%	20%
D	574	4%	0%	0%	0%	25%	20%
E	573	4%	0%	0%	0%	0%	20%
F	430	3%	0%	0%	0%	0%	0%
G	429	3%	0%	0%	0%	0%	0%
		Rs-q	0.6563	0.941	0.5259	0.306	0.1676

4.5.3 Stage 3: Trip Extraction and Classification

Identifying physical movements, or trip extraction, is a central objective within this analysis. It focuses on transforming consecutive, sparse call records into recognizable trips. A trip is delineated as a movement between spatially distinct locations, which are classified as Home, Work, and secondary significant locations.

The extracted trips are further categorized into three distinct segments, based on the combination of origin and destination, as illustrated in Figure 4.40. Movements that commence at home and conclude at work are designated as Home-Based Work trips. Conversely, trips originating from home and terminating at a significant location other than work are classified as Home-Based Other trips. Lastly, trips that do not start or end at home are categorized as Non-Home-Based trips. Table 4.7 provides further elucidation on this concept through the use of sample Call Detail Records (CDR), offering a clear depiction of how each trip type is distinguished based on its origin and destination.

Figure 4-40

Trip Classification Based on Origin-Destination Combinations

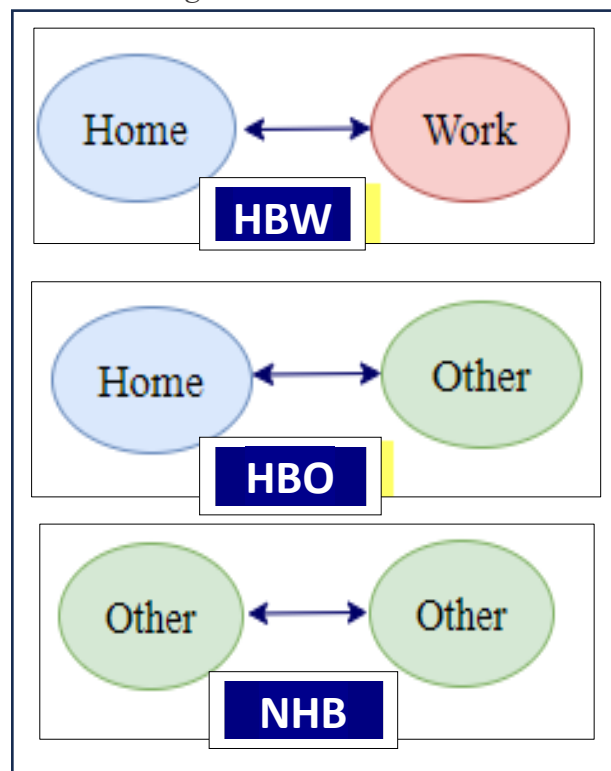


Table 4-8*Trip Type Classification Using Sample CDR Entries*

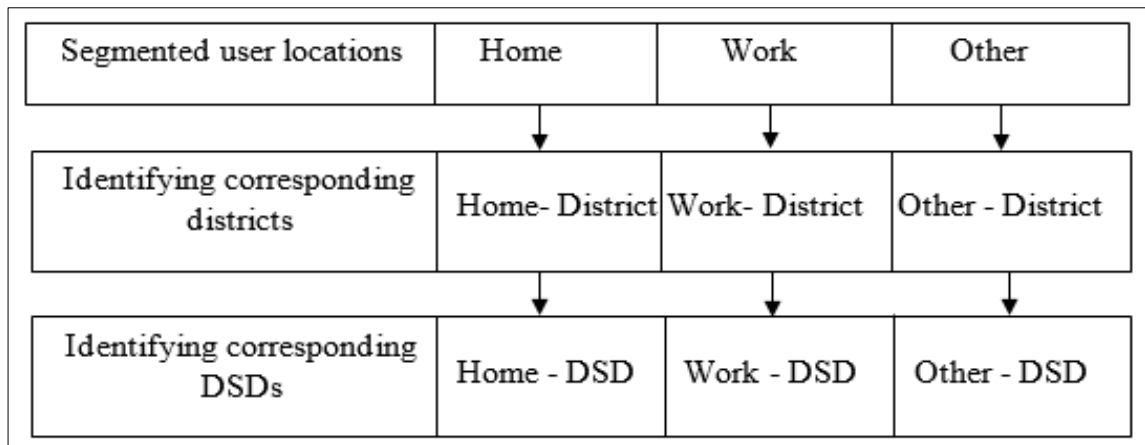
Cell #	Cell Type Identified as	Time of Appearance	
A	Work	01/05/2013- 15:30	
A	Work	01/05/2013- 16:40	
B	Home	01/05/2013- 18:10	} A Trip (HBW)
C	Significant location	01/05/2013- 19:30	} A Trip (HBO)
D	Significant location	01/05/2013- 20:15	} A Trip (NHB)
E	Non-Significant location	01/05/2013- 22:00	} Not a Trip

4.5.4 Stage 4: Estimation of Origin-Destination Matrices

Travel demand analysis necessitates the comprehension of aggregate trip patterns across various travel zones, a process known as OD (Origin-Destination) Matrix Estimation. This involves aggregating individual users' trips to discern the travel patterns between different travel zones during the time frame for Call Detail Records (CDRs), which are available. By mapping cell towers to their respective travel zones that span the geographic area covered by the CDRs, it becomes feasible to ascertain each trip's origin and destination traffic zones. This methodological approach enables the systematic identification of travel demand patterns, facilitating the analysis of travel behavior across the studied area.

Figure 4-41

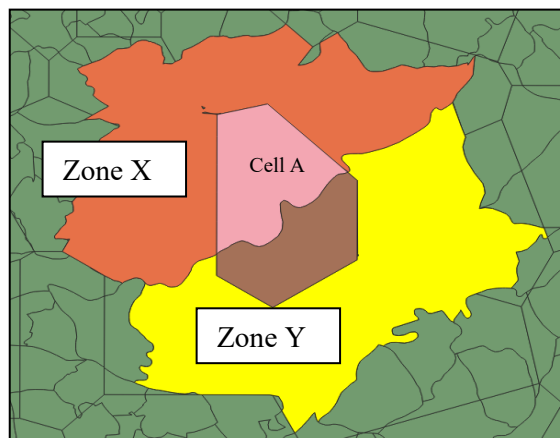
Spatial Mapping of Segmented User Locations to Administrative Zones



This study employs a detailed zoning approach (figure 4.41), incorporating two levels of spatial granularity: (a) administrative districts and (b) Divisional Secretariat Divisions (DSD). To facilitate this, a geospatial layer comprising cell towers, identified by their respective latitude and longitude coordinates, is overlaid onto another layer outlining the boundaries of districts and DSDs. This overlays process, executed using Geographic Information System (GIS) software, allows for precise mapping of cell towers within specific transport zones, as depicted in Figure 4.42. This methodology ensures that the spatial distribution of cell towers is accurately aligned with the defined transport zones, facilitating a more nuanced analysis of travel patterns within the study area.

Figure 4-42

Cell Tower Allocation to District and DSD



Due to the discrepancy between the boundaries of cell towers and the designated travel zones, an apportionment method is necessary to attribute trip generations and attractions to each zone accurately. Consequently, the contributions of trips originating from and destined for a particular cell tower are proportionally allocated based on the extent of overlap between the cell tower coverage area and the travel zone boundaries. This approach ensures that trip calculations reflect the spatial distribution and intensity of travel demand within and across the zones, thereby providing a more precise representation of travel patterns for the analysis.

Total Area of cell A = Area covered in Zone X ($A_{Zone X}$) + Area covered in Zone Y ($A_{Zone Y}$)

Assume that cell A generates T no. of trips,

Total no. of trips generated from = $T * [A_{Zone X} / (A_{Zone X} + A_{Zone Y})]$

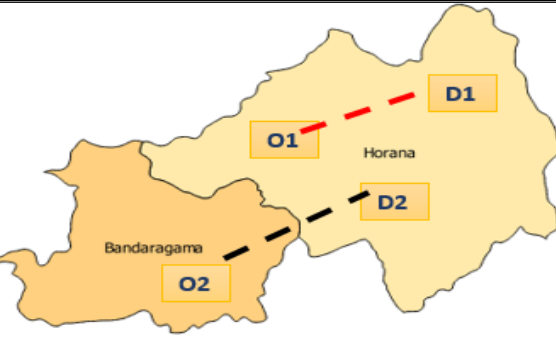
A that was initiated from Zone X

Total no. of trips generated from = $T * [A_{Zone Y} / (A_{Zone X} + A_{Zone Y})]$

A that initiated from Zone Y

Following the methodology above for assigning cell towers, the origins and destinations of trips are delineated as Districts and Divisional Secretariat Divisions (DSDs). The aggregation of trip origins and destinations by travel zones further allows for the classification of trips according to type, specifically as Home-Based Work (HBW), Home-Based Other (HBO), or Non-Home-Based (NHB) trips. This categorization facilitates the development of daily Origin-Destination (OD) matrices for each trip type. Moreover, each segment within these matrices is analyzed to discern travel patterns within (intra) and between (inter) districts and DSD levels. Table 4.8 illustrates the concept of intra- and inter-district/DSD trips. At the same time, Figure 4.43 summarizes the resulting OD matrices, offering a comprehensive overview of travel demand across the study area.

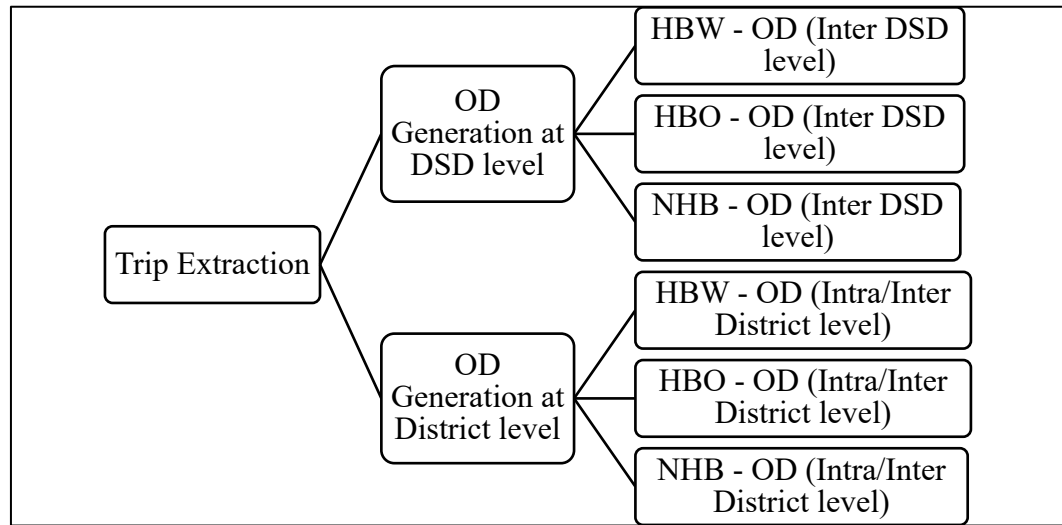
Table 4-9*Classification of Intra- and Inter-Zone Trips at District and DSD Levels*

Type of trip	Description
	Both Intra trip ODs (O1-D1) and inter-trips (O2-D2) ODs were generated at the district level.
	Both Intra trip ODs (O1-D1) and inter-trips (O2-D2) ODs were generated at the DSD level.

Note: Source, Developed by the Author

Figure 4-43

Trip Type Classification Across DSD and District Levels



4.5.5 Results and Validation

The comparison of trip proportions derived from Call Detail Records (CDR) data with those obtained from Household Travel Surveys (HVS) reveals a notable congruence, as illustrated in Table 4.9. This methodology enhances the accuracy of trip identification by proportionally allocating cell areas to relevant geographic zones rather than a simplistic allocation of cells to geographical areas. Adjusting the cell areas to match the specific geographic zones makes this approach more precise in capturing the spatial distribution of trips. It accounts for the variable densities and population concentrations across different zones, offering a nuanced representation of trip patterns. The alignment between the trip proportions from the CDR data and the HVS data underscores the effectiveness of this cell assignment methodology. It facilitates a deeper understanding of travel behaviors and supports rigorous analysis and planning in transportation and urban development.

Table 4-10*Comparison of Trip Type Proportions: CDR vs. Household Travel Survey Data*

Trip Type	Proportion of trips	
	CDR Data (Per Day for 400,000 sample)	HVS Data (Per Day for Western Province)
Home Based Work	422 (26%)	2700447 (27%)
Home Based Other	813 (51%)	5160050 (52%)
Non-Home Based	365 (23%)	2126235 (21%)

Figures 4.45 to 4.48 present the validation outcomes of Origin-Destination (OD) matrices categorized by purpose at both Intra-District and Inter-Divisional Secretariat Division (DSD) levels. Across all types of OD matrices, the results demonstrate a satisfactory R-squared value at these two geographical levels, indicating a reasonable level of accuracy in the data. Notably, the accuracy of Home-Based Work (HBW) trips diminishes at more disaggregated levels. However, among the various trip types, HBW trips exhibit the highest degree of correspondence. This notable match can likely be attributed to the higher regularity.

Predictability of home-based work trips, contributing to the observed similarity in the validation process.

