DATA ANALYTICS BASED MODEL TO ESTIMATE RIDE-SHARING POTENTIAL IN SRI LANKA

Nelum Suhashini Weerakoon

(148243X)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

December 2017

DATA ANALYTICS BASED MODEL TO ESTIMATE RIDE-SHARING POTENTIAL IN SRI LANKA

Nelum Suhashini Weerakoon

(148243X)

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

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

December 2017

DECLARATION

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

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

Signature: ………………………...   **Date:** ………………………
Name: Nelum Weerakoon

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

Signature of the supervisor: ………………………...   **Date:** ………………………
Name: Dr. Amal Shehan Perera

Acknowledgements

This research would not have been possible without the support of many people. I would first like to thank my thesis advisor Dr. Amal Shehan Perera for his support and guidance throughout. I would also like to thank all the academic and non-academic staff of the Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa Sri Lanka for their support.

My sincere thanks should go to LIRNEasia and Department of Logistics and Transportation, Faculty of Engineering, University of Moratuwa Sri Lanka for giving me their fullest support by providing me data. Also, thanks to all my friends who supported me.

Finally, I must express my very profound gratitude to my parents and my spouse for providing me with unfailing support and continuous encouragement. This accomplishment would not have been possible without them. Thank you.

Author

Nelum Weerakoon

Abstract

Traffic handling in cities is becoming a major issue worldwide. Everyone is in a hurry to go to work or deliver goods on time. The more the vehicular traffic, the higher the pollution. This is a waste of time, fuel and energy. Currently, there are many types of research being done to come up with a solution to reduce vehicular traffic.

Ride-sharing is one of the potential solutions to reduce the traffic congestion by reducing the number of vehicles entering a city. The idea of Ride-Sharing is to share rides to/from home/work daily, based on home/work locations. Identification of home/work locations is one of the major task in Ride-Sharing to identify potential ride-sharers. Identification of these locations can be done using CDR data. There are models and algorithms that have been introduced by several types of research to identify the home/work locations based on CDR data. However, this has not yet been implemented in Sri Lanka.

The idea of this study is to identify the potential of Ride-Sharing in Sri Lanka using CDR data. End-Point and En-Route Ride-Sharing are considered as the main Ride-Sharing options. Analysis was performed on data collected in 2012/2013 period for 41 cities of the Western Province of Sri Lanka. To identify the home/work locations, the hours between 21.00-05.30 was considered as home hour events and the hours between 10.00-15.00 considered as work hour events. As per the analysis based on the collected data it was identified that there are 72.94% potential ride-sharers and based on the transportation data it was identified that there are 38.43% Private transportation modes users in the selected cities/towns. Hence, it was identified, that there is a potential of implementing Ride-Sharing and it has a high impact on traffic congestion. The decision was mainly based on the number of vehicles entering the cities.

Keywords: Call Detailed Records, Ride-Sharing, Cluster Analysis, Carpooling, Cell Towers, Base Stations, Sri Lanka

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABM	Agent Based Modeling
ATS	Artificial Transportation Systems
CA	Cluster Analysis
CDR	Call Detail Record
CP	Carpooling
CPP	Carpooling Problem
CSV	Comma Separated Value
DSD	Divisional Secretariat Department
DSS	Decision Support System
ITS	Intelligent Transportation Systems
JVM	Java Virtual Machine
MSC	Mobile Switching Centers
RDD	Resilient Distributed Dataset
VKT	Vehicle Kilometers Travelled
VLR	Visitor Location Record

CHAPTER 1: INTRODUCTION

The introductory chapter provides the motivation for analyzing the potential of implementing Ride-Sharing in Sri Lanka along with research objectives, myths, and challenges. Section 1.1 elaborates the motivation of this study in terms of why the implementation of Ride-Sharing is important to Sri Lanka and how traffic can be controlled with ride-sharing. Sections 1.2 highlights the objectives of the research as to what are the key findings that the study focuses on. Section 1.3 enumerates the already identified myths and challenges in the Ride-Sharing models. There are several challenges when implementing Ride-Sharing and those need to be focused when implementing the Ride-Sharing concept. This section discusses those challenges and myths associated with them and how to overcome them. Finally, in section 1.4 summarizes the topics discussed in the chapter and outlines all organization of the rest of the study.

1.1 Motivation

With the development of urban areas, the traffic is becoming a huge issue in day to day life. For the past few decades, the number of vehicle usage has increased rapidly. People tend to use their own vehicles to travel to their day to day work based on the comfort, pride and most of all to perform their tasks on time. Everyone in the society is in a hurry to perform their day to day tasks; such as to report to their workplaces or reaching certain destinations on time, deliver the goods in business and so on. Traveling using the public transportation is tiresome and not much punctual. With these issues, people tend to use their own vehicles or personal taxi rides for traveling purposes as mentioned above. Due to these reasons, the number of vehicles on the road in peak hours has increased and this increases the traffic congestion. With the increasing traffic, performing a task on time has become problematic. Apart from that, this leads to many other issues as well. More the traffic is more the environmental pollution, and more waste of energy and fuel.

Currently, traffic congestion in Sri Lanka has become a huge issue during the peak hours on weekdays and in weekends in major cities. Traveling takes more time or a significant percentage of time than performing a given task. Even though public transportation is available such as trains and buses, availability and the capacity of those are again a question. In terms of capacity, there are more people to travel in a single bus or a train than available space during the peak hours. Also, the punctuality and the schedule of the public transportation media are problematic. Hence, traveling using public transportation is also not comfortable, punctual and safe even though they are cheap. This is a major issue all citizens face in day to day life. Due to these reasons, citizens tend to use their own vehicles or personal taxis to travel for their daily desires.

There are several ideas, concepts and methodologies developed during past few decades to come up with a solution to reduce the vehicular traffic congestion worldwide. There are some ongoing researches to find a solution to this problem. Ride-Sharing, Carpooling and Intelligent Traffic Handling Systems are some of the main research areas addressing this problem. The idea of this research is to propose a Ride-Sharing solution that would be derived from the above three areas in Sri Lanka. Ride-Sharing is one of the suitable solutions for the traffic congestion in Sri Lanka. This addresses the key requirements of people who are traveling using their own vehicles, such as comfort, punctuality, etc. Alternatively, it reduces the number of vehicles with a single rider on the road which indirectly reduces the traffic congestion.

1.2 Research Objectives

Intelligent Traffic Handling is a long-standing concept used in the past few decades to handle the vehicular traffic on busy roads using artificial intelligence (AI) concepts and methodologies. The basic idea of intelligent traffic handling is to identify the busy roads and peak hours of those roads to come up with a dynamic traffic plan. This is a part of the IBM Smart City concept as per IBM [27] and Srivastava et. al [28] as well. IBM has collected roads and traffic-related data with the use of the surveillance cameras. To occupy the data of surveillance cameras it is necessary to implement the cameras in identified locations which will be expensive and needs support from the government. Handling the traffic with an intelligent approach by these collected data can be done in several ways such as; giving the directions of less occupied roads via mobile phones to drivers while they are driving, handling the traffic control systems dynamically and another option is to alert the direction to this traffic police.

Carpooling and Ride-Sharing are some alternatives that can be used to reduce the vehicular traffic in an intelligent and indirect manner by reducing the number of vehicles entering into a city. This concept provides a solution for the parking problems in a city. Carpooling is the concept of sharing the cars of a group of people who has same origin and destination in day to day life. The companies can promote carpooling among their employees who live in neighborhoods or those who travel using the same routes daily. This is a concept that can be implemented in company environments simply so that all the employees find a way to travel to and from work daily without wasting their much time on roads. This topic will be further discussed in Chapter 2, Section 2.2 with published results.

Ride-Sharing is a wide area of carpooling. There, it is not necessary to stick to a certain set of people; so that anyone can share the ride with a person who is willing to share the ride. The idea of Ride-Sharing is two or more people share the ride based on the vehicle capacity from the origin to destination. It can be prearranged Ride-Sharing in terms that set of identified people share the ride daily by coming up with conditions that they have agreed. It also can be implemented as dynamic Ride-Sharing where people find the others who are willing to share the ride dynamically. The ride can be shared among known people as well as unknown people. The security, privacy, sharing the cost and identifying with whom to share the ride with;

are some of the main challenges coming up with this concept. The Ride-Sharing can be done using own vehicles of the participants or by sharing taxis.

There are several methodologies developed and proposed for Ride-Sharing. The idea of this study is to identify the potential of implementing Ride-Sharing and to implement it with more security. According to Cici et al. [3] Collecting CDR data which stands for Call Detail Records of cell phones to identify the locations of the people and cluster them into groups is one of the ideas in the study of Ride-Sharing. CDR data can be used to identify the home/work location of people. Cell phone networks are built using Base stations that are known as BTS. These BTSs are responsible for communicating cell phone devices with the phone network. The locations of the BTSs are defined using latitudes and longitudes. A cell phone in one area that can get the network coverage from one or more BTS. The BTS which gives the coverage to the phone depends on the geographical position and the current network traffic. When an individual makes a call, it will be routed through one of the nearest BTS. CDR is generated when a cell phone makes or receives a call or message. The time, date and the location of the call will be recorded as per the information based on the BTS. An entry of CDR consists of originating cell phone number, destination cell phone number, time and the date call is started, the duration of the call and the used BTS tower.

The data can be collected from CDR as mentioned above and then identify the home/work locations of people by considering the time spent in each location daily for an individual. There are some existing algorithms proposed and developed by several researchers which will be discussed in successive chapters. VLR data also can be used to identify the locations of individuals. VLR stands for Visitor Location Register which is a database handle by mobile companies to which has mobile communications networks. These networks are associated with Mobile Switching Centers (MSC). According to TelecomABC [18], the VLR consists of exact locations of the all mobile subscribers which are based currently in the service area of the MSC.

The main objectives of this study are,

- Identification of potential ride-sharers in the Western Province of Sri Lanka based on the CDR data gathered during the period of 2012/2013.
- Analyze the use of transportation modes in terms of Private/Public in the Western Province of Sri Lanka based on the transportation data gathered during the period of 2012/2013.
- Analyze the potential of implementing Ride-Sharing in the Western Province of Sri Lanka based on the identified potential ride-sharers and the transportation mode usage.
- Propose a dynamic Ride-Sharing model that can be used to implement.

The proposed Ride-Sharing model of this study will use CDR data to identify the home/work locations of people and cluster them into the groups based on the identified home/work location. Cluster analysis will be used for clustering purpose based on some identified rules. The study discusses the existing methodologies, concepts and ideas of intelligent traffic handling, carpooling and mainly on Ride-Sharing. The next few sections will describe the existing methodologies and their identified drawbacks and challenges.

The main objective of this study is to analyze the potential of implementing a Ride-Sharing model for Sri Lanka to reduce the daily increasing traffic congestion. The analysis was performed on CDR data gathered during the period of 2012/2013 in Western Province, Sri Lanka and transportation data of Sri Lanka gathered during the same period has been used as supportive evidence for the analysis to prove the need of implementing the Ride-Sharing concept. The study has come up with an Architecture that can be used to analyze CDR datasets to identify the protentional of Ride-Sharing, along with algorithms and methodologies to identify the home/work locations of CDR subscribers and potential ride-sharers. The study has proposed a Ride-Sharing model that can be implemented along with a pricing mechanism. The architecture, model and the algorithms suggested and used in the study can be used as a reference to solve similar problem worldwide.

1.3 Myths and Challenges

The section discusses the Ride-Sharing myths and challenges. It is vital to identify and understand the challenges that have faced by existing Ride-Sharing systems. Understanding those challenges helps to come up with a better solution.

Considering the traffic control car-sharing is one of the best options found to reduce the number vehicles entering to the city as explained in previous sections. The car-sharing can be categorized into several forms as given by Srivastava [10]. Those are car-sharing, car shuttle, and car-pooling. Car-sharing is known as a way of traveling where two or more people ride in a car. The concept of car shuttle is, when there is a designated driver and one or more people allowed to travel for monetary consideration. Car-pooling is where a group of people travel together from one location to another regularly by rotating the duty of taking their own vehicles. The participants will not change and the vehicles will be rotated.

In the carpooling the driver takes the extra overhead of time and the marginally cost. There are some myths presented by Srivastava [10] for carpooling.

- Myth1: finding riders and drivers can increase the carpooling,
- Myth2: giving money to the driver can increase the carpooling,
- Myth3: carpooling can be considered as the low-cost alternative for a car shuttle.

These myths can be taken into consideration when designing and modeling a carpooling system that can be used in real-time. As given by Srivastava [10] such a system can be started by considering some actions such as;

- Action1: identify a group of people that have similar traveling needs,
- Action2: a way to register to recognize carpools,
- Action3: respect a carpools' schedule,
- Action4: given an incentive to carpoolers at the source, destination, and en-route.

These myths and actions would be helpful for any carpooling system design to consider giving a proper and efficient solution.

The state-of-the-art of Ride-Sharing is described in detail by Furuhata et al. [15]. The adoption of Ride-Sharing is restricted because of the challenges that it has. It is vital to understand the challenges in the widespread use of Ride-Sharing and the nature of the development of the effective Ride-Sharing system. Furuhata et al. [15] have described and characterized the state-of-the-art of existing Ride-Sharing systems and some challenges to adoption to Ride-Sharing. There are matching agencies that handle the Ride-Sharing which matches the riders and the people who are willing to share their ride. As given in Furuhata et al. [15] there are two types of Ride-Sharing; unorganized Ride-Sharing which involves family, colleagues, friends, and neighbors. Basically, the people we know in day to day life. Organized Ride-Sharing is operated by the agencies that provide ride-matching opportunities for the participants without considering any historical records.

The providers of Ride-Sharing services can be classified as; Service Operators who provide Ride-Sharing using their own vehicles and drivers which is known as one-sided matching and matching agencies which provide a facility to match between individual driver and passengers who are willing to share the ride. This is known as two-sided matching. The main challenge identified by Furuhata et al. [15] for the matching agencies are to design a market mechanism that attracts both vehicle providers and the passengers in the market. Some patterns are identified for single passenger cases.

- Pattern1: is identical Ride-Sharing which is the pattern where the origin and the destination of both the driver and the passenger are same.
- Pattern2: is inclusive Ride-Sharing where both the origin and the destination of the passenger are on the way of the drivers' route.
- Pattern3: is called partial Ride-Sharing in which either of origin or destination location of the passenger is on the way of the drivers' route but not both.
- Pattern4: is called detour Ride-Sharing where both the destination and the origin of the passenger not on the original route of the driver.

However, taking a detour route covers both the pick-up and destination of the passenger. These are the main patterns that can be considered in any Ride-Sharing system. The state-of-the-art Ride-Sharing agencies should consider the above patterns when routing, scheduling and pricing the Ride-Sharing of the participants. As mentioned by Furuhata et al. [15] the two main taxonomic criteria for Ride-Sharing systems are primary search criteria and the target market. The primary search criteria consider; routing time, origin/destination pair and time, keyword/list submitted by the participants, and origin/destination pair and first-come-first-serve basis. While considering the target market some classifications provided by Furuhata et al. [15] are; an on-demand classification which is a casual, on-time and irregular trip for a relatively short distance which requires a real-time response. Commute classification is the Ride-Sharing between commuters who have regular work schedules and long-term relationships. The participants of such Ride-Sharing can take turns to use their vehicles. Long-distance classification is the Ride-Sharing that can take place for long distance which can be scheduled in advance with less restrictive requirements of meeting location and the time.

Above mentioned are the main target markets for Ride-Sharing. People need to travel daily to a certain distance or take long-distance rides for certain purposes occasionally. All these should be considered when developing a good Ride-Sharing algorithm. By considering the given patterns and the classifications of Ride-Sharing Furuhata et al. [15] have identified some classes of Ride-Sharing; dynamic real-time Ride-Sharing (Route and Time; On-Demand and Commute), carpooling (OD-Pair and Time; Commute), long-distance ride-match (OD-Pair and Time; Long-Distance), one-shot ride-match (OD-Pair and Time; Commute and Long Distance), bulletin-board (Keyword/List; Commute and Long-Distance), and flexible carpooling (OD pair and FCFS; Commute and Long Distance).

Furuhata et al. [15] describes some fundamental business functions of Ride-Sharing. As given there are three main business functions. Three business functions are planning, pricing and payments. Planning means the planning of the Ride-Sharing by identifying the classes given above and scheduling the Ride-Sharing accordingly so that it would not cause any impact on both the driver and the passenger. Next is pricing. There are several ways of pricing rideshare. The catalog pricing can be used where the driver has a limit price which is listed. The rule-based pricing can be used where the matching agency provides and specifies a cost calculation formula. Basically, these kinds of calculations can be defined as the standard rate per distance multiplied by the distance travelled. The other pricing scheme is the negotiation based pricing where the driver and the passenger can be negotiating the price and the matching agency is not involved in this. The final business function is payments. Payments can be done as direct payment or via a third party such as PayPal.

The Ride-Sharing services can be given in various ways. There are several types of matching agencies and Furuhata et al. [15] have given some such as; integrated services, coordination services, classified advertising and casual services. Integrated service provides all the three main business functionalities. Coordination services only support planning and pricing and provide the coordination between the participants. Classified advertising services only support for planning and leave the pricing to the participants. In casual services matching is based on the first-come-first-serve with predetermined meeting places and time. The pricing and the payment is fixed in such systems.

The identified challenges in dynamic Ride-Sharing are, pricing for dynamic Ride-Sharing, high dimensional matches, trust and the reputation and institutional design. Other than these challenges, Furuhata et al. [15] have mentioned that the privacy of ride-shares and the legal liability of the matching agencies are also some challenges that need to be considered when developing Ride-Sharing services. Some major challenges when developing a system for a Ride-Sharing agency are designing and attractive mechanism which takes price quotes, truthfulness and incentives for participants into consideration and developing a concierge-like ride-arrangements by considering the preferences and transportation modes. The most important challenge is building a solution that addresses all these challenges. In later chapters, solutions that can be used to overcome these challenges will be discussed.

1.4 Research Outline

This study discusses the potential of implementing Ride-Sharing in Sri Lanka with the use of a CDR Dataset and Base Station details. This is a data analytics based model. The above sections described the motivation behind the study and the research objectives in detail. Also discussed already are identified myths and challenges in Ride-Sharing. In this study, as mentioned in the previous sections, three algorithms are introduced to cluster the available BSTs, then to cluster people into a virtual location identified based on the BSTs and cluster them again based on same source and destinations that have the potential of sharing the rides.

This Chapter briefed the motivation of analyzing the potential of Ride-Sharing in Sri Lanka and the research objectives. It further discussed the myths and challenges that are identified in the Ride-Sharing concepts and what needs to be considered when implementing a Ride-Sharing model based on Srivastava [10] and Furuhata et al. [15].

The rest of the work will be organized as follows: Chapter 2 will discuss the existing work in terms of ridesharing, carpooling and intelligent traffic handling. It will further discuss the reason for choosing Ride-Sharing as the suitable concept to implement in Sri Lanka over Carpooling and Intelligent Traffic Handling to overcome the traffic congestion.

Chapter 3 discusses the CDR data and the Datasets used for this study and the reasons for choosing those for this study. Further, it will discuss the methodologies and algorithms used for this study and how they are improved and modified to match the Sri Lanka dataset and people. This methodology and algorithms section will elaborate the above three algorithms, Clustering Cell Towers, Identification of home/work locations and identification of potential ride-sharers along with the proposed Ride-Sharing algorithms along with a pricing mechanism. Then the chapter will describe the architecture and technologies used to analyze the data for this study. Finally, the chapter will discuss the benchmarking performed on the methodology and algorithms using the Nodobo Dataset.

Chapter 4 presents the analysis of the study using the three datasets; the actual CDR Dataset, the dataset of the Transportation usage of Sri Lanka and data gathered from a survey. Finally, the last Chapter, Chapter five will emphasize the conclusion of the study based on the analysis performed. Further, it will point the future work that can be performed in this study to be continued further and implemented in Sri Lanka. Finally, Chapter 5 will summarize the study and describe the key findings of the study with the conclusion on the potential of implementing Ride-Sharing in Sri Lanka. Also, the chapter consists of research limitations and future enhancements that need to be considered when modeling and implementing Ride-Sharing.

CHAPTER 2: EXISTING WORK

As mentioned in the Introduction, there are several methodologies, models, and applications introduced and implemented so far on Ride-Sharing and carpooling along with the idea of intelligent traffic handling. This chapter will discuss some of the existing methodologies and models such as agent-based model, dynamic Ride-Sharing models, and decision support systems. It will analyze more on a collection of data on individuals, methods to identify home/work locations, algorithms to implement Ride-Sharing, carpooling, and intelligent traffic handling and addressing the privacy concerns of the individuals. This Chapter will be further divided into Section 2.1 Ride-Sharing, Section 2.2 Carpooling and Section 2.3 Intelligent traffic handling and each of these sections will discuss different methodologies and models for different concepts. Finally, in Section 2.4, it will summarize the findings and highlights of the Chapter. Each section discusses the topic in detail with the selected existing work related to this study.

2.1 Ride-Sharing

The section discusses the existing methodologies and algorithms of Ride-Sharing in brief and some of the applications and models introduced.

Identifying the potential of developing a system is vital, before implementation. Cici et al. [2], [3] has come up with a methodology that quantifies and assesses the potential of Ride-Sharing. Their work has been done in two types of research and in the first paper their discussion is based on the quantifying the potential of Ride-Sharing only using the call description records and in the second research they have assessed it using both mobile and social data which is an enhancement to the first research by considering privacy concerns. The main idea of this work is to extract the mobility data from 3G call description records and relationships among the people using social network data. The social network data has used to identify the relationships between the people who have the potential for sharing the ride to make the Ride-Sharing more secure. They have introduced an efficient heuristic algorithm which is called as the ‘End-Point Ride-Sharing’. The steps taken in this algorithm are, first, to gather the home/ work locations of the users by using the CDR data and geo-tagged tweets (Social Network data). This social data is also used to identify the friendship between users. As the second step, they have developed a framework that can be used to match users who can share a ride.

Some of the constraints considered in this process are called Spatial, Temporal and Social. Spatial means share a ride with people who live/work in a certain distance, Temporal is sharing a ride with people who want to depart/arrive within a certain time window from actual and Social means share rides with directly known people or who have common friends.

Isaarcman et al. [16] have introduced a methodology to identify the important places of the users. This is a cluster based methodology. In the Isaarcman et al. [16] methodology clusters a user based on the recorded cell tower details to produce a list of places a user visits. Then the regression analysis is used on the ground truth users with their identified clusters and true important locations to determine the features that are used to distinguish the places such as home/work locations. The features they have considered are; (1) the number of days that the user appeared on the cluster, (2) the duration of user appearances on the cluster, (3) the rank of the cluster based on number of days appeared, (4) number of phone calls taken from 7PM to 7AM considered as home hour events, and (5) number of phone calls taken from 1PM to 5PM considered as work hour events.

Cici et al. [2], [3] have used Isaarcman et al. [16] methodology to develop their two ridesharing algorithms; End-Point Ride-Sharing and En-route Ride-Sharing. End-Point Ride-Sharing considers the people who live and work in common areas share the ride. This algorithm is used to identify the home/work locations of users by developing a practical algorithm using both CDR and Twitter data sets. A matching algorithm is used to identify the people who can share the rides by looking at the number of potential drivers, the capacity available in their vehicles and a formula to identify the minimum number of cars needed for Ride-Sharing.

In En-route Ride-Sharing algorithm, it is considered picking up passengers on the way to the destination of the driver when the vehicle is not fully packed. It is an iterative algorithm. The algorithm first runs the basic End-Points RS algorithm and then excludes from the solution; cars that get fully packed. Then order cars in decreasing order of passengers and start “routing” them across the urban environment using data from Google maps. When the currently routed car v meets a yet un-routed car v′, then v can steal passengers from v′ if it has more passengers than v. When a routed car gets fully packed it is removed from further consideration. Whenever a car with a single passenger is encountered the number of cars is reduced by one.

People normally hesitate to share rides with unknown people as it contains a level of risk. Therefore Cici et al. [3] have taken the social data filtering into consideration to identify the friends or the friends of friends to share the ride. They have selected some thresholds constraints based on both CDR and geo-tagged tweets to identify edges between users. When considering the CDR data they have considered the users who have at least one phone call connections as a friend or known person and so on. Based on the identified home/work locations and the social ties they have proposed a solution to match users for Ride-Sharing using the given algorithms. Among these two algorithms, they have concluded that the En-route Ride-Sharing considering the friends of friends is more effective.

