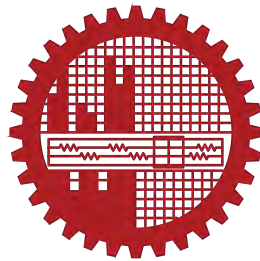


M.Sc. Engg. Thesis

DEVELOPMENT OF A NOVEL APPROACH FOR ESTIMATING TRAVEL TIME FROM MOBILE PHONE CALL DETAIL RECORDS

By
Md.Mahedi Hasan

Submitted to
Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science and Engineering

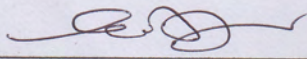


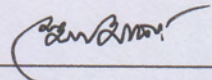
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka-1000

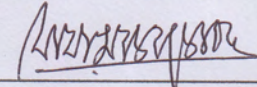
March 31, 2018

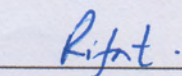
The thesis titled “Development of a Novel Approach for Estimating Travel Time from Mobile Phone Call Detail Records”, submitted by Md. Mahedi Hasan, Roll No. 0413052005, Session April 2013, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents. Examination held on March 31, 2018.

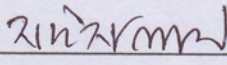
Board of Examiners

1. 

Dr. Mohammed Eunus Ali
Professor
Department of CSE
BUET, Dhaka-1205.
Chairman
(Supervisor)
2. 

Dr. Md. Mostofa Akbar
Professor & Head
Department of CSE
BUET, Dhaka-1205.
Member
(Ex-officio)
3. 

Dr. M. Kaykobad
Professor
Department of CSE
BUET, Dhaka-1205.
Member
4. 

Dr. Rifat Shahriyar
Assistant Professor
Department of CSE
BUET, Dhaka-1205.
Member
5. 

Dr. Salekul Islam
Associate Professor & Head
Department of CSE
United International University, Dhaka-1209.
Member
(External)

Candidate's Declaration

This is to certify that the work presented in this thesis entitled "**Development of a Novel Approach for Estimating Travel Time from Mobile Phone Call Detail Records**" is the outcome of the an investigation carried out by me under the supervision of Professor Dr. Mohammed Eunus Ali in the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka. It is also declared that neither this thesis nor any part thereof has been submitted or is being currently submitted anywhere else for the award of any degree or diploma.

Mahedi

Md. Mahedi Hasan
Candidate

Contents

<i>Board of Examiners</i>	iii
<i>Candidate's Declaration</i>	iv
Acknowledgements	ix
Abstract	x
1 Introduction	1
1.1 State-of-the-art	3
1.2 Objective of This Thesis	4
1.3 Summary of Results	5
1.4 Thesis Organization	5
2 Data	6
2.1 Study Area	6
2.2 CDR data	7
2.3 Real travel data	7
3 Literature Review	10
3.1 Traditional Approaches	11
3.2 Floating Car Data	12
3.3 Cellular Data	13
3.3.1 Passive Monitoring	14

3.3.2	Active CDR Data	14
3.4	Summary	16
4	Travel Time Estimation from static CDR	17
4.1	Tower-to-tower transient travel generation	18
4.2	Peak/off-peak hours segmentation	19
4.3	Determining actual travel time from transient travel time	22
4.4	Travel time estimation matrix construction	28
4.5	Summary	30
5	Travel Time Estimation from CDR Stream	31
5.1	Transient Travel Generation	33
5.2	Updating Hourly Travel Time Matrix	34
5.3	Summary	37
6	Results	38
6.1	Travel time estimation	38
6.2	Analysis of travel time matrix	39
6.3	Summary	40
7	Validation	49
7.1	Collecting travel time data	49
7.2	Estimated travel time vs. real travel time	51
7.3	Summary	52
8	Conclusion	55
	List of Publications	56

List of Figures

2.1	Frequency of calls per user	8
2.2	Example CDR data and locations of highlighted user	9
4.1	Block diagram of travel time estimation from static CDR	18
4.2	Intra-Cluster (Red Mark) and Inter-Cluster (Blue Mark) <i>t-travel</i> example	23
4.3	Empirical Evaluation of Number of Clusters	24
4.4	Major Locations within Dhaka City	25
4.5	<i>t-travel</i> to actual travel transformation	26
5.1	Block diagram of travel time estimation from Stream of CDR	32
5.2	CDR Stream to Dynamic Mapping	33
5.3	An example of 10 second sliding window	34
5.4	Formulation of <i>t-travel</i> using continuous query from dynamic map	35
6.1	Result Comparison of Mohakhali to Mirpur-1	41
6.2	Result Comparison of Kuril to Malibagh	42
7.1	Result Comparison of travel time between key locations	50
7.2	Result Comparison of travel time between any two minor locations	50
7.3	Result Comparison of Ramna to Mirpur-1	52
7.4	Result Comparison of Uttara to Shyamoli	53
7.5	Result Comparison of Mirpur-10 to Kawran Bazar	54

List of Tables

4.1	Peak/Off-Peak hour Estimation	21
4.2	Major Traffic Locations in Dhaka	27
6.1	Amount of data contributed in each step	38
6.2	Google travel time vs. off-peak estimated travel time from CDR	43
6.3	Morning peak-hour travel time vs. evening peak-hour travel time	44
6.4	Travel time from CDR stream comparison between 09 AM and 07 PM hour . . .	45
6.5	Travel time from CDR stream comparison between 02 AM and 07 PM hour . . .	46
6.6	Example of Measured Shortest Path based on travel time matrix(Peak as P, Morning as Morn and Evening as Eve)	47
6.7	Intra Zone Average Travel Speed in km/h(Peak as P, Morning as Morn and Evening as Eve)	48
7.1	The average travel time within 17 key locations	51

Acknowledgements

All the praises and thanks be to **ALLAH** Almighty, the Giver of bountiful blessings and gifts.

I would like to express my deep gratitude to my supervisor Professor Dr. Mohammed Eunos Ali for introducing me to the fascinating and prospective field of big data analytics. I have learned from him how to carry on a research work, how to write, speak and present well. I thank him for his patience in reviewing my so many inferior drafts, for correcting my proofs and language, suggesting new ways of thinking, leading to the right way, and encouraging me to continue my research work. I again express my indebtedness, sincere gratitude and profound respect for him for his continuous guidance, suggestions and wholehearted supervision throughout the progress of this work.

I would also like to thank my friends who gave me their support; in particular, to Suman Banerjee. I convey my heartfelt reverence to my parents and other family members for giving their best support throughout my work to overcome the tedium of repetitive trials to new findings.

Grameenphone Ltd., Bangladesh provided the data for research purpose. This research is partially funded by the research grant of ICT Division, Ministry of Post, Telecommunications, and Information Technology, Govt. of Bangladesh.

Finally, every honor and every victory on earth is due to Allah, descended from Him and must be ascribed to Him. He has endowed me with good health and with the capability to complete this work. I deeply express my sincere gratitude for the endless kindness of Allah.

Abstract

Traffic jam in Dhaka city, the capital city of Bangladesh and one of the most densely populated cities in the world, is one of the major problems for the commuters. The city dwellers have been experiencing intolerable traffic jam every day. Though the authority is trying to reduce the traffic congestion by building new roads and enforcing new rules for vehicles on the roads, unfortunately, they do not result in any visible improvement of the condition. The main reason is the inability to assess and predict the traffic condition of the road in real-time. Due to huge infrastructural cost, the city does not have the facility to collect traffic data by deploying sensors. In this thesis, we overcome this limitation by exploiting mobile phone call details record (CDR) data collected by mobile phone operators for billing purpose. We propose a methodology to estimate travel time between any two major locations of Dhaka city from the aggregated information of a high volume of users' mobile phone CDR data. Our experimental results based on real CDR dataset of 2.87 millions of users of the largest mobile operator, Grameenphone Ltd., show that we can effectively predict travel time between any two major junctions and any two minor locations of the city with an average accuracy of 87% and 76% respectively.

Chapter 1

Introduction

Nowadays traffic management in major cities has become a major concern with the increasing number of citizens living in cities and urban areas. Especially, cities of developing countries are struggling against the traffic congestion, which is the number one concern of the dwellers in many cities around the world. For example, traffic jam in Dhaka city, the capital of Bangladesh and one of the most densely populated cities in the world, is one of the major problems for the commuters in the city. The city dwellers have been experiencing intolerable traffic jam every day. It is affecting individuals in every aspect of their lives that include wastage of time and money.

To combat the traffic jam, the first and foremost thing is to assess the traffic in real-time and to predict the travel time in different routes. City planners can use this information to make a strategic plan on how to control traffic, and at the same time, it gives people choice to decide when to start their journey. The more individuals know about traffic, the easier it will be for them to avoid those traffic jams. Unfortunately, due to lack of infrastructure such as deployment of road and vehicle sensors, developing cities fail to collect real-time traffic data and thus are unable to fight against the traffic jam. In this thesis, we address this problem and propose a technique to estimate travel time by exploiting the mobile phone call detail records (CDR) data of millions of users in a densely populated area of Dhaka city.

The major challenge of estimating travel time based on the real-time traffic is the unavail-

ability of real traffic data. People’s commutes offer a vast amount of data. However, collecting these data is challenging. Traditional approaches collect data through roadside and household surveys (e.g., [15], [22], [14]). These approaches are costly and cannot capture real-time traffic due to the low frequency of update. Moreover, these techniques are prone to sampling biases and reporting errors. Modern cities deploy road mount detector or camera [36][40] to assess the traffic. However, these techniques also suffer from above limitations and low penetration rate.

The most recent approaches use wireless data from floating car known as floating car data (FCD) such as GPS traces from public transports or taxis (e.g., [12][34][39]). The FCD approaches are also limited by their limited coverage and low penetration rate, particularly in developing countries. Google also collects travel time data from their users’ smartphones; however, users are not willing to reveal their locations due to privacy concerns. Thus major limitations of this approach are the unavailability of the data and a low penetration rate of smart-phones in developing countries[25].

To overcome the above limitations, we resort to mobile users’ CDR data to estimate traffic and travel time, which is suitable for developing countries where mobile penetration rate is high and the cities are densely populated.

Due to the huge growth of mobile phone users and the availability of location traces of mobile users from their call records, recent research focuses on utilizing CDR data to mine interesting patterns of city dwellers. In recent years, CDR data has been used for mobility pattern extraction, human travel pattern visualization, route choice modelling, traffic model calibration, and traffic flow estimation (e.g., [38][13]).

Few recent approaches [8][13][31][17] measured traffic status using CDR data. However, these works are unable to predict travel time of roads, which is the most important parameter for analyzing traffic condition. The most relevant works [19],[18] with our study used *enriched* CDR data to estimate travel time in real time. Their work heavily relied on the signalling data generated from “idle” devices, or in other words devices that are not involved in call or data connection with the base station. These CDR data from “idle” devices may not be available in every country, especially in developing countries because of additional expenditure

to track a huge volume of idle phones. Moreover, they only considered CDR data which are generated from nearby important roads or highway. Another recent work [21] used CDR data to predict travel time between cities. This work does not consider a city's internal road network characteristics and people's mobility behaviour, and thus is not suitable for predicting travel time among different suburbs of the cities.