Figure 4-44

Validation of Trip Distribution derived from CDR data at Intra District Level

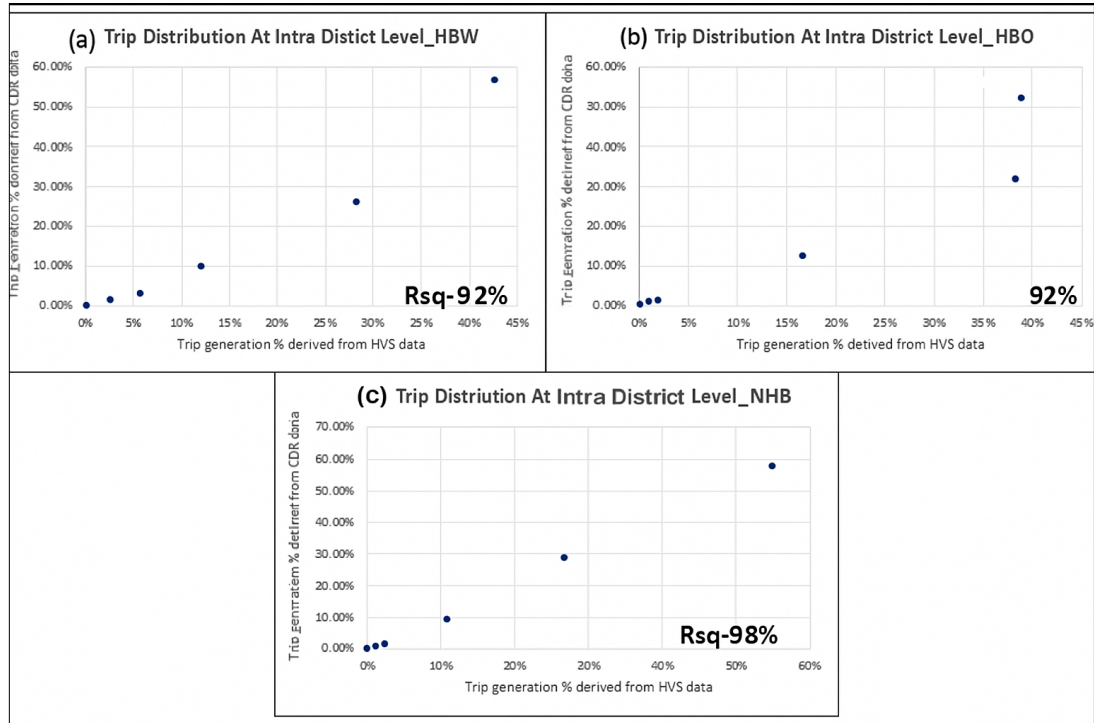


Figure 4-45

Validation of HBW Trip Distribution derived from CDR data at at DSD level

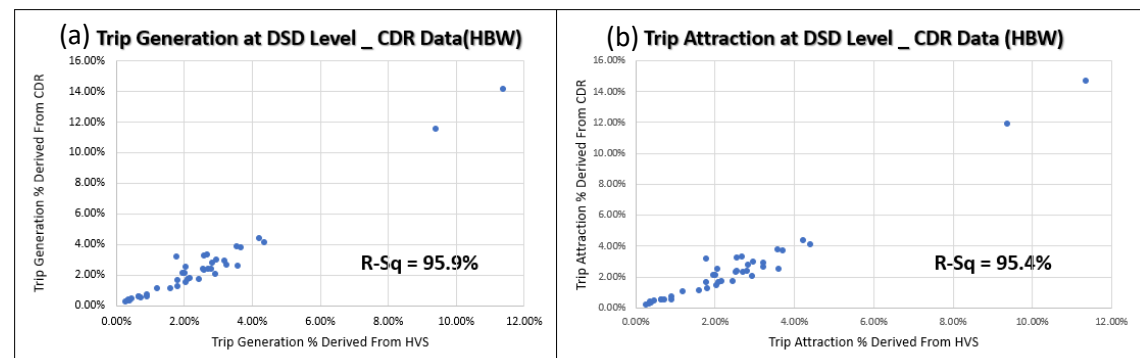


Figure 4-46

Validation of HBO Trip Distribution derived from CDR data at at DSD level

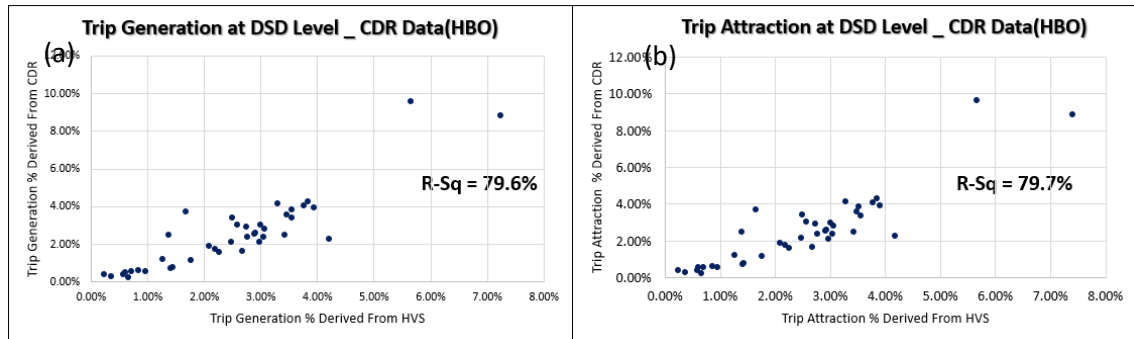


Figure 4-47

Validation of NHB Trip Distribution derived from CDR data at at DSD level

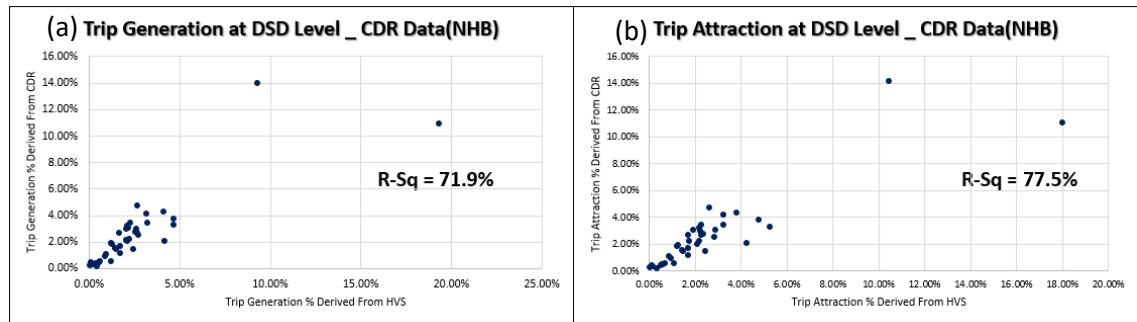
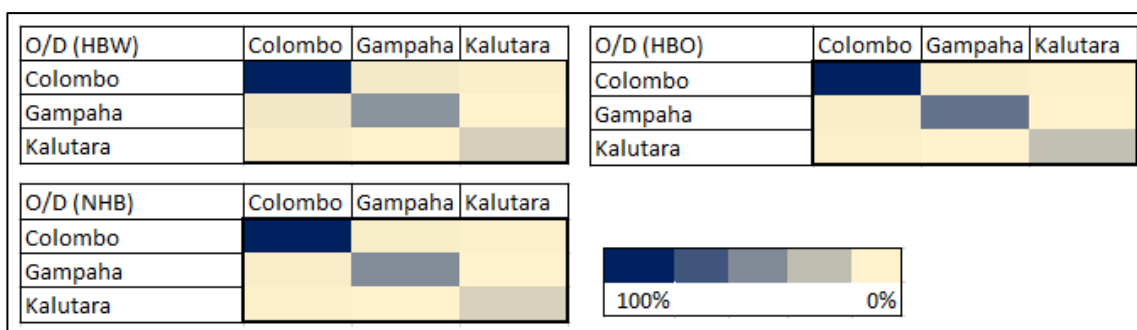


Figure 4.48 illustrates the intensity of Intra-District trips across three categories: Home-Based Work (HBW), Home-Based Other (HBO), and Non-Home-Based (NHB) trips. The diagram highlights the comparative intensity of trips occurring within each district. Notably, the Colombo District exhibits a significantly higher trip intensity for all three trip types when compared to the other two districts under consideration. This observation underscores the Colombo District's dominant role in trip generation and attraction across the categories examined.

Figure 4-48

Intensity map of Intra District Trips by Trip Type



Expanding Home-Based Work (HBW) trips, characterized by their regularity, into actual trip counts, is conducted straightforwardly. Table 4.10 shows this expansion process for a single cell within the district-level Origin-Destination (OD) Matrix. The expansion of trip counts is executed at the district level, encompassing inter-district and intra-district trips, and at the Divisional Secretariat Division (DSD) level for inter-DSD trips. This methodological approach ensures a comprehensive trip pattern analysis and counts across varying geographic scales.

Table 4-11

Expansion of HBW Trips to Actual Trip Counts Using Population Scaling

HVS Data		CDR Data	
Total Population Western Province:	P1	Total Population in Western Province as of the CDR sample:	P2
No. of Home-based Work trips from zone X to Y:	C1	No. of Home-based Work trips from zone X to Y for the considered period (Counting Weekdays only)	C2
		No. of Weekdays in the considered period	D
		No. of Home-based Work trips per day from zone X to Y for the expanded population**	$(C2/(P2*D)) * P1$

Table 4.11 presents the outcomes obtained at the district level, where an R-squared (R-Sq) value of 97% was observed, with a standard error of 19%. At the Divisional Secretariat Division (DSD) level, the R-Sq values for trip generations and attractions were recorded at 80% and 85%, respectively, with corresponding standard errors of 38% and 33%. This pattern indicates a decrease in model fit at more disaggregated levels, which can be attributed to inaccuracies in mapping tower zones to corresponding travel zones. Such inaccuracies tend to increase as the size of the travel zone becomes smaller, affecting the precision of the model's predictions.

Table 4-12

Comparison of Expanded HBW Trip Counts (CDR vs. HVS) at District Level

OD Pair (District level)	CDR Data expanded	HVS Data		
Colombo-Colombo	290039	1150559		
Colombo-Gampaha	18658	152107		
Colombo-Kalutara	10725	71792		
Gampaha-Colombo	21538	154907		
Gampaha-Gampaha	148854	763714		
Gampaha-Kalutara	206	3388		
Kalutara-Colombo	12076	72925		

Kalutara-Gampaha	73	3466		
Kalutara-Kalutara	68374	327588		
OD Pair (District level)	CDR Data expanded	HVS Data		
Colombo-Colombo	290039	1150559		
Colombo-Gampaha	18658	152107		
Colombo-Kalutara	10725	71792		
Gampaha-Colombo	21538	154907		
Gampaha-Gampaha	148854	763714		
Gampaha-Kalutara	206	3388		
Kalutara-Colombo	12076	72925		
Kalutara-Gampaha	73	3466		
Kalutara-Kalutara	68374	327588		

Figure 4.49 illustrates the distance distribution across the three categories of trips, revealing that trip distances range from 1 to 57 kilometers. Figures 4.50-4.52 present the hourly distribution of extracted Home-Based Work (HBW), Home-Based Other (HBO), and Non-Home-Based (NHB) trips. These figures indicate that the peak frequency for trip departures falls between 6 am and 7 pm, with the specific time of departure being used to determine the distribution.

Figure 4-49
Trip Distance Distribution Across All Trip Types

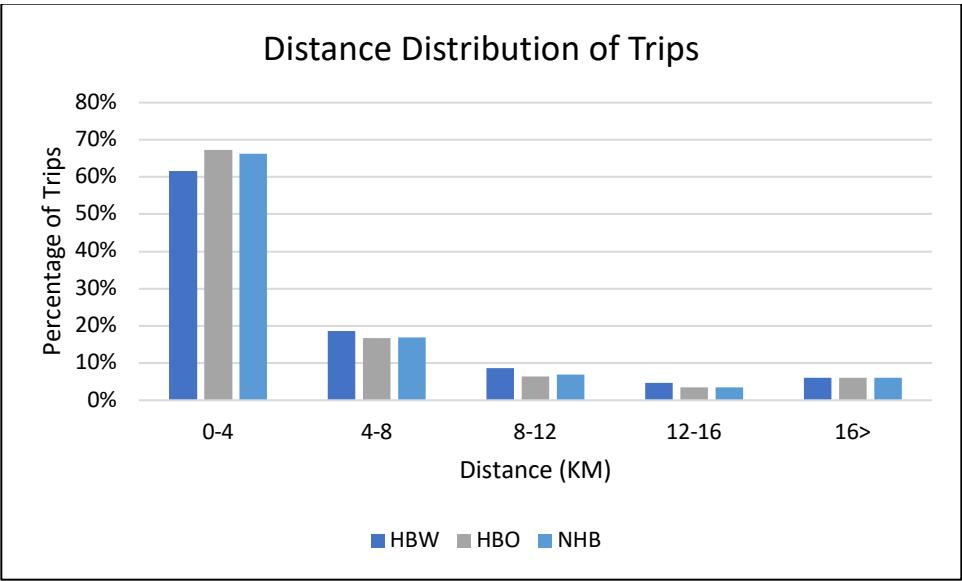


Figure 4-50

Hourly Distribution of HBW trips

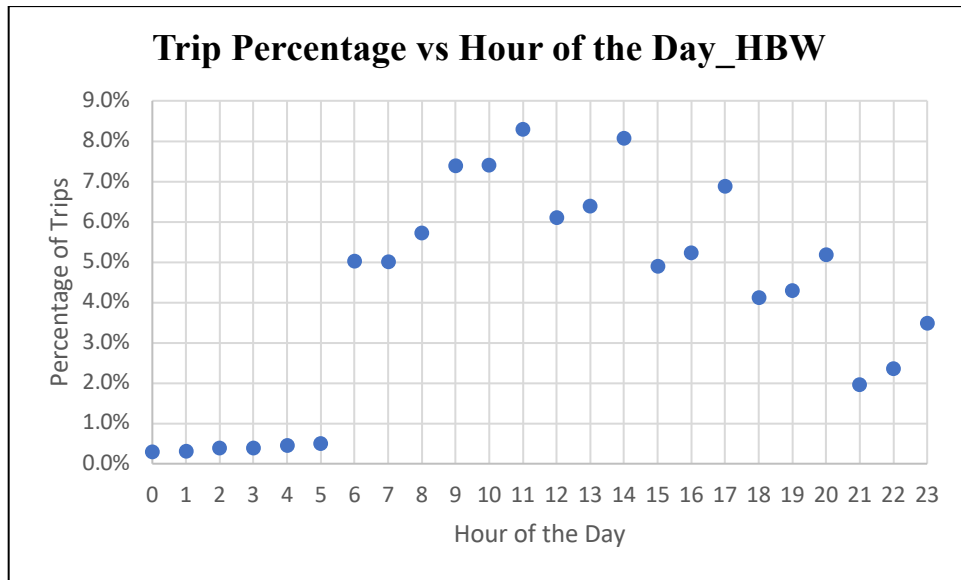


Figure 4-51

Hourly Distribution of HBO trips

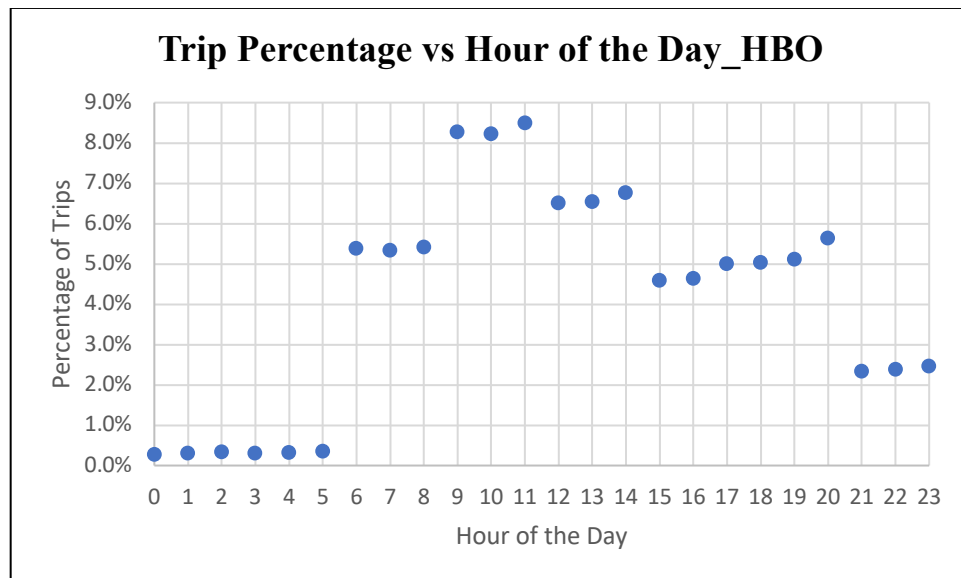
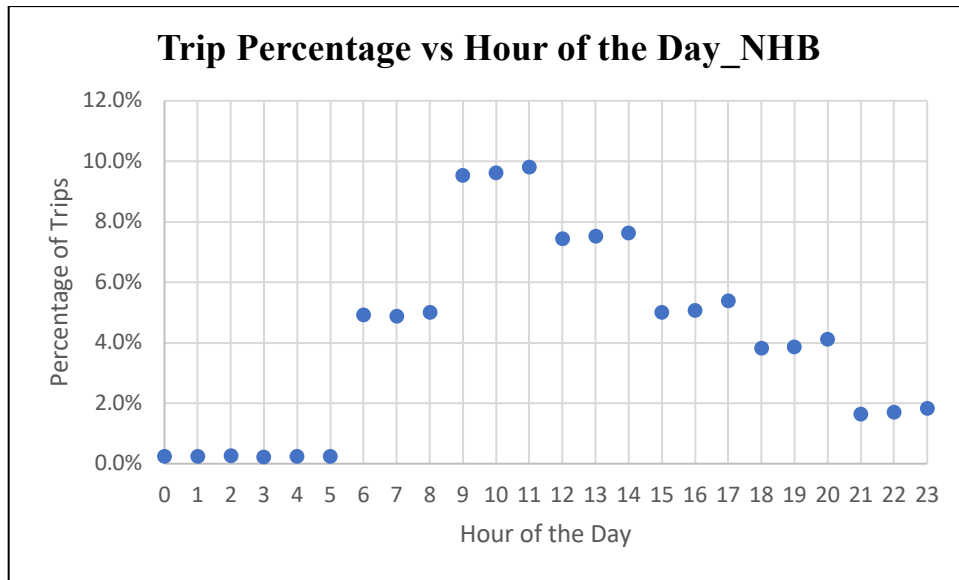


Figure 4-52

Hourly distribution of NHB trips



4.6 Conclusion

This chapter outlined the methodology for extracting and analyzing various components from call detail records (CDR), cell tower information, and other relevant data sources. It provided a comprehensive guide on the necessary procedures for data cleansing, modeling the coverage area of cell towers, and categorizing cells. This chapter's significant focus was categorizing trip types based on individual behaviors and examining the variability in time, distance, and the sampling of CDR trajectories. The chapter emphasized the importance of human activity data, whether real-time or historical, for applications aiming to understand or predict human behavior. Specifically, the utility of subscriber CDR trajectories as inputs for such applications was discussed.