As mentioned in previous sections Ride-Sharing can be planned or dynamic. There are some more concerns that need to be considered in dynamic ridesharing. Kleineret et al. [5] has presented a mechanism for dynamic Ride-Sharing. Their approach is adaptive to individual preferences of the participants. The proposed system has allowed to trade off the minimization of Vehicle Kilometers Travelled (VKT) with the overall probability of successful ride-shares. This is considered as an important feature when bootstrapping the system. This is a Dynamic Ride-share Solution based on the auction using a sealed-bid second price scheme. In this system the passengers can bid to increase their ranking which is visible to the drivers and the drivers can select the passengers according to their preferences. Their approach is named as DSR which is a Smartphone application for dynamic Ride-Sharing which operates in real-time. The presented system promotes Ride-Sharing among people with connections within social networks such as Facebook and Twitter. These connections are used to favor ride-share among users who know each other. The approach is; first the application should be launched on a Smartphone and then the user is asked to enter the goal location and a deadline and to place a bid. Once the bids are placed then the system determines the drivers who are online and compatible for the ride based on the length of the detour to be performed and whether the deadlines will comply. Then the passenger can select from the list of compatible drivers which are presented with the additional information like user ratings and the social network status. Finally, the drivers will receive the bid, which is considered as the binding commitment of the passenger.

After the ride passengers will get the opportunity to review the driver by giving a rating. In the driver’s perspective, this DRS system can be developed as a GPS car navigation system that is connected to the social network. The driver first specifies the goal location and the deadline and the system will compute the shortest path and provide the directions to the driver to reach the goal location. Then during the ride, the driver will be informed of the Ride-Sharing offers placed by passengers who are nearby his route vocally or in some other form.

Multi-Agent dynamic ridesharing systems are another way of implementing ridesharing. Kleineret et al. [5] have described that this system can be thought as a Multi-agent DRS system. Kleineret et al. [5] have considered that one driver can have one passenger to make the concept easier. An auction mechanism is used to provide a ranking for each driver for the bids received from the passengers. Agents can enter the system and announce their types. The passengers can additionally announce their bids that they are willing to pay at the end of the ride. The computation of dynamic matching assignments will be harder as there can be many requests and future requests are unknown. The study has used the deterministic rolling horizon approach to solve this issue. It computes the assignments by all known information within a planning horizon. The decisions are committed only when necessitated by a deadline. The computations are triggered when planning horizon grows with the new passenger deadline announcement which is depicted by the dotted lines in the Figure 2.1 Kleineret et al. [5].

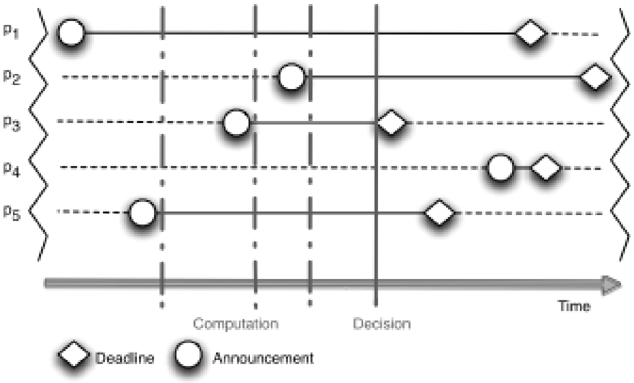


Figure 2.1: Deadline Announcements

One desirable property that the study has highlighted such a system is that it is an incentive compatible, which means that the user can state his type and preferences truthfully. According to the authors the presented DRS system has following properties which can be used for DRS system; allows more flexibility than state-of-the-art systems as it has given the freedom to the users to choose the Ride-Sharing partner, it is a robust system as it promotes truth-telling about most aspects and fairness in payments and allows to trade-off VKT saving with the probability of finding matches between drivers and passengers.

As mentioned above there are several ways of handling the ridesharing. Cici et al. [2], [3] have presented a heuristic algorithm and genetic algorithms can also be used.

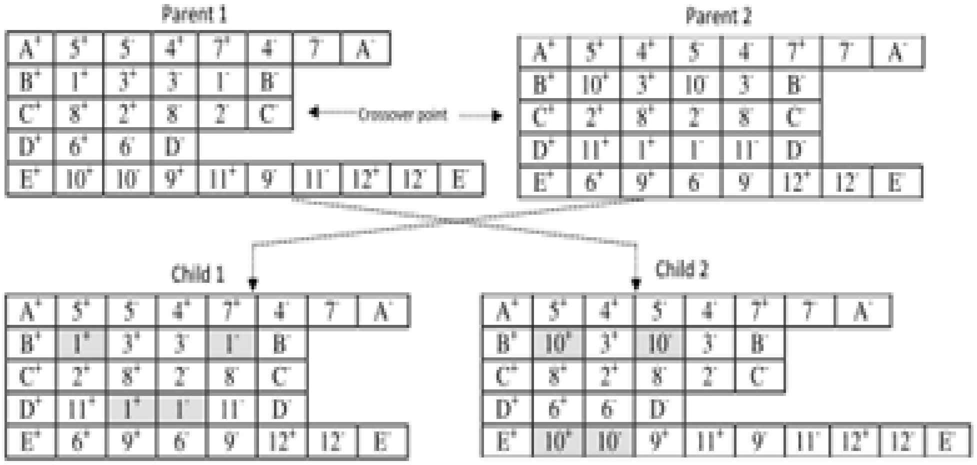
Herbawi and Weber [8] have presented an algorithm that is genetic and insertion heuristic to solve the dynamic ride matching problem with the time windows (RMPTW) in dynamic Ride-Sharing. In dynamic Ride-Sharing participants form Ride-Sharing with a short notice. This is considered as an optimization problem where the multi-criteria objective function has been optimized. The time window for each location will be defined by the earliest departure time and the latest arrival time specified by each participant. The identified criteria to optimize are; c1: The total distance of vehicles' trips should be minimized, c2: The total time of vehicles' trips should be minimized, c3: The total time of the trips of the matched riders' requests should be minimized and c4: The number of matched (served) riders' requests should be minimized.

It is given that the trip's time might include some waiting time. The proposed algorithm has adopted the dial-a-ride (DAR) by Agatz et al. [17] for RMPTW and modifies the definition of the objective function and has added additional constraints on it. The algorithm; first divide the day into a set of time periods. Then giving all requests and offers each day starts with executing a genetic algorithm to solve a static version of the problem. After the execution, they have utilized an insertion heuristic to give an online answer by updating the solution produced by the genetic algorithm in real-time for each newly received request/offer until the end of the period. Finally, all the non-matched requests by the genetic algorithm and the insertion heuristic will be stored for the future matching and to be the input for the genetic algorithm in the next period.

It is pointed out that the shorter the time, the more often the genetic algorithm will be executed resulting in better results on the cost of longer execution time. The genetic algorithm they have proposed is; schedule v number of routes, which is one route for each vehicle. Each route starts with a source point of the vehicle followed by a list of pickup and delivery points and ends with a destination point of the vehicle. The route is randomly selected to initialize from v routes and then the first two points to be inserted for the route are decided as the source and the destination of the vehicle.

Then select a random pair of pickup and delivery points from matching rider requests. The process will be repeated for all the routes.

They have considered the driver’s maximum distance constraints after inserting a new point and for the precedence constraint they have inserted the pickup point before the corresponding delivery point and have searched for a feasible insertion position for the delivery in the route to the right of its pickup point. They have come up with a single point crossover operator. When given two parent solutions they have selected a random crossover position (route index) and have made crossover by exchanging the routes among the parents starting from the selected crossover position in upwards fashion as depicted in Herbawi and Weber [8] Figure 2.2.



Parent 1

A^+	5^+	5^-	4^+	7^+	4^-	7^-	A^-		
B^+	1^+	3^+	3^-	1^-	B^-				
C^+	8^+	2^+	8^-	2^-	C^-				
D^+	6^+	6^-	D^-						
E^+	10^+	10^-	9^+	11^+	9^-	11^-	12^+	12^-	E^-

Parent 2

A^+	5^+	4^+	5^-	4^-	7^+	7^-	A^-
B^+	10^+	3^+	10^-	3^-	B^-		
C^+	2^+	8^+	2^-	8^-	C^-		
D^+	11^+	1^+	1^-	11^-	D^-		
E^+	6^+	9^+	6^-	9^-	12^+	12^-	E^-

Child 1

A^+	5^+	5^-	4^+	7^+	4^-	7^-	A^-
B^+	1^+	3^+	3^-	1^-	B^-		
C^+	2^+	8^+	2^-	8^-	C^-		
D^+	11^+	1^+	1^-	11^-	D^-		
E^+	6^+	9^+	6^-	9^-	12^+	12^-	E^-

Child 2

A^+	5^+	4^+	5^-	4^-	7^+	7^-	A^-		
B^+	10^+	3^+	10^-	3^-	B^-				
C^+	8^+	2^+	8^-	2^-	C^-				
D^+	6^+	6^-	D^-						
E^+	10^+	10^-	9^+	11^+	9^-	11^-	12^+	12^-	E^-

Figure 2.2: Single-point crossover operators

After the crossover operation, a repair operation is applied to the children so that it removes all the pickup and delivery points in the routes after the crossover position if they exist in any route before the crossover position. Also, insert the pickup and delivery points of rider’s requests which are not matched in the resulting children after each crossover. They have defined five route level mutation operators. If any mutation operation violates any given constraints it is rejected. Mutation operators are; push backward, push forward, remove-insert, transfer mutation and swap mutation.

Given a solution from the genetic algorithm, the insertion heuristic is designed to answer each new request/offer by modifying the solution when possible. When an offer is submitted by a driver the route is added to the solution and the heuristic tries to match a possible ride from the non-matched rider's requests. When a rider submits a request, heuristic tries to insert it to one of the available routes. Finding a match for request includes; finding all possible insertions for pickup and delivery points of the requests in all available routes and then find the best insertion point.

According to Herbawi and Weber [8], the best insertion point is the one that adds the minimum value to the heuristic objective function. The objective function is responsible for, minimize the total distance and time of the vehicles trips, minimize the total time of the trips of the matched rider's requests and maximize the number of matched riders' requests. Herbawi and Weber [8] have stated that the proposed algorithm can successfully solve the dynamic ride-matching problem by providing answers to Ride-Sharing requests in real-time and the algorithm is flexible and can be tuned between the solutions' quality and the responsiveness of the algorithm.

Other than using own vehicle it is possible to use taxi ridesharing. Set of people can share the same taxis to travel to their destination locations. Ma et al. [13] have proposed a taxi searching algorithm and a taxi scheduling algorithm based on the large-scale dynamic Ride-Sharing. The algorithm is named as 'T-Share'. The purpose of these algorithms is to serve dynamic queries of the customers as quickly as possible and reduce the total travel distance of the taxis. As mentioned by Ma et al. [13] these algorithms take small query processing time and help to save energy and greenhouse gas emission. Some of the challenges any such system like this would face are; the queries submitted by passengers and the locations of the taxis being highly dynamic and difficult to predict. In their system Ma et al. [13] have given the option to taxi drivers to leave the service at any time and join as they wish.

Passengers can submit their queries via mobile devices. As in many systems they have also considered one passenger to make the queries simple. Taxis can upload the status to an operation center when joined to the service and when passengers get in and get off. This is an idea that can be implemented for any Ride-Sharing system so that everyone can see the availability of the vehicle. This system has also maintained a spatiotemporal index of the joining taxis and processes the user queries faster. The Ma et al. [13] Figure 2.3 depicts the dynamic taxi ridesharing service framework developed and the Figure 2.4 depicts the Ma et al. [13] grid partition map and grid distance matrix used.

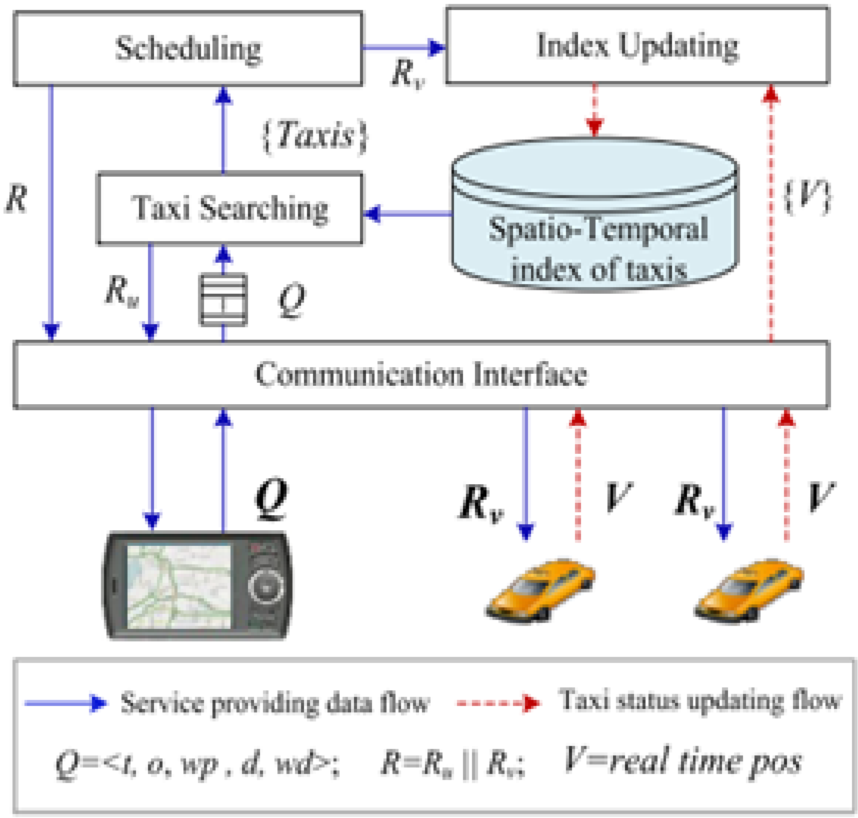
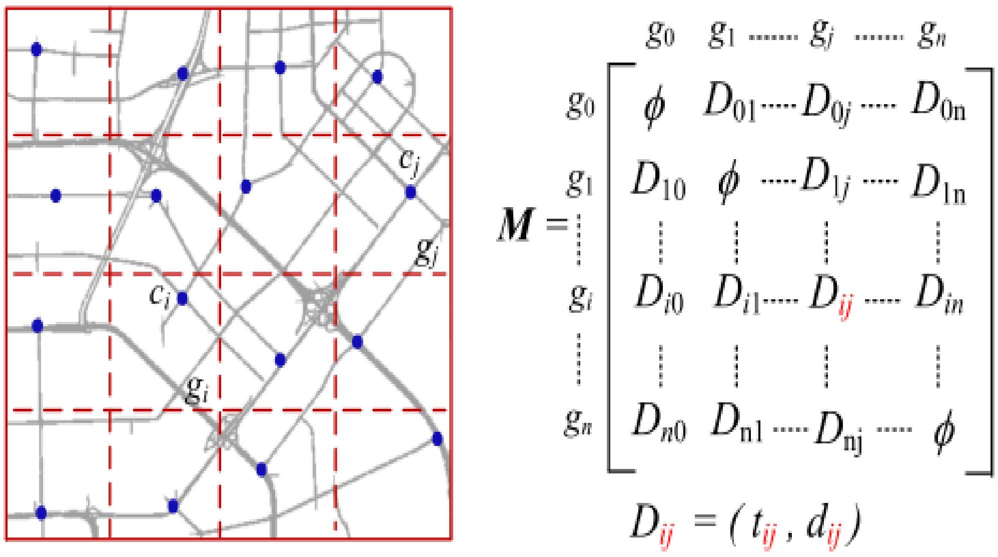


Figure 2.3: Framework of the dynamic taxi Ride-Sharing service


c_j
g_j
c_i
g_i
g_0 g_1 ...... g_j ...... g_n
$M =$
ϕ D_{01} ..... D_{0j} ..... D_{0n}
D_{10} ϕ ..... D_{1j} ..... D_{1n}
D_{i0} D_{i1} ..... D_{ij} ..... D_{in}
D_{n0} D_{n1} ..... D_{nj} ..... ϕ
$D_{ij} = (t_{ij}, d_{ij})$

A) Grid-partitioned map

B) Grid distance matrix

Figure 2.4: Grid partitioned map and the grid distance matrix

The aim of this taxi searching algorithm is to find the small set of taxis that would satisfy a user query which will serve the query with a small increase of the travel distance. If the number of taxis is huge the calculation of shortest path would be more expensive. The solution they have come to for this is to partition the road network into grids and create a grid distance matrix. This matrix provides the approximated distance between any two geographical locations in the road network which helps to find the shortest path. The purpose of their scheduling algorithm is to insert the schedule of a taxi which would satisfy a query with minimal additional travel distance. The insertion feasibility checks can be used when inserting these schedules. One way of doing the feasibility is the Lazy Shortest Path calculation where the triangle inequality and caching speed up the feasibility. The lower bound delays the shortest path calculation until that is really needed. It is mentioned the state-of-art shortest path algorithm can be used to speed up an online shortest path calculation.

Ma et al. [13] have proposed a simple yet effective pricing scheme as well. Some properties that are taken into consideration when building a pricing scheme are; taxi fare per mile is higher for multiple passengers and a single passenger and the taxi fare of shared distance is evenly split among the riding passengers. This technique makes everyone to pay a lesser amount than they occupy a single taxi when more people share the ride. This can be considered as a worthy pricing scheme methodology for Ride-Sharing.

2.2 Carpooling

Carpooling is a subsection of ridesharing as discussed in the previous chapters. This section describes the methodologies and applications use for carpooling. In this section, the existing algorithms and architectures are discussed in detail which can also be used in ridesharing.

The Cho et al. [1] has come up with a conceptual design of an agent-based interaction model for a carpooling application. The model discussed in this paper is a computation model which is used for simulating the actions and the interactions of autonomous agents and to analyze the factors related to areas such as infrastructure, behavior, and cost. In this application agents and profiles are used as the main components and social networks are used to initiate an agent communication model. To trigger the negotiation process between the agents a route matching algorithm and a utility function are used. As mentioned in the Cho et al. [1] an agent is a person who lives a study area and executes the daily schedule to fulfill his/her needs. A schedule is a list of activities in an order. As explained in the paper, during a simulation run there are several steps that agents need to follow such as, (i) setting the goal, (ii) scheduling activities based on the given resources and environment, and (iii) finally the execution of the schedule. In a carpooling application, an agent is willing to share a vehicle to reduce the travel time and the cost. Reducing the travel time is an important factor for carpooling participants.

The introduced communication and coordination aspects of Cho et al. [1] have a profile called ‘AgentProfile’ which is an XML based profile with attributes such as name, gender, household information, etc. According to the authors, this profile could contain a score based on feedback from other agents. Below Cho et al. [1] in Formula (2.1), they have come up with a way to identify the potential of carpooling between agents based on their common characteristics and interest.

$$CarpoolPotential(CP_n) = \{Location(L), SpatialRelevant(SR), Interests(I), Requirement(R)\} \quad (2.1)$$

For all the interacting agents should have a CP match and to identify whether there is a match between agents using the origin and the destination locations the match is called the Spatial Relevance factor. Cho et al. [1] Figure 2.5 depicts the spatial relevance factor.

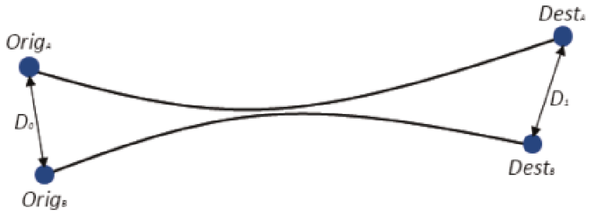


Figure 2.5: Spatial Relevance factor

The participation factor equation shows in Cho et al. [1], Formula (2.1).

$$pf = 1 - \frac{log_2(i)+1}{i}, AR(n)_i = AR(n)_{i-1} + (QoC(m) - T) \times \frac{AR(n)_{i-1} \times pf}{SRDist} \qquad (2.1)$$

The AR is the Agent Reputation being a potential carpooling candidate which is used to detect outliers and make the decision-making process easier. Participation factor ‘pf’ is used to consider the active participation of the agents, which is directly proportional to the change of AR. T is the reputation threshold. According to their algorithm, based on the quality of the carpool (QoC) feedback, the reputation value given will increase and decrease.

Cho et al. [1] have mentioned that negotiation is an important step in an agent-based model. There are some questions given in Cho et al. [1] that need to be answered during a negotiation phase such as; the issues over which negotiation takes place, the negotiation protocols that will be used and reasoning model which will be the agents employ. Given two agents identify the matching route when the following procedure has been introduced.

Tare et al. [11] has proposed a solution for carpooling using Android. In this application the passenger and driver are given the facility to rate and give feedback on each other. The system is designed to communicate with xampp server which has MySQL and PHP as cross platforms. This works in two-way communication between the driver and the passenger with a flexible environment. The purpose of selecting Android for the proposed system is; it is popular among users and less expensive. This is a vital point highlighted in the proposed system. Some of the major points taken into consideration in their goal of developing the systems are enhanced security for women passengers, high reliability due to real-time tracking, enhanced payment features, reviewing history and both driver and passenger can stay in touch with each other. These are some important points that need to be considered when developing a carpooling or Ride-Sharing system. The main modules of Tare et al. [11] system are driver and the passenger. They have identified the most important things needed to be there in a carpooling application and added to their developed system. The application should be deployed on both driver and passenger android phones, the database will allocate ids for both the driver and passenger. A central database which will manage other databases and control the activities and the rating and comments history should be displayed on mobile phones. Tare et al. [11] Figure 2.6 depicts the actual way the application would work once it is developed.

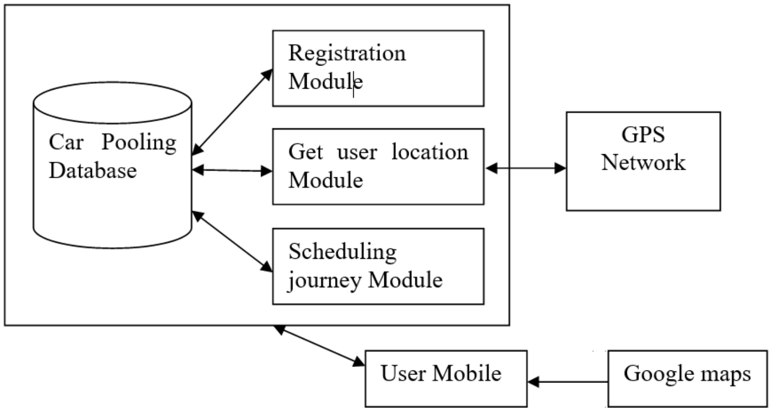


Figure 2.6: Actual Working of Application

The concept of decision support can be applied to carpooling and Ride-Sharing problems as well. When and where to carpool is a decision a person would take. Manzini and Pareschi [14] have proposed a decision support system for carpooling problem (CPP). It is a cluster based hierarchical model comprising an original decision support system (DSS). As discussed by Manzini and Pareschi [14] the carpooling problem can be categorized into two categories; Daily carpooling problem (DCPP) and the Long-term carpooling problem (LTCPP). As mentioned by Manzini and Pareschi [14] the computational complexity of Carpooling problem is NP-Hard. They have proposed a cluster-based approach for carpooling. This adopted approach has two phases; Cluster first and Route second.

This is based on three main activities or steps. In the first step, they have collected data from users, destination points, and the transportation network. The second step is mentioned as the first phase of adopted 2-phases heuristic approach for CPP which is based on cluster analysis (CA) and similarity indices. Then the participant's group into a homogeneous cluster called carpools. In the third step or the second phase, the aim is to identify the best set of routes to reach the destinations of grouped participants. Manzini and Pareschi [14] Figure 2.7 depicts the 2-phases approach for CPP.

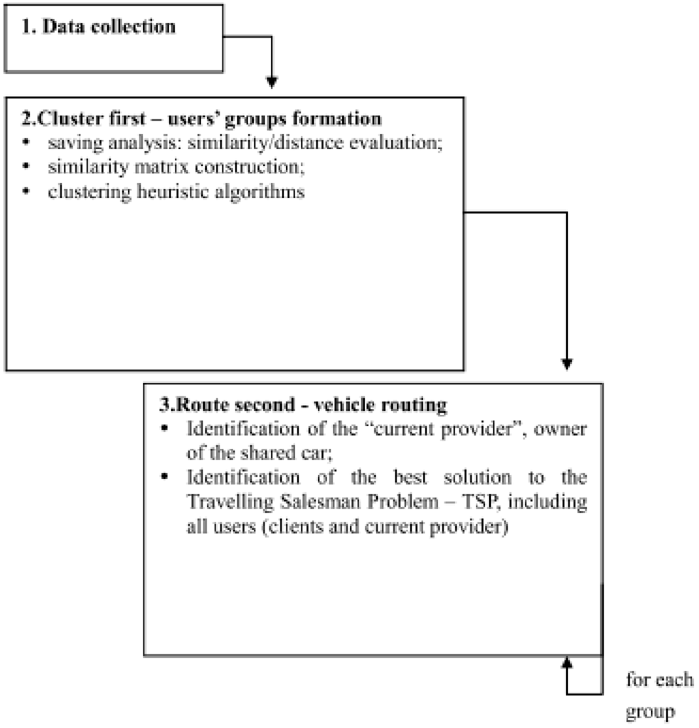

1. Data collection
2.Cluster first – users' groups formation
• saving analysis: similarity/distance evaluation;
• similarity matrix construction;
• clustering heuristic algorithms
3.Route second - vehicle routing
• Identification of the "current provider", owner of the shared car;
• Identification of the best solution to the Travelling Salesman Problem – TSP, including all users (clients and current provider)
for each group

Figure 2.7: Three steps - 2-phases approach to the CPP

In a decision support system of CPP some of the parameters that should be considered according to Manzini and Pareschi [14] are; Routing Strategy, Clustering Rule, Time-Based or Distance-Based Analysis and Age-Based Modifications. These parameters are some of the important parameters that should be considered in any decision support system for carpooling problem.