1.1 State-of-the-art

Traditionally, surveillance technologies such as camera, magnetic induction loop, Bluetooth scanner, etc., have been widely used to monitor the city traffic. Based on their data collection strategies, these techniques can be divided into two major categories: point based approach that counts a number of vehicles in the detection area, and distance based approach that takes the average speed and time between two designated monitoring points. Few recent studies focused on GPS based monitoring of vehicles, which are known as floating car data (FCD) techniques.

High deployment cost and low penetration rates are the two major bottlenecks of most of the above techniques (e.g., [5], [12], [34], [39], [20], [37]). Development of data-driven ITS can be found as a detailed survey in [40].

Google also tries to collect the travel time from Android users' mobile phones. However, since in developing countries smartphone penetration is low [25], and most users are not willing to disclose their locations, their approach is not applicable in our scenario [2]. In addition, algorithms used for estimating travel time from raw data is typically unavailable. Several recent approaches use cellular network data to find various traffic patterns in the city. In [28], the potentiality of mobile phone usage as traffic probes are analyzed and as per authors, compared with other alternatives, mobile phones consist of some appealing characteristics such as sample size, coverage, and cost.

Approaches use cellular network data to find various traffic patterns can be divided into two categories: CDR based approach, e.g., [38], and passive monitoring approach, e.g., [27].

CDR data has been used primarily for traffic flow statistics, origin-destination matrices determination, and understanding human mobility (e.g., [7],[10],[17],[30],[32],[33],[38]). [21] tries to identify travel time between two cities using CDR data. This approach fails to identify road traffic in a smaller region, i.e., between key locations inside the city. [9] estimated traffic intensity by counting the number of calls from mobile users of a specific region. [3] exploited double-handovers to estimate real-time road traffic from cellular network signalling. The result in study [5] shows a correlation between the traffic data obtained from mobile phone and magnetic loop detectors. But algorithm used in this study is undocumented and proprietary of a commercial company. Another algorithm is proposed in [23], where traffic status and speed estimated using low resolution positioning data gathered from cellular networks. But this approach is validated only by means of simulations.

Passive monitoring approaches rely on exchanged messages between mobile terminals and require additional monitoring infrastructure [27]. Additional monitoring infrastructure provides better accuracy [19],[9],[35], however, these infrastructure are absent in developing countries.

1.2 Objective of This Thesis

To address the limitations of existing studies, in this thesis, we propose a methodology to estimate travel times between key locations of Dhaka city from mobile users' CDR data generated by "active" devices. Our method also includes estimation of travel time between any two pairs of locations of Dhaka city by using travel time between any two key locations. We have used CDR data of 2.87 millions of users in Dhaka city collected by the largest mobile phone operator Grameenphone for a one month period between June 19, 2012, and July 18, 2012.

The main challenge of understanding the mobility from CDR data is that CDR data does not capture the continuous movement of a user, rather it only captures the location from where one makes a call or use the data connection. Hence, we devise a technique that considers only those consecutive CDR data that can be associated with the mobility of the user and the travel time in the real road network. Moreover, we apply a clustering technique to identify major road

junctions in the city and map the call transition between two cell-phone towers to two pairs of junctions in the city. To measure the travel time at different time zones, we also identify peak and off-peak hours from the CDR dataset. Our experimental results based on CDR dataset show that we can effectively predict travel time between any two major junctions of the city. We validate our results with real-time travel data collected through smartphones and find that our method can predict travel time with an average accuracy of 87%.

1.3 Summary of Results

In this thesis, we have the following contributions:

1. We are the first to propose an effective way to assess the travel time from mobile users' CDR data in a megacity of a developing country.
2. We exploit the road network travel time and mobile users mobility to estimate travel time between any pair of locations of Dhaka city.
3. We test our results with real-time traffic that shows a high prediction accuracy of our method.

1.4 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 describes the data set used in this paper. Chapter 3 focuses on literature review. Chapter 4 depicts the proposed methodology for estimating travel time from a static set of CDR data. And Chapter 5 presents another methodology for predicting travel time from the stream of CDR data. Chapter 6 shows the experimental results of our approach. Chapter 7 compares the accuracy of our approach with respect to real-time travel data collected through smart-phones. Finally, Chapter 8 concludes the thesis with the summary of the finding of this research and highlights the directions of future researches.

Chapter 2

Data

In this chapter, we provide some basic description of data set used. Definitions that are not included in this chapter will be introduced accordingly. We start, in Section 2.1, by giving some descriptions of our area of study. In Section 2.2 we describe attributes of CDR data. We devote Section 2.3 to describe real travel data which is used in result validation.

2.1 Study Area

The study area in this research is the central part of Dhaka city, where major roads of the city have been taken into consideration. The major road networks of Dhaka city consist of 67 nodes (road junctions) and 215 links (roads between nodes) which cover an area of about $300km^2$ and a population of about 10.7 millions [2]. Due to heavy traffic, the average speed of the motorized vehicle is approximately $17km/hr$. 38.3% trips in this area are found as non-motorized trips, and every day 2.74 trips are made by per person on average where 19.8% of them are making their trips by walking [2].

Approximately more than 90% area of Dhaka city has mobile phone penetration (66.36% is the national average) and the highest market shareholder mobile operator is Grameenphone Ltd. with 42.7million mobile phone subscribers nationwide.[1].

2.2 CDR data

The CDR dataset, provided by Grameenphone Ltd, consists of call records made by users over a one month period between June 19, 2012, and July 18, 2012. The dataset contains a total of 971.33 million anonymized call records. Figure 2.1 shows the frequency of calls for a different number of users over a one month period (top), and a number of calls vs. number of users in a randomly selected day (bottom).

Each entry in the CDR data contains the latitude and longitude of the Base Transceiver Station (BTS), caller's unique identification (anonymized), date, time and duration of their calls. Figure 2.2 presents a snapshot of the data (top). Figure 2.2 also shows the captured location in the CDR of a person (bottom) as she moves within the city.

2.3 Real travel data

Real world travel data among key locations of Dhaka city are also collected and used to validate the results of this research. GPS enabled smartphones are used to collect this data over a week within different time segments.

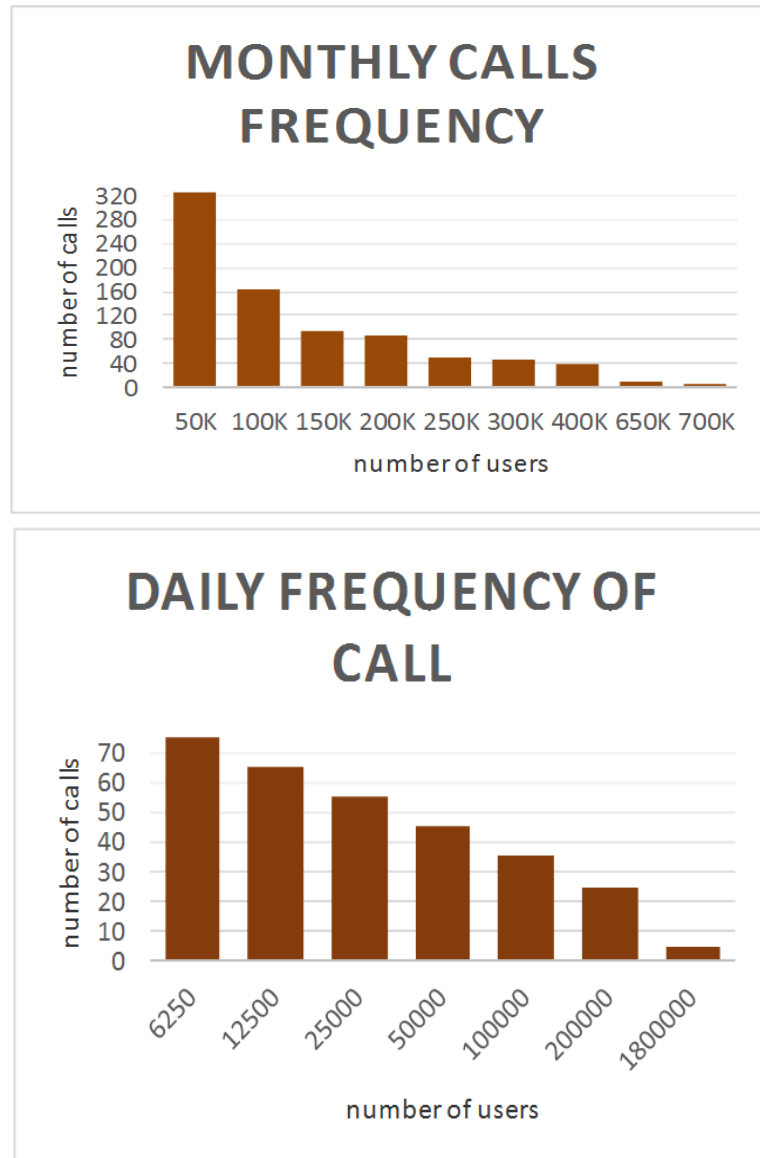


Figure 2.1: Frequency of calls per user

ID	Call Date	Call Time	Duration	Latitude	Longitude
AH03JAC8AAAbXtAId	20120701	09:34:19	18	23.8153	90.4181
AAH03JABiAAJKnPAa5	20120707	06:15:20	109	23.8139	90.3986
AAH03JABiAAJKnPAa5	20120707	09:03:06	109	23.7042	90.4297
AAH03JABiAAJKnPAa5	20120707	10:34:19	16	23.6989	90.4353
AAH03JABiAAJKnPAa5	20120707	18:44:53	154	23.6989	90.4353
AAH03JABiAAJKnPAa5	20120707	20:00:08	154	23.8092	90.4089
AAH03JAC5AAAAdAYAE	20120701	09:15:05	62	23.7428	90.4164
AAH03JAC+AAAcVKAC	20120707	08:56:34	242	23.7908	90.3753
AAH03JAC+AAAcVKAC	20120701	18:03:06	36	23.9300	90.2794
AAH03JAC5AAAAdAYAA	20120701	11:15:55	12	23.7428	90.4164