The narrative covered the initial two phases in the conventional process of travel demand estimation utilizing CDR data: the identification of Trip Generation and Attraction and the estimation of the Origin-Destination (OD) matrix. The results were validated against traditionally collected data with satisfactory accuracy. The forthcoming chapter will delve

into the network assignment phase of the conventional travel demand model, with a particular emphasis on extracting behavior related to movement. This will facilitate the prediction of travel paths between points of interest.

CHAPTER 5

NETWORK ASSIGNMENT

Effective traffic congestion management, public facility management, bus route optimization, and trip frequency scheduling are pivotal for cultivating a livable city. Traffic congestion, closely tied to human mobility, plays a vital role in ensuring the well-being of urban populations. It is critical to accurately determine flow capacities to prevent overinvestment and underinvestment in transportation facilities and planning.

Commuting trips, characterized by their regularity and repetition between an individual's home and workplace, are believed to contribute significantly to daily traffic congestion. According to the Comtrans survey, approximately 10 million trips occur on a typical weekday within the Western Province, of which approximately 2 million are home-based work trips. Thus, a thorough analysis and understanding of commuting trip patterns and associated individual behaviors are crucial for infrastructure planning and effective traffic management.

The adoption of big data in urban planning and transport management is increasingly becoming part of the local government's decision-making processes. Traditional methods for identifying people flow within a designated area, such as surveys or transport statistics, are

often costly and labor-intensive. In contrast, Call Detail Record (CDR) data offers an inexpensive and efficient way to capture people's aggregated movement within a city through spatio-temporal insights.

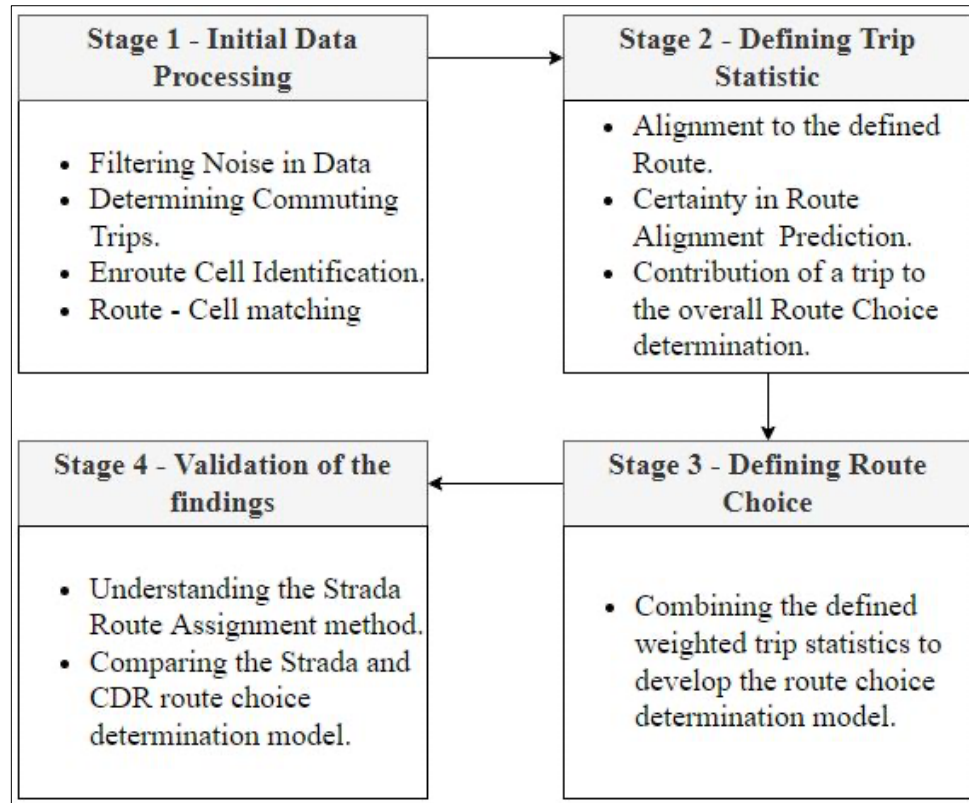
This thesis section aims to estimate the link flow of commuting trips by analyzing one month of CDR data. It maps out the commuting flow patterns across the study area for each trip during the specified period. Various trip statistics are defined based on user behavior, leading to developing a route choice model tailored to commuting trips. The model's accuracy is validated against traditionally collected data. The structure of this section is as follows: It begins with an explanation of the CDR data handling steps, including the extraction of various pieces of information from the dataset. It then details commuting trip mapping onto the network and establishes a comprehensive set of trip statistics. Subsequently, it describes the modeling of the defined concepts and concludes with an explanation of the validation process for the study period.

5.1 Methodology

This study outlines a multi-stage methodology for determining route choice, as depicted in Figure 5.1. The methodology is executed in four stages, each detailed in this section.

Figure 5-1

Methodology for Inferring Travel Routes Using CDR Data



5.2 Stage 1 – Initial Data Processing

Stage 1 of the methodology focuses on the initial processing of the data that is required for the route estimation.

5.2.1 Defining the Sample

In the preliminary preprocessing stage, approximately 11% of the data records were excluded due to instances of shared loads and incomplete information. The remaining 89% of the data then proceeded to the subsequent phase of analysis aimed at route choice determination. The initial step involves identifying primary locations, specifically the home and work locations of the callers, utilizing the labeled location data outlined in the preceding chapter. This process enables the identification of users' home and work locations. Those individuals with distinct home and work cells are categorized as daily commuters, forming a separate group for further analysis.

Following the initial processing and categorization of primary locations, the study advances to extracting commuter trips. Given the focus on understanding route choice among commuter

trips, it is crucial to pinpoint each caller's commute between their home and work locations. These identified commuter trips are then meticulously analyzed to discern route choices, conducted at both disaggregate and aggregate levels, to illuminate patterns of commuting behavior.

5.2.2 Enroute Cell Identification

Identifying enroute cells plays a pivotal role in determining the route choice for commuting trips. Enroute cells refer to the base stations users connect to while traveling to their work locations. This process utilizes two critical parameters, distance and time, to label enroute cells for each user's commuting trip.

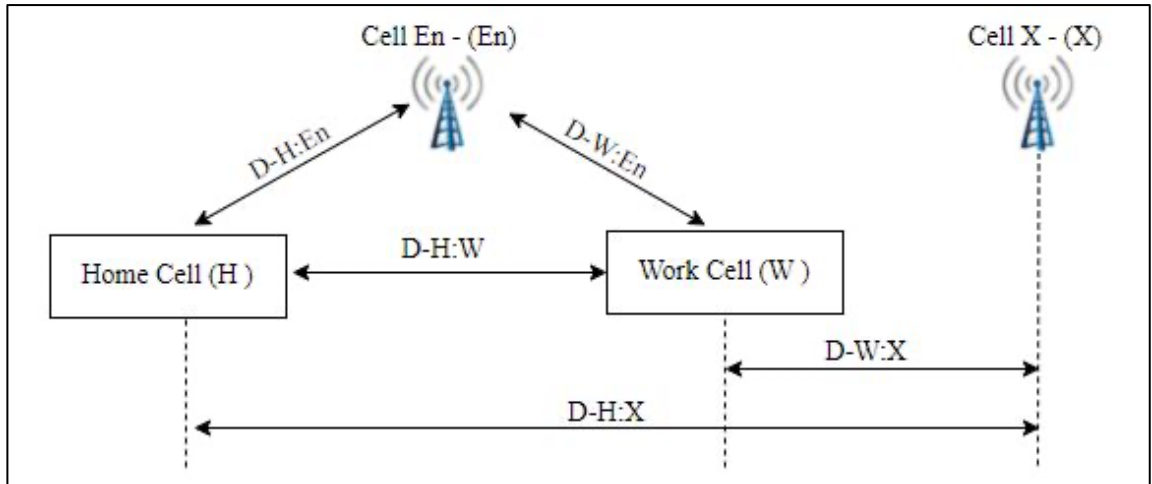
From a temporal perspective, a cell is designated as an enroute if a user connects to it within the timeframe extending from the last call from their home to the first call from their work location, or vice versa. This temporal bracket ensures the identified cell lies along the route between the user's home and work.

Spatially, the process considers the distances from the work and home locations to a potential enroute cell. A cell qualifies as an enroute if its distance between the work and home locations is less than the total distance between the user's home and work. This criterion guarantees that the enroute cells are reasonably located along the commuting path, avoiding any significant detour.

Figure 5.2 exemplifies this methodology by illustrating the exclusion of cell X from being considered an enroute cell due to its excessive distance from the home location, surpassing the direct distance between the user's home and work locations. By applying these spatial and temporal parameters, enroute cells are accurately identified, laying a foundational step for subsequent analysis in determining route choices for commuting trips.

Figure 5-2

Identification of Enroute Locations from Call Detail Records



In conclusion, to be defined as an enroute cell, spatial and temporal criteria should be fulfilled by the appeared cell.

C WL – last call from work, C WF – first call from work,

C HL – Last call from home, C HF – First call from home.

$$\text{Distance} \left\{ \begin{array}{l} C_H - C_{EN} \\ \text{and} \\ C_W - C_{EN} \end{array} \right\} > C_W - C_H \quad \text{Spatial parameter}$$

$$C_{EN} \left\{ \begin{array}{l} C_{WL} - C_{HF} \\ C_{HL} - C_{WF} \end{array} \right\} \quad \text{Time parameter}$$

Figure 5-3

Filtered User Set for Route Estimation Using CDR Data

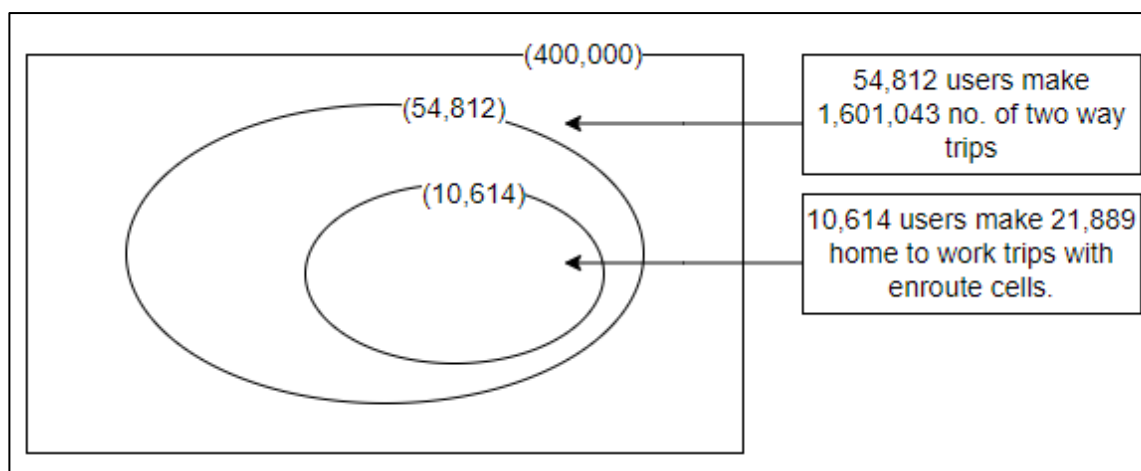


Figure 5.3 summarizes the process up to the point of enroute cell extraction. This summary shows that among the 10,614 users who undertook a total of 21,889 trips during the study period, trips with identified enroute cells were recorded. These commuting trips spanned across 993 Divisional Secretariat Division (DSD) pairs. Among these, trips with enroute cells were identified for 565 DSD pairs, and notably, 394 of these pairs reported more than 100 trips each.

5.2.3 Route-Cell Matching

The methodology for estimating the utilization of the road network by users during each trip relied on the spatial distribution of the previously defined enroute cells. This involved integrating the Voronoi cell layer with the road network layer to create connections based on a sequence of cells, which indicate the likely route the user takes. Specifically, the sequence of cell towers (route cells) connected by a user during their journey was mapped to identify the most probable path traveled within the road network. It is important to clarify that conventional map matching techniques, typically applied with GPS data, were not suitable here due to the sparse nature of cell data in each trip. A set of disaggregated trip statistics was devised to leverage the CDR data effectively. As detailed in the subsequent section, these statistics were formulated by synthesizing various parameters, such as connection sequences, cell appearances, and other pertinent factors.

This section presents a case study in which a randomly chosen subscriber's trips over four consecutive days are mapped onto a Voronoi layer (Figure 5.4-5.5)

Location-Based Mapping of Daily Commute Patterns

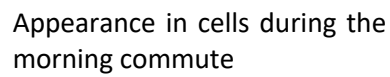
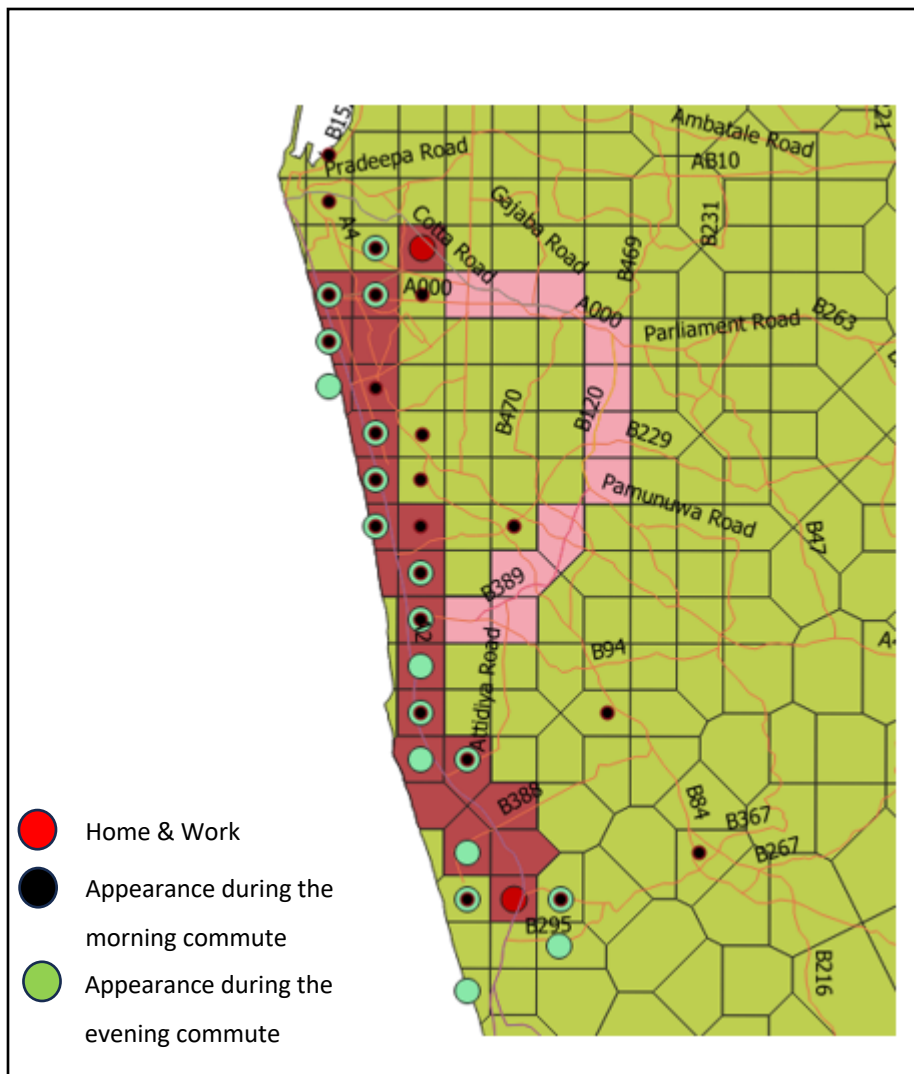


Figure 5-5

Mapping CDR derived locations onto the Most Probable Travel Path



5.3 Stage 2 – Defining Trip Statistics

Following the orderly data processing, the subsequent step involves gaining insight into the trip statistics utilized. Trip statistics are delineated at the disaggregate level, calculated for each trip a user makes, and the aggregate level, computed across all trips for each user. This approach ensures the extraction of the maximum amount of information from the available data.