2.3 Intelligent Traffic Handling

Intelligent traffic handling is one of the major topics discussed in society. With the development of urban areas and the increasing number of vehicles, the traffic is becoming a huge problem. The carpooling and ridesharing have been taken as alternative ways of rendering the traffic. This section describes the ways and methodologies that have been considered by researchers to handle the vehicular traffic by using artificial intelligence.

Due to the number of growing vehicles the traffic congestion, accidents, transportation delays and larger vehicle pollution emissions have increased. Figueiredo et al. [4] discuss intelligent traffic handling in by considering the achievements in this area in last few years. Developing more roads is not a solution to reduce the traffic congestion. Information Transportation Systems (ITS) considered as a global phenomenon. In ITS advanced communication, information and electronic technology are applied to solve the following transportation problems such as; traffic congestions, safety, transport efficiency and environmental conservation.

The conceptual model is given in below Figueiredo et al. [4] Figure 2.8.

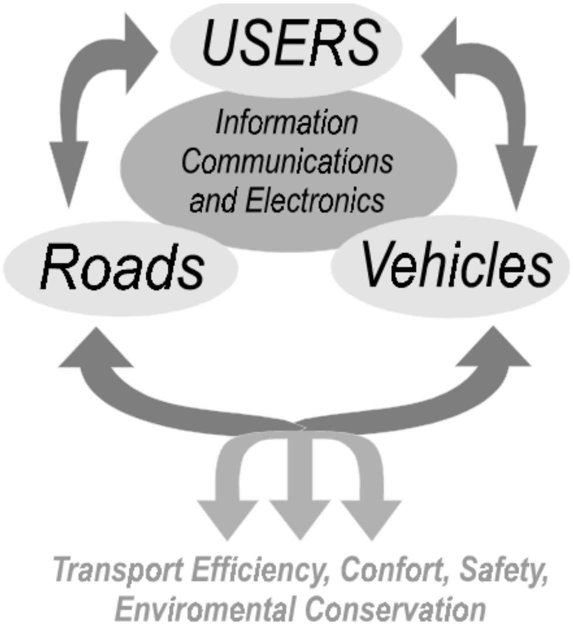

USERS
Information
Communications
and Electronics
Roads
Vehicles
Transport Efficiency, Confort, Safety,
Enviromental Conservation

Figure 2.8: ITS Conceptual Model

According to Figueiredo et al. [4], the purpose of ITSisto takes the advantage of appropriate technology to create more intelligent roads, vehicles, and users. The study is about technologies and the scientific aspects to develop the new systems which can solve some referred problems in transportation.

There are six major categories of ITS represented by Figueiredo et al. [4] with their global characteristics. The categories are as follows,

- Advanced Traffic Management Systems (ATMS) – collection of data team by monitoring traffic conditions, support system by using cameras, sensors, semaphores and electronic devices to help the system to operate, to manage and control the real-time traffic, and real-time traffic control systems which use the information from above two elements to change semaphores and send messages are the three main elements of ATMS.

- Advanced Travelers Information Systems (ATIS) – the purpose of these systems is to provide the real-time traffic information to the travelers, such as giving advice to drivers to select the most appropriate road to reach the destination.

- Commercial Vehicle Operations (CVO) – these are used to increase the efficiency and the safety of commercial vehicles and fleets. These systems include the technologies for traffic management, travelers' information and vehicle control management. Some of the technologies are; Automatic Vehicles Identification, Automatic Vehicles Classification, Automatic Vehicles Location, Pedestrian Movement Detection, Board Computers and Real Time Traffic Transmissions

- Advanced Public Transportation Systems (APTS) – these are used to improve the operations and efficiency of high occupation transports such as buses and trains with the use of electronic technologies. The automatic payment systems which use smart cards to store credits capture traveler details and journey profiles are also included with these.

- Advanced Vehicle Control Systems (AVCS) – these systems are coupled with sensors, computers and control systems to assist and alert the drivers on safety. Also take part in driving to improve the safety, and road system productivity by decreasing the congestions on roads and highways.

- Advanced Rural Transportation Systems (ARTS) – these systems are designed to address the problems arising in rural areas.

As given in the Figueiredo et al. [4] some of the major areas for future research of ITSs are simulation and modeling and fully automated systems. In the simulation and modeling traveler information, traffic management and driver steering behavior are some of the corresponding areas that should be taken into consideration.

Wang [6] has presented concepts, architectures, and application of parallel control and management for the intelligent transportation system. To conduct the operations of complex systems parallel control and management has been proposed as a new mechanism. The complex systems are the systems like transportation systems that are involved with complex issues in both engineering and social dimensions. The paper has described the basic concepts of ACP and artificial transportation systems (ATS) and the responsibility of those systems in ITS. According to Wang [6], the purpose of ACP approach is modeling, analysis, and control of complex systems. There are two main characteristics of the complex systems that are considered by ACP approach which appears in any system and involves human and social behaviors; Inseparability and Unpredictability.

The ACP approach consists of three steps; (1) Modeling and representation using artificial societies, (2) analysis and evaluation by Computational experiments, and (3) control and management through Parallel execution of real and artificial systems.

Wang [6] has pointed out that these characteristics lead to the few deductions. Those deductions are; the necessity of taking a holistic approach to dealing and modeling complex systems, there is no fix once-and-for-all solutions for problems in complex systems and no optimal solutions for complex systems.

The Proper's theory of reality is a base for ACP approach which says that the universe consists of three worlds; physical, mental and artificial. Wang [6] Figure 2.9 shows the Philosophical and Scientific foundation of the ACP approach.

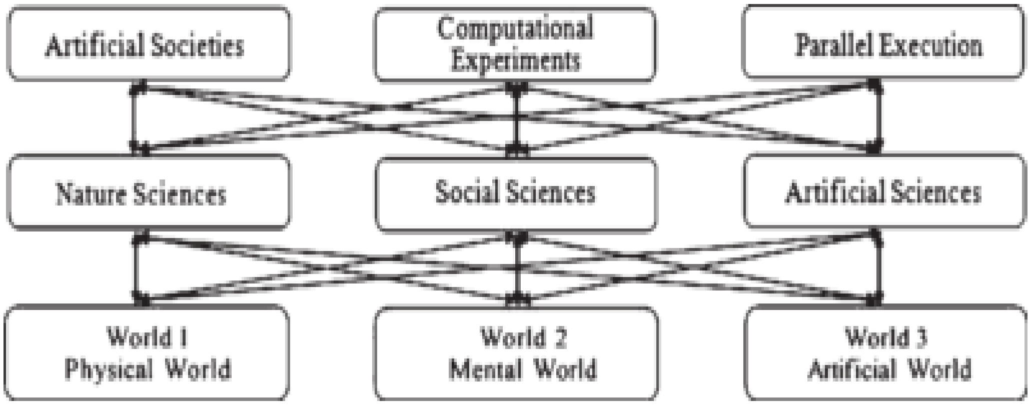


Figure 2.9: Philosophical and Scientific Foundation of the ACP Approach

Wang [6] has pointed to the need of ACP approach for ITS as the real-world transportation systems such as large-scale urban traffic systems consist of the characteristics of the ACP approach. It also says that the Cyber-Physical Systems (CPS) and cloud computing systems that are naturally embedded in this approach is an advantage. Lack of timeliness, flexibility, and effectiveness in current operation management systems in transportation is the motivation behind the use of ACP approach. The reliability and the performance of the current ITS technology will be significantly enhanced and improved by application of ACP. The cyber space-based parallelism could open a wide area of new applications scenarios in ITS. Some of those areas presented by Wang [6] are future driving in intelligent transportation spaces for integrated and better traffic management, vehicle safety, energy efficiency, reduced pollution and maintenance services, where vehicles, highways, roads, intersections, and operation centers are embedded with various types of intelligent spaces.

As given by Wang [6] there are differences between ATS and traffic simulations. Two major differences are objective and scope. The objective is the traditional traffic simulation which is to represent the true state of actual traffic. The objective of ATS is to build the live traffic in a bottom-up manner providing alternative ways for actual traffic activities. The scope is the traditional traffic simulation systems focus on direct traffic-related activities whereas the ATS must deal with a wide range of information and activities.

Proposed system architecture consists of few processes and systems. PtMS is an ACP based system architecture and process. Usually, parallel traffic control and management systems use more than one ATS. A PtMs consists of four major components, actual transportation systems, ATS, traffic operator and administrator training systems (OTSt), decision evaluation and validation systems (DynaCAS), traffic sensing control and management system (aDAPTS). The operation process is divided into three modes named as training, testing and operating.

- OTSt – known as Operator Training Systems for transportation is developed for learning and training mode operations for traffic operators and administrators

- DynaCAS– Known as Dynamic Network Assignments based on Complex Adaptive Systems, which are developed to design, conduct, evaluate and verify computational transportation experiments, detect existing and emerging traffic patterns and control and management of traffic systems.

- aDAPTS – known as Agent-based distributed and adaptive platforms for transportation systems provide supporting and environments to design, construct, manage and maintain autonomous agent programs for various traffic tasks and functions.

- Integrated traffic operation platforms provide five major functions such as Traffic data collection, Traffic information processing, Traffic analysis and evaluation, Traffic information services and Traffic control operations.

With the integration of above systems Wang [6] has presented a new control and management mechanism for parallel control and management for complex transportation systems which has integrated concepts and methods developed in AI, intelligent control, computation intelligence intelligent systems, intelligent spaces, complex systems, complexity theory, social computing and advanced computational technologies like agent programming and cloud computing.

Ramos et al. [7] have presented a telegeo-processing System for traffic and environment. The system includes prototypes for mobile urban traffic data acquisition with a GPS equipped vehicle, PDA application and wireless communications, and for a geo database with a related web application for Urban traffic and environment. The proposed framework is named as Intelligent Urban Traffic & Environment Operations (IUTEO). The framework was developed using Object-Oriented Modeling (OOM). The main objective of OOM is the development of models based on the real-world concepts, modularity, reusability, and extensibility. This framework is enclosed with five main subsystems; real system, data acquisition, database modeling and analysis platform, an application for municipalities, and communication, sensing & surveillance, information management & control.

The urban transportation system is mainly characterized by static and dynamic attributes such as traffic signals for given intersection, type of pavement for given road section, traffic counts for a specific road segment or travel delays from a route. These cannot be adequately explored by conventional databases. For a long time, the spatial data acquisition and integration has been one of the topics in main research. The continuous development of this area has been urged with the advances of GPS and wireless communication technologies as well as the increasing need of real-time information for intelligent traffic. Ramos et al. [7] Figure 2.10 depicts IUTEO framework and Figure 2.11 depicts the SysML package diagram.

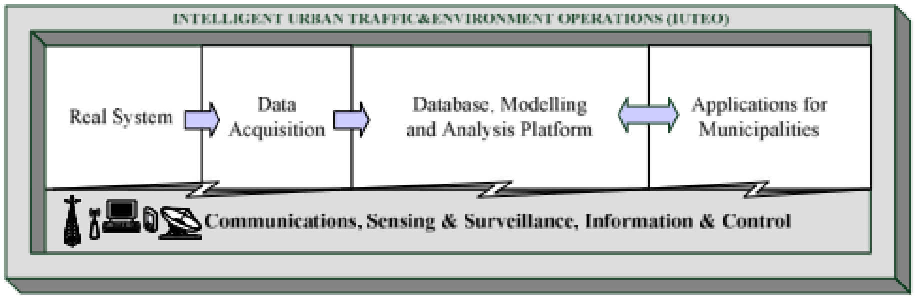

INTELLIGENT URBAN TRAFFIC&ENVIRONMENT OPERATIONS (IUTEO)
Real System
Data Acquisition
Database, Modelling and Analysis Platform
Applications for Municipalities
Communications, Sensing & Surveillance, Information & Control

Figure 2.10: IUTEO framework

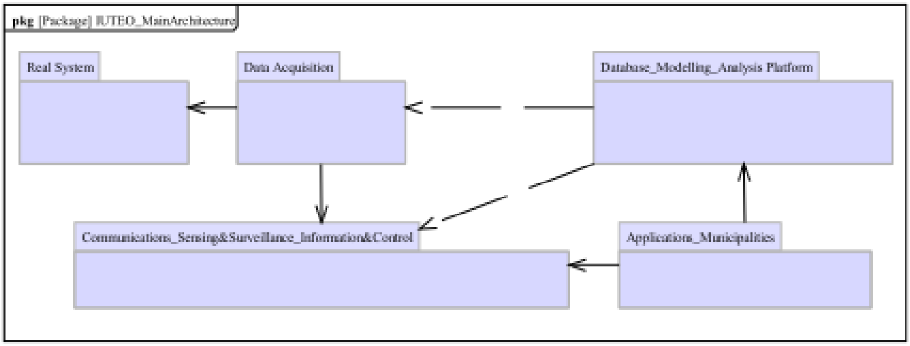

pkg [Package] IUTEO_MainArchitecture
Real System
Data Acquisition
Database_Modelling_Analysis Platform
Communications_Sensing&Surveillance_Information&Control
Applications_Municipalities

Figure 2.11: SysML package diagram for IUTEO framework

As given by Ramos et al. [7] telegeo-processing is formed of the association of remote sensing, spatial databases (GIS), and GPS and telecommunication systems to support real-time decision making. The infrastructure independent typical vehicle techniques use GPS to collect data. The procedure of collecting data is a GPS equipped vehicle driven in a traffic stream receiving data from road and traffic conditions and vehicle performance parameters. Then the data will be telecommunicated either offline or online to a traffic control center. After that, the data will be transformed with map-matching, data reduction, and data processing procedures and finally, it will be reported.

The main advantages of the integration of GIS/GPS given are; capacity to collect data in every second and from everywhere in the urban network, positional and other traffic data which can be stored automatically and used for real-time operations which also provides the ability to display the spatial data in a GIS environment which makes it easy to analyze and integrate with other relevant data.

The main disadvantage pointed out is, that there is a huge amount of data to process. Data transaction costs for real-time observation, and the biased data. The given integrated system has; a prototype for a Mobile system which is used for traffic data acquisition, a geo modeling and analysis system prototype to store and manipulate the collected data in GIS-T environment, and a public information system prototype to broadcast information to the public via WebGIS Application. Figure 2.12 depicts the Ramos et al. [7] GPS/GIS/PDA prototypes developed.

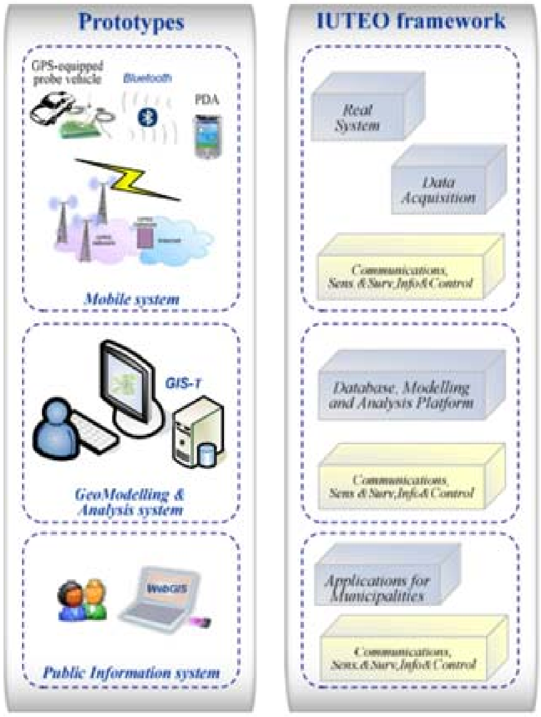

Prototypes
GPS-equipped probe vehicle
Bluetooth
PDA
Mobile system
GIS-T
GeoModelling & Analysis system
WebGIS
Public Information system
IUTEO framework
Real System
Data Acquisition
Communications, Sens & Surv, Info & Control
Database, Modelling and Analysis Platform
Communications, Sens & Surv, Info & Control
Applications for Municipalities
Communications, Sens & Surv, Info & Control

Figure 2.12: GPS/PDA/GIS prototypes for IUTEO

The telegeo-processing system has enabled the real-time data acquisition of traffic data and the distribution of the information gathered via WebGIS application. As stated by Ramos et al. [7] the proposed system will be beneficial for medium-sized municipalities with budget constraints to cost-effectively expand traffic and environment data collection coverage.

Halaoui [9] has presented the driving traffic problem and a solution for finding an efficient path between two points. It has described some available spatial databases, current solutions, and then have come up with a smart solution which is an extension of the previous solution using A* algorithm, a known artificial intelligent algorithm. The algorithm is named as the A* Traffic which was introduced by the same author's previous research. In the solution given by Halaoui [9] uses the time as the main factor in the graph representing the road network.

As given in Halaoui [9] a brief description of spatial databases; the spatial databases are the main databases that are used for geographical systems. It helps to store geographical information such as geometries, positions, coordinates and so on. Geographical information systems also known as GIS is a collection of hardware and software which helps to capture, manage, analyze and display all forms of geographical information. The main factors considered in GIS are distance, road situation, and road traffic. He has used two graph searching algorithms proposed by Russel and Noving [18]. They are Greedy best-first search and A* search algorithms. As explained by Halaoui [9] A* is a graph search algorithm proposed by Pearl and the main goal is to find the cheap cost graph path between two nodes in a graph using a heuristic function. Heuristic function has been used to minimize the selection list. This process has been identified as a very expensive and time consuming one and the A* Traffic is a variation to A* which can take traffic into consideration when computing the driving direction solution. The considered new factor of this algorithm is the average traffic value. In dynamic A* Traffic algorithm it assumes receiving online data when there is a related change. For an example if a road is categorized as 'heavy traffic' the online system will follow the process of; first calculating the average speed of the moving cars (AV) which does not exceed the max limit in the heavy traffic road. Then gets the distance (D) of the road where the heavy traffic exists. Finally sends T where,

$$T = (AV/D) * 60. \quad (2.2)$$

One advantages of the above algorithm is that it saves time when finding a path which guarantees a good solution but not an optimal solution. Another advantage is, that distance and speed are taken into consideration when applying time-weighted graphs which allows returning the fastest path than the shortest path. The main idea of applying such algorithms given by Halaoui [9] is to find a fast user solution for driving path in quickly and efficient manner.

In Cotton [12] it is said that the ITS can provide a city-wide visibility across the entire transportation network and the city services which depends on it to improve the responses on incidents. The cross-agency communication, the collaboration of incident response and infrastructure maintenance can support scenario planning in anticipation of natural disasters and other events by integrating the ITS with centralized databases platforms such as IBM’s Intelligent Operation Centers. IBM’s transportation product is developed as the software foundation for transportation module of the Intelligent Operation Center. Three main components of the product to support the key functions for historic and real-time views of traffic within transportation authority command center are; standards-based integration with traffic and road data capture systems, traffic event data modeling and storage and access to event data.

When ITS is integrated with a system like IBM Intelligent Operation Center the city managers gain a powerful city-wide view of traffic and most importantly, it shows how the traffic impacts on other city operations and services. This is a step ahead of traffic management and decision support with other city services. The output of these traffic prediction tools can be used by the authorities to identify the traffic patterns that impact on other activities.

2.4 Summary of Existing Work

The chapter discussed the existing work on Ride-Sharing, carpooling and intelligent traffic handling. From all these concepts Ride-Sharing is selected to develop as the solution for increasing vehicular traffic in this piece of work. Ride-Sharing handles the vehicular traffic in an indirect way by reducing the number of vehicles entering into a city. Handling traffic using artificial intelligence requires more work. It is hard to gather data for such systems. The algorithms can be implemented as described in existing solutions, but collecting data is difficult. It is not sufficient collecting vehicular density on a road on peak hours with CDR data. CDR data can be used to identify the density of an area at a given time. The density of an area does not mean that all those people are on the road. In urban cities, there are many buildings and offices with thousands of people. Hence the density will be higher in those areas during peak hours only. Surveillance cameras can be used to identify the traffic at a given time visually. Implementing surveillance cameras is expensive and need lots of image processing. Analyzing vehicular traffic only using CDR data is not a very good option.

The carpooling concept helps to reduce the vehicles by letting a set of people to share their day to day rides. The carpooling will mainly solve the parking issue of a company and at the same time it will reduce the number of vehicles entering into the city. When considering the Ride-Sharing it is not necessary to stick to a set of people as in carpooling. The relationships can be built using networks and allows sharing the rides with strangers. It is not safe to share the rides with totally unknown people. Considering the existing works discussed above we have some common methodologies in all three ways of handling the vehicular traffic. Ride-Sharing can be implemented as either planned or dynamic. Dynamic Ride-Sharing is more convenient as people do not have to stick only to a set of people. People can select with whom to share the ride based on their source/destination locations. In dynamic Ride-Sharing there is a high security concern. If an individual travel with a set of known people daily, gradually they get to know each other and safety will be guaranteed. In dynamic Ride-Sharing the issue is selecting the riders dynamically and no individual would like to travel with strangers. Hence a mechanism should be introduced to identify the relationships between ride-sharers before selecting an individual to share the ride. As mentioned in previous works social data can be used to identify the connections between people who are willing to share the rides.

As discussed in the above sections Ride-Sharing can be implemented as end-point Ride-Sharing as well as en-route Ride-Sharing. In end-point Ride-Sharing, the people who live/work in same locations tend to share rides and in en-route Ride-Sharing, the passengers who live/work on the drivers' route can join, based on the vehicle capacity. As mentioned before if the passenger sticks to a set of people rather than dynamically selecting the riders, if the driver would not go to work on a day, that person would be in trouble. Therefore, the Ride-Sharing services should allow the passengers to select a rider based on the time and destination they travel in a dynamic manner. As discussed in Cici et al. [3] and some other existing work, social network data can be taken into consideration to identify the relationships between people. When the matching agency systems give the potential Ride-Sharing options passengers and drivers can consider the list and select people to travel with, based on the relationships they have.

To analyze and identify the home/work locations of people who are willing to share the rides can be done using the cluster based analysis as mentioned in several existing works. This methodology seems to be a good method of identifying the potential people that can share the ride without increasing distance they must travel. CDR, GIS, GPA and PDA data can be used as the source data for this purpose as given above.

Dynamic Ride-Sharing methods, decision support systems, time windows, cluster analysis and distance matrixes are some of the most important points discussed in the presented existing works that would help to come up with better solutions for dynamic Ride-Sharing. Distance matrices can be used to identify the shortest path to get a passenger from the road networks to reduce the total distance of travel and to reduce the traveling cost.

As discussed, the planning, pricing, and payment are some of the main functions that need to be taken into consideration when building a Ride-Sharing service. By sharing the ride, it should be able to reduce the traffic congestion and the travel distance of a person should not be increased too much as well. The travel cost should be evenly shared between the passengers. As given in the Furuhata et al. [15] the patterns should be carefully identified and the passengers should be classified accordingly. The presented classes in Furuhata et al. [15] can be used for any Ride-Sharing system to implement a better solution as it has considered most important points. The challenges identified by Furuhata et al. [15] also very important so that the developers can be more focused to eliminate those challenges in future work.

Overall, the above given existing works give a basic idea of intelligent traffic handling, Ride-Sharing and carpooling along with the methodologies they have followed. Most important points are the drawbacks and the challenges they have identified in their systems and the concepts. These ideas and the methodologies can be used for the proposed system to provide a more accurate and efficient Ride-Sharing service.