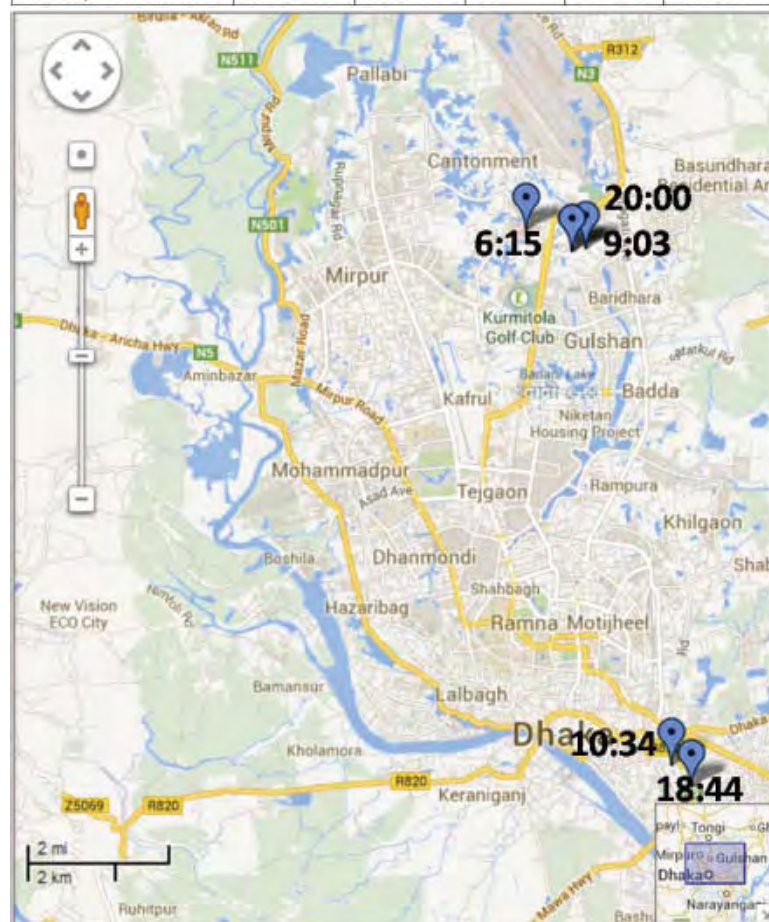


Figure 2.2: Example CDR data and locations of highlighted user

Chapter 3

Literature Review

This chapter discusses the related works which address travel time prediction. The travel time between two locations is a key constraint and descriptor of human mobility. For this reason, up-to-date information on travel time information is important for planning and governance of transportation as well as for travelers. Employing a relatively large amount of probes, the earlier bottlenecks can be reduced by directing traffic to other routes to mitigate congestion and provide more efficient routes to travelers. Monitoring road conditions and assessing access time to emergency services are instances where such information is useful. Therefore the importance of exact travel time is evident, but still, the limitations of such monitoring still exist in developing countries due to the insufficiency of necessary resources. By using information and communication technology (ICT) such monitoring and building an Intelligent Transportation System (ITS) are easily possible for addressing and mitigating transportation and congestion problems. An ITS relies on location-based information and obtains information about travel time, traffic condition and road incidents by monitoring and processing information of those locations of vehicles. Location-based information is generally used for monitoring real-time traffic and/or predicting road traffic status. Real-time road traffic is an open problem in cities that target the avoidance of road accidents, fuel, and energy consumption and unexpected delays. Predicting road traffic is an important functionality that can facilitate route planning and other traffic decision making processes. A detailed survey of latest developments of data-driven

ITS can be found in [40].

There are several approaches for estimating and monitoring travel time. Although some of these approaches provide highly accurate real-time estimates based on travel speeds and times, they typically require the installation of physical equipment (e.g. magnetic loop detectors) which makes them resource-intensive, or labor-intensive (surveys). There are two types of measurements: Point-based and Distance-based using traditional equipment, such as cameras, magnetic induction loops, microwave sensors, Bluetooth scanners, etc.

The most recent approach emerged in the past years driven by the spread of wireless technologies uses wireless data from floating car known as floating car data (FCD) such as GPS traces from public transports or taxies. FCD systems rely on mobile devices or onboard units (OBUs) equipped with positioning devices that actively report the vehicle location and speed to a specific server. Sufficient density of these probes produces traffic speed and intensity information after aggregating data from each probe.

Mobile phones are another potential source of information about human mobility due to high penetration rate. Due to the huge growth of mobile phone users and the availability of location traces of mobile users from their call records, recent research focuses on utilizing these records to find interesting patterns of city dwellers. In recent years, cellular data has been used for mobility pattern extraction, human travel pattern visualization, route choice modeling, traffic model calibration, and traffic flow estimation (e.g., [38][13]).

Therefore in Section 3.1 we discussed traditional approaches that have been used for solving transportation problems. Then Section 3.2 describes approaches that use floating car data to estimate travel information. And finally Section 3.3 describes approaches leveraging cellular data to analyze traffic.

3.1 Traditional Approaches

Traditional approaches offer two types of measurement, one is Point-based the other is Distance-based. Point-based approaches collect data through road side and household surveys (e.g.,

[15], [22], [14]). These approaches are not cost effective and cannot capture real-time traffic due to the low frequency of update. A large number of sensors must be installed in order to gain a realistic and complete view of traffic conditions using these approaches. Moreover, these techniques are prone to sampling biases and reporting errors. Distance-based approaches measure average speed and travel time for vehicles passing multiple detection zones. Like modern cities deploy fixed sensors, which can be either hard-wired or wireless like road mount detector or camera mounted on traffic lamps or road-sides or magnetic loop detectors [36][40][4], which provide information on the speed at which vehicles are traveling, as well as the capacity of a given road. They can also gather information on the type of vehicle, the distance between vehicles, pollution conditions, road surface conditions, and many others. Fixed sensors that are inductive, piezoelectric, or magnetic can be placed under the road, and those that use radar, ultrasound, or infrared with pyroelectric effect can be placed beside it. Each sensor is equipped with a control unit, battery system, solar panel, and transmission system. Distance-based approaches require vehicles to be identified and tracked. Hence, they may be prone to privacy constraints (e.g., license plate recognition) or to the limited representativeness of the probes (e.g., only vehicles equipped with DSRC, dedicated short-range communications, toll transponders). These techniques also suffer from the same limitations and low penetration rate [6].

There is another approach which exploits land usage pattern to derive a theoretical model to estimate the number of trips and their directionality. These approaches can be unreliable and can have financial and temporal costs.

3.2 Floating Car Data

With the spread of wireless technology, an approach has emerged which uses wireless data from continuously moving vehicle known as floating car data (FCD) such as GPS traces from public transports or taxies (e.g., [12][34][39]). FCD systems usually rely on mobile devices or onboard units (OBUs) equipped with positioning devices that provide vehicles location and

speed information to a pre-specified server. After data aggregation from each probe traffic speed and intensity are estimated depending upon the density of probe vehicles if it is high enough. Successful commercial examples of this systems are Google Maps (e.g., through their Android OS) or Here Maps (i.e., Nokia/Microsoft OS now owned by Audi, BMW, and Mercedes). A somehow improved approach of this systems use ordinary car drivers for contributing data in both an opportunistic crowdsourcing manner (e.g., a user allows its device to report its location to the service) but also a participatory crowdsourcing manner (e.g., a user reports on a map where a road was blocked). Successful examples of these systems include Google Maps, Apple Maps, Waze, Nokia's HERE maps, and Mapquest. The FCD approaches are also limited by their coverage and low penetration rate, particularly in developing countries. When FCD is collected from privately-owned cars, as in [12], the penetration rate becomes a limiting factor. The minimum amount of probe vehicles that allow for an accurate traffic status estimation has been extensively studied in literature [5], [16], [37] and depending on the reporting interval it can vary from 1 % to 5 % in highway scenarios and from 5 % to 10 % in urban scenarios. If FCD collected from taxis which is usually permitted to use dedicated lane in some cases are subjected to different speed limits than cars or buses. Google also collects travel time data from their users' smartphones; however, users may not be willing to reveal their locations due to privacy concerns. Thus major limitations of this approach are the unavailability of data and a low penetration rate of smart-phones in developing countries [25].

3.3 Cellular Data

In recent years, cellular networks have been used to monitor traffic status of a city and showed a valid alternative to replace FCD. Penetration of mobile devices is higher in developing countries than the penetration of smartphones. Thus a large amount of data is generated by mobile phone users on daily basis. There are two main approaches to estimate travel time from cellular data: call detail record (CDR) based and passive monitoring. Whenever a mobile user initiates or terminates a voice call or SMS/MMS or data connection, a CDR is generated and stored for

billing purposes. The format of CDR may differ from operator to operator, but the data always contain a users location and time information. On the other hand, passive monitoring is the observation of signaling message when idle users mobile terminal change network. The approaches that use passive monitoring are described at Subsection 3.3.1 and the approaches that use CDR data further are described in Subsection 3.3.2.

3.3.1 Passive Monitoring

A monitoring infrastructure is required to use passive monitoring approaches to tap the links of the cellular network and parse the signaling protocols [27]. The cost of monitoring installation, as well as the achievable accuracy and coverage, depends heavily on which network interface is monitored.

There are other works [19],[18] that are used for passive monitoring to estimate travel time in real time. Their work heavily relied on the signaling data generated from “idle” devices or devices those are not involved in a call or data connection with the base station. These CDR data from “idle” devices may not be available in every country especially in developing countries because of additional expenditure to track a huge volume of idle phones. Moreover, they only considered CDR data which are produced from nearby important roads or highway. For these reasons, passive monitoring may not always be a good option to estimate travel time.

3.3.2 Active CDR Data

Due to easy access to CDR data first used on study to understand human mobility [5] [13] [31]. The CDR data has also been used for traffic flow statistics, origin-destination matrices determination (e.g., [7],[10],[17],[30],[32],[33],[38]). The use of CDRs has been a matter of criticism recently due to characterizing human mobility [26]. There are few works which target real-time road traffic estimation where these works target nonreal-time applications. In [29] traffic estimation is found based on double-handovers. In [24], authors developed a system for detection of road traffic in real-time using mobile devices. The system was using low and high-

frequency motion detection and activity classification with precision and recall over 83% for traffic detection. A huge investment should be made in order to provide data using hardware infrastructure like cameras.

In [29], the feasibility of using mobile phones as traffic probes is analyzed. The authors mention that, compared with available alternatives, mobile phones offer some appealing characteristics such as sample size, coverage, and cost. In [5], a CDR dataset with cell handover information is used for measuring traffic speed and travel time across a highway segment of 14 km for several weeks. The results indicate a good correspondence between the cellular data and validation data from magnetic loop detectors. However, the study is limited to active users and is based on an undocumented proprietary algorithm from a commercial company. In [23] an algorithm is proposed, which uses low resolution positioning data (from cellular networks) to estimate traffic status and speed. This approach is validated only by means of simulations. In [9], the authors describe a real-time urban monitoring platform that uses mobile cellular data for the evaluation of statistical indexes based on monitoring the movement of mobile equipment. This platform can also be used to estimate the traffic intensity in specific regions of the monitored area by counting the number of calls that were made by mobile users over some time interval.

Another work [21] used CDR data to predict travel time between cities. This work does not consider a city's internal road network characteristics and people's mobility behavior, and thus is not suitable for predicting travel time among different suburbs of the cities.

In the present work, we introduced a new approach that includes, a semi-automatic algorithm to predict the travel time from mobile users' CDR data in a megacity of a developing country and an algorithm which enables real-time traffic monitoring from the stream of CDR generates from mobile users of a megacity of a developing country.

In this work, we consider the complexity of road network of a highly dense city like Dhaka and estimate travel time without using passive data which does not exist in developing countries.

3.4 Summary

This chapter discussed the existing approaches used to measure traffic status with their potential and limitation. Major approaches are fixed sensor or hardware based measurement, floating car data and cellular data. How cellular data provide an opportunity to overcome existing limitations to assess road traffic is also covered in this chapter.