(a) Alignment to the defined route (Disaggregated Measure)

(b) Certainty in route alignment prediction was inferred for each trip of the user (Disaggregated Measure)

(c) Contribution to route choice was ascertained per user (Aggregated Measure)

(a) Alignment of a trip to the defined route

The objective of the first statistic is to measure how closely a specific trip by a user aligns with the predefined route. This is achieved by dividing the number of enroute cells that fall along the predefined cell path by the total number of cells along this path. Consequently, this calculation produces a probability of alignment to the predefined route for each trip made by each user.

(b) Certainty in Route Alignment Prediction of a trip

The certainty in route alignment prediction was determined based on the frequency of connections to cells along the route for each trip. These connections acted as a weighting factor, assigning greater significance when more connections were observed. Figure 5.6 showcases the applied methodology, where a relative weightage for certainty in route prediction was calculated by dividing the number of calls originating from enroute cells along the path by the total number of calls made during the trip. Two previously mentioned metrics were integrated to calculate the probability of a user selecting a predefined route for commuting ($P(A)$), as illustrated in Figure 5.7. This derived probability was subsequently utilized as a critical input parameter for evaluating each user's contribution to inferring route choice.

(c) Contribution of a trip to Route Choice

Two metrics were employed to evaluate each user's impact on the route choice inference process. The first metric, the output probability, gauges the likelihood that a user will choose

a particular route for their commute. The second metric applies a weighting, calculated by dividing the number of trips substantiated by evidence by the total number of trips undertaken by the user.

Table 5.1 and Figure 5.8 offer an illustrative calculation for a single user to enhance comprehension. For simplicity, consider an ideal scenario. Suppose the designated path for route 1 encompasses six cells, and on day 1, the user connects with all six cells along route 1. This results in the first metric, alignment, being calculated as 1. Additionally, it is assumed the user makes ten calls, all connected to cells on route 1, and this behavior remains consistent across all five days. Thus, the alignment for each trip remains at 1, with each trip receiving a consistent weight based on the number of calls. By aggregating these calculated measures across all trips, we determine the probability of users opting for route 1 for their commute as 1. Consequently, this user's contribution to the route choice inference is assessed as complete, with evidence available for all five days, culminating in a final commuting route selection probability for route one as 1.

Figure 5-6

Steps in deriving the trip statistics(a) and (b)

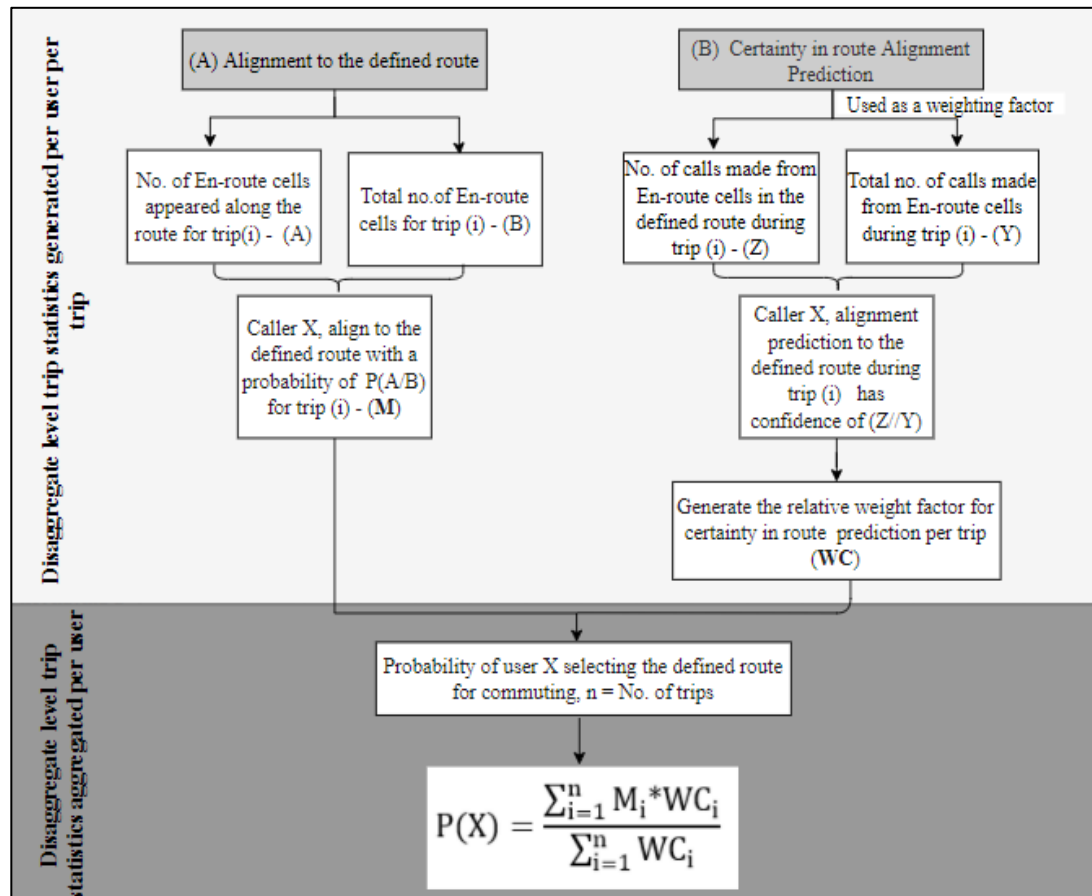


Figure 5-7

Steps in deriving the trip statistic (c)

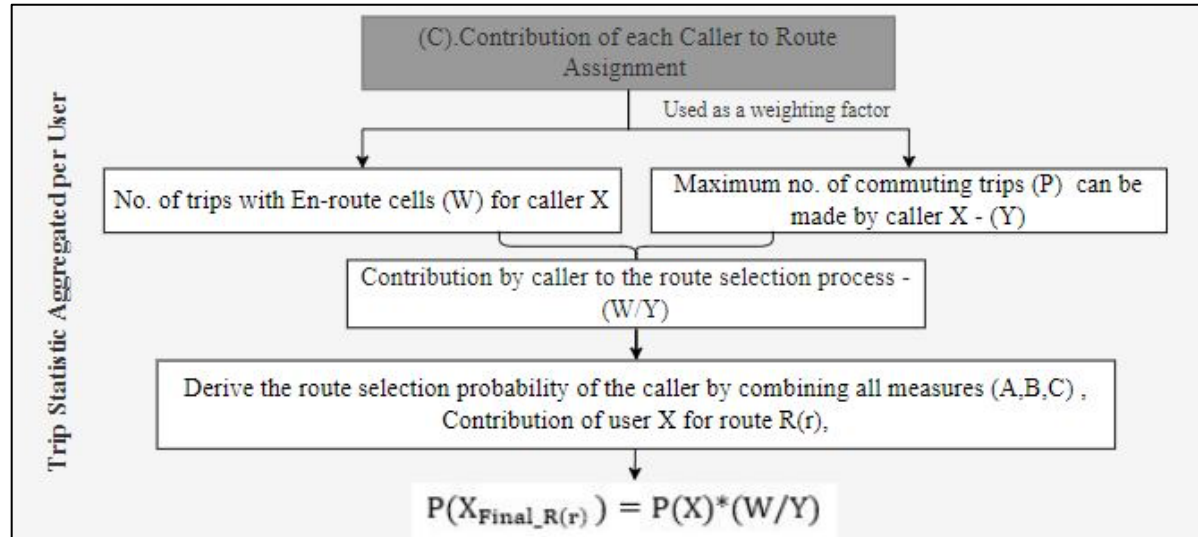


Figure 5-8

Route Choice Representation of a Single User

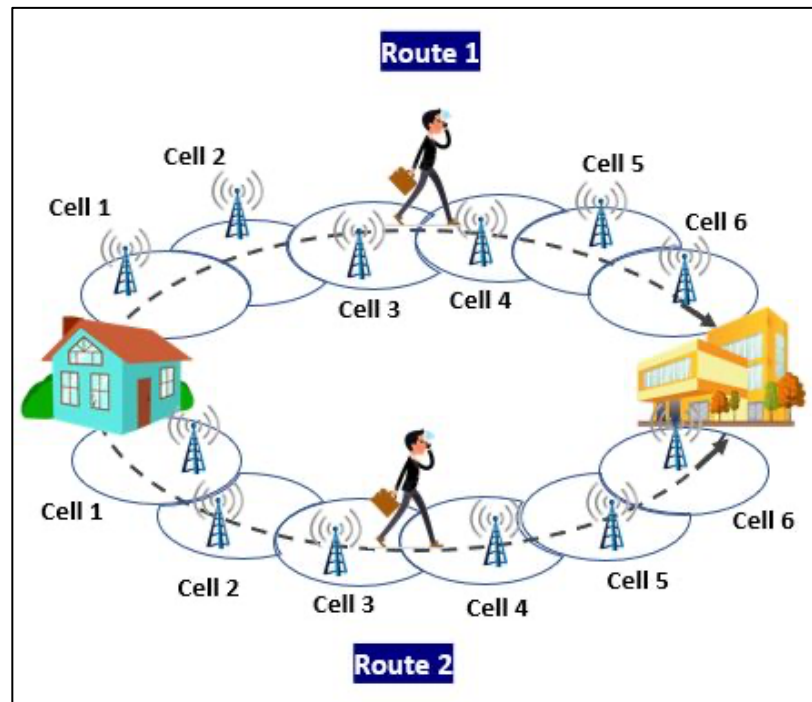


Table 5-1*Example Calculation of Determining Route Choice for a Single User*

Day No:	Route No:	No. of cells in the route connected to	Alignment to the route	Total no. of calls made	No. of calls made by connecting to the cells in the route	Certainty in route prediction	Weight for the trip
1	R1	6	6/6=1	10	10	10/10=1	1/5
2	R1	6	6/6=1	10	10	10/10=1	1/5
3	R1	6	6/6=1	10	10	10/10=1	1/5
4	R1	6	6/6=1	10	10	10/10=1	1/5
5	R1	6	6/6=1	10	10	10/10=1	1/5
Total=5							
Probability of the caller selecting a Route commuting $P(X)$				Contribution to the route choice $P(X_{R1})$			
The caller used route R1 to commute on all five days. $P(X) = (1 \cdot 1/5) + (1 \cdot 1/5) + (1 \cdot 1/5) + (1 \cdot 1/5) + (1 \cdot 1/5)$ $= 1$				$P(X_{R1}) = P(X) \cdot \text{Weighted average for route R1}$ $= P(X) \cdot \text{Total Weight/No. of days}$ $= P(X) \cdot 5/5$ $= P(X) \cdot 1$ $= 1$			

5.4 Stage 3- Determining the Route Choice

In conclusion, to ascertain the definitive route choice for each user, the calculated route selection probability is multiplied by the actual number of days the user has commuted to work. Following this, the proportion of trips taken via the considered route is determined, as outlined in Equation 5 and as depicted in figure 5.9. This approach assumes that the same probability of route selection applies to trips where enroute cells are not identified.

When (m) = No. of users among a particular DSD pair

(r) = No. of routes among a particular DSD pair

$P(X_{Final_R(r)})$ = Probability of user X selecting route (r) for commuting

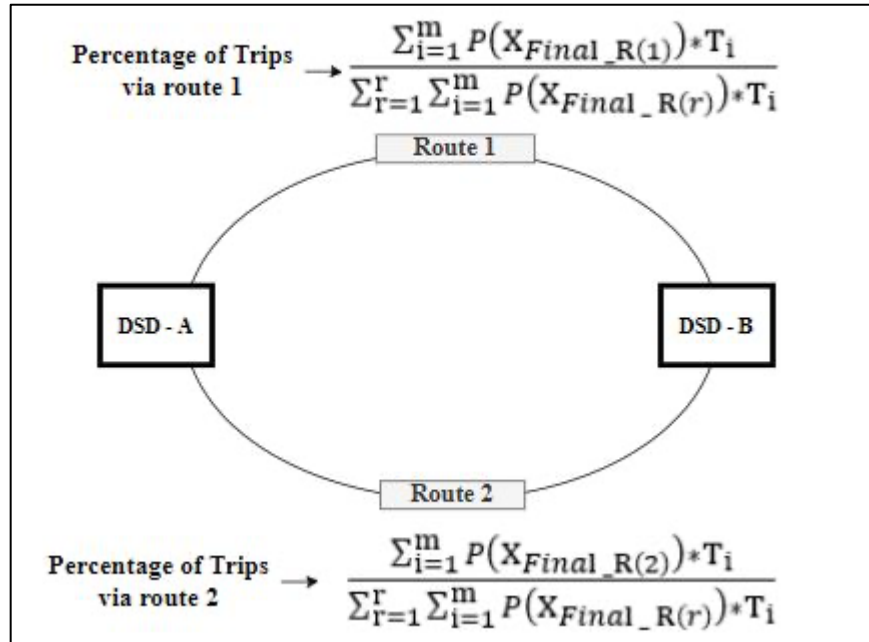
(T_i) = No. of days the user had traveled to work

The final route selection probability can be shown in the equation below,

$$\text{Percentage of Trips made via route } n = \frac{\sum_{i=1}^m P(X_{Final_R(n)}) * T_i}{\sum_{r=1}^r [\sum_{i=1}^m P(X_{Final_R(r)}) * T_i]} \text{-----(5)}$$

Figure 5-9

Route Choice generation function via CDR Data



5.5 Stage 4 – Validation using STRADA

5.5.1 Introduction to STRADA

The transport planning process encompasses several critical stages, ranging from analyzing current conditions to building demand forecasting models, identifying development initiatives, and assessing projected demand and proposed projects. To effectively support these varied stages, the STRADA software has been designed as a versatile tool. It features

around ten programmable modules, each configurable to various sequences and combinations, thus catering to the complex requirements of transport planning.

In the current phase of this study, we leverage the route assignment functionalities within the STRADA software to validate the route choice model derived from Call Detail Record (CDR) data. As elaborated in Chapter 3, transport modelers employ a range of trip assignment methods tailored to specific project needs. These methods include:

- (a) All or Nothing: This method loads trips from any origin zone to any destination zone onto a single, minimum-cost path between them.
- (b) User Equilibrium: Under this method, journey times on all used routes are equal and lower than those experienced by a single vehicle on any unused route.
- (c) Incremental Assignment: This process involves assigning fractions of traffic volumes in steps. Each step assigns a fixed proportion of the total demand based on the all-or-nothing assignment. After each step, link travel times are recalculated based on link volumes. If traffic exceeds the capacity of specific transport segments, congestion occurs, affecting travel time. This iterative procedure aims to converge towards a solution by minimizing transportation costs given a specific travel demand and the characteristics of the existing transportation network.

In the Comtrans study, comprehensive trip data for each household on a chosen day were meticulously gathered and subsequently scaled to reflect the broader population accurately. This process involved expanding the trips from each origin-destination (O-D) zone pair and categorizing them by their specific modes of travel. In the concluding phase of travel demand modeling, these expanded trips were systematically assigned to designated routes. The ComTrans master plan adopted a three-step network assignment methodology, rigorously tested, and successfully implemented in various contexts, as delineated in Table 5.2. It is important to recognize that multiple network assignment methods are available, and the choice of method should be carefully tailored to the characteristics of the scenario being modeled.

Table 5-2*Steps in Traditional Network Assignment*

Assignment number	Input	Method	Output
1	- 2013 Network file - Private transport OD Matrix. - Parameter file: 4 modes	User Equilibrium	Private mode assigned IRE
2	- 2013 Network file - Public transport OD Matrix. - Private mode assigned IRE as the initial assignment. - Parameter file: 1 mode	Transit Assignment	Public travel assigned IRE
3	- 2013 Network file - Private transport OD Matrix. - Public travel assigned IRE as the initial assignment. - Parameter file: 4 modes	User Equilibrium	All mode assigned IRE

In the Strada software, the route assignment process requires the input of three essential files, as specified in the initial assignment. Figure 5.10 showcases the user interface where these files are uploaded. An initial load file is uploaded only when available, as demonstrated in the second and third assignments.