CHAPTER 3: METHODOLOGY

This chapter will discuss the datasets used for this study and how the data was extracted to match the scenarios. This will further discuss the algorithms used to identify the home/work locations of the mobile subscribers and to classify the Mobile BSTs to identify the areas to clusters of the uses. The rest of this Chapter will be organized as; Section 3.1 discusses what is CDR Data and how to use this data to model Ride-Sharing. Section 3.2 discusses the methodologies and algorithms suggested, followed and used for the study. This section has been broken into four subsections based on the different methodologies and algorithms. Those methodologies and algorithms are, clustering cell towers, identification of home/work locations, identification of Ride-Sharing potential and proposed algorithms for Ride-Sharing along with a pricing mechanism. Section 3.3 describes the three main datasets used for this study and the Survey published for known participants to identify the willingness to share the rides. Further, Section 3.4 discusses the architecture used for the analysis and Section 3.5 discusses the technologies used. Finally, Section 3.6 discussing the benchmarking performed on methodologies and algorithms based on a sample dataset.

3.1 CDR Data

For this study Cell-Phone Data are used to identify the Home/Work locations of the people and to analyze the potential of Ride-Sharing in Sri Lanka. CDR stands for the Call Detailed Records. These CDRs are generated from the BTS tower which is also known as Base Transmission Station. BTSs are facilitating the wireless communication between the cell phone networks and the cell phone equipment. Figure 3.1 depicts an image of a BTS tower.

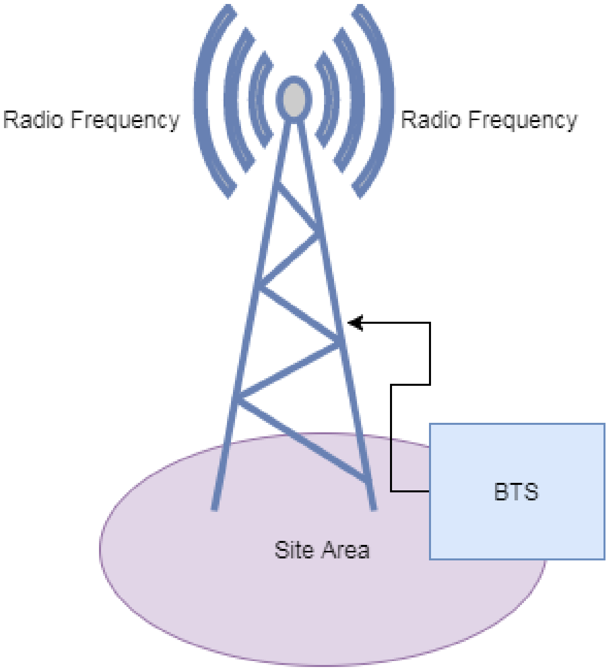


Figure 3.1: Cell Tower Base Station (BST)

A BTS is a wireless communication station installed at a fixed location. Their geographical locations can be expressed by longitudes and latitudes. The area covered by the BTS tower is known as a Cell. The area of a cell varies from urban areas to rural areas. In urban areas, there are many BTS towers available and the network coverage is high. Hence, one BTS needs to cover only few hundred square meters, whereas in rural areas, number of BTS towers are lower and one might need to cover 2-3 square kilometers. At a given moment, one or more BTS towers can provide coverage to a cell phone device. The process of assigning a BTS is; when a person makes a call or any other telecommunication transaction based on the network traffic at the given moment and the location of the person, a BTS will be assigned. Then the call will be routed via the assigned BTS. Figure 3.2 depicts how the BTS towers and cells are situated in an area.

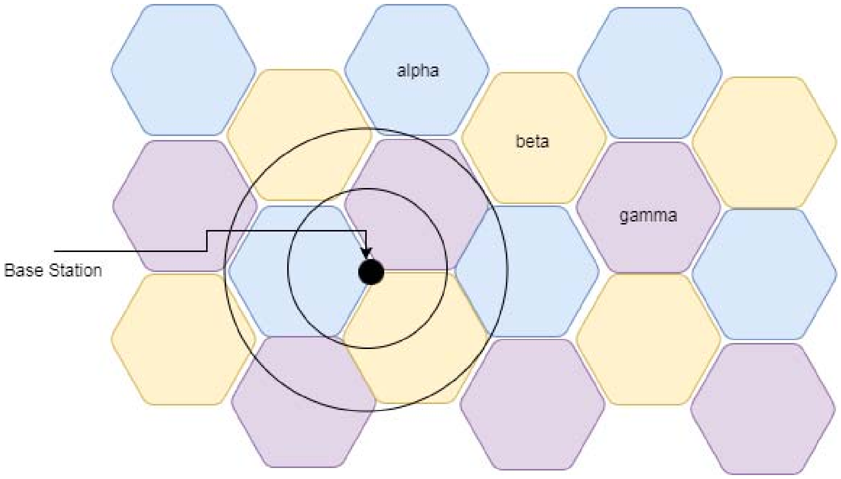


Figure 3.2: Cell Representation

Whenever a person makes or receives a call or any other telecommunication transaction such as SMS, MMS, a CDR will be generated. This CDR consists of the information of the BTS tower that provides the coverage. Information includes the data/time and the location BTS. As per Cici et al. [2], [3] the main fields of the CDR includes, (1) the source cellular phone number, (2) the destination cellular phone number, (3) the date/time of the call started, (4) the call duration and (5) BTS tower id of the originating cell phone device or both BTS towers ids of originating and destination cell phones. It is not possible to capture the exact location of the cell phone user as the location id. Hence, the location is recorded as same as the location of BTS tower.

3.2 Methodology and Algorithms

This section explains the followed methodologies to identify the potential of Ride-Sharing in the Western Province of Sri Lanka along with algorithms used to extract and identify the home/work locations of the subscribers in the CDR Dataset.

The methodology of this study follows three steps to identify the potential of Ride-Sharing. The first step is to cluster the cell towers available in the dataset which includes all the cell towers in the Western Province of Sri Lanka. The second step is to identify the home/work locations of all the mobile network subscribers in the dataset in terms of the cell towers. The third step is to cluster all the subscribers based on the cell tower clusters and identify home/work locations in terms of virtual locations identified in the cell tower clustering algorithm. Rest of this chapter describes the methodologies and the algorithms used in steps mentioned above.

3.2.1 Clustering cell towers

The first step of the methodology was to cluster the cell towers. As mentioned in the previous sections the actual dataset consists of all the available cell towers of the Western Province of Sri Lanka. As mentioned in the previous chapters, there can be many cell towers available in a location and when an individual performs a mobile network event such as making a call or sending an SMS, the event will be routed via one of those nearest cell towers in the area. It is not guaranteed that the same cell tower will be assigned to an individual always. The assignment of a cell tower will be done based on the network traffic of the tower at a moment. This is explained more in the section CDR data. This means that the calls/ messages sent from one individual can be routed via different cell towers in the same area. Hence, cell tower clustering to identify an area of an individual is more important for this study before identifying the home/work locations of the mobile network subscribers. This area identified from the cell tower clustering will be called as a virtual location in the rest of the study. This virtual location consists of almost all the cell towers that can be assigned to an individual. Based on the virtual locations, home/work locations of the individuals will be identified.

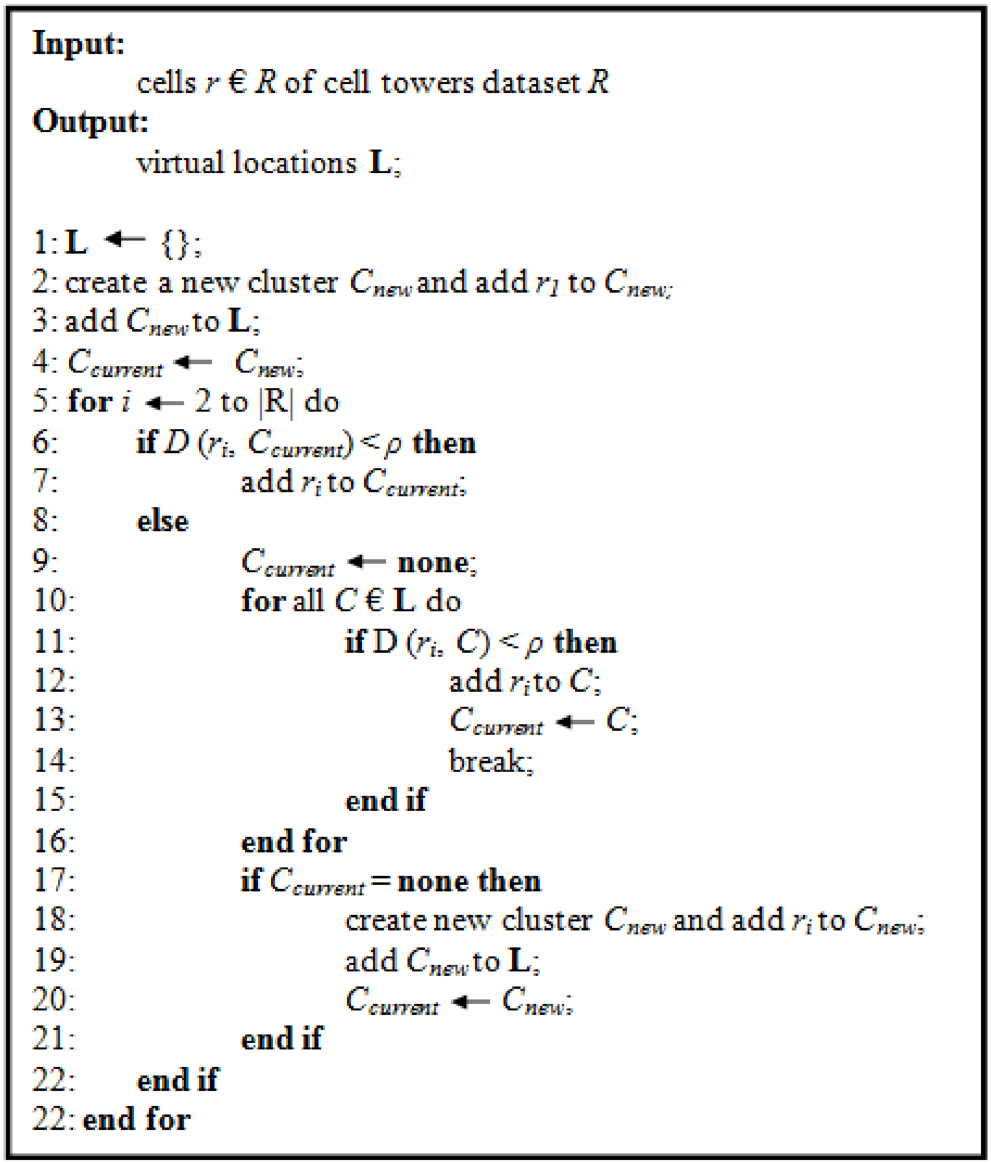
To build up an algorithm to identify these virtual locations, first, a threshold radius was defined. The defined threshold radius was 1km (1000m) based on Maldeniya et al. [29], [30]. As per Maldeniya et al. [29], [30] this diameter is chosen as an appropriate trade-off between the level of special resolution and the reduction of noise due to localization errors particularly in areas with very high tower density. The basic idea is to identify all the cell towers located in the radius of 1km and cluster them to a virtual location id which is named as 1km_cell. The algorithm is developed based on the algorithm introduced by Isaarcman et al. [16] which is again based on the Hartigan's leader algorithm [19]. The leader algorithm is a clustering algorithm that does not require a predefined number of clusters before starting the clustering. As per the Han et al. [20], most of the clustering algorithms such as K-Mean algorithm require a desired number of clusters before starting the clustering of a dataset. When the dataset is large as 5-10 GB of size as the dataset used for this study, it is not practical to identify the number of clusters in advance. Hence, an algorithm that defines the clusters dynamically will be the most suitable option for a study as such.

In the leader algorithm the first cell tower of the list will be as the centroid of the first cluster and then for all the other subsequent cell towers, it checks if the cell tower falls within the threshold radius of the centroid of an existing cluster. The threshold radius considered here is defined as ρ and it is 1000m or 1km. If the cell tower does not fall within an existing cluster, then a new cluster will be created. This is how the leader algorithm works as per the Hartigan et al. [19]. The distance between two cells will be calculated by considering their geographical locations d. The distance calculation of the study was based on Haversine distance calculation formula which shows in Formula (3.1). Mercator projection can also be used to calculate the distance as Sri Lanka is located close to the equator as per Gall [32] and Tobler [33]. However, this study used data preprocessed by LIRNEasia and they have used Haversine formula to calculate the distances between cell towers.

$$\mathbf{d} = 2r \sinh^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi 2 - \varphi 1}{2}\right) + \cos(\varphi 1)\cos(\varphi 2)\sin^2\left(\frac{\lambda 2 - \lambda 1}{2}\right)}\right) \quad (3.1)$$

The explanation of the Formula (3.1) is, d: is the distance between the two points (along a great circle of the sphere; see spherical distance), r: is the radius of the sphere which is the threshold radius selected, $\varphi 1$: is the latitude of the cell tower 1 and $\varphi 2$: is the latitude of cell tower 2, in radians, $\lambda 1$: is the : longitude of cell tower 1 and $\lambda 2$: is the longitude of cell tower 2, in radians. If the distance d is within the threshold radius ρ then a cluster is created by calculating the midpoint of the two clusters.

For this study, algorithm is developed as aforementioned. With the first cell tower in the dataset, the first cluster was created and makes this tower as the centroid of the created cluster. Then each subsequent cell tower was checked if that falls within the 1km distance of the first cell tower and if it falls within the threshold radius then those towers are merged together and the midpoint of those cell towers is taken as the centroid of the cluster. This centroid is the virtual location of the virtual cell tower identified from the algorithm. If there is another cell tower in the dataset that falls within a 1km radius, of the newly created cell location then again those are merged and the midpoint will be calculated again. Likewise, algorithm repeats. If the next cell tower does not fall within the 1km radius of already created virtual location then a new cluster is created and the algorithm repeats until all the cell towers find a virtual cell tower location. This new cell tower id is named as 1k_cell and included into the dataset. This algorithm provides a cell area with larger coverage when there are many cell towers together. This algorithm clusters all the cell towers in the dataset into Virtual Locations. Once the home/work locations of the available subscribers in the dataset are identified from the algorithm described in the Section 3.2.2, all these home/work locations will be again masked with Virtual Location id of the cell tower which is called the 1K_Cell_Id in the study. These virtual locations will be used to create the source/destination key of each subscriber in the dataset and based on the source/destination key the potential ride-sharers will be identified. This will be described in detail in Section 3.2.3. The pseudo code in Figure 3.3 depicts the algorithm.



```
Input:
        cells $r$ € $R$ of cell towers dataset $R$
Output:
        virtual locations L;

L ← {};
create a new cluster $C_{new}$ and add $r_1$ to $C_{new;}$
add $C_{new}$ to L;
$C_{current}$ ← $C_{new;}$
for $i$ ← 2 to |R| do
    if $D$ ($r_{i,}$ $C_{current}$) < $\rho$ then
        add $r_i$ to $C_{current;}$
    else
        $C_{current}$ ← none;
        for all $C$ € L do
            if D ($r_{i,}$ $C$) < $\rho$ then
                add $r_i$ to $C$;
                $C_{current}$ ← $C$;
                break;
            end if
        end for
        if $C_{current}$ = none then
            create new cluster $C_{new}$ and add $r_i$ to $C_{new;}$
            add $C_{new}$ to L;
            $C_{current}$ ← $C_{new;}$
        end if
    end if
end for
```

Figure 3.3: Cell tower clustering algorithm

As mentioned in earlier chapters, although the urban areas consist of many cell towers close to each other, in rural areas distance between two towers are high. Hence, when choosing a threshold radius, selecting a value that gives coverage of a larger area is needed. As mentioned in the Isaarcman et al. [16] and Maldeniya et al.

[29], [30], have selected this threshold radius after experimenting with a range of radius values due to the same reasons mentioned earlier.

3.2.2 Identification of home/work locations

In this study, home/work locations are identified as the most common locations people travel more frequently in day to day life. In Isaarcman et al. [16], they have considered all the activities people might perform to identify all the favorite locations they tend to travel. However, this study is limited to identify only home/work locations of the mobile phone subscribers. This section describes how the algorithm is developed to identify the home/work locations in this study.

The dataset was first separated into two subsets based on the time the CDR data is recorded. The time based on the assumption that the nighttime data is recorded from the home location and the daytime data is recorded in the work locations. In the existing work chapter, it was mentioned that according to the Cici et al. [2], [3] these periods were selected as 19:00 – 7:00 as time spent in home location and 1 pm to 5 pm as the time spent in the work locations. According to them, this is the most accurate timelines to identify the home/work locations of people. For this study, the subsets of data were created as Nighttime and Daytime based on Maldeniya et al. [29], [30]. The nighttime was selected as 21:30 to 05:30 of each day, where it is analyzed that a person is most likely to be at home. The Daytime was selected as 10:00 to 15:00 of each day where it is analyzed that a person is most likely to be at work. This timeline varies from the timelines mentioned in Cici et al. [2], [3]. This was identified as the most suitable timelines for Sri Lanka and considers the people who are working for shifts as well. Then the algorithm counts the number of records for each mobile phone subscriber for each cell during the nighttime and the daytime. Once these counts are taken, the algorithm identifies the cell_id that appears most times (mode) as the home location by considering the subset with nighttime records and identifies the cell_id that appears most times (mode) as the work location by considering the subset with daytime records. Then the two subsets are merged so that the home location cell_id and the work location cell_id for each subscriber are listed. To ensure that this dataset contains only Western Province data, a filtration has been run to filter out any subscribers who have a work or home location other than Western Province. Figure 3.4 depicts the graphical representation of the algorithm used identify the home/work locations of the subscribers in the CDR dataset.

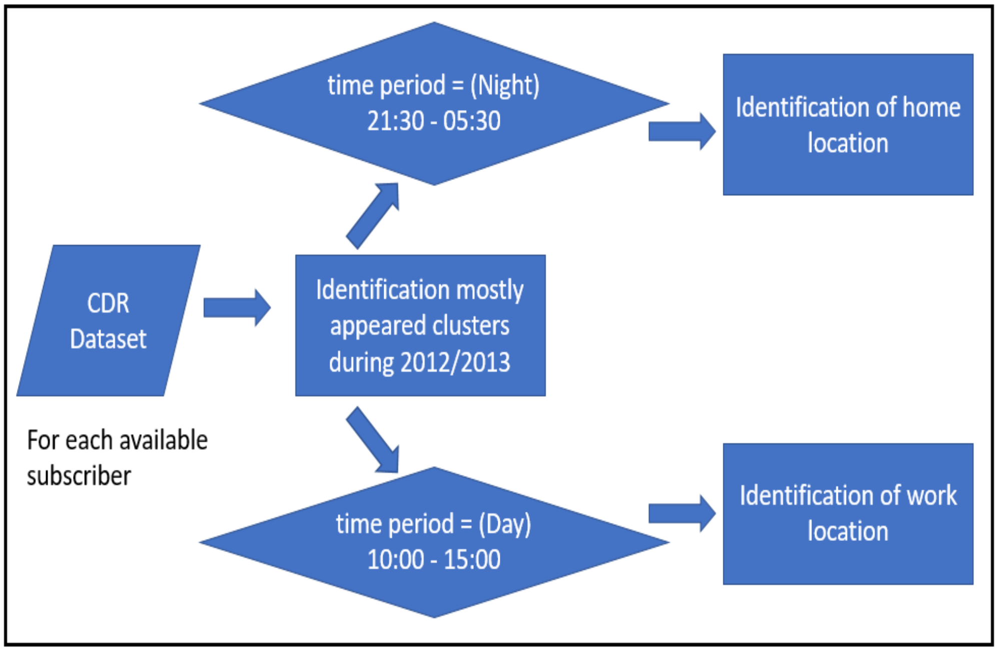

time period = (Night)
21:30 - 05:30
Identification of home location
CDR Dataset
Identification mostly appeared clusters during 2012/2013
For each available subscriber
time period = (Day)
10:00 - 15:00
Identification of work location

Figure 3.4: Identification of home/work locations of subscribers

3.2.3 Identification of potential ride-sharers

This section describes the algorithm developed to identify the potential of Ride-Sharing. As mentioned in the previous two sections; clustering cell towers identified a list of virtual locations of cell towers based on the 1km threshold radius and the identification of home/work location has identified the home cell_id and work cell_id of the mobile phone subscribers. In the cell tower clustering algorithm explained in Section 3.2.1, a virtual cell_id is created for all the cell towers in the dataset based on the 1km radius. As the next step; the second dataset with the home/work cell_ids of the mobile phone subscribers, home_cell_id and the work_cell_id will be replaced by the relevant virtual cell_id of the home_cell_id and work_cell_id. Then the algorithm runs to identify the subscribers with same home_virtual_cell_id and the same work_virtual_cell_id. These are the subscribers who have a potential to share the ride. Figure 3.5 depicts the graphical representation of the algorithm to identify potential ride-sharers.

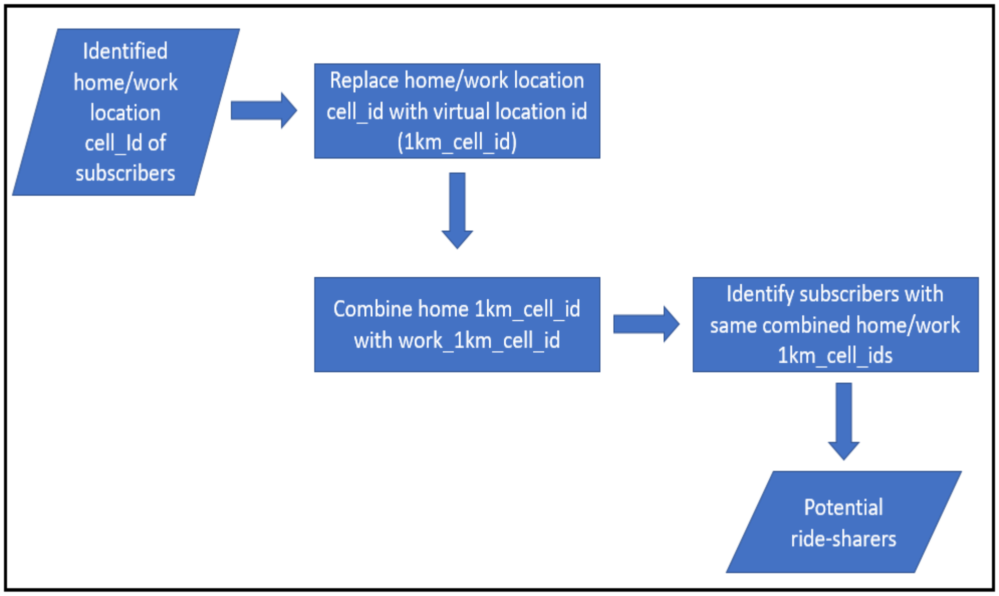

Identified home/work location cell_Id of subscribers
Replace home/work location cell_id with virtual location id (1km_cell_id)
Combine home 1km_cell_id with work_1km_cell_id
Identify subscribers with same combined home/work 1km_cell_ids
Potential ride-sharers

Figure 3.5: Identification of Potential Ride-Shares

A filtration runs to remove the subscribers whose work_virtual_cell_id is equal to home_virtual_cell_id. These subscribers are most likely to be people who stay at home or have their offices at home. To make the more accurate and valuable output, these filtrations are used. Figure 3.5 depicts the methodology of identifying potential ride-sharers.

3.2.4. Proposed Ride-Sharing models

As impacted by Cici et al. [2], [3]; this study proposes two Ride-Sharing algorithms that can be implemented based on the analysis of potential Ride-Sharing in Sri Lanka. The two algorithms are End-Point Ride-Sharing algorithm and En-Route Ride-Sharing algorithm.

The End-Point Ride-Sharing algorithm is sharing the rides with the neighbors; the people who are located close by. As mentioned in the previous sections, the study clusters the subscribers based on the 1k_cell_id which is the identity of the virtual location by clustering the cell towers within 1km radius area. This algorithm helps to identify the subscribers who live close by. The proposed algorithm has few steps to identify End-Point ride sharers. First; it should identify the potential drivers and let the drivers route their cars along with the capacity for passengers. Second; the geographical distance between the driver and the potential passenger should be identified based on the original cell tower geographical locations. These locations will be denoted by longitude and latitude. Third; identify the corresponding distance between the work location of the driver and the passenger. A maximum distance for both the distances of home locations and work locations of the driver and the passenger should be defined. The distance between the driver and the passenger from both home/work locations should be less than this maximum distance to share a potential ride with a driver. Fourth; the departure time and the arrival time of the home/work location of both the driver and passenger should be identified. The arrival and the departure time from home/work locations of drivers and passengers also should define a maximum difference. Again, the differences between the arrival/departure time from home/work locations of both driver and the passenger should be less than the maximum time differences defined to be able to share the ride with the driver. Finally, should check the capacity availability of the vehicle of the selected driver. Figure 3.6 depicts the End-Point Ride-Sharing.