Chapter 4

Travel Time Estimation from static CDR

Figure 4.1 shows the block diagram of our methodology for travel time estimation from static CDR data (from secondary storage). Our method takes raw CDR data as input. Our key insight of the proposed methodology is that the travel time of a source to a destination can be approximated from a user's two consecutive calls when the call is made from two distant locations and the time difference between these two calls is proportional to the travel time. Thus, the main challenge is to identify those CDR pairs that can contribute to the travel time. In the first phase of our approach, we apply a set of pruning rules to eliminate CDR pairs that cannot be a part of contributing the travel time. Then, from the selected set of CDR pairs, we approximate travel time for each CDR pair, which we call transient travel (*t-travel*) time. In the next phase, we partition all *t-travels* into three-time segments: off-peak, morning-peak, and evening peak hours. At this stage, we also identify key city locations (key road junctions) by analyzing CDR data. By combining *t-travels* (and their corresponding origin-destination pairs) with the underlying road network among key locations, we estimate the actual travel time between any pair of key city locations. Using travel time between any pair of key city locations we also estimate travel time between any location pair.

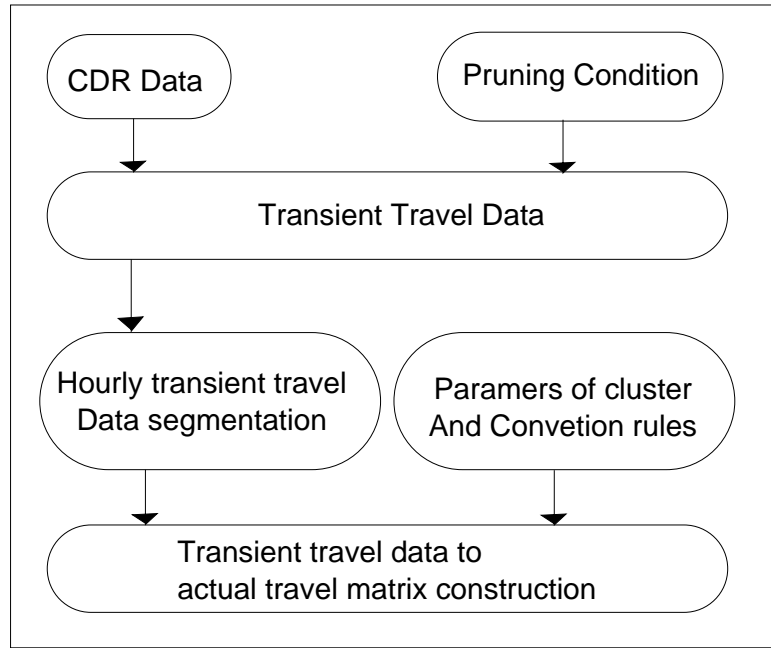


Figure 4.1: Block diagram of travel time estimation from static CDR

The process is summarized in subsequent sections. In Section 4.1, we discuss transient time generation between BTS (towers). In Section 4.2, we describe the methods to determine peak and off-peak hours and to partition t -travels into these peak/off-peak categories. After that, in Section 4.3 we estimate the actual travel time from transient travel time for the identified peak/off-peak categories. Finally, we generate the matrix of travel time for all pairs of key locations in the city and use it further to estimate travel time between any pair of locations of Dhaka city.

4.1 Tower-to-tower transient travel generation

First, we group CDR records of each user and sort these in ascending order of the call time. At this stage, we use the BTS location, from where the user made the call and the time of the call to determine transient travel (t -travel) time. If a user makes two consecutive calls from two different BTSs and the time difference of these two calls is proportional to the traveled distance between these two BTSs, then we extract this time difference as t -travel.

Since CDR data contain sparse and irregular records of a user’s movement [11], we need to filter out those records that cannot contribute to t -travel. For example, a user’s first location may be observed at location P_1 at time 12:30 and the next location may be recorded at location P_2 at time 20:45 and in this case. Since no record is available for a long duration, these two consecutive records cannot be used for estimating t -travel. Other than irregular records, sometimes the operator switches the call to an adjacent BTS in order to balance the workload. In such cases, we also may get some misleading t -travel.

To overcome the above limitations and to reduce the number of false trips from CDR data, we have used two pruning conditions. The first condition is that the duration of a t -travel cannot be more than 130 minutes, as the travel time between any two key locations of our observed area in Dhaka city is generally less than two hours. In our second condition, we eliminate a t -travel if the travel duration between two BTSs is less than 5 minutes. This condition helps us to eliminate records involving tower switching for load balancing purpose.

Thus, we consider t -travel records whose travel times lie between 5 minutes and 130 minutes. Note that, this parameter is city dependent, and thus to analyze CDR data of other cities, the parameter needs to be adjusted.

t -travels obtained from CDR data largely vary in different part of the day. This is due to peak and off-peak traffic hours of the city. Thus, in the next section, we identify peak and off-peak hours from these t -travels.

4.2 Peak/off-peak hours segmentation

In this section, we automatically identify peak and off-peak hours of the city based on our identified t -travel, and then partition all t -travels of a day into different time segments based on the identified peak and off-peak information.

First, we partition t -travel data on an hourly basis. Thus, we have a total of 24 partitions for 24 hours of the day. Each t -travel contains one origin-destination (OD) pair. For each hour time slot, there may be hundreds of t -travels for each OD pair. Let us assume that there are

total tc_i number of t -travels for i th hour of time slot. For the time slot i , let us assume min_i be the total number of t -travels that represent minimum travel time among all t -travels for any OD pair. That is, to compute min_i , we first determine the minimum t -travel for each OD pair, and then we check whether this minimum t -travel occurs in the time slot i and count the number of minimum t -travels in the time slot i .

Finally, we compute a ratio of occurrence of minimum travel time counts min_i to total travel time count tc_i for each hour time slot i . This ratio gives a measure of peak and off-peak hours. If the value of the ratio is high then the time slot hour is denoted as off-peak. On the other hand, if the value of the ratio is small, then the corresponding time slot hour is considered as a peak-hour. The intuition behind choosing this metric is as follows. If for a particular time slot, the number of travels with minimum t -travel is large, then it is highly likely that the corresponding time slot is an off-peak hour.

So, the key insight of measuring off-peak hours is that the higher the value of the ratio of minimum occurrence to total occurrence of an hour, the higher the probability that the hour is an off-peak hour. The opposite is true for peak hours.

Table 4.1 shows the computed ratio for each hour time slot. First nine rows (ID 01-09) show the off-peak hours, the middle five rows (ID 10-14) represent medium-peak, and the last ten rows (ID 15-24) show peak-hours of Dhaka city. Our analysis shows that in Dhaka city the most probable peak hours are 09AM-01PM and 05PM-09PM, and the most probable off-peak hours are between 11PM and 07 AM. From the analysis, we further divide peak hours into two categories, namely morning peak hours (09AM-01PM) and evening peak hours (05PM-09PM). All t -travel data is divided into these three categories for the next step of processing.

t -travel data obtained from CDR do not always reflect real traffic scenarios due to the incompleteness of the recordings of user movements. Thus, we need to filter out data using some heuristics based pruning.

In our first level of pruning, we check whether the speed for a t -travel is realistic or not. We filter out t -travels whose speed does not match with motor vehicle's speed. By considering the traffic jam situation in Dhaka city, motor vehicles normally can travel at a speed range

ID	HOUR	RATIO	FREQUENCY
01	04:00 - 04:59 AM	7842 / 50488	15.53%
02	03:00 - 03:59 AM	4782 / 34774	13.75%
03	02:00 - 02:59 AM	5319 / 48641	10.94%
04	05:00 - 05:59 AM	21520 / 215363	9.99%
05	01:00 - 01:59 AM	8479 / 94215	9.00%
06	06:00 - 06:59 AM	32645 / 425359	7.67%
07	12:00 - 12:59 AM	16183 / 212670	7.61%
08	11:00 - 11:59 PM	22853 / 345533	6.61%
09	07:00 - 07:59 AM	27912 / 505545	5.52%
10	10:00 - 10:59 PM	31255 / 674994	4.63%
11	08:00 - 08:59 AM	25546 / 643787	3.97%
12	03:00 - 03:59 PM	39291 / 999870	3.93%
13	02:00 - 02:59 PM	38887 / 1025169	3.79%
14	04:00 - 04:59 PM	34547 / 942852	3.66%
15	09:00 - 09:59 AM	27713 / 850075	3.26%
16	01:00 - 01:59 PM	32610 / 1006672	3.24%
17	10:00 - 10:59 AM	31409 / 1017175	3.09%
18	09:00 - 09:59 PM	29233 / 950041	3.08%
19	05:00 - 05:59 PM	27564 / 963108	2.86%
20	12:00 - 12:59 PM	30424 / 1072673	2.84%
21	11:00 - 11:59 AM	29447 / 1056864	2.79%
22	06:00 - 06:59 PM	22222 / 991133	2.24%
23	08:00 - 08:59 PM	23445 / 1109712	2.11%
24	07:00 - 07:59 PM	21843 / 1162863	1.88%

Table 4.1: Peak/Off-Peak hour Estimation

between 3km/h and 40km/h. Thus, city-specific fine-tuned maximum speed bound leads us to a more accurate filtering.

In our second level of pruning, we filter out a *t-travel* if the shortest road network distance between the OD pair of the *t-travel* deviates by a threshold distance from the Euclidean distance of that OD pair. This is because, if the road network path differs significantly from the Euclidean path, then there is a tendency to introduce additional errors in the travel time estimation due to the complex and possibility of many alternative routes. We have observed that if the deviation between the Euclidean and the shortest road network distances of an OD pair is within 25 percent of the Euclidean distance, then the *t-travel* data reflects realistic travel time in most of the cases.

At the final stage of pruning, if for an OD pair and for a particular time segment (off-peak/morning-peak/evening-peak), we do not have a sufficient number of *t-travels*, we omit those *t-travels* and the corresponding time-segment of the OD pair. In our case, we have observed that if we have at least five *t-travel* data for a particular time segment of the OD pair, then we can accurately estimate the travel time for that OD pair.

After three levels of pruning, for each time segment and for each OD pair, we take the average of all *t-travels* to estimate the travel time between the origin and the destination locations of the OD pair for that particular time segment. Note that we consider only weekdays travel data as the difference in travel time between weekdays and weekends is significant, and weekends have much less traffic. In Dhaka city, Sunday to Thursday are considered as weekdays.

4.3 Determining actual travel time from transient travel time

Our ultimate objective is to estimate travel time between any pair of major city locations and then estimate travel time between any pair of the city location. Thus, we first identify these key city locations from our CDR data. Then we map our estimated *t-travel* for OD pairs with these key locations, to get our actual travel time between any pair of key locations of the city.