The process begins with uploading the network file. It is crucial that the network, or the link properties contained within, are accurately aligned with the scenario being simulated. The Initial OD (Origin-Destination) file follows a specific format, as illustrated in Figure 5.11. The first column lists the origins, the second column lists the destinations, and the third column specifies the number of trips between each origin-destination pair. The final input is the parameter file, which encompasses diverse attributes, as depicted in Figure 5.12. This

structured approach ensures that the route assignment process is comprehensive and tailored to the specific demands of the simulated scenario.

Figure 5-10

Initial Interface of STRADA

Note: Source: STRADA Software

Figure 5-11

Sample OD format in STRADA

AOD2* Matrix Manipulator 3.511/13/2022 12:28:26 PM			
47	1	2	Pax Trips
1	1	115696	
1	2	0	
1	3	0	
1	4	0	
1	5	0	
1	6	0	
1	7	0	
1	8	0	
1	9	0	
1	10	3464	

Note: Source: STRADA Software

The assignment methodology underwent modifications to include Call Detail Record (CDR) inputs, aiming to validate the inferred route choices from CDR data against actual travel patterns. Initially, the first two network assignments followed the conventional approach. However, input parameters were adjusted for the third assignment to address the validation objectives specifically.

Figure 5-12

Parameter interface of STRADA

The screenshot shows the 'Init' dialog box of the STRADA software. It has several tabs: General, Zone, Time Value, QV, Turn Controls, Output Option, and Others. The 'General' tab is selected. Inside, there are several sections:

- General:** Four input fields: 'No. of Links' (3990), 'No. of Nodes' (2598), 'No. of Zones' (476), and 'No. of Modes' (5).
- Type of Assignment:** Four radio buttons: 'Incremental', 'User Equilibrium' (which is selected), 'System Optimum', and 'Time of Day UE'.
- Assignment Option:** Five checkboxes: 'Route Search by Mode' (checked), 'Impedance between Zones' (checked), 'Peak / Average' (radio buttons, 'Peak' is selected), 'Link Traffic Direction' (checked), 'Minimum Route Information' (checked), and 'Initial Loading' (checked).
- for Equilibrium Assignment only:** Three input fields: 'Maximum Iteration' (0), 'Limit for Equilibrium Check (%)' (0.0010), and 'Damper Parameter(Damp)' (1.000). Below these is a formula: $T_c = T_c - I + Damp(T_c - T_c - I)$.

On the right side of the dialog, there are 'Save' and 'Close' buttons.

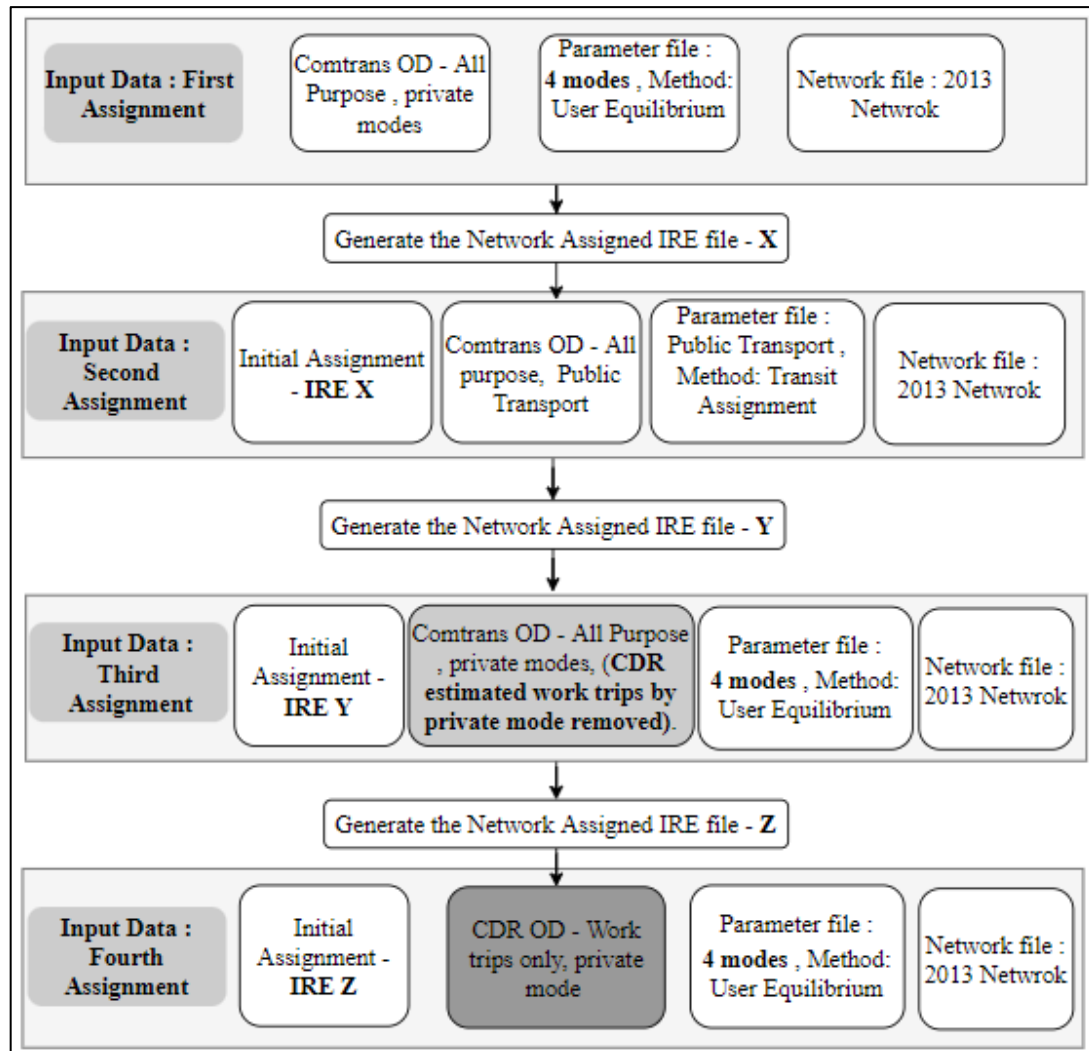
Note: Source: STRADA Software

In the third assignment phase, as illustrated in Figure 5.13, commuting trips undertaken by private transportation modes were selectively excluded from the input origin-destination (OD) matrix. An additional assignment step was introduced, incorporating the commuting trips estimated from Call Detail Record (CDR) data. This step was designed to precisely delineate the travel routes for each trip between specific Divisional Secretariat Division (DSD) pairs.

For this CDR-based trip assignment, an underlying assumption was made that the mode composition of the CDR-derived trips mirrored that of the trips estimated through traditional methods. This assumption was crucial in enabling a direct comparison and subsequent validation of the route choices inferred from CDR data against those derived from existing datasets.

Figure 5-13

Network Assignment Methodology with CDR Data and STRADA,



5.5.2 Results and Validation

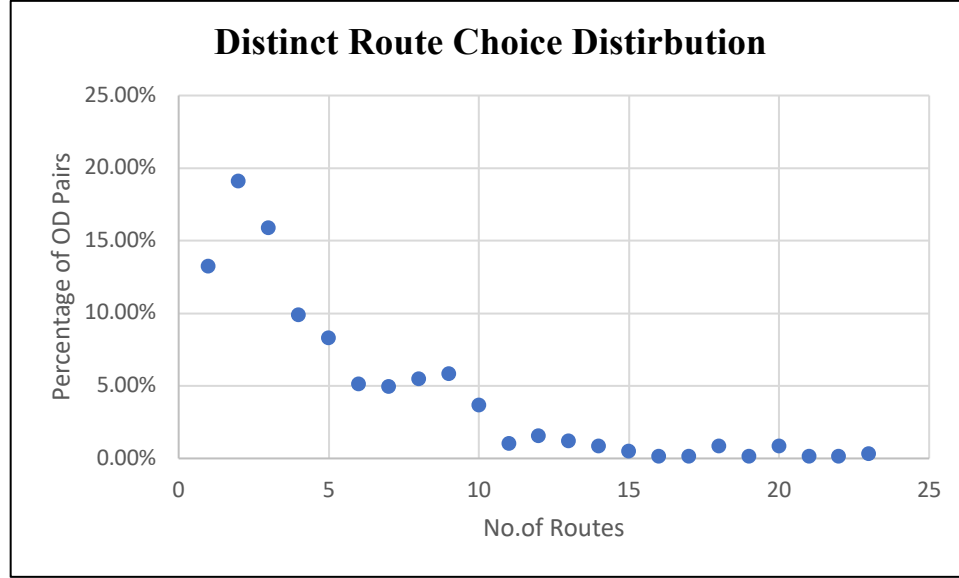
The dataset identified 2,209 origin-destination (OD) pairs, among which commuting trips occurred between 993 pairs. For 565 of these DSD pairs, enroute cell data were available, and notably, 394 pairs recorded more than 100 trips. The assignment of trips between these 394 OD pairs to the road network was carried out utilizing the methodology previously outlined.

Following the assignment process in STRADA, the observed number of route choices ranged from 1 to 20, with the Kalutara to Colombo DSD pair exhibiting the highest diversity, presenting 23 different route choices. Conversely, 52 OD pairs were identified with the

minimum diversity, each offering a single route choice. Figure 5.14 illustrates the distribution of these distinct route choices across the dataset.

Figure 5-14

Distribution of the Route Choices



Removing Implausible Routes

To refine route identification accuracy and mitigate discrepancies arising from the lower resolution of CDR data, STRADA's approach to defining routes was adjusted. Originally, STRADA considered any variation in at least one link between origin-destination (OD) pairs as defining a separate route. This method, however, could lead to inaccuracies when contrasted with the CDR method due to its lower resolution.

To overcome this challenge, a strategy was implemented to amalgamate similar routes. This was accomplished by employing the Jaccard similarity index. This measure calculates the similarity between two sets based on the ratio of intersection size to union size, as denoted in equation 5. By applying this index, routes with considerable overlap in common links were identified and subsequently merged.

$$JI(C, D) = \frac{|C \cap D|}{|C \cup D|} \text{-----}(5)$$

Where,

JI = Jaccard Index,

C = Cells in Route “C”

D = Cells in Route “D”

This adjustment aimed to enhance the match between the STRADA-generated routes and those inferred from CDR data, thus improving the overall accuracy of the route choice analysis by accounting for the inherent limitations of the CDR resolution.

Routes exhibiting over 80% similarity were consolidated for a comparative analysis between CDR and STRADA assignments. This threshold of 80% was chosen arbitrarily to identify routes with substantial overlap. Subsequently, trips were allocated to these plausible routes. Following this assignment, several steps were undertaken to compare the outcomes with data obtained from household visit surveys.

Accordingly, for all DSD pairs OiDj, where $i \neq j$,

$$\text{With CDR : } 1 = \sum_{n=1}^m \left[\frac{R_{n_CDR}}{[\sum_{i=1}^n R_{n_CDR}]} \right] , \quad \text{With HVS : } 1 = \sum_{n=1}^m \left[\frac{R_{n_Strada}}{[\sum_{i=1}^n R_{n_Strada}]} \right]$$

R_{n_CDR} / R_{n_STRADA} - No. of Trips via route n with CDR/STRADA

m - No. of Routes identified for the considered DSD pair (n=1,2,3...m)

i – Origin DSD (1,2,3, ...47)

j – Origin DSD (1,2,3, ...47)

For example, assume three routes between the origin (i) and destination (j), then compare them as shown in Figure 5.15.

Figure 5-15

Sample Route Assignment comparison

$$\begin{array}{c}
 1 = \left(\frac{R_1 \text{ Strada}}{R_1 \text{ Strada} + R_2 \text{ Strada} + R_3 \text{ Strada}} \right) + \left(\frac{R_2 \text{ Strada}}{R_1 \text{ Strada} + R_2 \text{ Strada} + R_3 \text{ Strada}} \right) + \left(\frac{R_3 \text{ Strada}}{R_1 \text{ Strada} + R_2 \text{ Strada} + R_3 \text{ Strada}} \right) \\
 \updownarrow \text{Considered as 1 sample unit} \quad \updownarrow \text{Considered as 1 sample unit} \quad \updownarrow \text{Considered as 1 sample unit} \\
 1 = \left(\frac{R_1 \text{ CDR}}{R_1 \text{ CDR} + R_2 \text{ CDR} + R_3 \text{ CDR}} \right) + \left(\frac{R_2 \text{ CDR}}{R_1 \text{ CDR} + R_2 \text{ CDR} + R_3 \text{ CDR}} \right) + \left(\frac{R_3 \text{ CDR}}{R_1 \text{ CDR} + R_2 \text{ CDR} + R_3 \text{ CDR}} \right)
 \end{array}$$

As illustrated in Figure 5.16, the analysis encompasses 394 Divisional Secretariat Division (DSD) pairs. An illustrative case within the data shows that for a specific route (n) between the DSD pair OiDj, STRADA allocated 66% of the trips, whereas the CDR method allocated 76% of trips from location i to j.

The computed correlation coefficient between the STRADA and CDR allocations stands at 0.78, signifying a statistically significant relationship. A hypothesis test was conducted to assess the robustness of this correlation. The outcome led to rejecting the null hypothesis in favor of the alternative hypothesis, affirming a statistically significant linear relationship between the STRADA and CDR data at the 95% confidence level.

Outliers were initially identified through visual inspection during the analysis. To validate these findings, further outlier detection was conducted using R software, which highlighted several DSD pairs—such as Bandaragama–Kelaniya, Dodangoda–Kaduvela, Gampaha–Dehiwala, Homagama–Panadura, Horana–Kelaniya, Kolonnawa–Dompe, Mirigama–Dehiwala, Negombo–Ratmalana, Padukka–Attanagalla, and Panadura–Kolonnawa—as exhibiting unusually low volumes of work-related trips compared to other pairs. Upon removing these outliers from the dataset, the overall correlation coefficient improved to 0.83, as illustrated in Figure 5.17. Additionally, Figure 5.18 presents a visual comparison between traditional route assignments and those derived from CDR data.