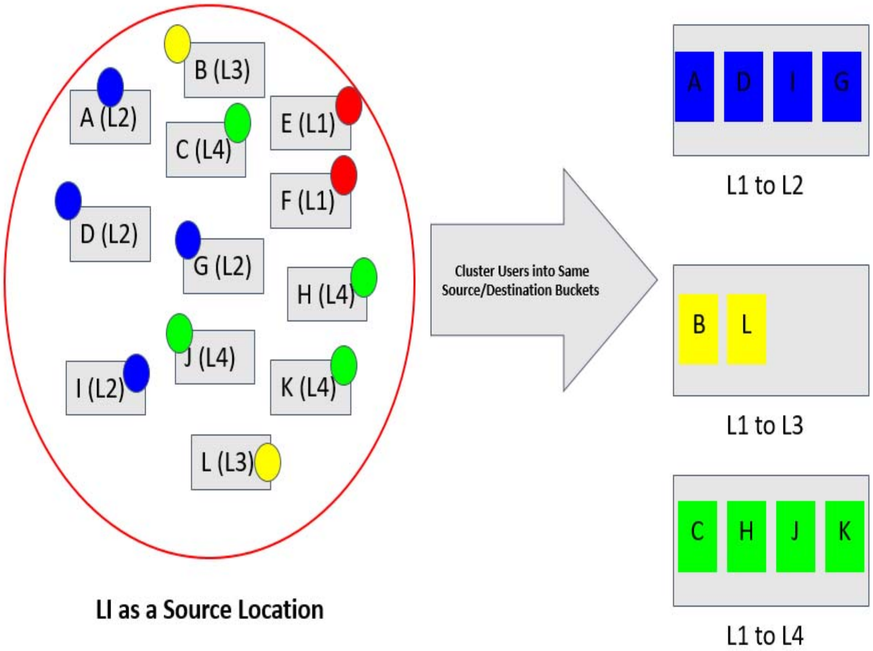


Figure 3.6: End-Point Ride-Sharing

En-Route Ride-Sharing is sharing the ride with the people who are on the route of the driver from home to work and vice versa. To identify the En-Route ride-sharers the nearest 1k_cell_towers can be considered. If the vehicle does not get filled from the current cluster then it is possible to find the passengers from the nearest clusters on the same route of the driver. To implement this algorithm two steps can be followed. First; run the previously mentioned basic End-Point Ride-Sharing algorithm. Second; if the vehicle still has the capacity to share the ride, route the car again to find more passengers on the route. Figure 3.7 depicts En-Route Ride-Sharing.

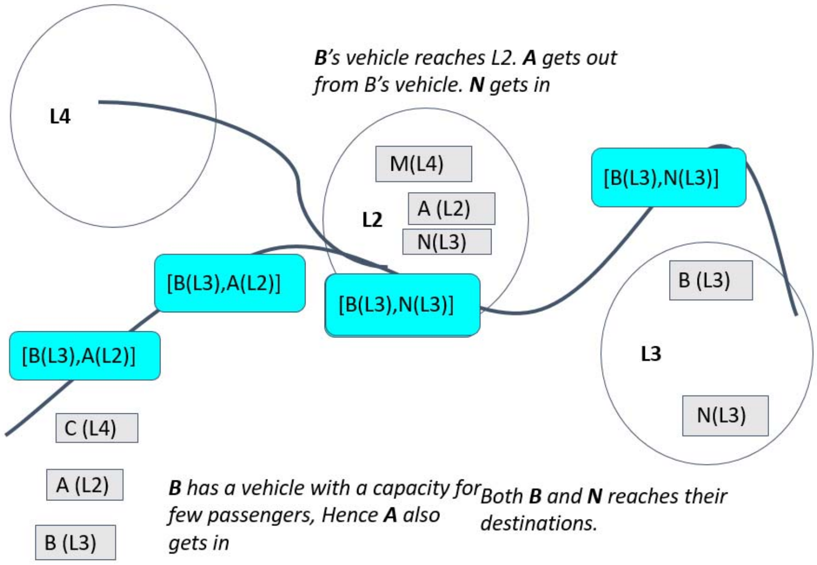

B's vehicle reaches L2. A gets out from B's vehicle. N gets in
L4
M(L4)
A (L2)
L2
N(L3)
[B(L3),N(L3)]
[B(L3),A(L2)]
[B(L3),N(L3)]
B (L3)
[B(L3),A(L2)]
L3
C (L4)
N(L3)
A (L2)
B (L3)
B has a vehicle with a capacity for few passengers, Hence A also gets in
Both B and N reaches their destinations.

Figure 3.7: En-Route Ride-Sharing

Considering both the above-mentioned scenarios, a dynamic Ride-Sharing model can be proposed considering vehicle owner's perspective and the passenger's perspective.

Model based on the vehicle owner:

1. Register the details; Source Location, Destination Location, Departure Time, Vehicle Capacity and the details of the vehicle and the owner.
2. Route the vehicle and pick up the passenger who is closer to the source location and the departure time.
3. If the passenger is willing to share the ride with other passengers, then route the vehicle again including the details of the passenger already picked up. All the other passengers should be picked within a threshold radius such as 1000m (for End-Point Ride-Sharing) and destination location within the driver's route.

4. The third step can be repeated until the vehicle is full.

Model based on the passenger:

1. Register the passenger details along with the Source Location, Destination Location, Departure Time, the expected Arrival Time and willingness to share the ride with other passengers.
2. Search for a vehicle and confirm an available vehicle based on the details.

A pricing model can be proposed for this model as below.

1. Mark the driver's source and the destination locations along with the total distance.
2. Identify the source and the destination location along with the total distance of each passenger. The source and the destination locations of all the passengers should be within the driver's source and the destination.
3. The price will be shared as follows,
    a. $D_{driver:}$ driver's total travel distance
    b. D_{pi}: passenger's total travel distance
    c. Identify the shared distance between each passenger and the driver.
    d. Identify the shared distance of each passenger.
    e. Then share the amount based on the identified shared distances.

For these proposed models, drivers and the passengers should share their details with each other. Some of those details are pickup locations, payment methods and identity information. This is the reason that Ride-Sharing implementation should consider on privacy and the security of the ride-sharers. Also, as mentioned above a proper pricing mechanism and a payment methodology should be implemented, so that the ride-sharers do not have to share their payment details with each other.

3.3 Datasets

For this study, three main datasets are used along with a result set of an online survey. The first dataset is used for benchmarking the algorithms and methodologies for this study. The second dataset is the actual CDR dataset of the Western Province of Sri Lanka received from leading mobile companies of Sri Lanka to identify which is used to identify the home/work locations of the subscribers as well as the subscriber who has the potential to share the rides. The third dataset contains the transportation data of Sri Lanka which has been collected for a survey conducted by Department of Transport and Logistics, Faculty of Engineering, University of Moratuwa Sri Lanka in the year of 2012/2013 JICA [26]. This third dataset was used as a supportive evidence for the analysis of this study to prove the need of Ride-Sharing in Sri Lanka. The first dataset used for benchmarking was the Nodobo-2011-01-v1 McDiarmid et al. [22] and Bell et al. [23]. Nodobo dataset was gathered during a study of mobile phone usages of 27 high-school students. This includes data from September 2010 to February 2011. Table 3.1 shows the number of records that consist of the Nodobo Dataset.

Table 3.1: Number of data in Nodobo Dataset

Record Type	**Number of Records**
Call Records	13035
Message Records	83542
Presence Records	5292103

The size of this dataset was 1GB. In this dataset, there were database tables as Calls and Messages, Cell Towers, Devices, Presences, Users and Wifis. The detailed descriptions of the dataset and the tables will be described below.

The table Calls and Messages consist of fields; other_id, number, duration and length. Other_id is the id of the other user on the call (NULL if not in the study); the number is the phone number of the other end of the call/message (related: Users#number); duration is the length of the call-in seconds, and the length is the number of characters in the message. The table Cell Towers consists of the fields; cell id and lac. Cell id is the GSM base transceiver station CID and lac is the location area code. The next table was named as Devices and it consists of fields; imei and mac. Imei is the blank for this release of the data; and mac is the Bluetooth MAC (related: Presences#mac). Table Presences consists of other_id, mac, Bluetooth_class and name fields. Field other_id is the user_id of the detected device (NULL if not in the study); mac is the Bluetooth MAC (related: Devices#mac); bluetooth_class is the reported class of the device, and the name is the human-readable name of the device. Users table consists of name and number fields. Field name has the value "Anonymous" for this release of the data; and the number is the phone number of the studying user (related: Calls#number, Messages#number). The last table was Wifis which consists of fields' ssid and bssid. Field ssid is the human-readable name of the base station; and bssid is base station MAC.

To achieve the anonymity, they have altered some of the fields in the dataset. Those fields are Call #number, Message #number, User #number, Device #mac, Presence #mac, Wifi #bssid, Presence #name, CellTower #cellid, and CellTower #lac. They have made each of real value of these fields' maps 1:1 to a randomly generated anonymous value. The process followed to generate these values are, the phone number has been altered by a random number with the same number of digits. For an example, if original number is 3 or more digits, keep the original first 2 digits. MAC address is replaced by a 12 random hex digits. Bluetooth name/Wifi ssid are altered by a random sequence of dictionary words. A same number of words as original name Cell ID and LAC are altered by random number with the same number of digits. This Nodobo dataset is used to benchmark the analysis and the developed algorithms to use as a POC to receive the actual dataset of Sri Lanka.

The second dataset was the actual dataset of part of Sri Lanka. The analysis of this study was done for the Western Province of Sri Lanka. The sample dataset of the Western Province is selected to analyze the potential of implementing Ride-Sharing in Sri Lanka without analyzing a dataset gathered for the whole country. The dataset was gathered for the years 2012-2013. This includes almost 10 Million subscribers in the Western Province who have subscribed for all the cellular service providers available. The Western province is the urbanist province in Sri Lanka where the capital of the country is also situated. Most of the workplaces of Sri Lanka are established in the Western Province including the leading government agencies and departments. Hence most of the workforce of Sri Lanka is available in this province during the working days during peak hours. Also, most of the people who are working in this province have temporary residencies in the Western Province on working days due to travel difficulties between different provinces. Hence, this dataset covers most of the workforce in Sri Lanka.

This CDR Dataset consists of the fields; CALL_DIRECTION_KEY: which indicates whether it is an incoming or an outgoing call; DEVICE_NAME: which is the model number of the device; ANUMBER: the anonymized identifier for the particular phone number that places the call; OTHER_NUMBER: the anonymized identifier for the other phone number in a voice call; CELL_ID: this is a unique identifier for the cell; CALL_TIME: Date & time field formatted as YYYY-MM-DD:HH:MM:SS; DURATION: this is measured in seconds and assumes that the maximum number of seconds is 60 x 60 x 24 x 30 = 2,592,000 seconds. The process followed to create the unique key identifier of the Cell created was, when the CALL_DIRECTION_KEY equals 2 (i.e. outgoing call) then the CELL_ID corresponds to the cell utilized by the party represented by ANUMBER; and when the CALL_DIRECTION_KEY equals 1 (i.e. incoming call) then the CELL_ID corresponds to the cell utilized by the party represented by the OTHER_NUMBER.

As explained in the above paragraph the anonymity and the privacy of the subscribers are highly maintained in this dataset. The CDR data used for this study has been taken from around 10000 BTS towers available in the Western Province of Sri Lanka owned by different cellular access providers. The dataset of this BTS towers is used to identify the locations of the subscribers for clustering purposes. Hence the actual dataset used consists of BTS tower dataset and the CDR Dataset of the mobile subscribers. This BTS towers dataset mainly consists of the fields, cell_id, site_id, longitude, latitude, Province, District and there are more fields which were not used for this study. Both datasets are used to identify the locations and to analyze the potential of Ride-Sharing in Sri Lanka. The list of regions identified from these two Datasets is shown in Table 3.2 and Figure 3.8 shows the marked map of the Western Province Sri Lanka with all the areas considered in this study.

Table 3.2: List of cities in actual Dataset of the Western Province of Sri Lanka

District	**Region based on Base Station Locations**
Colombo	Colombo North Colombo south Dehiwala Hanwella Homagama Kaduwela Kesbewa Kolonnawa Maharagama Moratuwa Padukka Ratmalana Sri Jayawardanapura Kotte Thimbirigasyaya

District	Region based on Base Station Locations
Gampaha	Attanagalla Biyagama Divulapitiya Dompe Gampaha Ja-Ela Katana Kelaniya Mahara Minuwangoda Mirigama Negombo Wattala
Kalutara	Agalawatta Bandaragama Beruwala Bulathsinhala Dodangoda Horana Ingiriya Kaluthara Madurawala Mathugama Millaniya Palindanuwara Panadura Wallawawita

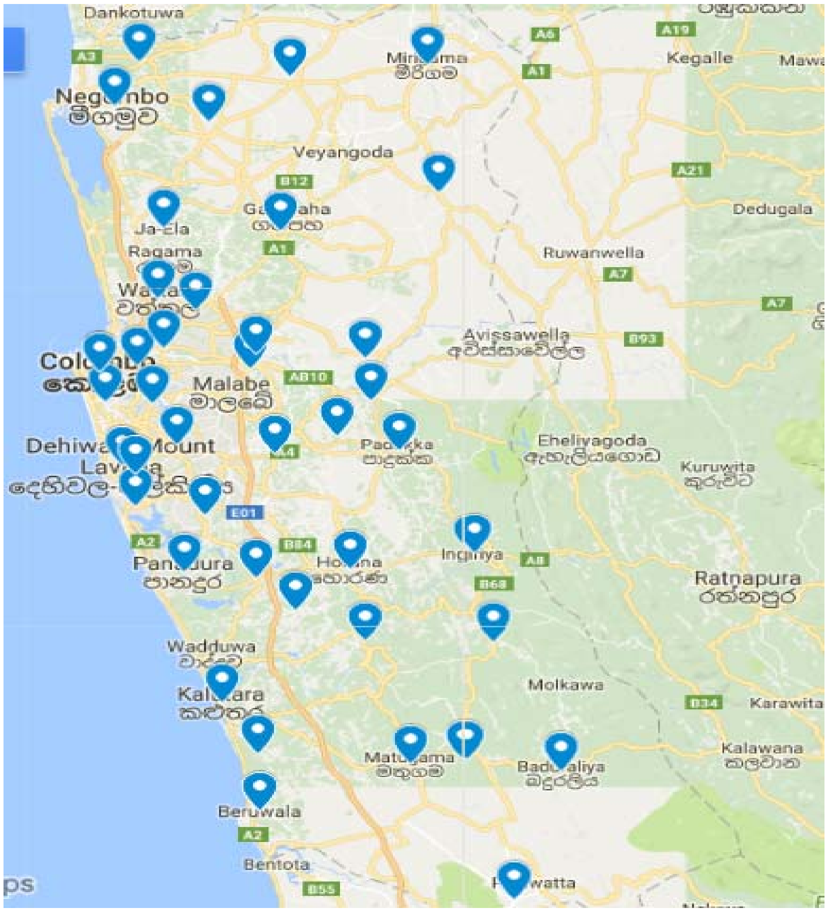


Figure 3.8: Cities of BTS Towers

The next sections of this chapter will describe the algorithms used for clustering, identification of home/work locations and Ride-Sharing.

As mentioned above, the third dataset contains transportation details of Sri Lanka and the study has used it as an evidence to prove the need of Ride-Sharing in Sri Lanka. This dataset contains the number of vehicles used between identified traffic zones. According to this dataset, each Divisional Secretariat is divided into several traffic zones. For an example Thimbirigasyaya Divisional Secretariat in Colombo District, Western Province of Sri Lanka is divided into 22 traffic zones in this dataset.

Basically, a traffic zone represented in this dataset is an area where there is a daily traffic congestion. The dataset has considered 18 different transportation modes used by Sri Lankans. Table 3.3 provides the transportation modes considered in JICA [26] and the way those modes are categorized for this study. Categories used for this study are Private/Public/Ride-Sharing. Some of the modes in the JICA [26] dataset is not applicable for this study and they are marked as N/A in Table 3.3.

Table 3.3: Categorization of transportation modes considered for this study

Mode Identity	Transportation Mode	Transportation Category
1	Walking only	N/A
2	Walking to/from bus stop/railway station	N/A
3	Bicycle	N/A
18	Others	N/A
4	Motorcycle	Private
5	Three-wheeler (private use)	Private
6	Car/Jeep/Van	Private
7	Pickup	Private
8	Three-wheeler (hired)	Private
9	Taxi (car/van)	Private
10	Taxi (Nano)	Private
14	Non- A/C bus (private)	Public
14	Non- A/C bus (SLTB)	Public
16	A/C bus	Public
17	Railway	Public
11	Employee transport	Ride-Sharing
12	Staff service	Ride-Sharing
13	School bus/van	Ride-Sharing

This dataset has provided a number of vehicles used to travel from one traffic zone to another daily by the transportation modes mentioned in Table 3.3. This study has used only data of the Western Province of Sri Lanka.

Also, a small survey was published among 53 known participants to identify the willingness to share the rides. It is more important to check whether people have an intention to share their rides in their day to day travels before implementing the concept. The sample size was decided based on the actual population considered for the study. These will be further discussed in Chapter 4, Section 4.3.

3.4 Architecture

This section briefly describes the architecture used for the analysis of this study. Figure 3.9 depicts the high-level architecture.

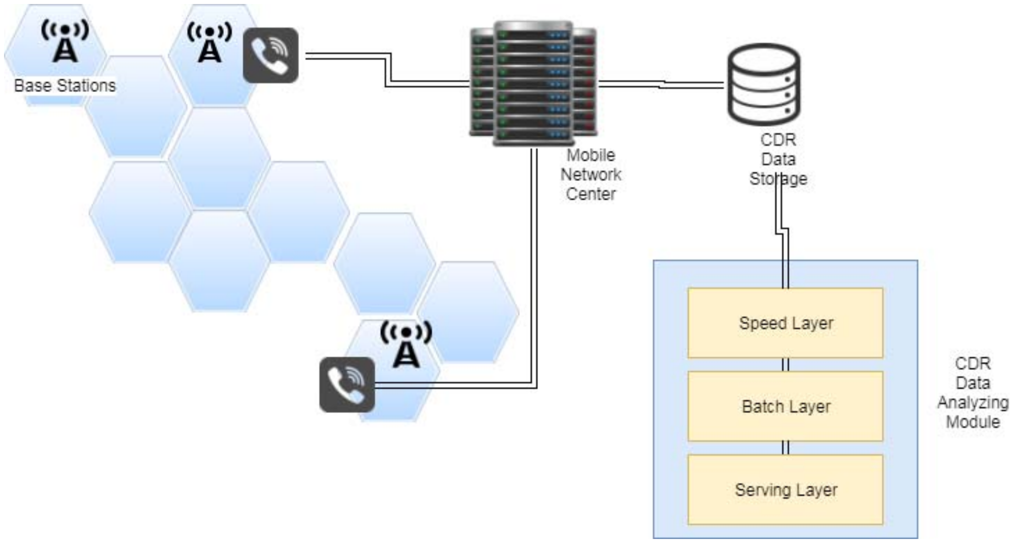


Figure 3.9: High Level Architecture

As described in the previous sections, CDR data is captured by the cell towers which are also known as BTS towers. This CDR data then transferred and stored in a data storage system at the mobile network center. For this study, data was gathered from different GSM Mobile Access Providers. Storing mechanism used by different GSM Mobile carriers are different from each other. These storing mechanisms are proprietary and not exposed to the outer world. These CDR data are fetched real or near real time by the CDR data analyzing module. These fetching mechanisms also differ from GSM carrier to carrier which again is not exposed due to security reasons.

Then the CDR data is fetched into CDR analyzing module. This module is based on the lambda architecture. The lambda architecture is the architecture designed for data-processing which is capable of handling large-scale datasets with the use of both batch processing and stream processing methods. Inside the CDR analyzing module, there are three layers; Speed Layer, Batch Layer, and Serving Layer. At the Speed layer, fetched data are processed and converted to CSV file formats. At the batch layer, converted data are filtered and batch processed for clustering to create processed CSV files. In this study, the Cell Towers Dataset and the CDR Dataset with specific fields are taken at the batch layer. Also, the algorithms to identify the cell tower virtual locations, identification of home/work locations of the mobile subscribers and identification of potential of Ride-Sharing are performed at this layer. Once the files are created, these are fed into Serving Layer to analyze the insights and create graphs, queries etc. Figure 3.10 depicts the High-Level Architecture of the CDR Data Analyzer System.

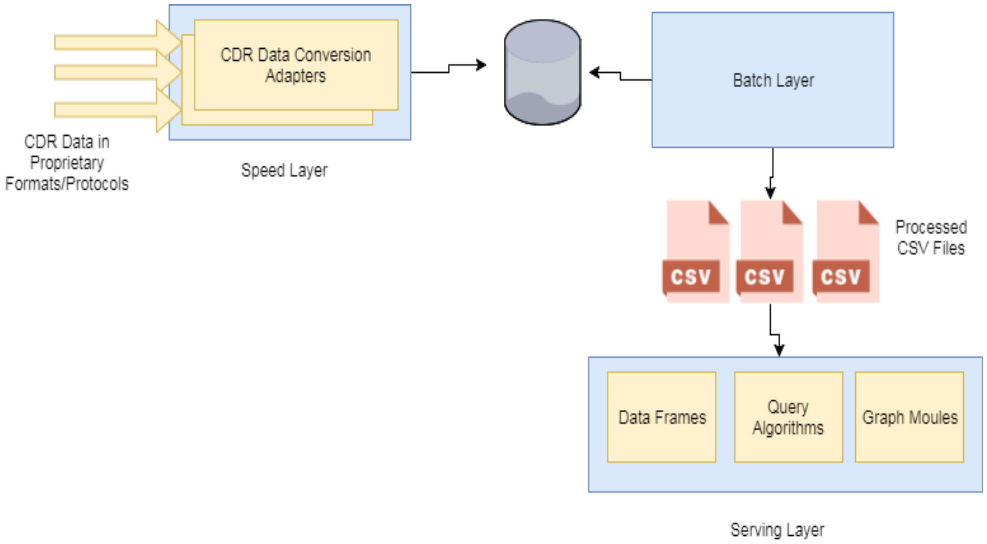


Figure 3.10: CDR Data Analyzer System – High-Level Architecture

3.5 Technologies

Data processing and analyzing of the data was done on Apache Spark Framework. The CDR Data Analyzer module depicted in the Architecture is the module that is directly associated with this study. The batch layer and the serving layer are developed on Apache Spark framework. This is an open-source cluster-computing framework. It is also known as a fast, in-memory data-processing engine. This consists of APIs which allows executing streaming, machine learning or SQL workloads that require fast and iterative access to datasets. Spark runs on Apache Hadoop YARN. This feature of Spark, allows data scientists from different locations to work on the single shared database. This framework consists of a core and a set of libraries. This core acts as the distributed execution engine and the platform for distributed ETL application development is provided by Java, Scala and Python APIs [21].

For this study Spark framework was selected as this is a study performed on a large-scale dataset of CDR. Since the clustering algorithms and data processing needed to be done fast and iteratively Spark was the most suitable framework as mentioned above being a fast, in-memory data-processing engine. For the application development, Scala APIs were used. Scala was selected as it is fast and moderately easy to use on top of Spark Framework as per [21].

Scala has three main data structures; Dataframes, Datasets and RDD. Dataframes are mainly used for this study as a dataframe is an immutable distributed collection of data which are organized into named columns. RDD cannot be organized into named columns [21]. Dataframe is more like tables in a relational database. In Scala a Dataset is either a strongly-typed or an untyped. As per [21] conceptually a Dataframe can be considered as an alias for a collection of generic objects *Dataset[Row]*. A Row is here a generic untyped JVM object. As this study requires to analyze the available CDR and Cell Towers dataset and create datasets in different aspects on the run, Dataframes are used. To create the Dataframes, Lambda expressions were used as they are more efficient and easy and no additional class loading and compilations are needed. Hence the development of the algorithms is done using more efficient, easy to use and faster technologies and data structures. For analyzing and graphing R is used.

3.6 Benchmark methodology and the algorithms

As mentioned in the Section 3.2, Nodobo dataset was used to benchmark the methodology and the algorithms. The dataset consisted of CDR data of 27 users. Out of these 27 users 14 were identified as subscribers with same home/work locations and who have potential to share the ride by following the methodology prepared for this study and the algorithms. Identification of the home/work locations of the subscribers was done using the methodology discussed in Section 3.2.2, identification of home/work locations. All their work cell_id was identified as identical as they are high school students who studied at the same high school. For this study, departure time and the arrival time were not considered to check whether these users leave and arrive at the same time. This is mainly focused to identify the potential subscribers to share the rides based on the home/work locations. Few modifications were done on the output of the dataset of the home/work identification algorithm to fill out the missing entries in this dataset. Once the missing entries are filled, the algorithm to identify subscribers with same home/work locations was executed and the Table 4 shows the outcome of the algorithms. Modifications of the algorithms were done based on the outcomes of this dataset and the algorithms were rerun several times to fine-tune and to receive an accurate output. Subscribers with same cell_id for the home/work location were removed from the list and Table 3.4 shows the number of users identified, who have same home/work locations. Table 3.4 has excluded the routes with single subscriber.