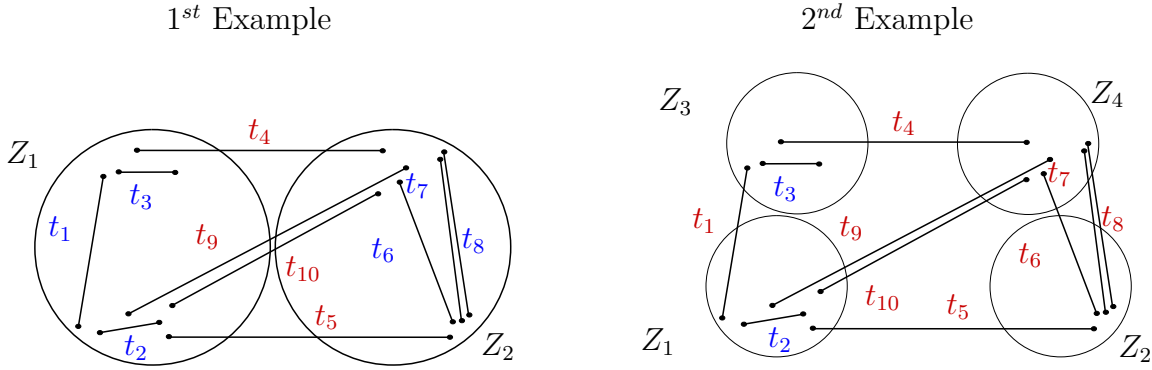


Figure 4.2: Intra-Cluster (Red Mark) and Inter-Cluster (Blue Mark) t -travel example

In order to estimate travel time, it is necessary to identify major traffic points of the city first. In the previous section, we discussed that every t -travel data consists of an origin and a destination location. These two locations are actually the location of mobile phone tower.

To identify all major traffic points (or in other words, *key locations*), we take BTS (tower) locations from all t -travel data as an input. Next, we consider towers as points in a 2D space and apply k -means algorithm. We determine the value of k through an empirical evaluation with the CDR data.

Let us explain our approach using two examples shown in Figure 4.2. In the first example, there are two clusters/zones, Z_1 and Z_2 , and in the second example there are four clusters/zones Z_1 , Z_2 , Z_3 and Z_4 . There are ten t -travels in both examples where intra zone travels (intra t -travel) are marked in blue and inter zone travels (inter t -travel) are marked in red. A higher number of inter t -travels yields more accurate estimation of travel time between key locations pairs. We see in the figure that the second example has more inter t -travels than the first example. Actually, with the increase of the number of clusters, the total amount of inter t -travels also increases (and the total number of intra t -travel decreases). However, with the increase in the number of clusters, the number of inter t -travels between clusters/zones pairs decreases. We see in Figure 4.2 that in the first example there are four inter t -travels (t_4, t_5, t_9, t_{10}) between Z_1 and Z_2 whereas in the second example there are only three inter t -travels between Z_2 and Z_4 . However, without sufficient amount of inter t -travels between each key locations pairs, accurate estimation of travel time between those two locations is also

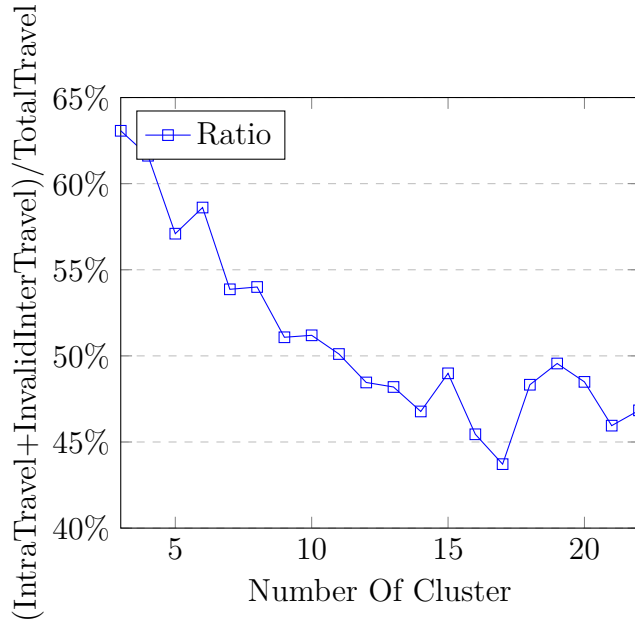


Figure 4.3: Empirical Evaluation of Number of Clusters

not possible. Therefore a tradeoff is required between the number of clusters and the amount of valid inter t -travels. Hence to model this tradeoff as a ratio, which can be expressed as follows.

$$Ratio = \frac{count(intraTravel) + count(InvalidInterTravel)}{count(totalTravel)}$$

The above ratio essentially captures the ratio between the amount of valid inter t -travel and the total amount of t -travel. For a better prediction accuracy in the final result, less amount of intra t -travel and invalid inter t -travel are desired. The smaller the ratio, the better the cluster. Let us now assume that at least three inter t -travels are required between each key locations pairs to compute travel time. Then, for the first example, this ratio is $(6+0)/10$ or 60% and for the second example, the ratio is $(2+5)/10$ or 70%. In this example, the first partition is more desirable than the second one.

To determine the optimal value of k we compute the ratio while varying the value of k from 3 to 22, and found that $k = 17$ yields the best ratio. Figure 4.3 shows the simulation results.

The centroids found from k -means algorithm are further adjusted to nearby road junction to obtain the key locations of Dhaka city. Key locations of Dhaka city are shown in Figure 4.4

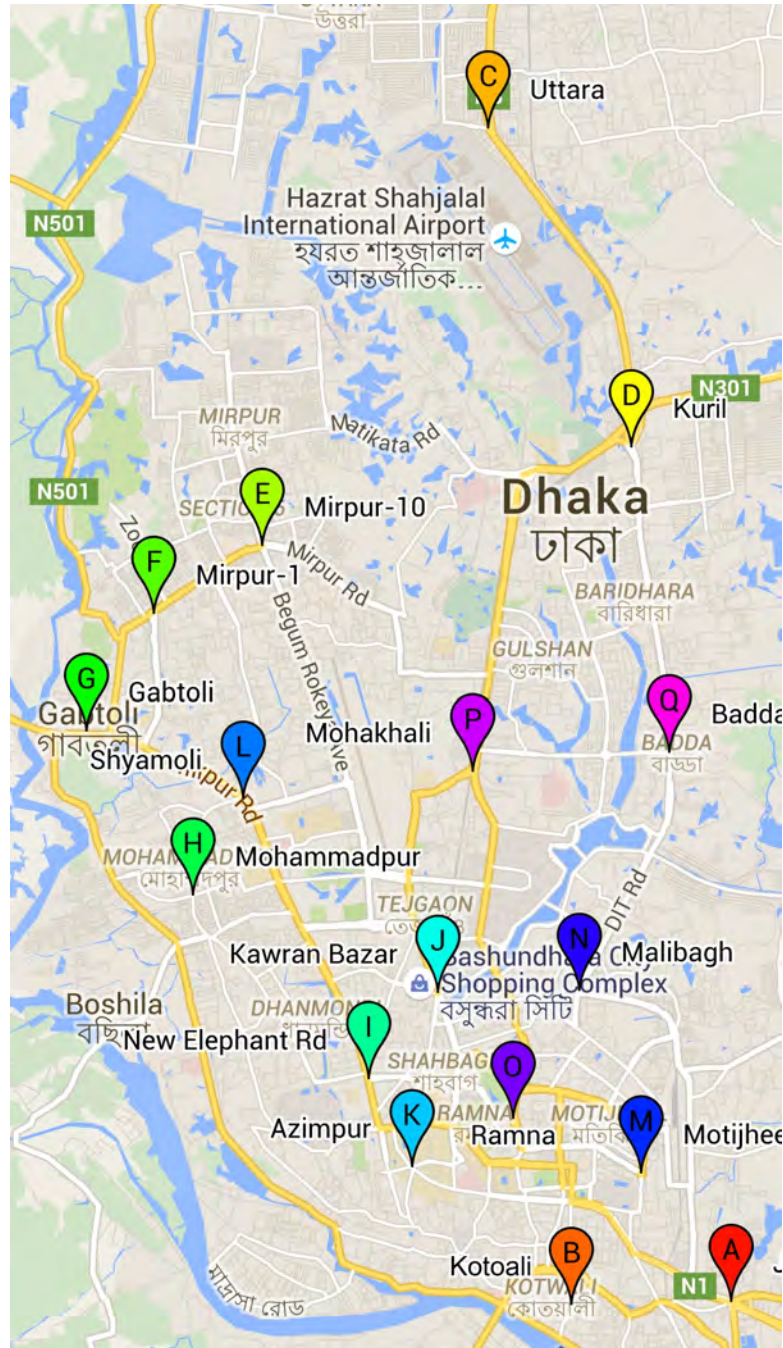
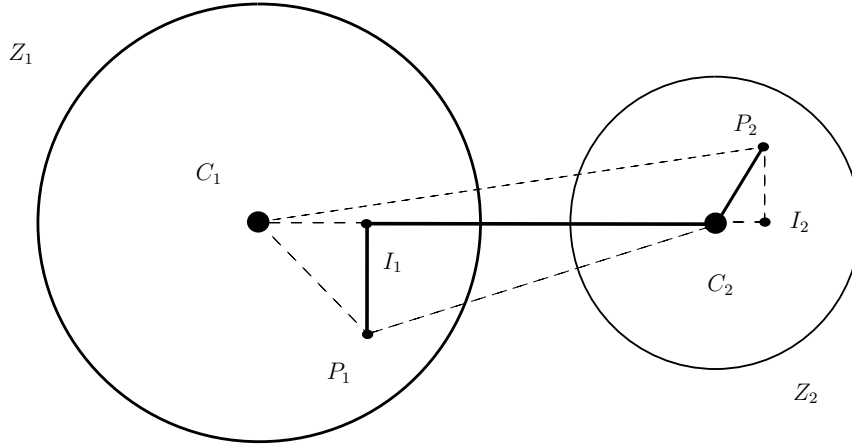


Figure 4.4: Major Locations within Dhaka City

Figure 4.5: t -travel to actual travel transformation

and their latitude and longitude are shown in Table 4.2.

Note that, this is the only step where some sort of manual input is required. However, these key locations can be identified once and reused for any subsequent data analysis. After identifying all major traffic locations, OD pairs of every t -travel data are mapped to one of those 17 clusters/zones.

To estimate the travel time between two key locations from t -travels, we take the following steps. Let us explain these steps using an example shown in Figure 4.5. Figure shows a t -travel data for an OD-pair, where P_1 and P_2 are two end-points of this OD pair. P_1 is associated to Zone 1, Z_1 and P_2 is associated to Zone 2, Z_2 . Here C_1 and C_2 are two key locations representing Zone 1 and Zone 2, respectively.

Let $dist(C_1, C_2)$ be the distance between C_1 and C_2 , and $dist(P_1, C_2)$ be the end distance between P_1 and C_2 . Let $dist_{network}(P_1, P_2)$ and $time(P_1, P_2)$ be the road network distance and travel time from P_1 to P_2 , respectively. Then the travel time between C_1 to C_2 , denoted as $time(C_1, C_2)$, can be computed as follows:

$$time(C_1, C_2) = dist(C_1, C_2) \times \frac{time(P_1, P_2)}{dist_{network}(P_1, P_2)} \quad (4.1)$$

To compute the network distance from P_1 to P_2 , we use Algorithm 1. Essentially, Algorithm 1 finds the shortest road network distance through key points C_1 and C_2 . In Algorithm 1, d is the network distance between P_1 and P_2 , $dist_{network}(P_1, P_2)$. I_1 and I_2 are two intersection

Index	Name	Latitude	Longitude
01	JatraBari	23.710174	90.434322
02	Kotoali	23.709804	90.411932
03	Uttara	23.860675	90.400266
04	Kuril	23.819763	90.420328
05	Mirpur-10	23.807099	90.368584
06	Mirpur-1	23.798385	90.353371
07	Gabtolli	23.783352	90.343925
08	Mohammadpur	23.762357	90.358840
09	Elephant Road	23.738767	90.383529
10	Kawran Bazar	23.749862	90.393174
11	Azimpur	23.727450	90.389548
12	Shyamoli	23.774750	90.365875
13	Motijheel	23.726638	90.421696
14	Malibagh	23.750079	90.413001
15	Ramna	23.733680	90.403714
16	Mohakhali	23.778174	90.398093
17	Badda	23.780661	90.425633

Table 4.2: Major Traffic Locations in Dhaka

points, where perpendicular lines from P_1 and P_2 to line C_1C_2 meet, respectively.