Figure 5-16

Route Assignment comparison with CDR Vs STRADA

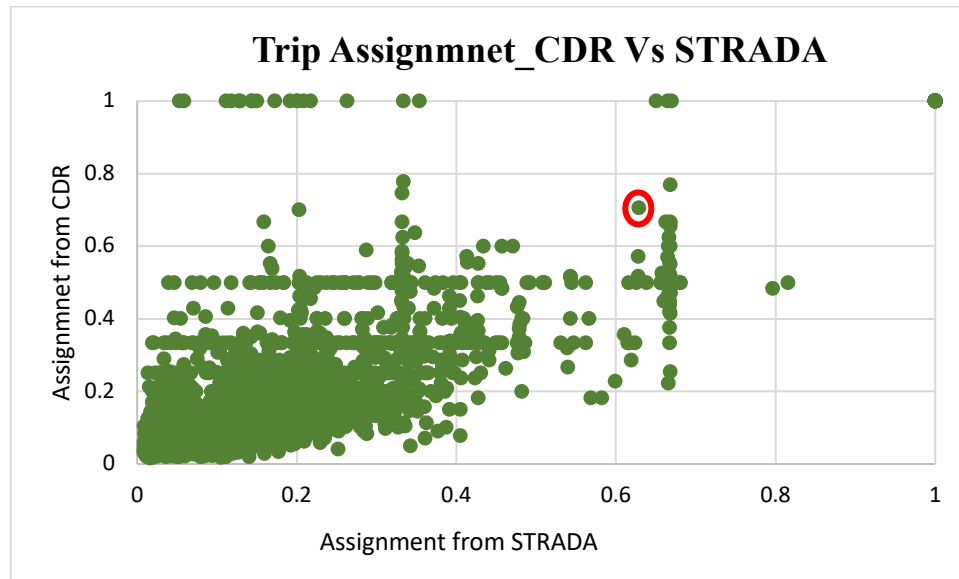


Figure 5-17

Route Assignment comparison of CDR Vs STRADA after outlier detection

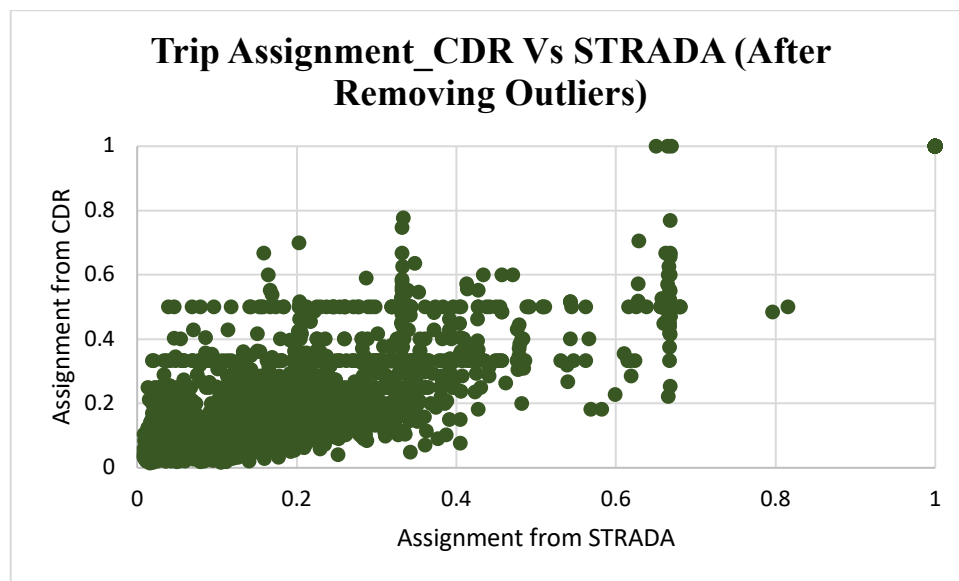
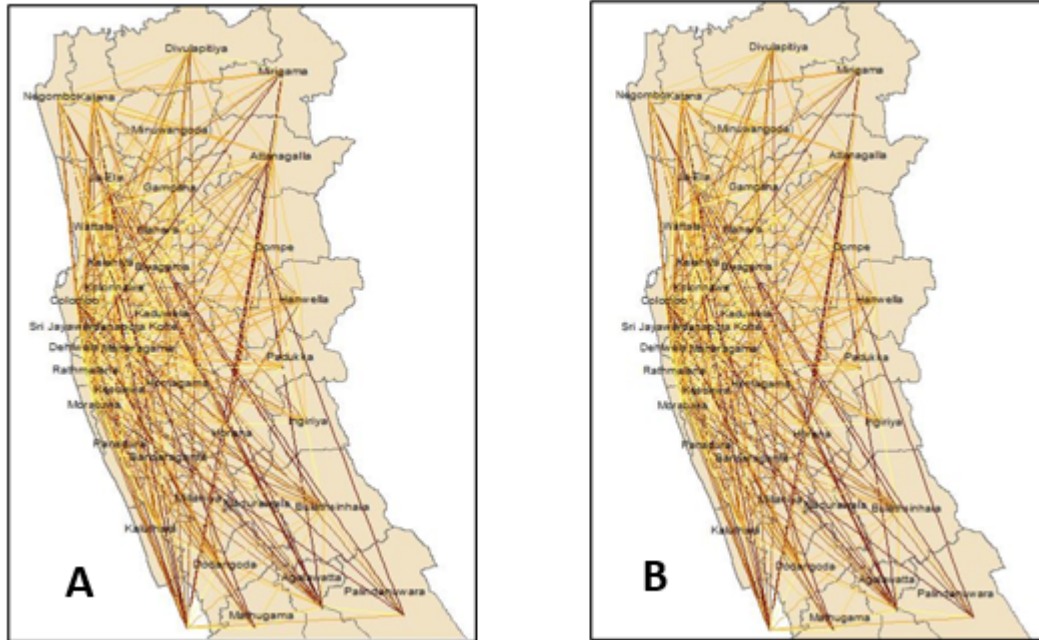


Figure 5-18

Route assignment from CDR Data (A) and Traditional Data (B)



5.6 Summary and conclusion

This chapter introduces an innovative methodology designed to identify subscribers' commuting routes between designated areas of interest, leveraging Call Detail Record (CDR) trips to trace the observed journeys from home to work locations. A CDR commute trip is defined as the documented movement of a subscriber from their residential to their occupational location. This methodology innovatively utilizes specific trip statistics derived from CDR data to align CDR trips with the existing road network, facilitating the assignment of commute trips to this network through a novel approach. The efficacy of this methodology was validated by comparing the route assignments generated via this new method with those obtained through conventional network assignment techniques. The comparison yielded a satisfactory validation level, underscoring the effectiveness and potential applicability of the proposed methodology for future trip assignment endeavors.

CHAPTER 6

SUMMARY AND CONCLUSION

This study significantly contributes to large-scale mobility estimation using call detail records (CDR) from mobile telephony, offering a methodology that facilitates applications such as location labeling, trip estimation, and travel route identification. Throughout the development of this methodology, the research addresses a range of challenges, including the extraction of mobile features, managing computational complexity, navigating localization uncertainty, and safeguarding privacy.

The dissertation begins with a succinct review of applications to monitor human activities. It progresses to discuss various strategies for extracting meaningful spatiotemporal data and provides an overview of contemporary research that employs such data in four-step demand modeling processes. Furthermore, it assesses the potential of leveraging a modern mobile telephony network as a robust platform for sensing and analyzing human activity patterns, corroborating these findings through validation with ground-truth data.

6.1 Utilizing CDR for the Four-Step Demand Modeling Process

Utilizing Call Detail Records (CDR) in the four-step travel demand modeling process leverages the extensive data contained within CDRs to analyze and comprehend travel patterns, behaviors, and demands.

In the trip generation phase, CDRs offer invaluable insights into the spatial and temporal patterns of mobile phone usage, facilitating the estimation of trip origins across various locations. By capturing the origins and destinations of individuals' calls and text messages, CDRs shed light on trip generation by analyzing patterns such as call and text frequencies, call durations, and timestamps. This analysis aids in estimating trip volumes originating from distinct zones or areas and identifies variations in travel behavior influenced by time of day, week, or other factors. Through the examination of call patterns and user mobility, CDRs contribute to enhancing the accuracy of trip-generation models.

During the trip distribution step, CDRs provide insights into the destinations of mobile phone users, offering a deeper understanding of travel patterns and assisting in modeling the flow of trips between different zones. This data is instrumental in determining the likelihood of trips between various origins and destinations. CDRs enhance trip distribution analysis by revealing the spatial distribution of communication events, thus identifying prevalent travel patterns and the flow of trips between regions or zones. This facilitates the estimation of trip attraction and generation for specific destinations and origins while highlighting temporal variations in trip distribution, such as peak travel periods and congestion hotspots.

Although CDR data may not directly reveal the mode of transportation used, they indirectly inform mode choice analysis when combined with other data sources, such as transportation surveys or GPS data. Analyzing movement speeds inferred from CDR data makes it possible to distinguish between walking, cycling, and motorized modes. However, the current study needs to address this step due to limitations in data frequency.

CDR data also contribute to traffic assignment models by offering insights into travel patterns and route choices. Analyzing the timestamps and locations of calls and messages enables estimating travel times between different zones or segments of a transportation network. This information is crucial for calibrating and validating traffic assignment models and understanding the distribution of traffic flows across various routes and road segments.

Incorporating CDRs into the four-step travel demand modeling process thus enables a more comprehensive understanding of travel behavior, enhancing the accuracy and reliability of travel demand forecasts. By capitalizing on the spatial and temporal information in CDR data, transportation planners and researchers can improve the precision and detail of their models, leading to more effective transportation planning and decision-making processes.

6.2 Validation with HVS Data

The validation of findings derived from Call Detail Record (CDR) data is a crucial process for assessing the accuracy and reliability of the information obtained. Ensuring the validity of these findings is essential for their application in further research and planning activities. As discussed in previous chapters, the analysis was conducted in three primary stages: Trip Generation/Attraction, Origin-Destination (OD) Estimation, and Network Assignment.

In the Trip Generation phase, an initial comparison was made between the estimated trip rates and purposes derived from the CDR data and those obtained from Household Visit Survey (HVS) data. This comparison aimed to verify the consistency of trip generation metrics across both data sources.

During the OD Estimation phase, the spatial patterns of trip distribution, as determined by the four-step model using CDR data, were compared with the trip distribution observed in the HVS data. This step assessed whether the distribution of trips between different zones or areas was consistent or if discrepancies in trip flow patterns were evident.

At the Network Assignment phase, the four-step model's traffic flow patterns and route choices were compared with the observed traffic flow patterns derived from the HVS data. This comparison aimed to determine if the traffic assignment results based on CDR data closely resembled the actual traffic flow patterns observed in the HVS data.

By conducting these comparisons, it is possible to assess the level of agreement and accuracy between the data generated from the four-step model using CDR and the HVS data. Given the high similarity observed at the Divisional Secretariat Division (DSD) and District levels and an acceptable model fit at each step, the four-step model employing CDR data provides a reliable and valid representation of travel behavior. This conclusion underscores the potential of CDR data as a valuable resource for transportation planning and the study of travel behaviors.

6.3 Application of CDR-based Findings

The utilization of Call Detail Records (CDR) significantly enriches travel demand modeling by providing detailed insights into trip characteristics such as origins, destinations, and frequencies. As elucidated in Chapter 5, CDR data facilitates the estimation of trip rates, identification of trip purposes, and understanding of temporal variations in travel behavior. This enhances the precision of trip generation models and ensures a more accurate representation of actual travel patterns.

Furthermore, trip spatial distribution patterns derived from analyzing CDR data allow travel demand modelers to delineate common travel corridors, estimate trip flows between zones, and discern variations in travel behavior across different regions. As mentioned previously,

this detailed analysis augments the accuracy of trip distribution models, contributing to the design of more efficient transportation networks.

Chapter 5 discusses the utility of CDR data in providing comprehensive information about individual movements within the network, including travel times, origins, and destinations. This enables transportation modelers to refine traffic assignment models through the calibration and validation against observed CDR-derived traffic patterns, thereby improving the representation of real-world traffic conditions within the models.

Moreover, CDR data illuminates individual route choices and preferences, offering insights into preferred routes, route-switching behavior, and the influence of various factors such as congestion levels and time of day on route selection. This valuable information supports the development of route choice models that accurately reflect real-world routing decisions.

Integrating CDR-based findings with real-time or historical traffic data further enhances dynamic traffic assignment models. By incorporating CDR data, transportation modelers can more accurately depict evolving traffic conditions, adapt route assignments to changing travel demand patterns, and capture real-time traffic flows.

Additionally, Chapter 5 highlights how analyzing the temporal variations, duration, and frequency of communication events captured in CDR data can help identify peak travel periods, analyze travel patterns over time, and understand travel behavior variations across different user segments. This aids in fine-tuning travel demand models to better capture dynamic travel behaviors.

In summary, the application of CDR-based findings significantly advances the field of travel demand modeling by improving its accuracy, detail, and real-world applicability. Leveraging CDR data fosters a data-driven approach to transportation modeling, supporting more informed decision-making in transportation planning, infrastructure development, and policy formulation.

6.4 Summary of Key Contributions of the Study

This study has made several important contributions to the field of transport modeling using Call Detail Records (CDRs), particularly within the context of data-scarce environments like

Sri Lanka. One of the primary contributions is the development of a robust trip detection framework capable of extracting individual-level mobility patterns from temporally sparse and noisy CDR data. By introducing a speed-based filtering technique to address the Load Sharing Effect (LSE) and employing a presence-pattern-based method for identifying stay points, the study enhances the reliability of inferred trips. Building on this foundation, the research proposes a novel travel demand modeling framework that transforms raw CDR data into structured inputs for conventional transport models. This includes the construction of origin-destination (OD) matrices, classification of trip types (such as home-based work and non-home-based trips), and the incorporation of route assignment—a component often overlooked in previous CDR-based studies due to spatial resolution limitations.

A significant methodological advancement of this work is the ability to infer likely travel routes based on tower transition frequencies, temporal patterns, and network proximity, enabling CDR data to be integrated into the route assignment phase of transport models. Additionally, the study introduces a labeling mechanism for classifying trip purposes, thereby aligning CDR-based insights more closely with activity-based modeling principles. To further understand the heterogeneity in user behavior, a segmentation framework was developed that categorizes users based on their caller activity levels and trip frequencies. This behavioral segmentation helps identify biases in CDR datasets and supports more nuanced mobility analysis.

Empirical validation forms another critical contribution, with the CDR-derived results being rigorously compared to Household Visit Survey (HVS) data across three districts in Sri Lanka's Western Province. This validation effort demonstrates the reliability and limitations of CDR-based travel demand estimates and provides confidence in the applicability of the proposed framework. Moreover, the study refines heuristic criteria for identifying geographic and socioeconomic boundaries within the dataset, ensuring that the analytical methods are contextually relevant to the Sri Lankan setting. Overall, the research offers a scalable, cost-effective alternative to traditional survey methods and contributes a comprehensive, adaptable framework for transport analysis in data-constrained environments.

6.5 Limitations of the Study

Adopting Call Detail Records (CDR) for transportation analysis presents a pioneering method, especially in Sri Lanka, where research in this domain is nascent. The need for more related literature, particularly focusing on developing countries such as Sri Lanka, poses challenges for conducting an extensive literature review. Previous research efforts have predominantly centered on developed nations, where socioeconomic conditions, cultural practices, and infrastructural development may need to align with those in Sri Lanka. It is critical to recognize that the direct transposition of methodologies or parameters from studies conducted in dissimilar contexts may lead to diminished accuracy due to these variances.

One significant limitation of CDR data is its reliance on the geographical positioning of cell towers, which inherently lacks the spatial resolution necessary for capturing detailed travel behaviors at a granular level. This limitation is particularly pronounced in capturing movements within neighborhoods or rural areas characterized by wider spacing between cell towers, leading to potential misclassification of intra-cell tower movements as non-movements.

Additionally, CDR data does not inherently distinguish between the modes of transportation used during trips. The challenge of differentiating modes of travel—such as walking, cycling, or motorized transport—based solely on CDR data constrains the development of precise mode choice models. The current study, acknowledging this limitation, does not attempt to determine mode choice.

Furthermore, CDR data must provide direct insights into the purposes behind individual trips. Although this study categorizes trips into three broad purposes (home-based work, home-based other, and non-home-based), a comprehensive understanding of trip purposes—such as work, leisure, or shopping—is essential for nuanced travel demand modeling. Deriving accurate trip purposes necessitates additional data sources or the application of inferential assumptions.

In summary, while using CDR data in transportation analysis offers novel insights, especially in a developing country context like Sri Lanka, it is accompanied by inherent limitations that must be carefully considered. Addressing these challenges requires innovative approaches and

the integration of supplementary data sources to enhance the accuracy and applicability of transportation models developed from CDR data.

6.6 Future Research

Future research avenues leveraging Call Detail Records (CDR) data in transportation studies present promising directions, particularly in mode detection and classification, inferring trip purposes, dynamic origin-destination matrix estimation, multi-modal travel behavior analysis, and predictive modeling for travel demand.

Enhancement of Mode Detection and Classification Algorithms: Advanced machine learning techniques, including deep learning and ensemble methods, hold the potential to significantly improve the accuracy of mode detection and classification using CDR data. Such advancements could contribute to more refined mode choice models and a deeper understanding of travel behavior patterns.

Development of Methods to Infer Trip Purposes: It is recommended that innovative approaches that leverage additional contextual information, such as temporal patterns, user profiles, and points of interest data, to accurately infer trip purposes from CDR data be investigated. Enhanced insights into travel motivations could support more targeted transportation planning and policy-making.

Estimation of Dynamic Origin-Destination Matrices: Exploring methodologies that enable the estimation of dynamic origin-destination matrices using CDR data is another promising research direction. Such methods should account for uncertainties and variations in travel patterns influenced by factors like time of the day, seasonality, and special events. Accurate dynamic origin-destination estimation can improve transportation demand forecasting and infrastructure planning.