Table 3.4: Home/Work routes identified on Nodobo Dataset

Home Cell Id	**Work Cell Id**	**Number of Users**
11059	25224	4

Home Cell Id	**Work Cell Id**	**Number of Users**
55275	25224	3
16506	25224	2
41479	25224	2
48024	25224	3

The outcome showed in Table 3.4 provided confidence in applying the same algorithms and the methodologies on the actual CDR Dataset. Even though the number of users was small in this Dataset, it had more than one Million records of CDR data. Hence, identifying 14 out of 27 subscribers as potential ride-sharers provided a better insight on the algorithm and the methodology on improvements and benchmarking. This was used to fine-tune the algorithms and the methodology to apply the same to the actual dataset.

3.7 Summary

This chapter discussed what is CDR data and how CDR data can be used to analyze the potential of Ride-Sharing in Section 3.1. Section 3.2 discussed the methodologies and algorithms used for the study. As mentioned in Section 3.1 there can be many BTSs or Cell Towers located in major cities/towns, but in urban areas, there are only few towers based which cover a large area. Hence, an algorithm to cluster Cell Towers is used to cluster the towers that can be assigned to an individual located in an area. This algorithm was explained and described in Section 3.2.1 how the clustering was done based on the pre-identified threshold radius of 1000m. Those identified Cell Tower clusters were given a Virtual Location id and named a 1KM_Cell_Id.

Section 3.2.2. discussed how the home/work locations of the mobile data subscribers captured in the CDR dataset were identified. Then Section 3.2.3 explained the identification of potential ride-sharers based on the identified home/work locations of the subscribers by using the Virtual Locations identified in the Cell Tower clustering algorithm in Section 3.2.1. All the Cell Towers within the Virtual Location were first renamed by the 1KM_Cell_Id to identify the potential ride-sharers. Section 3.2.4 explained the proposed Ride-Sharing models, End-Point Ride-Sharing and En-Route Ride-Sharing. Both these models are explained in detail. Also, the section has described the key data and processes in terms of both the driver and the ride-sharer.

Section 3.3 discussed the Datasets used for the analysis for this study. There are three main datasets, actual CDR dataset of the Western Province of Sri Lanka, Dataset of Transportation Usage in the Western Province of Sri Lanka and the Survey data performed to identify the willingness of Ride-Sharing. Apart from those datasets, another small CDR dataset call NODOBO was used to benchmark the algorithms and the methodologies used for the study before using them on the actual datasets. How the benchmarking is done was explained in Section 3.6 in detail.

The architecture used to analyze the data for this study was explained in Section 3.4. The architecture used for this study was based on the Lambda Architecture which is a data processing architecture designed to handle large datasets with the use of batch and stream processing methods. Then in Section 3.5 technologies used to analyze the data was explained in detail.

CHAPTER 4: ANALYSIS

This chapter presents the analysis of the study based on the data. Section 4.1 discusses the analysis done on the actual CDR Dataset of the Western Province of Sri Lanka by executing the methodology and the algorithms used for this study. Then, in Section 4.2, the Dataset of Transportation details gathered in the period of 2012/2013 was analyzed to identify the usage of public and private transportation in Sri Lanka. This section will discuss the need for introducing and implementing Ride-Sharing based on the current transportation data. Section 4.3 analyzes the data gathered from an online survey published for 50 volunteers as mentioned in Chapter 3 to analyze the ground truth of the willingness to share the rides. Finally, in Section 4.4 a summary of the analysis will be presented.

4.1 Analysis of CDR Dataset of Western Province

As mentioned in previous chapters the actual dataset consists of CDR data of the Western Province of Sri Lanka collected from different mobile access providers. As mentioned in the previous chapter, the dataset used to benchmark the algorithms were used to identify the data to be used and to be removed from the outputs prior to analyzing the potential of Ride-Sharing in the actual dataset. Table 4.1 shows the Number of CDRs and number of unique phones in the Western Province Dataset.

Table 4.1: Number of data in the Western Province Dataset

Metric	Counts
Total Unique Phones	4152611
Total Unique CDRs	10 Million

The study was performed in several steps. First, the identification of the home/work locations was done and the resulted output provided a larger amount of home/work locations. Then for each subscriber's home/work locations were replaced by the 1K cell_id as mentioned in Chapter 3. Then a source/destination key was created by combining the 1K cell_ids of home and work locations of each subscriber in the dataset. This provided the potential routes for Ride-Sharing and the number of subscribers who share the same source/destination key. Table 4.2 shows an example of how the source/destination route key is created.

Table 4.2: Home/Work Route identification

Home 1K_Cell_Id (Source)	14620
Work 1K_Cell_Id (Destination)	14375
Route (Source/Destination)	14620_14375

This source to destination route is named as a sub route in this study. There are many sub-routes identified in all three districts of the Western Province of Sri Lanka. Table 4.3 provides the number of subscribers who have the potential to share a ride from home to work and vice versa and the number of sub-routes identified. This excludes the subscribers who are not going out for work and includes the subscribers who work in the same area. The areas are the nearest city identified for 1K_cells. Table 4.3 shows the number of records in terms of subscribers, number of identified routes (source/destination routes) from the analysis, and the number of cities/areas based on Districts.

Table 4.3: Number of Data in the Western Province Dataset

Metric	**Number of Records**
Number of Subscribers	4152611
Identified sub-routes	329509
Number of Areas Considered	41
Number of Areas in Colombo District	14

Metric	Number of Records
Number of Areas in Gampaha District	13
Number of Areas in Kalutara District	14
Number of Identified Potential Ride-Sharers	3029073

As mentioned above to perform a better analysis further these 1K_cell_ids were classified into areas by identifying the nearest City area where these 1K_cells are located. Figure 4.1 shows the areas which contain the majority of the home locations. Figure 4.1 shows all top cities of the 41 areas mentioned above.

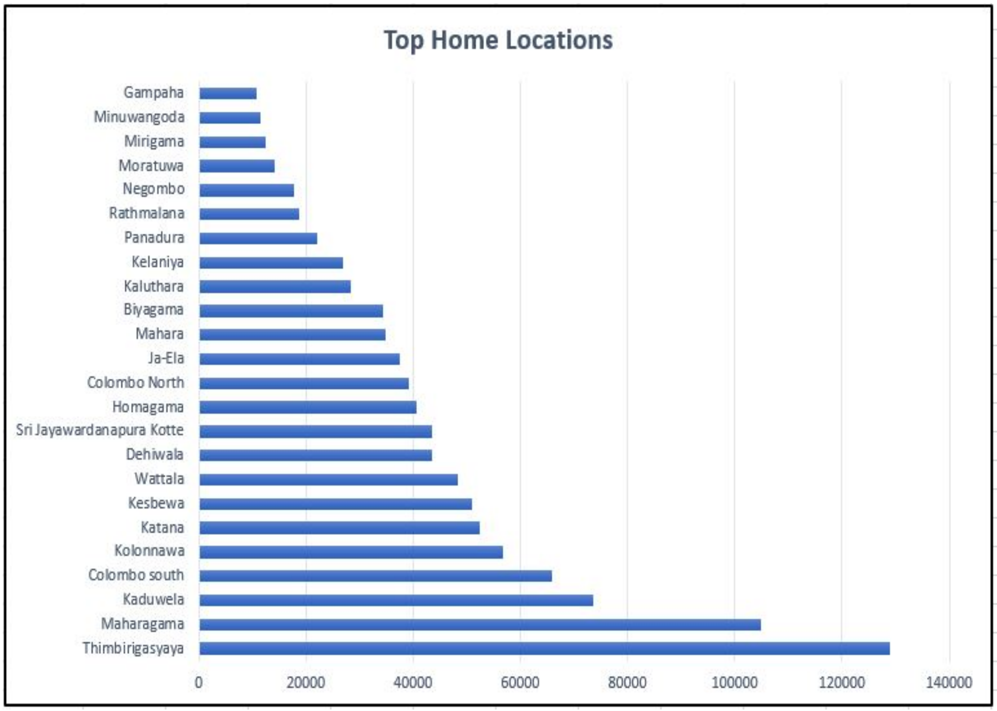


Figure 4.1: Top areas with top home locations

Out of all the available areas, Figure 4.1 shows the areas with a high population which is more than 10K home locations situated in the Western Province of Sri Lanka as per the CDR data analyzed. As mentioned in the previous chapters, the analysis was performed for data collected for the 2012/2013-year period for all the mobile subscribers who have appeared in selected areas within the time selected. This data includes temporary citizens and migrants. No specific filtration performed to identify migrants and temporarily located people. But excludes non-working subscribers and those who work in the same area. Basically, the subscribers with the same cell_id for both home/work locations are excluded. Out of the areas shown in Figure 4.1, Thimbirigasyaya, Maharagama, Colombo South and Kaduwela are identified as the high residence areas in the Western Province as per the collected dataset.

As mentioned previously there are many sub-routes identified in all 41 areas considered in the Western Province for the analysis. Those sub-routes were then categorized based on a number of subscribers for each sub-route as shown in Table 4.4 and Figure 4.2. The number of sub-routes identified with less than 25 subscribers is around 300532. As this is a large number, it does not show in Figure 4.2.

Table 4.4: Number of routes with a specific number of subscribers

Number of Subscribers per route	**Number of Routes**
< 25	300532
25 < 50	15095
50 < 100	7601
100 < 250	4397
250 < 500	1385
> 500	499

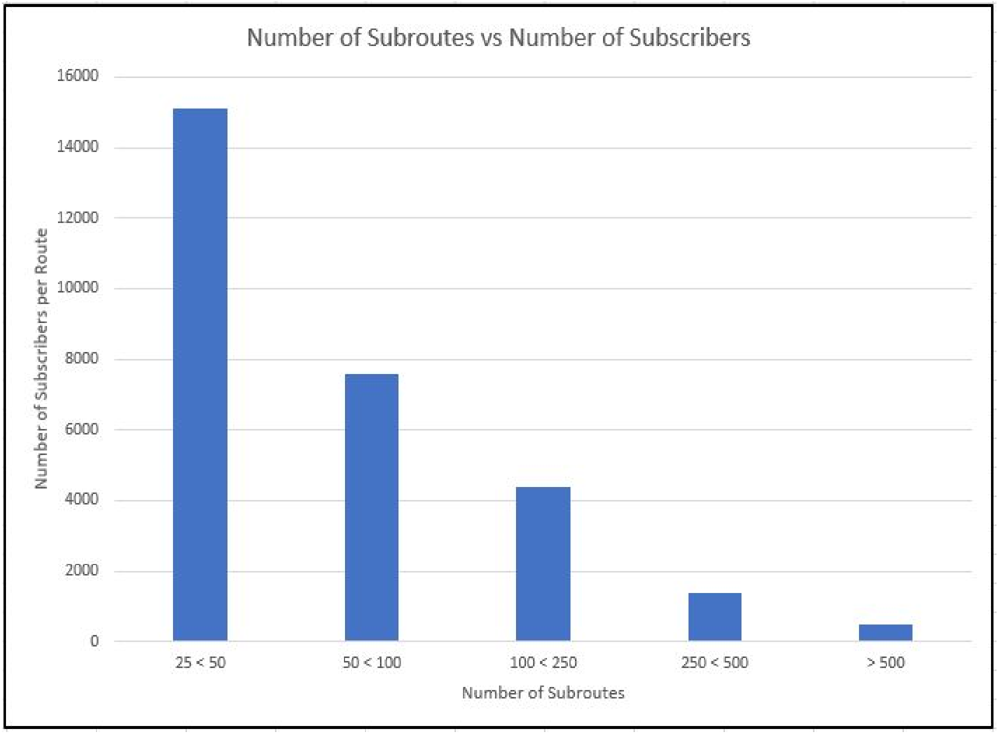


Figure 4.2: Number of subscribers ranges vs number of routes

The above graph shows the number of sub-routes along with the number of subscribers per route. As shown in Figure 4.3 there are many routes with 26-50 subscribers. Even though this is a small number of users, as the routes are defined with the 1km radius, 26-50 people with same source and destinations can be considered as potential ride-sharers. If at least few of these users, use their personal vehicles with one or two people to travel day to day, Ride-Sharing can reduce the number of vehicles in these sub-routes substantially. Also in Figure 4.2, it shows there are many sub-routes with more than 50 subscribers with same source-destinations. Especially, there are a significant number of sub-routes with more than 250 people. These routes with more than 250 subscribers can make a huge impact to reduce the number of vehicles by sharing the rides. These are the routes with high potential of Ride-Sharing in the Western Province.

All the above-identified sub-routes were again classified by the areas and identified the number of subscribers travels daily to and from in terms of areas. There are 55 top source/destination routes identified area wise. All these source/destination routes include sub-routes identified as above. Those 54 source/destination routes are shown in Figures 4.3-4.8 along with the number of subscribers each.

Figure 4.3 depicts the routes starting from Dehiwala, Gampaha, Minuwangoda, Moratuwa, Negombo, Panadura and Ratmalana. This does not include all the routes starting from these areas, but all the routes with top hits. All these depicted source/destination area routes have more than 10000 subscribers. Out of the below source/destination routes, Dehiwala to Thimbirigasyaya has many subscribers moving daily.

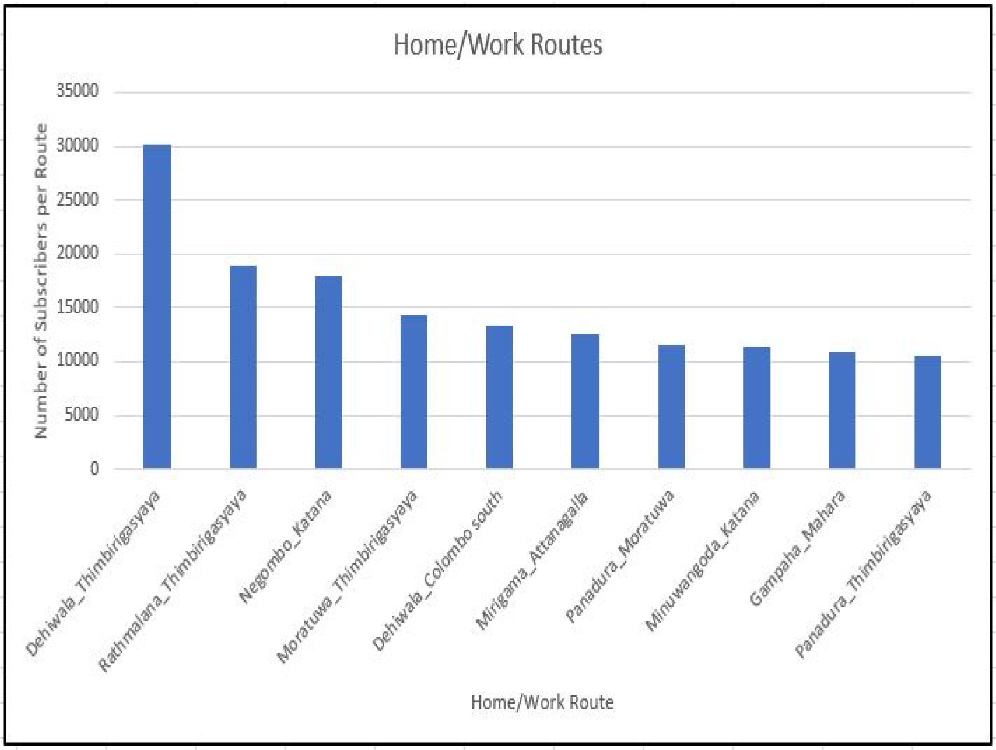


Figure 4.3: Number of subscribers per each source/destination route of selected cities - 1

Figure 4.4 shows the source/destination routes starting from Colombo North and

Colombo South. This again does not show all the routes but top hits. Out of below five routes depicted, Colombo North to Colombo South and Colombo South to Thimbirigasyaya have a significant amount of people moving daily. Colombo South to Thimbirigasyaya has more than 40000 daily travelers.

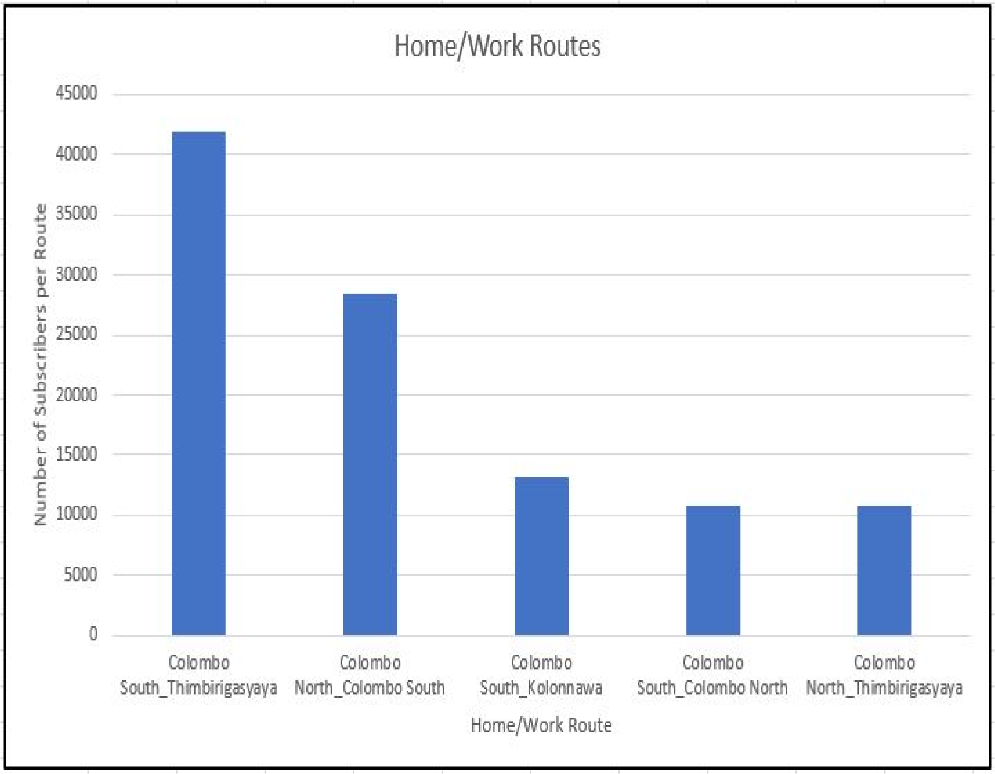


Figure 4.4: Number of subscribers per each source/destination route of selected cities - 2

Figure 4.5 shows the number of subscribers traveling daily from Biyagama, Kaduwela, and Kelaniya. Among those also the destination with the highest number of subscribers is Thimbirigasyaya and the route is Kaduwela to Thimbirigasyaya. There are more than 15000 daily travelers from Kaduwela to Thimbirigasyaya.

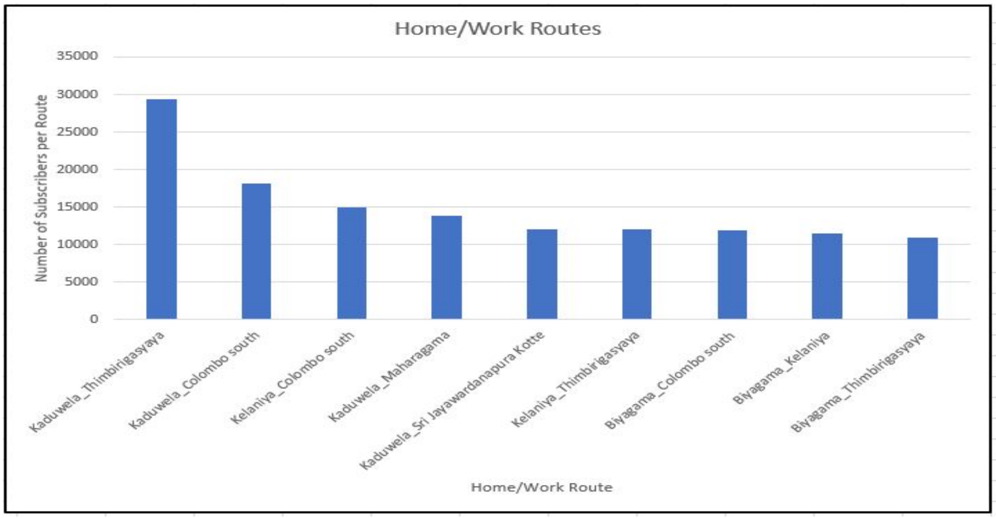


Figure 4.5: Number of subscribers per each source/destination route of selected cities – 3

Figure 4.6 shows the number of subscribers traveling from Homagama, Kesbewa and Maharagama areas. These are high residential areas in Colombo District. Out of the below 12 routes, Maharagama to Thimbirigasyaya has more than 30000 daily travelers.

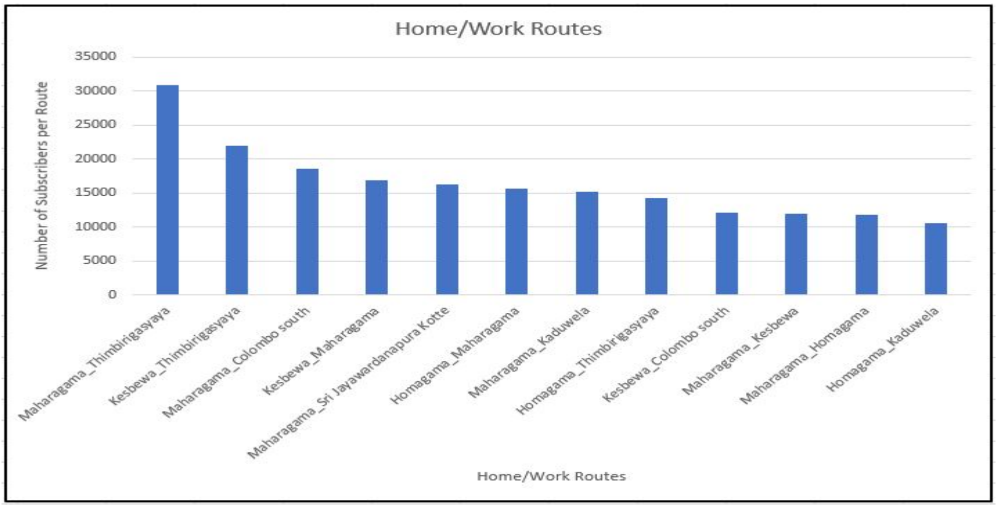


Figure 4.6: Number of subscribers per each source/destination route of selected cities - 4

Figure 4.7 shows the number of subscribers traveling from Ja-Ela, Katana, Mahara and Wattala. These areas are high residential areas in Gampaha District. Out of the below 12 source/destination routes, Katana to Negombo has more than 20000 daily travelers and Wattala to Colombo South has almost 20000 daily travelers.

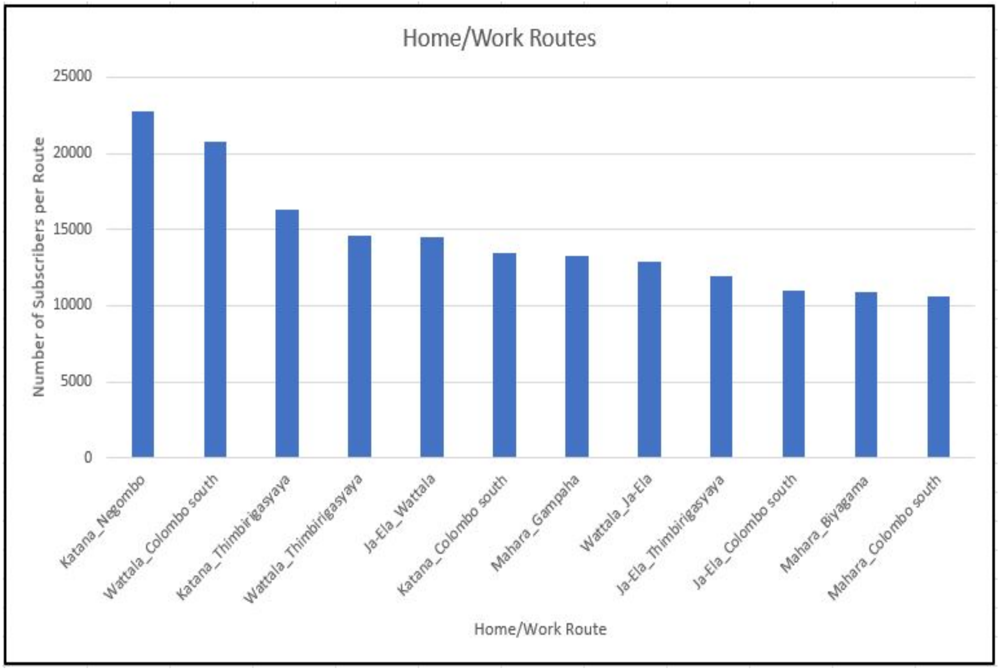


Figure 4.7: Number of subscribers per each source/destination route of selected cities - 5

Finally, Figure 4.8 shows the number of travelers from Thimbirigasyaya area. As mentioned previously, Thimbirigasyaya is identified as the highest residential area in the Western Province and there are many travelers from Thimbirigasyaya to different areas daily. Out of those 6 source/destination routes are showed in Figure 4.8, the top route with the highest number of travelers is, Thimbirigasyaya to Colombo South. There more than 60000 daily travelers from this source to destination.