Algorithm 1 Actual network distance estimation of a t -travel data

```

1: function ACTUALNETWORKDIST( $t$ -travel)
2:    $d \leftarrow \text{dist}(P_1, C_1) + \text{dist}(C_1, C_2) + \text{dist}(C_2, P_2)$ 
3:   if  $\text{dist}(P_1, C_2) < \text{dist}(C_1, C_2)$  then
4:      $d \leftarrow d + \text{dist}(P_1, I_1) - \text{dist}(I_1, C_1) - \text{dist}(P_1, C_1)$ 
5:   end if
6:   if  $\text{dist}(P_2, C_1) < \text{dist}(C_1, C_2)$  then
7:      $d \leftarrow d + \text{dist}(P_2, I_2) - \text{dist}(I_2, C_2) - \text{dist}(P_2, C_2)$ 
8:   end if
9:   return  $d$ 
10: end function

```

The key intuition to estimate $\text{dist}_{\text{network}}(P_1, P_2)$ is that, for each point P_1 (or P_2), we take the shortest possible path to connect it to C_1C_2 line. After computing this network distance, we scale t -travel duration to actual travel duration as per formula is given above. This process is then applied to all t -travel data to get the actual travel time.

4.4 Travel time estimation matrix construction

For each pair of the zone, there could be multiple numbers of actual travel data that generated from the previous step. We take the average of these actual travel times for a particular time segment (off-peak/morning-peak/evening-peak) to estimate the actual travel time between two key locations, represented as the centroid of two zones. By applying, the above technique for every pair of zones, we generate the travel time matrix. We also generate intra zone travel speed of different time segment for each zone by using those t -travel which lies within the same zone boundary. We take average here too in case of multiple t -travel exists within the same zone. Intra zone travel speed is required for travel time estimation between any locations

pair. To explain the estimation of travel time between any two locations of Dhaka city let us use Figure 4.5 again. Suppose we want to estimate travel time between P_1 to P_2 , denoted as $time(P_1, P_2)$. The $time(P_1, P_2)$ could estimate as below:

$$time(P_1, P_2) = shortestDist(P_1, C_1C_2) \times travelSpeed(Z_1) + time(C_1, C_2) \times \frac{partialDist(P_1, P_2)}{dist(C_1, C_2)} + shortestDist(P_2, C_1C_2) \times travelSpeed(Z_2) \quad (4.2)$$

Let us assume that $shortestDist(P_1, C_1C_2)$ is the shortest distance between location P_1 to line C_1C_2 , $travelSpeed(Z_1)$ is intra zone travel of Z_1 that we obtain previously and $partialDist(P_1, P_2)$ is the part of distance from C_1C_2 path that overlap with the path between P_1 and P_2 that is obtained using Algorithm 1. Meaning of all other terms in this equation are same as before. Thus, we can obtain $partialDist(P_1, P_2)$ using Algorithm 2.

Algorithm 2 Partial distance estimation of a t -travel that lies between the path of two centroids

```

1: function PARTIALDIST( $t$ -travel)
2:    $d \leftarrow dist(C_1, C_2)$ 
3:   if  $dist(P_1, C_2) < dist(C_1, C_2)$  then
4:      $d \leftarrow d - dist(C_1, I_1)$ 
5:   end if
6:   if  $dist(P_2, C_1) < dist(C_1, C_2)$  then
7:      $d \leftarrow d - dist(C_2, I_2)$ 
8:   end if
9:   return  $d$ 
10: end function

```

We use a different formula to estimate travel time between two minor locations than the formula for major locations, as traffic congestion has a lesser effect on intra zone travel time than that of inter-zone travel time. To reflect this attribute into the equation of travel time estimation between two minor locations we consider the intra zone travel speed. Therefore both

intra zone travel speed table and travel time matrix of key locations are required to estimate travel time between any two locations in Dhaka city.

4.5 Summary

In this chapter, we depict a methodology to measure key locations of Dhaka city and identify peak off-peak hour categorization from the dataset. Finally, we construct travel time matrix between key locations.

Chapter 5

Travel Time Estimation from CDR Stream

In our previous chapter, we have discussed travel time prediction from static CDR data. In this chapter, we are going to discuss real-time travel time estimation from the stream of CDR data. Our method takes a stream of CDR data as input. Our key insight of the proposed methodology is that the travel time of a source to a destination can be approximated from a user's two consecutive calls when the call is made from two distant locations and the time difference between them is proportional to the travel time. Thus, the main challenge is to identify those CDR pairs that can contribute to the travel time in real time. Figure 5.1 shows the block diagram of our methodology for travel time estimation from Stream of CDR data. For each CDR at first, we generate a transient travel (*t-travel*) time using the concept of the sliding window with a predefined maximum window range. With a dynamic mapping of CDR from the stream, whenever a transient travel can formulate, it goes for further steps. Actual travel time is then calculated using *t-travel* from the previous step. Finally, hourly travel time matrix is modified by actual travel time following some pre-defined rules. The process is summarized in a subsequent section. In Section 5.1, we discuss the formulation process of *t-travel*. After that in Section 5.2 how to formulate actual travel time from *t-travel* and update hourly travel

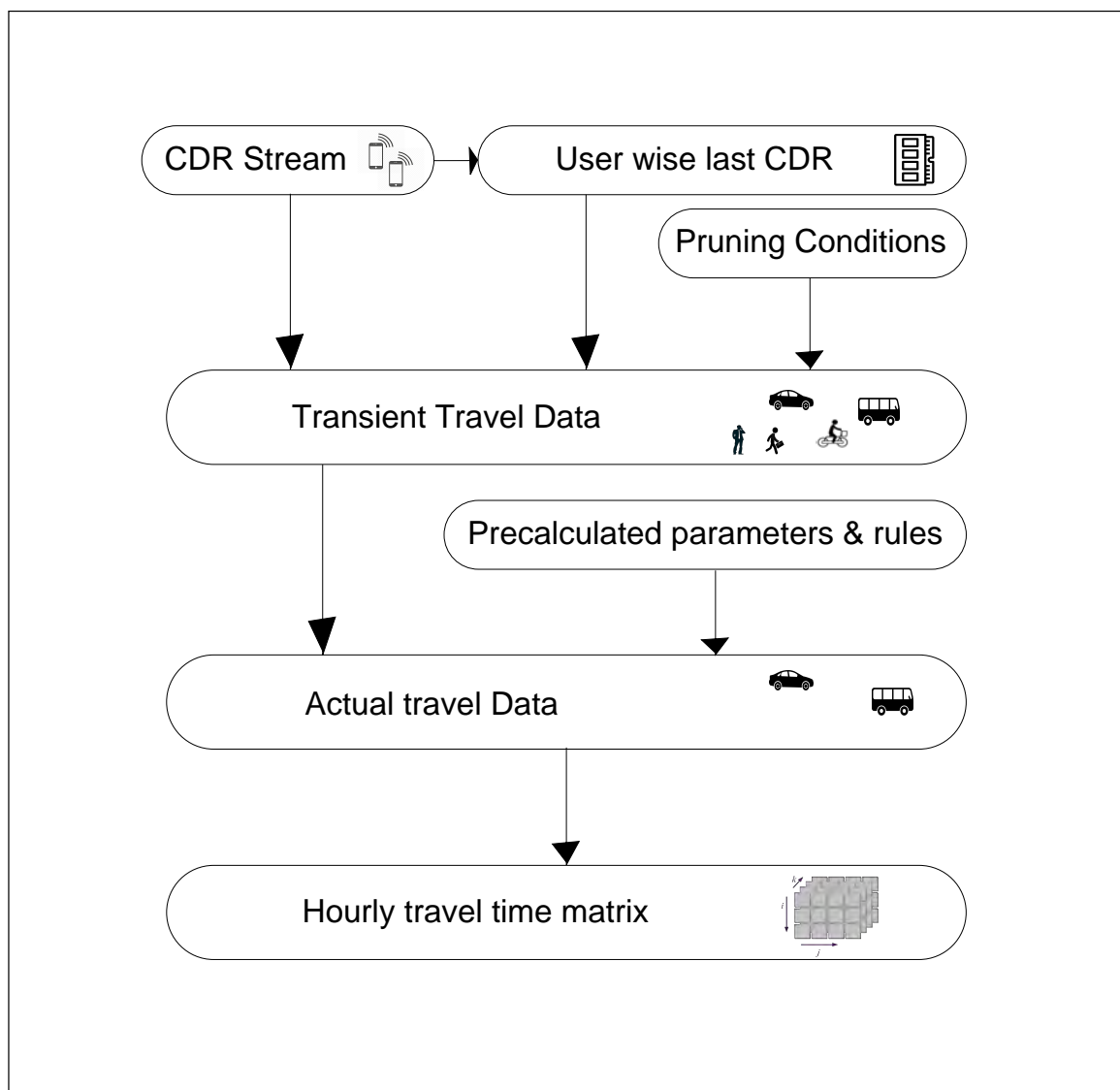


Figure 5.1: Block diagram of travel time estimation from Stream of CDR

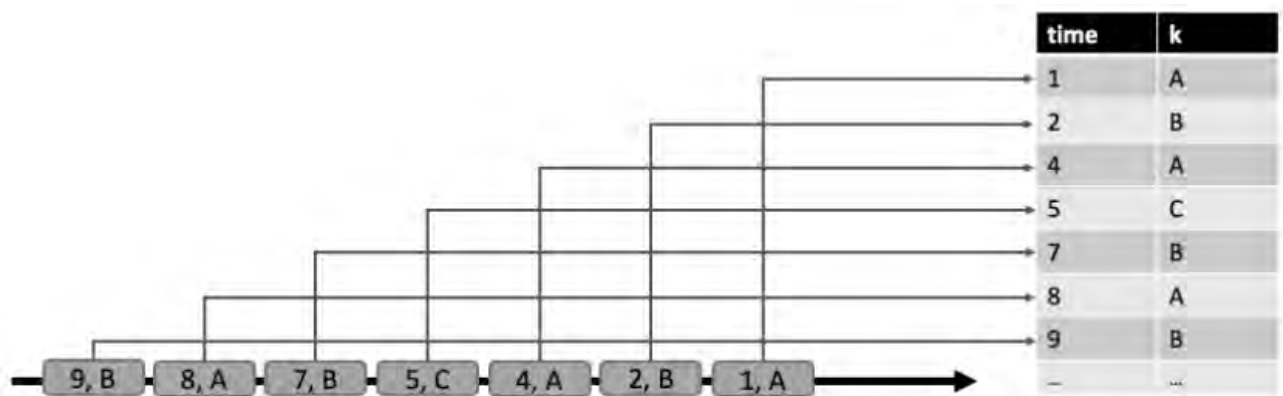


Figure 5.2: CDR Stream to Dynamic Mapping

time matrix is discussed.