Analysis of Multi-Modal Travel Behavior: Integrating CDR data with other datasets, such as public transportation records or social media data, offers opportunities to analyze multi-modal travel behavior comprehensively. This research direction could inform the development of integrated transportation systems and enhance transportation accessibility.

Predictive Models for Travel Demand Forecasting: Further investigation into predictive models that utilize CDR data, alongside machine learning algorithms and data fusion techniques, could enhance the accuracy of future travel demand forecasts. Incorporating additional data sources, such as weather conditions or event schedules, may enrich these models. Additionally, exploring the application of CDR data in optimizing transportation network operations, including dynamic traffic signal control or demand-responsive transit systems, represents a valuable area for future research.

Collectively, these research directions underscore the potential of CDR data in advancing transportation studies, offering insights that could lead to more efficient, responsive, and user-centered transportation systems.

REFERENCES

- Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1), 3–27. <https://doi.org/10.1080/10630731003597306>
- Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, 240–250. <https://doi.org/10.1016/j.trc.2015.02.018>
- Aniello, S., Donna, D., Cantelmo, G., & Viti, F. (2015). *A Markov chain dynamic model for trip generation and distribution based on CDR*. June, 3–5.
- Arai, A., & Shibasaki, R. (2013). Estimation of human mobility patterns and attributes analyzing anonymized mobile phone CDR: Developing real-time census from crowds of greater dhaka. *CEUR Workshop Proceedings*, 1136.
- Bollverk, O., & Hadachi, A. (2023). *CDR-based trajectory reconstruction of mobile network data using transformers*. In 2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (pp. 1–6). IEEE. <https://doi.org/10.1109/MT-ITS56129.2023.10241747>
- Bonnel, P., Hombourger, E., & Olteanu-raimond, A. (2015). Passive mobile phone dataset to construct origin-destination matrix : potentials and limitations. *Transportation Research Procedia*, 11, 381–398. <https://doi.org/10.1016/j.trpro.2015.12.032>
- Breyer, N., Gundlegård, D., & Rydergren, C. (2018). Cellpath Routing and Route Traffic Flow Estimation Based on Cellular Network Data. *Journal of Urban Technology*, 25(2), 85–104. <https://doi.org/10.1080/10630732.2017.1386939>
- Bwambale, A., Choudhury, C. F., & Hess, S. (2017). Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*, June, 1–11. <https://doi.org/10.1016/j.jtrangeo.2017.08.020>
- Chang, D. M., & Dresser, G. B. (1990). *a Comparison of Traffic Assignment Techniques*.

<https://static.tti.tamu.edu/tti.tamu.edu/documents/1153-3.pdf>

- Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. (2015). Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities. *Transportation Research Record*, 2526, 126–135. <https://doi.org/10.3141/2526-14>
- Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J. C., Huens, E., Van Dooren, P., Smoreda, Z., & Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and Its Applications*, 392(6), 1459–1473. <https://doi.org/10.1016/j.physa.2012.11.040>
- Demissie, M. G., & Antunes, F. (2016). Phone Data : A Case Study of Senegal. *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1–6.
- Demissie, M. G., Phithakkitnukoon, S., & Kattan, L. (2019). Trip Distribution Modeling Using Mobile Phone Data: Emphasis on Intra-Zonal Trips. *IEEE Transactions on Intelligent Transportation Systems*, 20(7), 2605–2617. <https://doi.org/10.1109/TITS.2018.2868468>
- Demissie, M. G., Phithakkitnukoon, S., Kattan, L., & Farhan, A. (2019). Understanding human mobility patterns in a developing country using mobile phone data. *Data Science Journal*, 18(1), 1–13. <https://doi.org/10.5334/dsj-2019-001>
- Democratic Socialist Republic of Sri Lanka Ministry of Transport, Japan International Cooperation Agency, & Oriental Consultants Co., Ltd. (2014). *Urban transport system development project for Colombo Metropolitan Region and suburbs: Urban transport master plan final report (Main report)*. Ministry of Transport, Sri Lanka.
- Dong, H., Ding, X., Shi, Y., Jia, L., Qin, Y., & Chu, L. (2014). *Urban Traffic Commuting Analysis Based on Mobile Phone Data*.
- Fekih, M., Bellemans, T., Smoreda, Z., & Bonnel, P. (2020). A data - driven approach for origin – destination matrix construction from cellular network signalling data : a case

- study of Lyon region (France). In *Transportation* (Issue 0123456789). Springer US. <https://doi.org/10.1007/s11116-020-10108-w>
- Ghosh, N., Sarkar, U., & Nagesh, P. (2023). Review on Application of Call Details Records (CDRs) Data to Understand Urban Mobility Scenarios for Future Smart Cities. In A. Majumdar, M. Debnath, & S. Das (Eds.), *Urban Commons, Future Smart Cities and Sustainability* (pp. 831–845). Springer. https://doi.org/10.1007/978-3-031-24767-5_36
- Gundlegård, D., Rydergren, C., Breyer, N., & Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95, 29–42. <https://doi.org/10.1016/j.comcom.2016.04.015>
- International Telecommunication Union. (2022). *Guidelines for the use of mobile phone data for official statistics: Handbook*. ITU Publications. https://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-HDB-2022-Mobile.1-pdf-E.pdf
- Iqbal, S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin – destination matrices using mobile phone call data. *TRANSPORTATION RESEARCH PART C*, 40, 63–74. <https://doi.org/10.1016/j.trc.2014.01.002>
- Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying important places in people’s lives from cellular network data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6696 LNCS, 133–151. https://doi.org/10.1007/978-3-642-21726-5_9
- (JICA), J. I. C. A., & ORIENTAL CONSULTANTS CO., L. (2014). *URBAN TRANSPORT SYSTEM DEVELOPMENT PROJECT FOR COLOMBO METROPOLITAN REGION AND SUBURBS*.
- Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38, 122–135. <https://doi.org/10.1016/j.trc.2013.11.003>

- Jiang, S., Ferreira, J., & Gonzalez, M. C. (2016). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, 3(2), 208–219. <https://doi.org/10.1109/tbdata.2016.2631141>
- Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2505821.2505828>
- Khan, F. H., Ali, M. E., & Dev, H. (2015). A hierarchical approach for identifying user activity patterns from mobile phone call detail records. *Proceedings of 2015 International Conference on Networking Systems and Security, NSysS 2015*. <https://doi.org/10.1109/NSysS.2015.7043535>
- Kujala, R., Aledavood, T., & Saramäki, J. (2016). Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Sci*. <https://doi.org/10.1140/epjds/s13688-016-0067-3>
- Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6). <https://doi.org/10.1371/journal.pone.0096180>
- Lind, A., Hadachi, A., & Batrashev, O. (2017). A new approach for mobile positioning using the CDR data of cellular networks. *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, 315–320. <https://doi.org/10.1109/MTITS.2017.8005687>
- Li, Y., Ran, Z., Tsai, L., & Williams, S. (2023). Using call detail records to determine mobility patterns of different socio-demographic groups in the western area of Sierra Leone during early COVID-19 crisis. *Environment and Planning B: Urban Analytics and City Science*, 50(5), 1298-1312. <https://doi.org/10.1177/23998083231158377>
- Luo, T., Zheng, X., Xu, G., Fu, K., & Ren, W. (2017). An improved DBSCAN algorithm to detect stops in individual trajectories. *ISPRS International Journal of Geo-Information*, 6(3), 1–16. <https://doi.org/10.3390/ijgi6030063>

- Luo, X., Zhou, Y., Yang, Y., & Wu, S. (2020). Research on Home and Work Locations Based on Mobile Phone Data. *Journal of Physics: Conference Series*, 1486(5). <https://doi.org/10.1088/1742-6596/1486/5/052013>
- Lwin, K. K., Sekimoto, Y., & Takeuchi, W. (2018). Estimation of hourly link population and flow directions from mobile CDR. *ISPRS International Journal of Geo-Information*, 7(11). <https://doi.org/10.3390/ijgi7110449>
- Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin–destination matrices obtained from CDR data. *Sensors (Switzerland)*, 19(20), 1–17. <https://doi.org/10.3390/s19204470>
- Mamei, M., Colonna, M., & Galassi, M. (2016). Automatic identification of relevant places from cellular network data. *Pervasive and Mobile Computing*, 31(April 2014), 147–158. <https://doi.org/10.1016/j.pmcj.2016.01.009>
- Munghanya, W., Phithakkitnukoon, S., Demissie, M. G., Kattan, L., Veloso, M., Bento, C., & Ratti, C. (2019). Constructing Time-Dependent Origin-Destination Matrices with Adaptive Zoning Scheme and Measuring Their Similarities with Taxi Trajectory Data. *IEEE Access*, 7, 77726–77737. <https://doi.org/10.1109/ACCESS.2019.2922210>
- Naing, N. N. (2003). Determination of sample size. *Malaysian Journal of Medical Sciences*, 10(2), 84–86.
- Ortúzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport. In *Modelling Transport*. <https://doi.org/10.1002/9781119993308>
- Oxford Business Group. (2023). *A competitive market helps form a dynamic telecoms sector in Sri Lanka*.
- Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., & Ratti, C. (2010). Taxi-aware map: Identifying and predicting vacant taxis in the city. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6439 LNCS, 86–95. https://doi.org/10.1007/978-3-642-16917-5_9

- Ricciato, F., Lanzieri, G., Wirthmann, A., & Seynaeve, G. (2020). Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, 68, 101263. <https://doi.org/10.1016/j.pmcj.2020.101263>
- Rodrigues, C., Veloso, M., Alves, A., Bento, C. (2023). Using CDR Data to Understand Post-pandemic Mobility Patterns. In: Moniz, N., Vale, Z., Cascalho, J., Silva, C., Sebastião, R. (eds) Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science(), vol 14116. Springer, Cham. https://doi.org/10.1007/978-3-031-49011-8_35
- Sakamane, P., Phithakkitnukoon, S., Smoreda, Z., & Ratti, C. (2020). Methods for inferring route choice of commuting trip from mobile phone network data. *ISPRS International Journal of Geo-Information*, 9(5), 1–23. <https://doi.org/10.3390/ijgi9050306>
- Santi Phithakkitnukoo. (2022). *Urban Informatics Using Mobile Network Data Travel Behavior Research Perspective*. Springer. <https://doi.org/10.1007/978-981-19-6714-6>
- Shaadya Ismail. (2018). Traffic speed down to a crawl not just in the city. *Sunday Times*.
- Socialist, D., Of, R., Lanka, S. R. I., & Transport, M. O. F. (2014). *Democratic Socialist Republic of Sri Lanka Ministry of Transport Urban Transport System Development Project for Colombo Metropolitan*. 5.
- Tongsino, L., & Muangsin, V. (2018). Exploring Home and Work Locations in a City from Mobile Phone Data. *Proceedings - 2017 IEEE 19th Intl Conference on High Performance Computing and Communications, HPCC 2017, 2017 IEEE 15th Intl Conference on Smart City, SmartCity 2017 and 2017 IEEE 3rd Intl Conference on Data Science and Systems, DSS 2017, 2018-Janua*, 123–129. <https://doi.org/10.1109/HPCC-SmartCity-DSS.2017.16>
- Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *TRANSPORTATION RESEARCH PART C*, 1–16.

<https://doi.org/10.1016/j.trc.2015.04.022>

US Census Bureau. (n.d.). *Understanding the Margin of Errors and the Coefficient of Variance in the American Community Survey*.

Valerio, L., Conti, M., & Passarella, A. (2018). Energy efficient distributed analytics at the edge of the network for IoT environments. *Pervasive and Mobile Computing*, 51, 27–42. <https://doi.org/10.1016/j.pmcj.2018.09.004>

Vanhoof, M., Reis, F., Smoreda, Z., & Ploetz, T. (2018). *Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art*. <http://arxiv.org/abs/1808.06398>

Vieira, M. R., Frías-Martínez, E., Bakalov, P., Frías-Martínez, V., & Tsotras, V. J. (2010). Querying spatio-temporal patterns in mobile phone-call databases. *Proceedings - IEEE International Conference on Mobile Data Management*, 239–248. <https://doi.org/10.1109/MDM.2010.24>

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87(December 2017), 58–74. <https://doi.org/10.1016/j.trc.2017.12.003>

Wang, M. H., Schrock, S. D., Vander Broek, N., & Mulinazzi, T. (2013). Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data. *International Journal of Intelligent Transportation Systems Research*, 11(2), 76–86. <https://doi.org/10.1007/s13177-013-0058-8>

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012). Understanding road usage patterns in urban areas. *Scientific Reports*, 2. <https://doi.org/10.1038/srep01001>

Yang, P., Zhu, T., Wan, X., & Wang, X. (2014). Identifying Significant Places Using Multi-day Call Detail Records. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2014-Decem*, 360–366. <https://doi.org/10.1109/ICTAI.2014.61>

- Yin, M., Sheehan, M., Feygin, S., Paiement, J. F., & Pozdnoukhov, A. (2018). A Generative Model of Urban Activities from Cellular Data. *IEEE Transactions on Intelligent Transportation Systems*, 19(6), 1682–1696. <https://doi.org/10.1109/TITS.2017.2695438>
- Zagatti, G. A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C. J., Albert, M., Gray, J., Antos, S. E., Burci, P., zu Erbach-Schoenberg, E., Tatem, A. J., Wetter, E., & Bengtsson, L. (2018). A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. *Development Engineering*, 3, 133–165. <https://doi.org/10.1016/j.deveng.2018.03.002>
- Zhao, Z., Shaw, S. L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9), 1738–1762. <https://doi.org/10.1080/13658816.2015.1137298>

Appendix A – Sample Representation from each DSD

	DSD_name	Population	Ratio of Population	Calculation	Sample required from each DSD
Base DSD	Madurawala	34,381	1	1*2330	2330.00
	Colombo	323,257	9	9*2330	21907.12
	Kolonnawa	191,687	6	6*2330	12990.63
	Kaduwela	252,041	7	7*2330	17080.82
	Homagama	237,905	7	7*2330	16122.82
	Hanwella	113,807	3	3*2330	7712.70
	Padukka	65,267	2	2*2330	4423.14
	Maharagama	196,423	6	6*2330	13311.58
	Sri Jayawardhanapura kotte	107,925	3	3*2330	7314.08
	Thimbirigasyaya	238,057	7	7*2330	16133.12
	Dehiwala	88,962	3	3*2330	6028.95
	Ratmalana	95,506	3	3*2330	6472.44
	Moratuwa	168,280	5	5*2330	11404.33
	Kesbewa	245,232	7	7*2330	16619.37
	Negambo	142,136	4	4*2330	9632.55519
	Katana	235,291	7	7*2330	15945.66854
	Divulapitiya	144,506	4	4*2330	9793.170065
	Mirigama	164,580	5	5*2330	11153.58483
	Minuwangoda	178,331	5	5*2330	12085.48995
	Wattala	175,525	5	5*2330	11895.32736
	Ja-Ela	201,521	6	6*2330	13657.076
	Gampaha	197,667	6	6*2330	13395.89046
	Attanagalla	179,565	5	5*2330	12169.11812
	Dompe	154,005	4	4*2330	10436.91719
	Mahara	207,782	6	6*2330	14081.3839
	Kelaniya	137,339	4	4*2330	9307.462552
	Biyagama	186,585	5	5*2330	12644.86344
	Panadura	182,285	5	5*2330	12353.45249
	Bandaragama	109,236	3	3*2330	7402.922544
	Horana	113,364	3	3*2330	7682.67706
	Ingiriya	53,896	2	2*2330	3652.53134
	Bulathsinhala	64,600	2	2*2330	4377.941305
	Madurawala	34,381	1	1*2330	2330
	Millaniya	52,176	2	2*2330	3535.966958
	Kalutara	159,697	5	5*2330	10822.66397
	Beruwala	164,969	5	5*2330	11179.94735
	Dodangoda	63,960	2	2*2330	4334.568512
	Mathugama	81,286	2	2*2330	5508.751345
	Agalawatta	36,669	1	1*2330	2485.057735
	Palindanuwara	50,801	1	1*2330	3442.783223
	Walallavita	54,628	2	2*2330	3702.138972
Total					398861.02

Appendix B – Household Survey Questionnaire

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

1st. TRIP		Official Use			
(1) Origin (Please put ☒)		Landmark/ Famous Building near by:			
☐ 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)		House No. and Street Name:			
(2) Origin Type Please Select from Table 1		Village/ Town :			
Others ()		GN division :			
(3) Departure Time : ☐ A.M. ☐ P.M.		D/S division :			
In case of Colombo Municipality, Please fill in Postal Zone (01-15)		District			
(4) What was your Trip Purpose? (Please put ☒)		Official Use			
☐ 1. To Work ☐ 4. Business ☐ 7. Others (specify _____) ☐ 2. To School ☐ 5. Private Matters ☐ 3. Shopping ☐ 6. To Home					
(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-Summary of the Literature Review

Summary of OD Matrix Estimation

	Reference	Methodology	Data	Key findings
1	(Ghosh, Sarkar, and Nagesh 2023)	Identifies users' home and work zones by analyzing the most frequented cell tower locations during nighttime and daytime, respectively. It also distinguishes between stationary and pass-by locations. Based on this, Origin-Destination (OD) flows are computed. The study applies a population expansion factor to scale up sample-based OD flows to represent the entire population, enabling broader applicability of the results. The accuracy of these expanded OD flows is assessed using log-linear and gravity models, comparing both inter- and intra-district flows and trip distance distributions.	Anonymized CDRs from Senegalese mobile phone users, Timeframe: 2-week period, Includes: Voice and SMS data.	Estimates OD flows using limited-resolution CDR data by identifying significant user activity zones. Log-linear and gravity models help evaluate the accuracy of expanded OD matrices.