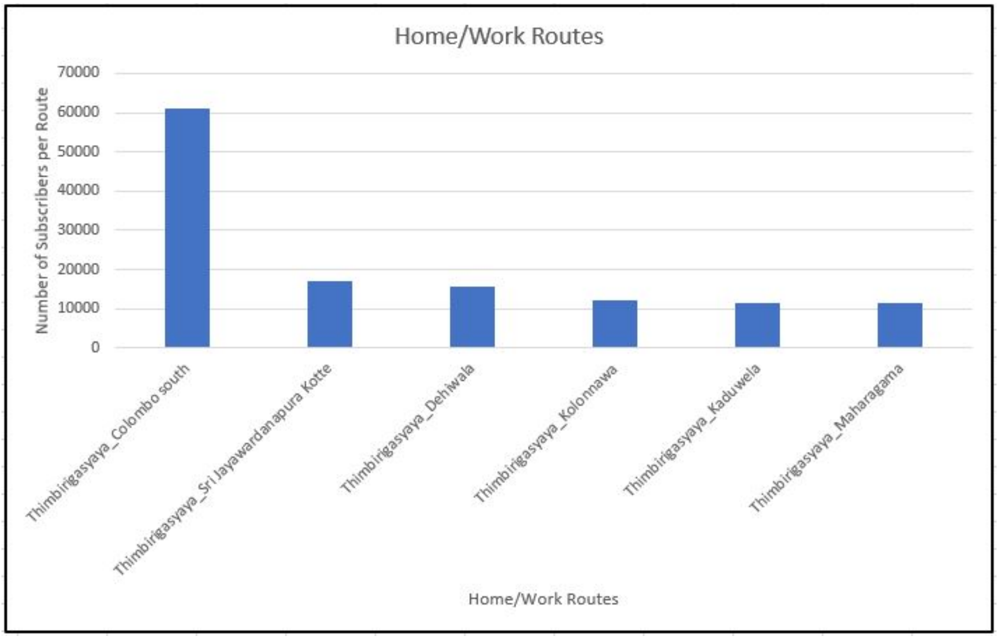


Figure 4.8: Number of subscribers per each source/destination route of selected cities – 6

By looking at the Figures 4.3-4.8, it is identified that the residential area of a highest number of subscribers is Thimbirigasyaya. According to the Figure 4.1, more than 120000 subscribers are living in this area. Also, according to all the above figures, the highest number of workforce travels to Thimbirigasyaya area from many different areas daily. From the selected top source/destination routes identified, Thimbirigasyaya is identified as the most common destination for many routes. Apart from Thimbirigasyaya, Figures 4.3-4.8 show the source/destination routes which have more than 10000 daily travelers.

Between above-identified sources and destinations which have more than 20000 daily travelers, there is a high demand for transportation. Hence, there is a high potential of implementing Ride-Sharing between such areas. However, Ride-Sharing is needed if usage of private transportation mode to travel between source/destination is high. If there are sufficient public transportation modes available and the majority of people are using those, there is no need of implementing a concept such as Ride-

Sharing in such areas. Next section presents the analysis of transportation modes used in the Western Province of Sri Lanka and number of vehicles which travel between identified areas daily. The section will provide the number of private transportation modes and public transportation modes used in day to day chaos.

As per the data analyzed, out of the 4152611 subscribers considered in the selected 41 towns, 3029073 are identified as potential ride-sharers. As a percentage this is 72.94% of the considered population.

4.2 Analysis of Transportation Statistics of Sri Lanka

This section provides the outcomes of the analysis performed on transportation modes and the number of vehicles associated with those identified transportation modes to travel between two areas daily in Sri Lanka. The analysis was performed on a Survey conducted by the Department of Transport and Logistics, Faculty of Engineering, University of Moratuwa, Sri Lanka. For this study only, data of the Western Province is extracted and analyzed. Chapter 3, Section 3.2, described the nature of the dataset and the Table 4.3 provided the transportation modes identified in Sri Lanka and how they are categorized to perform the analysis for this study. The analysis of transportation is performed in the order of, First; the district wise analysis and Second; Divisional Secretariat area wise analysis.

The district wise analysis was done for three Districts in the Western Province of Sri Lanka, Colombo/Kalutara/Gampaha as the CDR data was also analyzed only for these three districts. The modes of transportation identified from the Survey JICA [26] was categorized into three main categories in this study as Private Transportation, Public Transportation, and Ride-Sharing. The categorization of this is described in Table 4.3 Chapter 3, section 3.2. The transportation modes and the number of vehicles used for each transportation mode between these three Districts in the Western Province of Sri Lanka was analyzed and showed in Figure 4.9.

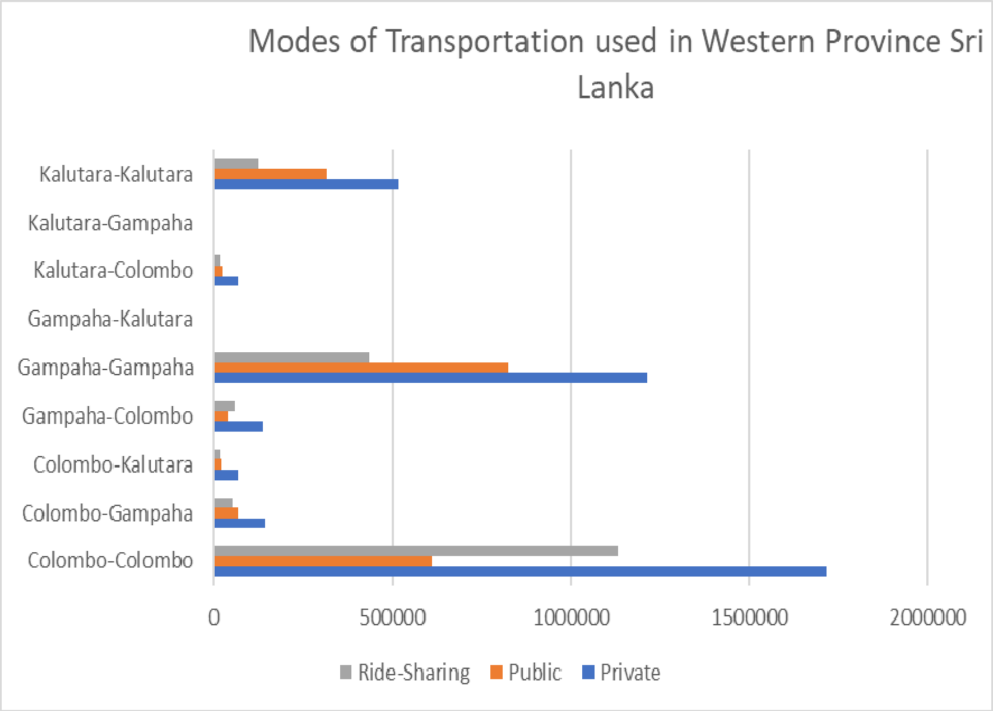


Figure 4.9: Modes of Transportation Used in the Western Province of Sri Lanka between districts

As per the Figure 4.9, it clearly shows that the number of private transportation used between each district is higher than the number of public transportation used. The Ride-Sharing category represents the Employee Transport Service provided by the employers, Staff Services and the School Vans. These three are identified as already used Ride-Sharing methods which is basically using the concept of En-Route Ride-Sharing. Based on Figure 4.9, usage of this transportation mode is also higher than the usage of public transportation by the travelers in their day to day life. This provides an insight that people are willing to share the rides. However, the usage of private transportation is still higher than both the other categories.

If there is a proper methodology implemented to choose Ride-Sharing dynamically, there is a high tendency that people would select Ride-Sharing for their daily traveling purposes based on Figure 4.9. Figure 4.10 shows the percentage of each transportation category used between each district and it provides a clear picture of the high usage of private transportation modes.

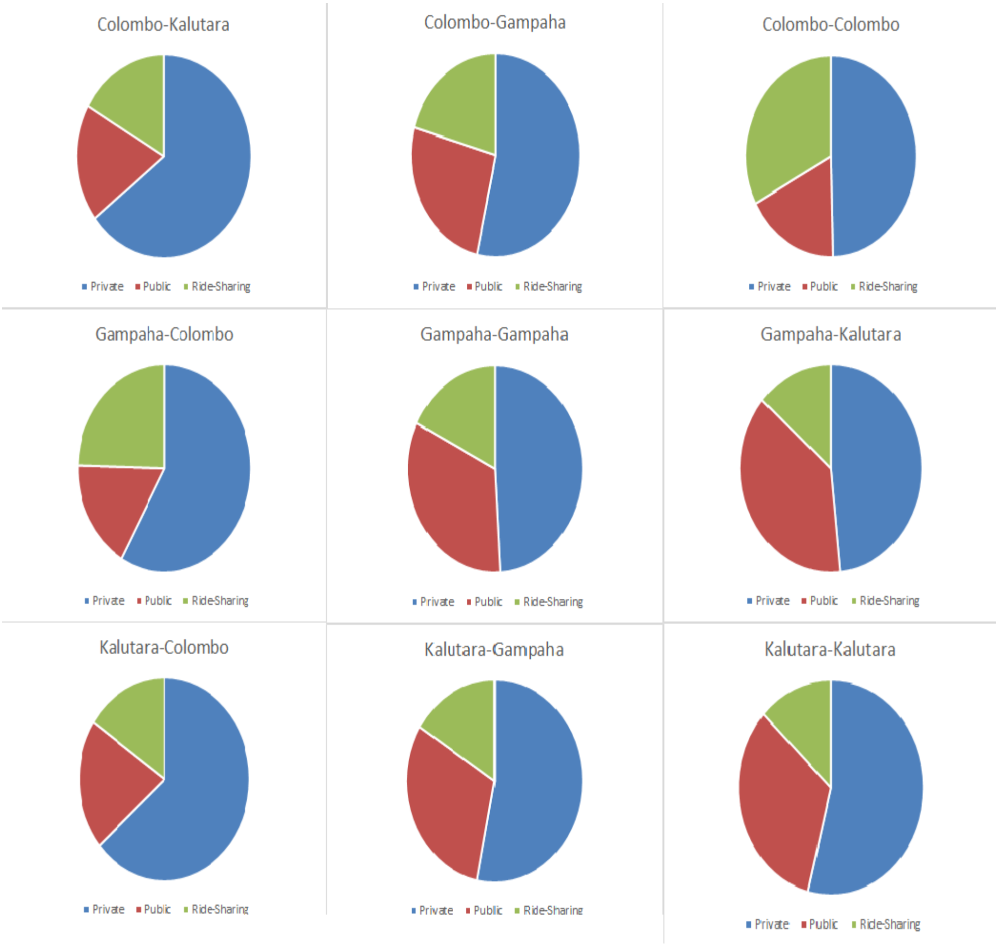


Figure 4.10: Percentage wise representation of Modes of Transportation Used in the Western Province of Sri Lanka between districts

The next analysis was done to identify the public/private transportation used between selected areas and those were highlighted in Section 4.1 which has more than 10000 daily travelers. In Section 4.1, it was discussed about the number of travelers between 54 source/destinations. Out of those 54, 14 source/destination routes were selected for this analysis with more than 10000 travelers each.

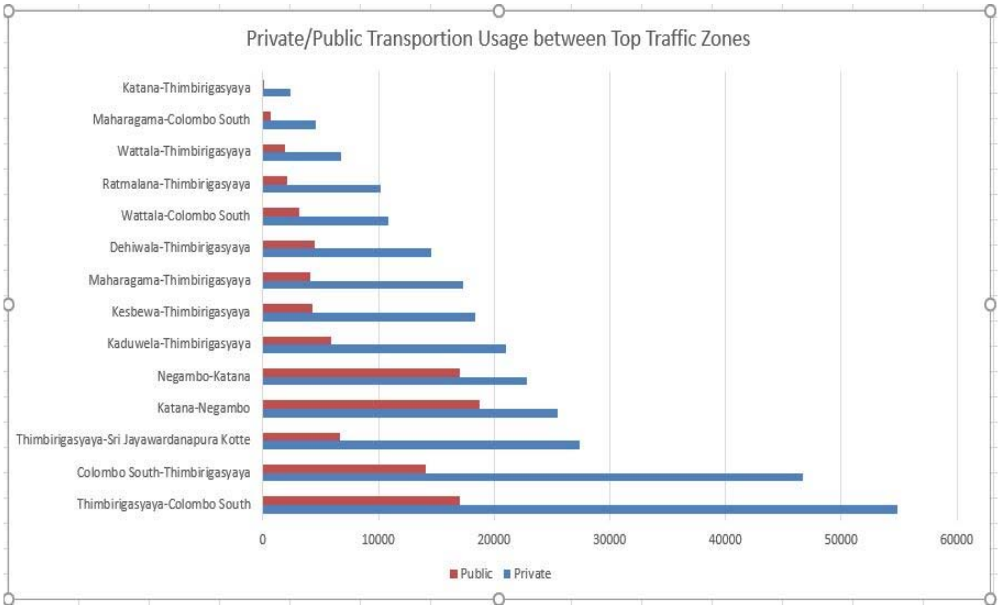


Figure 4.11: Private/Public transportation usage between top traffic zones in the Western Province of Sri Lanka

Figure 4.11 evidently shows that between the selected 14 source/destination routes there is a high usage of private transportation modes compared to public transportation modes. On each of the Figure 4.11, it shows source/destination routes having a higher number of private transportation mode usage than public transportation modes.

Table 4.5 shows the metrics considered for this analysis. As per the metrics shown in Table 4.5, usage of private vehicles is 38.43%, usage of public vehicles is 18.93% and usage of the Ride-Sharing option is 18.28%. As mentioned in the previous chapters, private school vans and buses are also considered as the Ride-Sharing category.

Table 4.5: Metrics used for Analysis

Metric	**Count**
Number of people considered	10086093
Number of private vehicle users	3876448

Metric	Count
Number of public vehicle users	1909474
Number of ride-share users	1843943

Considering Figures 4.9, 4.10, 4.11 and Table 4.5, it is evident that the reason for high traffic congestion in the identified traffic zones is the high usage of private transportation modes. The percentage calculated for private transportation usage based on the data collected for the survey is 38.43%. This is a high percentage compared to the public transportation usage. Hence, it shows that there is a demand to reduce the number of vehicles entering these traffic zones to minimize the traffic congestion. An implementation of Ride-Sharing on these areas can be a solution to reduce the number of private vehicles entering into the cities.

Providing an option to share rides dynamically would highly support to reduce the number of vehicles traveling daily on the routes shown in Figure 4.11. This will directly help to reduce the traffic congestion. The next section will provide ground truth data gathered from 50 known people living in Colombo suburbs to identify the willingness of sharing the rides.

4.3 Analysis of Ground Truth Dataset

An online survey was published for more than 53 known volunteers around Colombo suburbs to identify the willingness to share the rides in their day to day travels from home to work and vice versa. This sample size was calculated based on the Statistical Analysis as per [31] for a population of 4152611 subscribers in the CDR dataset. Table 4.6 shows values considered to calculate the sample size. Expected sample proportion was decided based on the results of the analyzed CDR data and the results of the Nodobo dataset that was used for benchmark methodology.

Table 4.6: Values considered to calculate sample size for the Survey

Population	4152611

Margin of Error	10%
Confidence Level	90%
Expected Sample Proportion	75%

Based on the selected values calculated sample size was 51 and the actual survey was done using 53 participants. All these participants were professionals who work and live in the Western Province. The target of this survey was to identify the willingness to share their rides daily. Figure 4.12 shows the percentage of female/male participation.

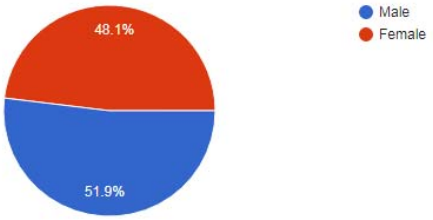


Figure 4.12: Survey data – percentage of gender participation

Both female and male participants were included in this survey and the number of participants from both genders is almost the same. The above Figure 4.12 shows the number of participants of both genders in percentage. Figure 4.13 shows the transportation modes they are using in their day to day travels.

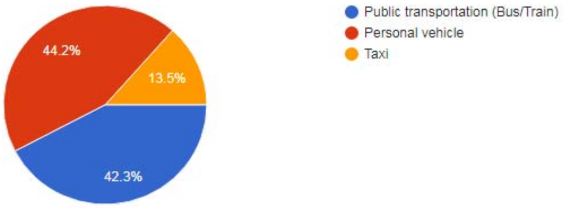


Figure 4.13: Survey Data – Mode of daily transportation

As per the Figure 4.13, there are three main transportation methods used by these volunteers. They are public transportation (buses/trains), personal vehicle and taxis. Taxis again is an individual transportation mode in Sri Lanka. The highest percentage of participants in the survey is using their personal vehicle as the daily transportation mode. The Figure 4.14 shows the willingness to share their rides to and from work.

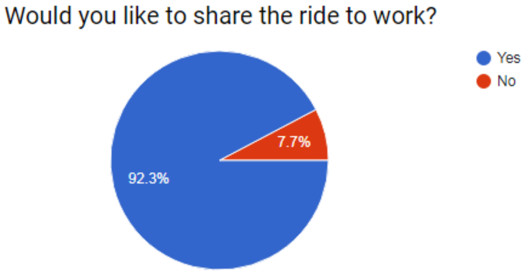


Figure 4.14: Survey Data – Willingness on Ride-Sharing

Out of the more than 50 participants of the survey, more than 92% people like to share their rides to and from work. This shows a high potential of implementing Ride-Sharing in Sri Lanka.

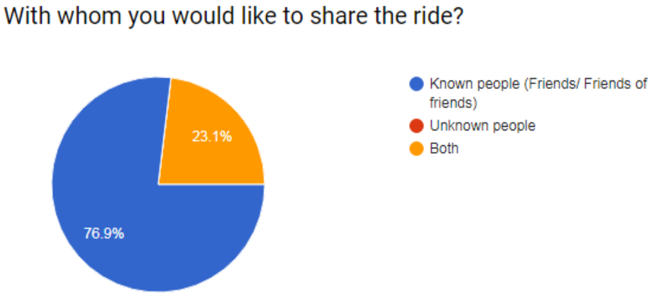


Figure 4.15: Survey Data – Participants preference of sharing the ride

The Figure 4.15 shows with whom they are willing to share the rides. The above graph shows that people are concerned about the privacy and the security. This is a challenge when implementing a Ride-Sharing model. Most of the people have a habit of sharing their rides with known people. This collected ground truth data helped to identify the areas that need to be considered when implementing a Ride-Sharing model.

4.4 Summary of the Analysis

As mentioned at the beginning of this chapter, the chapter described and analyzed the datasets collected. The Section 4.1 highlighted how the sub-routes and daily using main source/destination routes were identified. Also explained was how the number of users was identified in each identified route by running the benchmarked algorithms and the methodology. This section showed top residential areas and the top work areas identified based on the CDR dataset analyzed. The percentage of the potential subscribers that can share the rides in their day to day travels is 72.94% based on the analysis performed on CDR dataset. Out of all the areas available in the dataset, 41 towns with high population were selected with 4152611 subscribers. This percentage has proved that there are a significant number of potential ride-sharers. The same can be later analyzed and implemented island wide.

In section 4.2, the study analyzed the transportation data of Sri Lanka during the period of 2012/2013. This analysis was done to identify the modes of transportation used in Sri Lanka in day to day travels. Then they were categorized into three categories namely; Public, Private and Ride-Sharing as described in Chapter 3 Section 3.2. According to the analysis performed on this transportation data, the percentage of Private transportation usage for day to day travels was identified as 38.43% and Public transportation usage as 18.93%. The identified percentage of already used Ride-Sharing options is 18.28%. Already using Ride-Sharing options are; Employee Transportations, Staff Services, and School Vans. Comparatively to Private transportation usage, it is higher than using public transportation as per the dataset analyzed.

The last section of this chapter brought out the ground truth analysis of willingness to share the rides based on the data collected from a published online survey for a calculated sample size of 53 survey participants. The survey was conducted to analyze the willingness of people to share the rides of home to work and work to home daily. The expected proportion of the survey was 75% based on the calculated sample size and the outcome was 92.3%.

All three datasets provided results that prove, that there is a high potential for implementing Ride-Sharing in Sri Lanka. CDR dataset provides that, there is a high percentage of subscribers who share the same home/work locations who have the potential to share the rides. Then the transportation dataset provided an evidence, based on the analyzed data that there is a high percentage of people using Private transportation modes to travel daily over Public transportation modes. Using private transportation modes mostly increase the traffic as the number of vehicles entering a city/town is higher. The number of vehicles entering a city/town should be minimized to decrease the traffic congestion. If the ride can be shared among the people who tend to use their private transportation modes based on the available vehicle capacities, then the number of vehicles will be reduced. Reducing the number of vehicles entering will decrease the traffic congestion of the city/town. Finally, the data collected from the survey proved that there is a willingness to share the rides. If a high percentage of people who are using their private transportation modes and travel daily are willing to share their rides, then it proves there is a potential to implement Ride-Sharing and decrease the number of vehicles on the road. The conclusion of the analysis will be further discussed in Chapter 5, Section 5.1.

CHAPTER 5: CONCLUSION

This chapter contains three sections as Conclusion, Summary and Future work. Section 5.1 provides the conclusion of the whole study which shows how the potential of Ride-Sharing is analyzed. This section will describe the results of the analysis performed on the datasets and how those results will prove the potential of sharing the rides. Further, the section discusses the key findings of the study. Section 5.2 summarizes the work done in this study. Also, it will discuss how implementing Ride-Sharing will impact on reducing the high traffic congestion. Section 5.3 explains the limitations of the study and they were addressed. Section 5.4 points out the enhancements that need to be considered in the future when implementing a Ride-Sharing model in Sri Lanka or any other country.

5.1 Conclusion

This study was to identify the potential of implementing Ride-Sharing in Sri Lanka by considering the areas where the majority of the workforce is in Sri Lanka. The study further discussed a methodology and algorithms that can be used on any CDR dataset to identify the home/work locations of mobile data subscribers and cluster the subscribers who share the same source/destination for traveling in day to day life. The study has proposed a Ride-Sharing model that can be implemented as a solution for the increasing traffic congestion in Sri Lanka.

The Analysis was performed on two main datasets and a survey conducted for 53 known participants. One dataset consisted of CDR data collected for the year period of 2012/2013 and this was the main dataset analyzed for this study. The second dataset was transportation data collected for same year period 2012/2013 and this was used as a supportive dataset to prove the need of Ride-Sharing. Both the datasets were collected for the same year period and for this study it was only data of the Western Province that was considered which includes data from three main districts, Colombo, Kalutara and Gampaha.

CDR dataset considered for this study consisted of data of all the areas in the Western Province of Sri Lanka. Hence, to increase the accuracy of the analysis, 41 towns/cities with a higher population was selected based on the subscriber population identified in the CDR dataset. These selected towns are shown in Table 3.2. There were 4152611 subscribers identified in these selected 41 towns and out of those, 3029073 were identified as potential ride-sharers based on the methodology and algorithms described in Chapter 3. The percentage of identified potential ride-sharers based on population considered is 72.94%. This is a significant percentage. But this does not mean all these identified potential ride-sharers are using private vehicles for traveling from home to work and vice-versa in their day to day lives.

To identify the Public/Private transportation usage in the Western Province of Sri Lanka analysis was performed on the Transportation data based on JICA [26]. Transportation data analysis was also performed for same 41 towns/cities selected for the CDR data analysis. Based on the dataset, 18 transportation modes were identified and categorized into three categories as Private, Public and Ride-Sharing which was shown in Table 3.3. Ride-Sharing category considered the identified already existing Ride-Sharing options such as Office Transports, Staff Services, and School Vans. As per the analysis, the percentage usage calculated for Private transportation was 38.43%, for Public transportation was 18.93% and for Ride-Sharing was 18.28%. As per the analyzed data Private transportation usage shows a significantly higher number of percentage compared to Public transportation usage in these selected 41 towns/cities.