5.1 Transient Travel Generation

In CDR stream, each record contains a timestamp as its generation DateTime. At first step, each CDR from the stream is mapped into a temporary hashmap order by their timestamp. Figure 5.2 shows the general idea of how a stream dynamically mapped into memory. In Figure 5.2 the head of the arrow represents the starting point of the stream and each record in the stream contains a timestamp (time) as the number and a key (k) as a letter (UserId in our case) which is used to find the record in future. Each record is inserted into the dynamic map in ascending order of time. This map continuously holds last one second's data in order to maintain memory constraint. This one-second range is like a window of from stream of CDR. This window continuously slides by one second and mapped CDRs found within that window into memory. An example of time sliding window is shown at Figure 5.3 where the window is set for last ten-second records.

In next phase for each sliding window, a continuous query applies to this dynamic map and all further insertion, update or deletion operation from the query applies to another map. A user id is the key of the other map. This map keeps last CDR of those users that are found within a maximum time boundary. The time limit set into max_range(in second) which is same

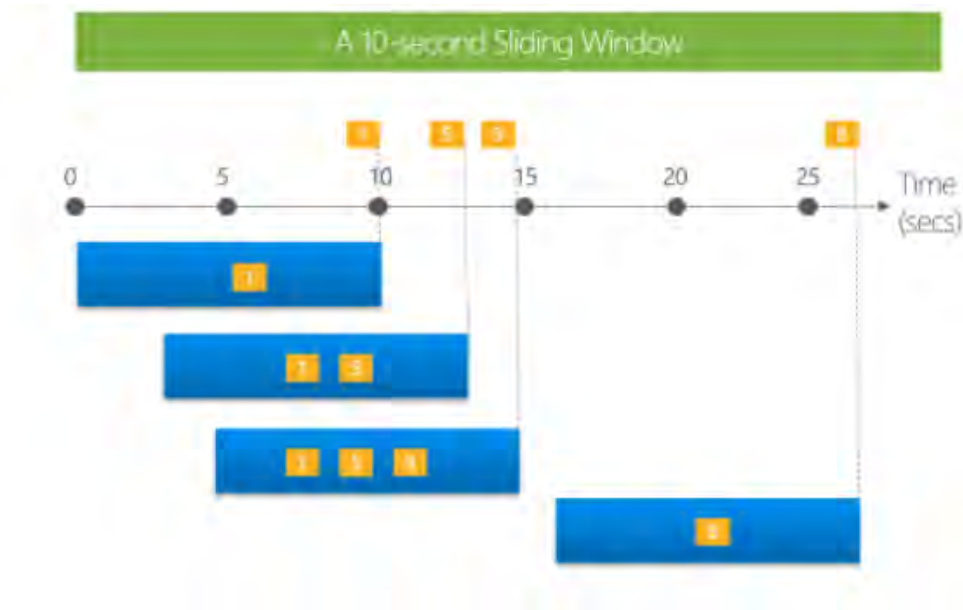


Figure 5.3: An example of 10 second sliding window

as maximum travel duration, discussed at Section 4.1. By using this two map, now we are able to generate transient travel or *t-travel*. The process of generation *t-travel* from a dynamic map using continuous query is depicted in Figure 5.4. In Figure 5.4 it is noticeable that only update operations are producing *t-travel* for next stage. The reason behind this is that to form a *t-travel* we need a pair of CDR which is found only in case of the update operation. When sliding widow advance by one-second delete operation takes place for oldest CDRs at second hashmap. This is how the memory size does not exceed a certain limit and this limit depends upon a number of unique user and number of CDR generated within the maximum travel duration we are allowing. And this limit is the minimum limit to formulate *t-travel* from the stream of CDR.

5.2 Updating Hourly Travel Time Matrix

To generate actual travel time we need to apply further filtering to *t-travel* produced from the previous step. *t-travel* within the minimum and maximum travel speed limit (described at Section 4.1) will be taken into consideration for further processing. Also only *t-travel* with the

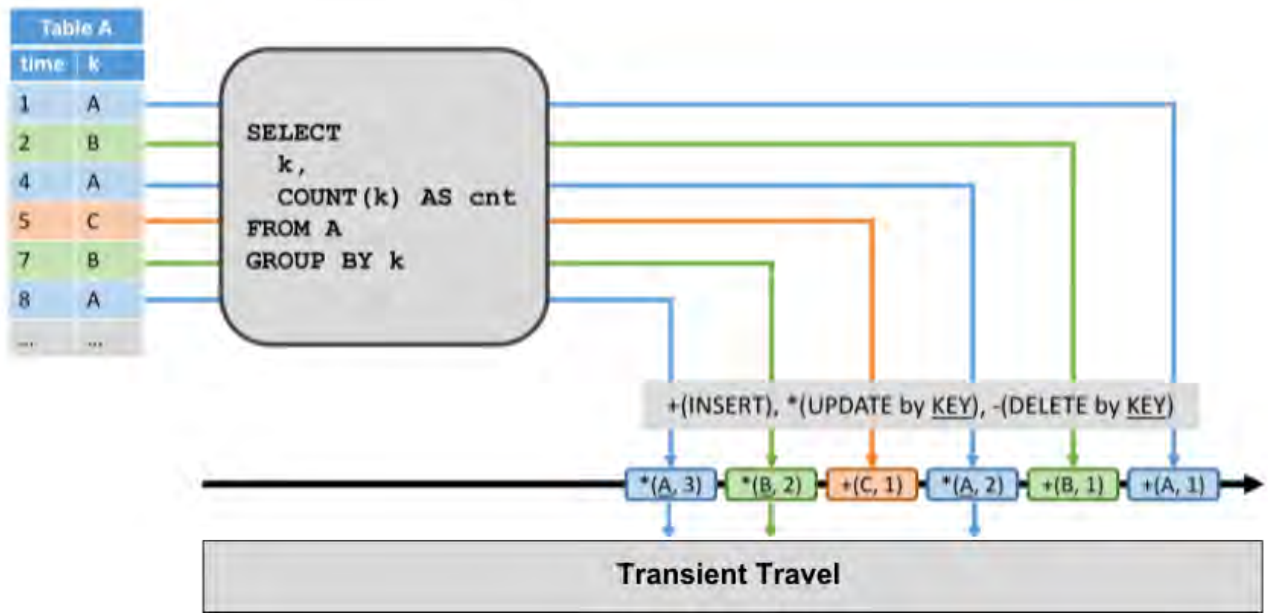


Figure 5.4: Formulation of *t-travel* using continuous query from dynamic map

valid source-destination pair will be taken into consideration to travel time estimation.

To estimate the travel time between two key locations from *t-travels*, we take the steps which are same as our previous method. Let us explain these steps using an example shown in Figure 4.5. Figure shows a *t-travel* data for an OD-pair, where P_1 and P_2 are two end-points of this OD pair. P_1 is associated to Zone 1, Z_1 and P_2 is associated to Zone 2, Z_2 . Here C_1 and C_2 are two key locations representing Zone 1 and Zone 2, respectively.

Let $dist(C_1, C_2)$ be the distance between C_1 and C_2 , and $dist(P_1, C_2)$ be the end distance between P_1 and C_1 . Let $dist_{network}(P_1, P_2)$ and $time(P_1, P_2)$ be the road network distance and travel time from P_1 to P_2 , respectively. Then the travel time between C_1 to C_2 , denoted as $time(C_1, C_2)$, can be computed as follows:

$$time(C_1, C_2) = dist(C_1, C_2) \times \frac{time(P_1, P_2)}{dist_{network}(P_1, P_2)} \quad (5.1)$$

To compute the network distance from P_1 to P_2 , we use Algorithm 1. Essentially, Algorithm 1 finds the shortest road network distance through key points C_1 and C_2 . In Algorithm 1, d is the network distance between P_1 and P_2 , $dist_{network}(P_1, P_2)$. I_1 and I_2 are two intersection points, where perpendicular lines from P_1 and P_2 to line C_1C_2 meet, respectively.

The key intuition to estimate $dist_{network}(P_1, P_2)$ is that, for each point P_1 (or P_2), we take the shortest possible path to connect it to C_1C_2 line. After computing this network distance, we scale t -travel duration to actual travel duration as per formula is given above.

Each actual travel time from previous section update 17×17 travel time matrix. There is twenty-four travel time matrix for each hour of a day and each cell of a matrix contain latest travel time for that source-destination with its last updated date-time. To update the corresponding matrix from actual travel time we use Algorithm 3. In Algorithm 3, $curTTime$

Algorithm 3 Travel time matrix updating from each actual travel time

```

1: function UPDATETRAVELTIMEMATRIX( $rValSigFact$ )
2:    $thld \leftarrow curTTime \times (1 + rValSigFact \times DateDiff(newTTime, curTTime))$ 
3:   if  $thld < curTTime$  then
4:      $curTTime \leftarrow newTTime$ 
5:   end if
6: end function

```

indicate current travel time value between two key locations stored in travel time matrix and $newTTime$ represent the travel time obtained from the stream using our calculation. We use the term $rValSigFact$ (factor of recent value's significance) while calculating $thld$ (threshold value) in Algorithm 3. That how quickly the impact of recent changes in travel time between any source-destination on travel time matrix depends on the value of this $rValSigFact$. The greater the value of $rValSigFact$ the quicker the matrix change with recent value and vice-versa. In our case we take 0.005 as the value of $rValSigFact$. Route-specific $rValSigFact$ could also apply which might in some cases, e.g. routes under long repair/under construction process. We could also use a mapping table for this factor so that some regular basis changes (e.g. seasonal impact) on routes could adjust automatically. This hourly segmented travel time matrix could feed to any dashboard for monitoring purposes or something similar.

5.3 Summary

In this chapter, we have measured hourly travel time matrix using sliding window and dynamic table concept from the stream of CDR. We also introduced a rule to update travel time matrix from a new travel time, extracted from the stream of CDR. We provided some future guideline for an automatic way to replace manual factoring in our calculations.

Chapter 6

Results

In this section, we present the results of our travel time estimation from CDR data and then analyze the estimated travel time.

6.1 Travel time estimation

The mobile phone network comprises of 1350 towers within the study area. Our data set consists of one month of CDR data. We have filtered out weekend data from this data set.

From these data, we have extracted 16399597 transient travel (*t-travel*) data for 80733

Step Name	Data Amount
Total User	6926973
Total Data	971328700
After Positional Displacement Prune	28228258
After Exclusion of Weekend's data	16399597
Distance Ratio within 25%	80733

Table 6.1: Amount of data contributed in each step

distinct origin-destination (OD) pairs. Table 6.1 shows the amount data contributed to each step of our methodology.