2	(Phithakkitnukoo, 2022)	Initially, the method identifies users' home and work zones by analyzing the most frequented cell tower locations at night and during the day. It also detects stationery and pass-by locations. Following this, Origin-destination (OD) flows are calculated based on the gathered information. Finally, the study expands the data using a population expansion factor, linking sample user findings to the general population.	Anonymized CDRs from Senegalese mobile phone users for two weeks (voice and short message service data).	Examine various log-linear and gravity models for treating total flow (inter and intra-district flows), as well as trip distances. OD flow derived from sample users are enlarged to the general population and compared.
---	-------------------------	---	--	---

3	(Fekih et al., 2020)	Trips are identified from users' spatiotemporal data by applying a minimum stationary time criterion to define activity zones. An The origin-destination (O-D) matrix is then created and scaled to represent the entire population using a scaling factor.	2G and 3G signalling records of over two million anonymous mobile phone users were recorded in June 2017 for one day.	Introduce a workflow of steps to generate origin-destination (O-D) matrices from 2G and 3G cellular network signalling data. Derive an expansion factor to represent the total population based on the number of residents estimated by the cellular signalling data
4	(Mamei et al., 2019)	Trips are identified from cell movement patterns or inferred habitual routes, emphasizing frequently visited locations. These are aggregated to construct the Origin-Destination (O-D) matrix.	CDR of 20,000 users for three months	Compute time-based matrices and routine-based matrices associated with the home-work commute from CDR data and project those matrices onto the actual road network.

5	(Bwambale et al., 2017)	The model predicts individuals' demographic group affiliation using phone usage patterns, applying these predictions as weights in a latent class model to generate trips. Each class corresponds to distinct socio-demographic groups.	Call logs and the GSM cells data (mobility data) of 158 participants with known demographics.	Propose a hybrid trip generation framework using a small representative section of willing users and calibrate that to the full CDR population.
6	(Kujala et al., 2016)	Computed smoothed inter-observation time distributions generate the maxima and finally detect the peak corresponding to typical travel time.	Twenty-five anonymized mobility data sets. Each set contains 300,000 mobility traces for two weeks.	Introduced a method for extracting typical travel times between cities from CDR-based mobility data.
7	(Gundlegård et al., 2016)	Trips were defined using a sequence of cell paths, specific origins, and destinations generated from CDR data. Identified trips were scaled using time-dependent travel demand scaling.	Calls and SMS of 50,000 users for two weeks	Propose a computational pipeline including trip extraction, travel demand estimation, and link travel flow estimation

8	(Aniello et al., 2015)	Identified the traffic zones using k-means clustering analysis. Then, the traffic between different clusters is analysed to identify the zone type by evaluating the traffic call density for each time window. Markov Chains theory evaluates the demand propagation among the different zones.	Hourly antenna to-antenna traffic voice every month for one year and CDR of 300.000 of randomly selected users.	Develop a model that enables to predict hourly CDR variations in time and space, which act as a proxy for the trip distribution.
9		Trips are deduced from the Spatio-temporal trajectory of devices through the stationarity within a Location Area. Trips are then aggregated in an origin-destination matrix and compared with traditional data (census)	Paris region inhabitants of 11 days, consisting of 1.5 billion logs	the traditional and CDR OD matrices have little in common. Further, Estimation of the location of the home and place of work or study of the individuals from mobile phone data is essential to improve the matrices produced significantly.

10	(Alexander et al., 2015)	Records are first converted into clustered locations where users engage in activities for an observed duration. These locations are inferred to be home, work, or other depending on observation frequency, day of week and time of day. Trips are then constructed for each user between two consecutive daily observations. These trips are multiplied by expansion factors based on the population of a user's home Census tract and evaluated the average daily trips.	Triangulated CDR of 2 million uses for two months.	Estimate the average daily origin-destination trips from triangulated mobile phone records.
----	--------------------------	--	--	---

11	(Iqbal et al., 2014)	CDR is used to generate tower-to-tower transient OD matrices for different periods. These are then associated with corresponding nodes of the traffic network and converted to node-to-node transient OD matrices. The actual OD matrices are derived by scaling up these node-to-node transient OD matrices.	6.9 million users CDR data over a month. Video data collected from 13 key locations of Dhaka city network over three days for validation	Generate node-to-node t-ODs and present that the flows between adjacent nodes are substantially higher than those between non-adjacent nodes with CDR data.
12	(Dong et al., 2014)	Use the K-means clustering algorithm to divide the traffic zone. The attribution of traffic zones is identified using mobile phone data. Then, the OD distribution and the commuting travel is analysed.	CDR data of 10 million users for a daytime on Friday.	Cell phone base station data can be used to identify OD characteristics, such as such as real-time traffic, the inflows, outflows, and the increment of each period.

Mobility Pattern Recognition and Significant Location Identification

	Reference	Methodology	Data	Key findings
1	(Rodrigues et al., 2023)	Analysis of population mobility patterns by identifying movements and stops at the cell antenna level using anonymized Call Detail Records (CDRs). Comparison of mobility data from two time periods: just after lifting COVID-19 measures and the same period the following year.	Two anonymized mobile phone CDR datasets covering Portugal, spanning the months immediately following the first wave of the pandemic and the same months in the subsequent year.	Significant differences in mobility patterns were observed between the two periods, demonstrating the impact of COVID-19 restrictions and their easing. The study highlights the effectiveness of using CDRs to analyze population mobility changes during pandemic contexts.
2	(Li et al., 2023)	Combined radius of gyration, movement regularity, and motif types to analyze mobility patterns from CDR data. Applied K-Means clustering on these mobility metrics to identify behavior types, then correlated the results with census-based socio-economic indicators.	Lower socio-economic groups maintained higher mobility before and after lockdowns, exposing them to greater COVID-19 risk. Higher-status groups reduced mobility significantly and did not return to pre-lockdown levels. Study shows CDR can reveal socio-demographic differences	CDR data from Africell covering February to April 2020 in the Western Area of Sierra Leone; active users (92%) from 3 sample weekdays; census data for socio-economic classification.

			in response to health crises.	
3	(Luo et al., 2020)	Mobile phone CDR is combined with road network data to identify Trajectories aim to decrease base station positioning errors. Next, the trajectory regularity is measured, and home and work locations are estimated using the regularity.	A CDR data set of 350,000 users for six months.	Propose a framework to identify home and work locations.
4	(Ricciato et al., 2020)	Analyse CDR in a general and modular methodological structure. Geo-location modules: Decrease the spatial uncertainty. Followed by space-time interpolation and density inference.	All individuals signaling records across the whole Belgium for a single day.	Develop a methodological framework for processing CDR to estimate the spatial density of the present population.

5	(Vanhoof et al., 2018)	Evaluate Home Location identification using five different algorithms based on the frequency of calls, using the maximal no. of distinct days, Time window method, Spatial perimeter method and a combination of spatial perimeter and time window methods.	A CDR data set of 18 million users for six months.	Construct a framework to assess the spatial uncertainty related to the application of home detection algorithms.
6	(Yin et al., 2018)	Inferred the activity patterns, including primary, secondary, and heterogeneous activity transitions, using an unsupervised generative state-space model.	A CDR data set of 1 million users for a month.	Develop input-output hidden Markov models to infer travelers' activity patterns from CDRs.

7	(Tongsinoot & Muangsin, 2018)	The work location detection method uses both CDR and G-CDR data to find a grid cell that has regular usage The user is in the grid cell based on the number of days and hour slots. The home location detection method is based on detecting sleeping time and then, I found the last call before bedtime and the first call after waking up. If there is high usage at any grid within ± 4 hours from the bedtimes, the possibility of being the home location is high.	A CDR data set of 4,340,833 users for three months.	Develop a method to exploit daily aggregated internet usage data (G-CDR) in addition to work and home location identification.
8	(Zagatti et al., 2018)	Meaningful locations of callers were identified using a non-supervised learning algorithm. The identified locations were distributed in a grid cell to aggregate the individual data and to create OD matrices of weighted connections between home and work locations.	A CDR data set of 2 billion records for three months.	Investigate night and daytime population densities and commuting patterns.

9	(F. Wang & Chen, 2018)	Locational uncertainty is addressed by clustering spatially close cell towers, which could also be represented by clusters containing towers that frequently channel call activities. Oscillations were treated using a time window-based method.	Triangulated CDR of one million users for a 1-month period.	Propose a data processing framework to address two key issues associated with the CDR data: locational uncertainty and oscillation.
10	(Lind et al., 2017)	Developing an algorithm allows estimating the location of mobile subscribers by knowing only the network coverage cell to which they are connected using an enhanced Kalman filter featured with integrated mobility models.	GPS and CDR data of 271 users. ** Period	Propose a new approach for estimating the location of mobile subscribers within the coverage area of a mobile network.
11	(Mamei et al., 2016)	Cluster CDRs together based on the spatial parameter and weigh clusters based on time to determine those associated with frequented places.	A CDR data set of 15 million users for 12 months.	Propose a method to identify frequented places from cellular network data.

12	(Jiang et al., 2016)	The phone users' home location was initially detected, and statistically representative samples were selected from the parsed CDR data. Derived an expansion factor for the filtered phone observations by combining processed phone data and the census data.	3.17 million anonymous mobile phone user's data for a two-week period.	Extract individual mobility patterns comparable to the activity-based approach, including identification of population, daily trip and motif scaling.
13	(Khan et al., 2015)	Raw CDR data are used as the input in the initial layer and the output from each consecutive layer is used as new input combined with the original CDR data in the next layers to find more detailed facts like traffic density in different areas on working days and holidays.	CDR Data of 6.9 million users for one month	Proposed a hierarchical analytical model where CDR Dataset is used to find facts on the daily life activities of urban users in multiple layers.

14	(Çolak et al., 2015)	Labelled home- and work-based according to frequency and time of day. Then, by using the labels (home, work, or other) of consecutive stays, trips were assigned to purposes such as home-based work. Next the resulting trip pairs were expanded for the total population from census data.	A CDR data set of 2.8 million users for 5 months and a triangulated CDR data set of 2 million users for two months.	Shows that trip correlation is significantly higher when aggregated to larger polygons, and CDR data are not representative of the population for inferring trips between smaller census tracts or Traffic Analysis zones.
15	(Yang et al., 2014)	Infer the trajectories using the stop and move state of users. Then adopt DBSCAN clustering algorithm to look for significant places using multi-day trajectories.	A CDR data set of 10 million users for 78 days.	Proposes a novel method for mining trajectory and identifying significant places.
16	(Kung et al., 2014)	Individuals' travel portfolios were constructed using spatial-temporal data, and the work and home locations were defined. Compute the commute distance as the great circle distance between the home and work locations, determining the morning/evening commute times.	Examine five different CDR data sets of 5 countries with varying periods and sample rates	Identified the work-home locations and shows that the home-work time distributions and average values within a single region are largely independent of commute distance or country despite substantial spatial and infrastructural differences.

17	(Järv et al., 2014)	Activity spaces of all users were measured using the number of activity locations that each user possessed. Using that, the mean monthly spatial behaviour was measured.	CDR of 1310 mobile phone users for one year	Explored that there are limited monthly variations in the number of activities locations but significant variations in the sizes of individual activity spaces. Discovered how seasonality affects the monthly variation in individual spatial behaviour' and individual monthly spatial behaviour is predominantly explained by intrapersonal variability.
18	(Arai & Shibasaki, 2013)	Demographic attributes were estimated using interview data as training data. Moreover, CDR is used as test data employing SVM. Along with SVM, demographic attributes of CDR are separately estimated employing non-negative matrix factorization (NMF) and semi-supervised machine learning.	CDR data of 7 million people for six months. Interview-based Origin-Destination survey which includes demographic attributes and one-day travel-activity records of 75,000 people.	Develops real-time Census of Greater Dhaka from CDR represents the population composition of Greater Dhaka and is labelled with demographic attributes such as sex, age groups, and occupational types

19	(Isaacman et al., 2011)	Critical location identification methodology is based on clustering and logistics regression.	A CDR data set of 100,000 users for 78 days.	Identify meaningful locations, such as home and work.
20	(Vieira et al., 2010)	The regular-expression query language is employed on CDR data that allows for any combination of spatial and temporal predicates with constraints, including variables like the layout of the areas covered by the cellular towers and "areas" that label places of interest	CDR data of 800,000 users for six months.	Presents a Spatio-Temporal Pattern System (STPS) to query spatio-temporal patterns in large CDR databases.
21	(Ahas et al., 2010)	Initially, home and work anchor points were identified by assessing the proportion of days spent at an anchor point. Secondary Anchor Points were also identified using the same process of prioritizing appearance.	CDR data of 0.5 million anonymous respondents for 12 months.	Develop a model to identify meaningful places for mobile telephone users, such as home and work anchor points.

Route Assignment

	Reference	Methodology	Data	Key findings
1	(Bollverk & Hadachi, 2023)	Dual-Transformer framework: first Transformer performs map matching on GPS traces, second Transformer reconstructs road segments from sparse cell-level data.	Synthetic CDR dataset generated from GPS traces of two individuals in an urban area.	Demonstrates feasibility of using Transformers for CDR-based trajectory reconstruction. Performance acceptable but limited by data sparsity and population generalizability.
2	(Sakamane et al., 2020)	Methodology based on interpolation of route waypoints, the shortest distance between a route choice and mobile usage locations, and Voronoi cells that assign a route choice into coverage zones.	12 months of CDR data of 110,213 users	Propose a methodology to use CDR data to infer individual route choices for commuting trips.

3	(Breyer et al., 2018)	Consider all trips made between the specific origin and destination. The cell paths of the trips are sorted by their number of occurrences. For every OD-pair, only the five cell paths that have been detected, and most often, they are used, and a route is estimated for each cell path. The flow is assigned to the routes based on the proportion of occurrence.	320,000 unique users CDR data for two weeks	Propose a method that aims to reconstruct the used route through a transportation network from call detail records (CDRs)
4	(Lwin et al., 2018)	Generate hourly OD pairs of individuals and found the shortest paths between the pairs. Converted the Cell ID locations to the nearest road nodes and assigned the trips.	One week of CDR data of 20 million subscribers.	Estimate hourly link population and flow directions based on mobile-call activities of origin-destination pairs
5	(Toole et al., 2015)	OD Flows are assigned to road networks constructed from OpenStreetMap data using an incremental traffic assignment algorithm	Three weeks of call detail records (Triangulated data) made by all users of a particular carrier	Present an online, interactive visualization platform to communicate the network assignment,