To analyze the ground truth data as explained in Chapter 4.3 a survey was conducted for 53 known participants to identify the willingness to share the ride in their day to day travel from/to home/work. As mentioned in the Chapter 4.3 sample size 53 was selected based on the subscriber population considered in CDR dataset and the values shown in Table 4.6. Out of the 53 participants of the survey, 92.3% have marked as willing to share the rides and 57.7% are using personal vehicles and taxis to travel from/to home/work daily.

Hence, by considering the results of all three analysis it is possible to state that there exist 72.94% potential ride-sharers in these selected 41 towns/cities in the Western Province of Sri Lanka. Out of those 72.94% of potential ride-sharers, it is most likely that 38.43% are using Private transportation modes to travel from/to home/work daily based on the analysis performed on transportation data. As per the survey data, it is possible to state that out of these 38.43% ride-sharers who are using Private transportation modes to travel, 92.3% are willing to share their rides. This can be depicted in Figure 5.1.

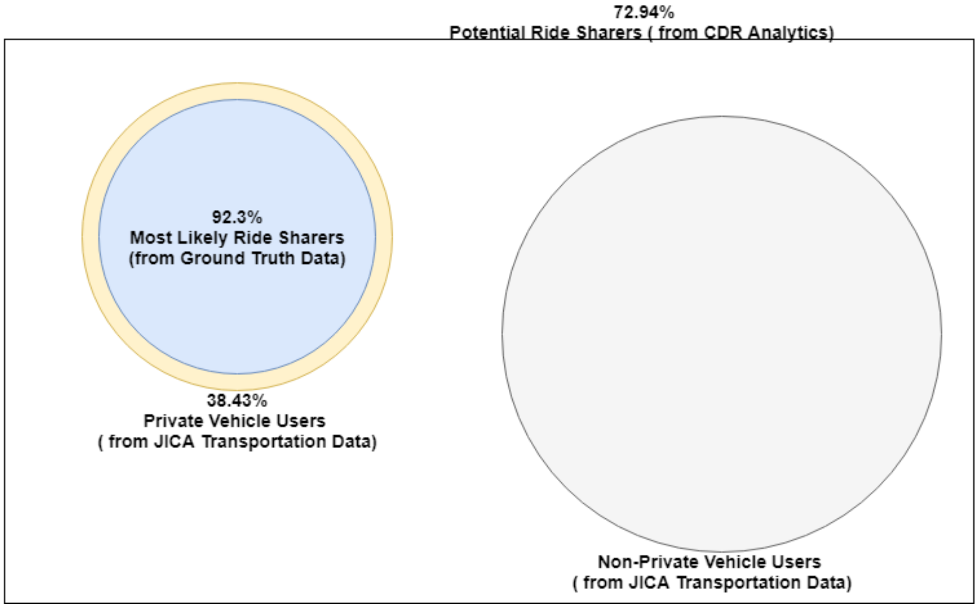

72.94%
Potential Ride Sharers (from CDR Analytics)
92.3%
Most Likely Ride Sharers
(from Ground Truth Data)
38.43%
Private Vehicle Users
(from JICA Transportation Data)
Non-Private Vehicle Users
(from JICA Transportation Data)

Figure 5.1: Potential Ride-Sharers

Based on the above-explained and depicted analysis there is a high chance that the implementation of Ride-Sharing in Sri Lanka will be a success in Sri Lanka. But, as the survey analysis was performed among a selected set of people the actual proportion of willing to share the rides can be less 92.3%. However, it is visible that there is a potential of implementing Ride-Sharing in Sri Lanka based on the analyzed data and the Figure 5.1. If there is potential to implement the Ride-Sharing, then there is a high potential to reduce the incoming vehicular traffic of the cities and it will help to reduce the high traffic congestion.

5.2 Summary

This study was to analyze the potential of implementing Ride-Sharing in Sri Lanka. Ride-Sharing is the concept of sharing the day to day rides with people who travel between same source/destination locations daily. The motivation of analyzing the potential of Ride-Sharing in Sri Lanka is to address the issue of daily traffic congestion and to find out an option for people to travel to and from home comfortably and securely. This is identified as an important concept to analyze and implement due to increasing traffic congestion and high fuel consumption in Sri Lanka, which indirectly help to reduce the traffic congestion by reducing the number of vehicles entering the city.

The study was carried out to identify and analyze the existing work available in the concept of Ride-Sharing worldwide and identify methodologies, algorithms that can be used to establish a ride-sharing model in Sri Lanka. As pointed out in the Chapters, Introduction and Existing Work, there are many models, methodologies, and algorithms introduced and implemented worldwide. In the existing work chapter, it was explained that the ride-sharing is another section of carpooling. The concepts of Carpooling, Intelligent Traffic Handling and Ride-sharing are described in detail in Chapter 2 along with the currently used algorithms, methodologies, and models. Carpooling basically can be used with people who are well known or working in the same company. However, to practically implement Intelligent Traffic Handling, it is required to have technical equipment such as surveillance cameras and so on. Ride-Sharing is a combination of both the concepts. This is the reason for selecting Ride-Sharing as a potential solution to control traffic congestion in Sri Lanka. Ride-

Sharing allows sharing the rides between known and unknown people who have same source/destination locations. The key research papers which inspired this work are Cici et al. [2], [3], Isaarcman et al. [16] and Widhalm et al. [24].

Chapter 3 explained what is CDR data and how it is collected and can be used to analyze the potential of Ride-Sharing. There are three main datasets used for this study along with data collected from a Survey published for 53 known participants to analyze the ground truth data on willingness to use Ride-Sharing. First dataset was a CDR dataset gathered of 27 high school students in USA which is called Nodobo dataset McDiarmid et al. [22]. This dataset was used to benchmark the algorithms and methodologies used to identify the potential of Ride-Sharing in this study. Then the actual CDR dataset of Sri Lankan mobile data subscribers was used to analyze the potential ride-sharers in Sri Lanka. This dataset consisted of Subscribers who are living and working in the Western Province of Sri Lanka for a period of one year from 2012-2013. This had more than 10 million CDRs along with the cell tower details of the Western Province Sri Lanka. Analysis was done in three steps, first the cell towers were clustered into virtual locations by a threshold radius of 1km. Then in the second step subscribers' home/work locations were identified based on the time they are appearing in each cell tower throughout the year. Then these identified home/work locations were again replaced by the 1k_cell_id, the virtual location identified in the first step. Then in the last step cluster the subscribers who have same source/destination routes daily based on the number of hits on the routes. Dataset of Transportation Statistics of Sri Lanka gathered and analyzed during the period of 2012/2013 used to analyze the Public/Private transportation mode usage in Sri Lanka. This dataset consisted of the transportation modes used in Sri Lanka in day to day life, identified traffic zones based on the Divisional Secretariats and the number of vehicles in each transportation mode that travels between DSDs. Furthermore, proposed algorithms of Ride-Sharing have been described in this Chapter. The proposed algorithms were End-Point Ride-Sharing and En-Route Ride-Sharing. All the used datasets, algorithms and the methodologies are described in detail. The study has provided a clear picture of the architecture used for this analysis and how the data was analyzed. In addition, the technologies used for this analysis and the reasons for selecting those technologies are also well explained.

Chapter 4 has presented the details of the analysis with supportive graphs and tables. The cell tower clustering algorithm and the algorithms to identify the source/destination routes provided many source/destination routes which are shared by many mobile data subscribers in their day to day travels. As there were many source/destination routes identified, for a better visualization, sources and destinations were again classified using the nearest town/city in the analysis. Out of all the towns/cities available 41 towns/cities were selected for the analysis based on the high population of subscribers identified from CDR dataset. Then again the number of users who are traveling from same source town/city to same destination town/city daily were identified. For an example number of travelers between Kelaniya and Thimbirigasyaya. From these such routes, 54 routes with a high number of subscribers were selected to visualize and prove there are many potential users in identified areas who can share the rides. It was the top home and work locations based on the nearest city of the cell towers in the Western Province and the source/destination travel based on the cities daily that were highlighted. It was identified that there are 72.94% potential ride-sharers in the selected 41 towns/cities. Then using the Transportation dataset, the transportation usage was analyzed in the selected areas and it was identified that the Private transportation modes usage in these 41 towns/cities is 38.43%. Private transportation mode usage was comparatively higher than the public transportation mode usage as per the analyzed data. This proves that between most of the cities in the Western Province of Sri Lanka, the usage of Private Transportation modes is higher, which ultimately proves that there are many vehicles which travel between these areas. Hence implementing Ride-Sharing and providing an option for dynamic Ride-Sharing will help people to reduce using their private vehicles. As per the survey performed amongst 53 known people, it is identified that 92.3% participants are willing to share the rides. If rides are shared the number of vehicles on the road will be reduced. If the number of vehicles is reduced, it will solve the high traffic congestion and stop the wastage of energy significantly. Hence, the conclusion in Chapter 5 was developed based on Chapter 4: Analysis and it was identified that there is a potential of implementing Ride-Sharing in the Western Province of Sri Lanka and can be carried out across the country.

Security, privacy, and pricing schemes are the concerns that need to be implemented along with Ride-Sharing. These are the concerns users consider if they are to share their rides in day to day rides. As per the survey, it was highlighted that people tend to share their rides with known people rather than with totally unknown people. Hence, a Ride-Sharing model addressing these concerns will be a useful implementation in Sri Lanka. As mentioned in Chapter 4 Section 4.1, there are identified cities/towns with a high population of daily in and out travelers. This can be considered as an insight where the public transportation can be improved.

The study introduced a methodology and an architecture that can be used to analyze the potential of Ride-Sharing using any CDR dataset. The algorithms, methodology, and architecture used for the study can be made use of to identify the home/work locations of subscribers in any CDR dataset and identify the potential ride-sharers. The analysis proved that there is potential for implementing Ride-Sharing in Sri Lanka with the proof of supportive transportation data and ground truth data used for the analysis. The study provided the conclusion based on the analysis performed on all datasets used, that implementing Ride-Sharing in Sri Lanka will help to reduce the increasing traffic congestion in the city areas.

5.3 Research Limitations

There were several limitations faced during the analysis of the study. The first limitation was the timeline and the study had to be scoped down to analyze only the potential of implementing Ride-Sharing in Sri Lanka. Data was not provided as raw data and data processing algorithms were given to the Data Provider LIRNEasia and the algorithms were executed in their environment and the datasets received in CSV format to process and identify the potential ride-sharers.

This study considered only the subscribers who are subscribed to the mobile service providers and this does not mean all the citizens in the Western Province of Sri Lanka are considered. Also, there was no methodology used to analyze whether one subscriber has more than one mobile data connection. The study has not excluded the migrants in those areas within the period considered. Transportation data based on the research JICA [26] and the transportation data is captured only between major cities/towns. There was no validation performed on this data. Hence, the conclusion provided in this study is based on these limitations.

5.4 Future Work

As per the literature review performed and the survey performed to gather the ground truth data, there are few major areas identified as future work when implementing the Ride-Sharing in Sri Lanka. It is identified that people are more concerned about whom they travel with. As explained in Chapter 2 Existing Works, in the papers of Cici et al. [2], [3], Facebook and Twitter data have been used to identify the connections between the ride-sharers. As mentioned in the previous section and the Analysis Chapter, people are willing to travel with friends or friends of friends. Basically, people are willing to travel with known people, people that they know directly or indirectly. Most of the people believe that traveling with completely unknown people is a risk. As per the Cici et al. [2], [3], they have considered identifying friends or friends of friends of ride-sharers due to this reason.

Security and privacy are the most concerned points around the world. Hence, when implementing Ride-Sharing anywhere in the world, it is important to consider and address them. In Sri Lanka also, Facebook data and Geo Tweet data can be used to identify the connections between people. This study used anonymized CDR data to identify the potential of Ride-Sharing. If the actual CDR data can be used to implement the Ride-Sharing, the real mobile details can be used to identify the connections between the ride-sharers. However, if the Ride-Sharing is implemented as an application, the users can be given the option to identify the potential ride-sharers for them along with connection via Facebook and Twitter.

A pricing mechanism should be provided to share the fare of the ride. This can be implemented as dynamic Ride-Sharing, taxi Ride-Sharing, and fixed Ride-Sharing. In dynamic Ride-Sharing, the passengers and drivers can be changed day by day and pricing mechanism is more important. In fixed Ride-Sharing, once the potential drivers and passengers are identified, they will fix to share the rides on a daily basis and it is possible for them to arrange a method to share their fares. In taxi Ride-Sharing, the taxi company itself can implement a mechanism to divide the fare amount among the passengers based on the distance they are traveling. Likewise, a proper pricing mechanism should be arranged for each type of these Ride-Sharing methods. However, the study has provided a basic pricing model that can be improved and implemented.

However, people are more concerned about the privacy, security and pricing. Hence, it is important to identify a way to guarantee their privacy and the security when they share their rides. An option should be given to identify the connection between the ride-sharers always. A methodology should be identified and developed as a further enhancement to improve the privacy, security and pricing of the rides when implementing a Ride-Sharing model.

Reference List

[1] S. Cho, A. Yasar, L. Knapen, T. Bellemans, D. Janssens, and G. Wets, “A Conceptual Design of an Agent-based Interaction Model for the Carpooling Application”, in *Proc. 1st International Workshop on Agent-based Mobility, Traffic and Transportation Models, Methodologies and Applications*, Canada, June 2012.

[2] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris, “Quantifying the Potential of Ride-Sharing using Call Description Records”, in *Proc. 14th Workshop on Mobile Computing Systems and Applications,* February 2013.

[3] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. “Assessing the Potential of Ride-sharing Using Mobile and Social Data: A Tale of Four Cities.”, in *Proc. ACM Int. Joint Conf. Pervasive UbiComp*, 2014.

[4] L. Figueiredo, I. Jesus, J. A. Tenreiro Machado, J. R. Ferreira, and J. L. Martins de Carvalho, “Towards the development of intelligent transportation systems”, in *Proc. IEEE Intell. Transp. Syst. Conf.*, Oakland, CA, 2001, pp. 1206–1211.

[5] A. Kleiner., B. Nebel, and V. Ziparo, “A mechanism for dynamic ride sharing based on parallel auctions”, in *Proc. 22nd Int. Joint Conf. Artificial Intelligence*, 2011.

[6] F.-Y. Wang, “Parallel control and management for intelligent transportation systems: concepts, architectures, and applications”, *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630-638, Sep. 2010.

[7] A. Ramos, J. Ferreira, and J. Barceló, “An Integrated GPS/PDA/GIS Telegeoprocessing System for Traffic & Environment”, *Journal of Systemics, Cybernetics and Informatics*, vol. 7, no. 6, 2009, pp. 47-53.

[8] W. Herbawi and M. Weber.“A genetic and insertion heuristic algorithm for solving the dynamic ridematching problem with time windows”, in *Proc. ACM Int. Conf. Genetic Evol. Comput.*, 2012.

[9] H. Halaoui. “Intelligent Traffic System: Road Networks with Time-Weighted Graphs”, in *Proc. International Journal for Infonomics*, vol. 3, no. 4, Dec 2010.

[10] B. Srivastava. “Making Car Pooling Work – Mythsand Where to Start”, in *Proc. 19 ITS World Congress Semantic Cities Workshop*. 2012.

[11] S. Tare, N. Khalate and a. Mahapadi, "Review Paper on CarPooling Using Android Operating System-A Step Towards Green Environment”, in *Proc. International Journal of Advanced Research in Computer Science and Software Engineer*ing, vol. 3, no. 4, pp. 54-57, 2013.

[12] B. Cotton.“Intelligent Urban Transportation Predicting, Managing, and Integrating Traffic Operations in Smarter Cities”, 2014.

[13] S. Ma, Y. Zheng, O. Wolfson. “T-Share: A Large-Scale Dynamic Taxi Ridesharing Service”, in *Proc. of ICDE*, 2013.

[14] R.Manzini, A. Pareschi."A Decision-Support System for the Car Pooling Problem", *Journal of Transportation Technologies*, 2012.

[15] M. Furuhata, M. Dessouky, F. Ordez, M.-E. Brunet, X. Wang, and S. Koenig, “Ridesharing: The state-of-the-art and future directions”, *Transportation Research Part B*, vol. 57, pp. 28–46, 2013.

[16] S. Isaacman, R. Becker, R. C´aceres, S. Kobourov,M. Martonosi, J. Rowland, and A. Varshavsky,“Identifying Important Places in People’s Lives from Cellular Network Data.”, in *Proc. Pervasive Computing*, June 2011.

[17] Agatz, N. A., Erera, A. L., Savelsbergh, M. W.,and Wang, X. “Dynamic ride-sharing: A simulation study in metro Atlanta”, *Transportation Research Part B: Methodological* 45, 9 (2011), 1450 – 1464.

[18] Telecom ABC. (2005). VLR. Available At http://www.telecomabc.com/v/vlr.html

[19] J. A. Hartigan. *Clustering Algorithms*. New York: John Wiley & Sons, 1975.

[20] J. Han, M. Kamber and J Pei. "Cluster Analysis Basics: Concepts and Methods", *Data Mining Concepts and Techniques*, 3rd ed. USA: Morgan Kaufmann, 2012, pp.443-494.

[21] Apache Spark. (2017, Nov. 10). *Spark Overview* [Online]. Available: https://spark.apache.org/docs/2.2.1/

[22] A. McDiarmid, S. Bell, J. Irvine, and J. Banford, "Nodobo: Detailed Mobile Phone Usage Dataset", Dept. Electronics & Electrical. Eng., Strachcylde Univ., Glasgow.

[23] S. Bell, A. McDiarmid and J. Irvine, "Nodobo Capture: Mobile Data Recording for Analysing User Interactions in Context", Dept. Electronics & Electrical. Eng., Strachcylde Univ., Glasgow, 2011.

[24] P. Widhalm, Y. Yang, M. Ulm, S. Athavale and M. González, "Discovering urban activity patterns in cell phone data", New York, 2015.

[25] N. R.Chopde and M. K.Nichat, "Landmark Based Shortest Path Detection by Using A* and Haversine Formula," in *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, no. 2, p. 5, 2013.

[26] JICA, "Urban Transport System Development Project for Colombo Metropolitan and Suburbs", 2014.

[27] IBM, "IBM Intelligent Transportation Solution for Active Traffic Management", USA.

[28] B. Srivastava and a. Ranganathan, "Tutorial: Traffic Management and AI – IBM Research", July 2014.

[29] D. Maldeniya, S. Lokanathan and A. Kumarage, “Origin-Destination Matrix Estimation for Sri Lanka Using Mmobile Network Big Data”, Sri Lanka, 2015.

[30] D. Maldeniya, S. Lokanathan, A. Kumarage, G. Kreindler and K. Madhawa, “Where did you come from? Where did you go? Robust policy relevant evidence from mobile network big data”, LIRNEasia, Sri Lanka, 2015.

[31] J. Han, M. Kamber and J Pei. “Cluster Analysis Basics: Concepts and Methods”, *Data Mining Concepts and Techniques*, 3rd ed. USA: Morgan Kaufmann, 2012, pp.443-494.

[32] Rev. J. Gall, “Use of cylindrical projections for geographical astronomical, and scientific purposes”, in *Scottish Geographical Magazine*, 2008, vol 1, no 4, pp 119-123, doi: 10.1080/14702548508553829.

[33] W.R. Tobler, “A Proposal for an Equal Area Map of the Entire Work on Mercator’s Projection”, in *The American Cartographer*, 2013, vol 5, no 2 pp 149-154, doi: 10.1559/152304078784022827.

[34] R. Nisbet, G. Miner, K. Yale, J. F. Elder, and A. F. Peterson, *Handbook of statistical analysis and data mining applications*, 2nd ed. London: Academic Press, 2018.

LIST OF APPENDICES

Appendix A – Identified Top Home/Work Locations

These are the counts of home locations considered for the Figure 4.1.

Area	Home Location Count
Biyagama	34342
Colombo North	39192
Colombo south	65993
Dehiwala	43579
Gampaha	10868
Homagama	40733
Ja-Ela	37485
Kaduwela	73489
Kaluthara	28281
Katana	52501
Kelaniya	26987
Kesbewa	51012
Kolonnawa	56787
Mahara	34800
Maharagama	104840
Minuwangoda	11464
Mirigama	12530
Moratuwa	14247
Negombo	17850
Panadura	22144
Ratmalana	18844
Sri Jayawardanapura Kotte	43427
Thimbirigasyaya	128983
Wattala	48261

Appendix B – Top Home/Work Routes Identified

These are the source/destination routes considered for Figures 4.3 – 4.8.

Route	Potential_user_cout	Home Count
Biyagama_Colombo south	11865	34342
Biyagama_Kelaniya	11559	
Biyagama_Thimbirigasyaya	10918	
Colombo North_Colombo south	28448	39192
Colombo North_Thimbirigasyaya	10744	
Colombo south_Colombo North	10777	65993
Colombo south_Kolonnawa	13233	
Colombo south_Thimbirigasyaya	41983	
Dehiwala_Colombo south	13394	43579
Dehiwala_Thimbirigasyaya	30185	
Gampaha_Mahara	10868	10868
Homagama_Kaduwela	10658	40733
Homagama_Maharagama	15721	
Homagama_Thimbirigasyaya	14354	
Ja-Ela_Colombo south	10968	37485
Ja-Ela_Thimbirigasyaya	11981	
Ja-Ela_Wattala	14536	
Kaduwela_Colombo south	18174	73489
Kaduwela_Maharagama	13816	
Kaduwela_Sri Jayawardanapura Kotte	12076	
Kaduwela_Thimbirigasyaya	29423	
Kaluthara_Kaluthara	28281	28281
Katana_Colombo south	13410	52501
Katana_Negombo	22783	
Katana_Thimbirigasyaya	16308	
Kelaniya_Colombo south	14924	26987
Kelaniya_Thimbirigasyaya	12063	
Kesbewa_Colombo south	12137	51012
Kesbewa_Maharagama	16914	
Kesbewa_Thimbirigasyaya	21961	
Kolonnawa_Colombo south	30473	56787
Kolonnawa_Thimbirigasyaya	26314	
Mahara_Biyagama	10898	34800

Route	Potential_user_cout	Home Count
Mahara_Colombo south	10647	
Mahara_Gampaha	13255	
Maharagama_Colombo south	18618	104840
Maharagama_Homagama	11815	
Maharagama_Kaduwela	15195	
Maharagama_Kesbewa	12000	
Maharagama_Sri Jayawardanapura Kotte	16336	
Maharagama_Thimbirigasyaya	30876	
Minuwangoda_Katana	11464	11464
Mirigama_Attanagalla	12530	12530
Moratuwa_Thimbirigasyaya	14247	14247
Negombo_Katana	17850	17850
Panadura_Moratuwa	11578	22144
Panadura_Thimbirigasyaya	10566	
Rathmalana_Thimbirigasyaya	18844	18844
Sri Jayawardanapura Kotte_Colombo south	14289	43427
Sri Jayawardanapura Kotte_Thimbirigasyaya	29138	
Thimbirigasyaya_Colombo south	61127	128983
Thimbirigasyaya_Dehiwala	15620	
Thimbirigasyaya_Kaduwela	11638	
Thimbirigasyaya_Kolonnawa	12294	
Thimbirigasyaya_Maharagama	11304	
Thimbirigasyaya_Sri Jayawardanapura Kotte	17000	
Wattala_Colombo south	20791	48261
Wattala_Ja-Ela	12902	
Wattala_Thimbirigasyaya	14568	

Appendix C – Survey Results

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Female	Personal vehicle	Yes	Both
Male	Personal vehicle	Yes	Both
Female	Public transportation (Bus/Train)	Yes	Both
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Public transportation (Bus/Train)	Yes	Both
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Taxi	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Both

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	No	Known people (Friends/ Friends of friends)
Female	Personal vehicle	No	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Both
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Female	Public transportation (Bus/Train)	Yes	Both
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	No	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Both
Male	Taxi	Yes	Both
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Both
Male	Personal vehicle	Yes	Known people (Friends/ Friends of friends)

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Female	Taxi	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Both

Gender	Current transportation mode to work	Would you like to share the ride to work?	With whom you would like to share the ride?
Female	Taxi	Yes	Known people (Friends/ Friends of friends)
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Female	Taxi	Yes	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Known people (Friends/ Friends of friends)
Female	Taxi	No	Known people (Friends/ Friends of friends)
Female	Personal vehicle	Yes	Both
Male	Public transportation (Bus/Train)	Yes	Known people (Friends/ Friends of friends)
Male	Taxi	Yes	Known people (Friends/ Friends of friends)
Male	Personal vehicle	Yes	Both