As per our findings in Table 4.1, we have partitioned all *t-travel* data into two groups: off-peak, and peak hours. After applying a different level of pruning, we have selected 3443910 and 6559830 *t-travels* for off-peak and peak hours, respectively. This result indicates that among all *t-travel* data, approximately 21% represents off-peak and 40% represents peak *t-travels*. This is a reasonable finding a number of travels made during off-peak hours are much less than that of peak hours.

To process stream of CDR data all travel data partitioned into 24 parts, each for every hour of a day. We order same data set in ascending order by recorded date-time and use it as a source of a stream of CDR.

6.2 Analysis of travel time matrix

Table 6.2 shows the estimated travel time and Google Map API's predicted travel time (of March 2014) for 14 pairs of key locations (we omit the other three pairs due to the space limit for the Table placement). Similarly, morning peak-hours and evening peak-hours estimated travel times are shown in Table 6.3. Intra zone travel time for each time segment is shown at Table 6.7. Diagonal entries in Table 6.2 and Table 6.3 represent travel time within a zone, which are marked as not measured (NM). In both tables, some cell is marked as DI, which denote data insufficiency to estimate the travel time. For example, travel time from Kuril (index-04) to Gabtoli (index-07) at Table 6.3 is marked as DI. This happens because all 289 (17x17) routes may not have a direct connection through major junctions in Dhaka city.

However, we can apply all pairs shortest paths algorithm in our resultant matrix with proper road network information, to fill these data insufficiency (DI) gaps and find best alternative routes with regards to travel time. For example, Kuril to Gabtoli travel time can be obtained by combining the routes of Kuril to Mohakhali (index-16) and to Mohakhali (index-16) to Gabtoli route. This route will take 51 (23+28) minutes to travel in peak hour.

We show some examples of alternative routes obtained by applying all pairs shortest paths algorithm at Table 6.6. Results from all pairs shortest paths show that on average 66% route has a better alternative route. Here in the case of long-distance travel, our assumption was that the user travels through the shortest path; but in real life scenario, this travel may also pass through some alternative zone/junction. We have also found that the estimation of travel time between two long distant key locations can be estimated more accurately than that of the travel time between intermediate zone. We can use information on alternative routes to uniformly distribute traffic among all roads of the city, which can be effective in reducing a significant amount of travel time.

Figure 6.1 and 6.2 is two examples that show how to travel time changes with changes of time and with the amount of CDR data. With a greater amount of CDR data smoother result can be achieved while less number of CDR data will produce less reliable results. Figure 6.1 and 6.2 also indicate that travel time in some time segment is always low and travel time in some segments of time is always high for most of the route.

After completing the processing of a stream of CDR, we got 24 travel time matrix. Table 6.5 represents estimated travel time of 2AM and 7PM and Table 6.4 represents estimated travel time of 9AM and 7PM.

6.3 Summary

We represent travel time matrix of the peak, off-peak hour produced from a static source of CDR dataset as well as travel time matrix of 24 hours of a day from the stream of CDR with a figure showing how result quality improve with the increase in the amount of data in this chapter. Google's travel time matrix and some example of alternative path prediction from estimated travel time matrix are also presented in this chapter.

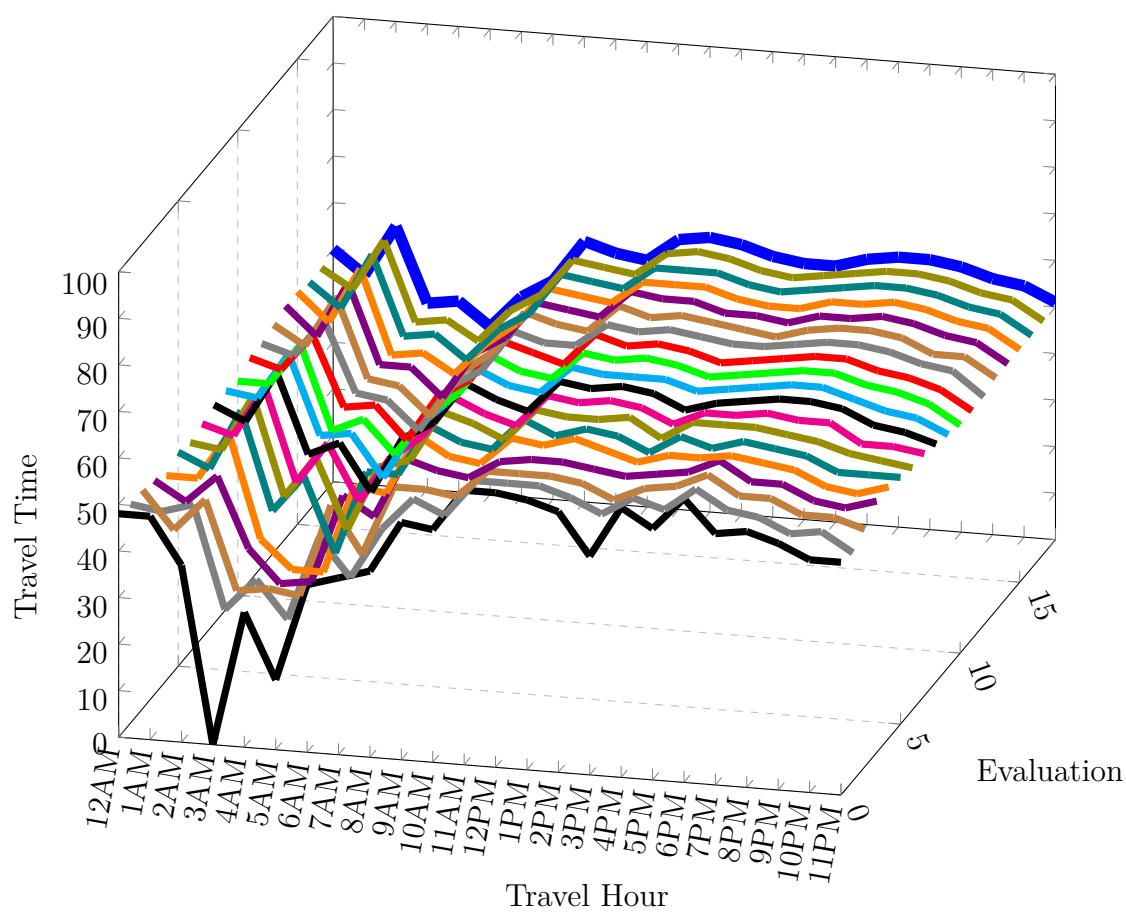


Figure 6.1: Result Comparison of Mohakhali to Mirpur-1

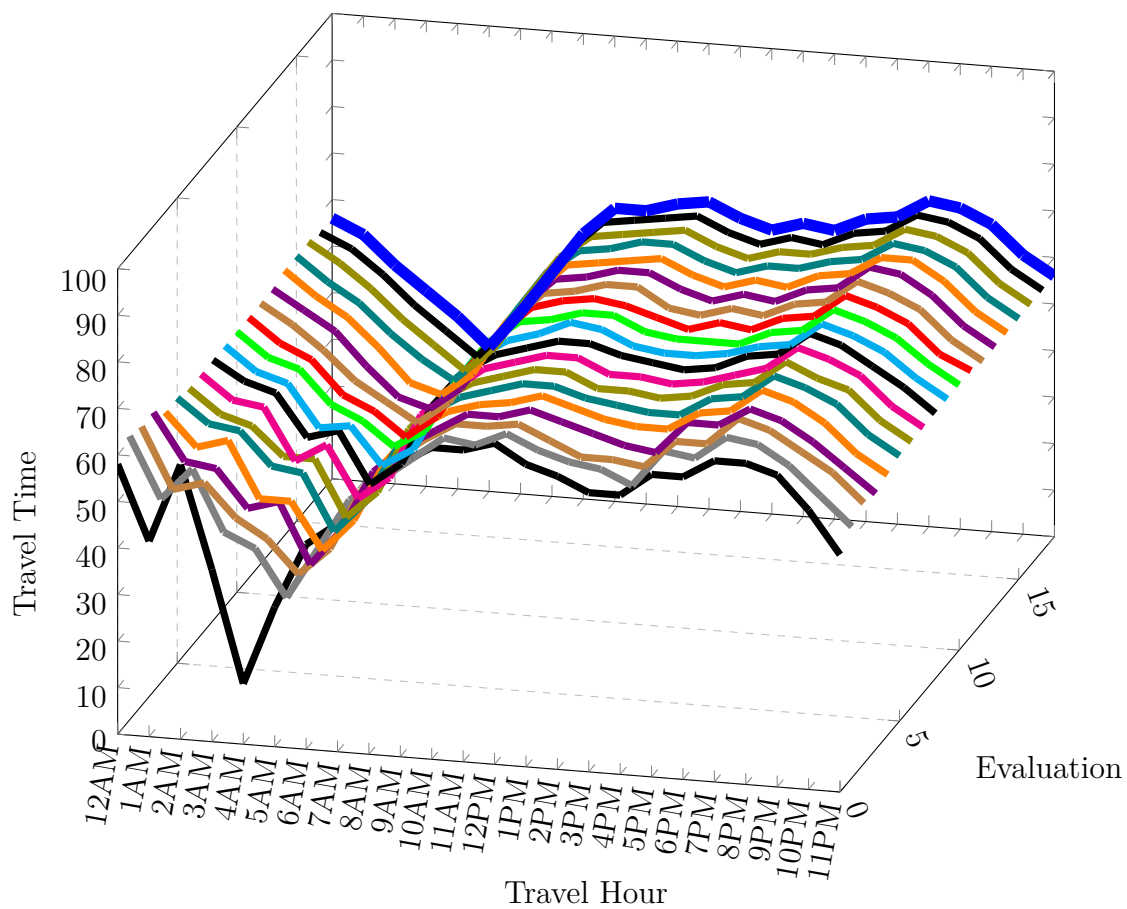


Figure 6.2: Result Comparison of Kuril to Malibagh

	01	02	03	04	05	06	07	08	09	10	11	12	13	14
01	NM	09/19	44/92	34/66	29/65	32/55	30/71	23/51	12/34	13/38	08/56	24/57	09/21	19/31
02	09/15	NM	43/85	32/69	30/65	34/67	29/62	22/42	15/28	14/31	10/09	25/50	08/10	17/27
03	37/90	42/90	NM	13/29	DI	28/33	35/55	33/62	33/65	28/62	36/79	32/41	36/73	26/61
04	30/63	36/67	15/25	NM	17/19	DI	DI	27/41	26/55	21/36	29/59	DI	30/48	19/36
05	29/68	32/69	DI	18/22	NM	03/10	10/22	16/35	18/43	15/34	23/56	11/23	25/56	23/45
06	31/53	33/73	34/39	DI	05/11	NM	06/09	12/26	18/39	19/36	22/55	07/14	29/49	28/34
07	31/76	28/66	40/52	31/41	11/26	09/10	NM	08/18	19/33	19/34	22/43	07/16	30/57	28/44
08	23/55	22/42	36/60	27/37	15/30	14/25	10/18	NM	13/11	14/14	14/24	04/08	23/35	23/23
09	12/32	14/27	34/64	25/43	19/33	22/42	18/30	12/11	NM	05/06	03/11	12/18	12/22	16/11
10	13/37	16/29	30/56	20/33	15/30	18/36	17/30	11/12	04/05	NM	07/13	10/17	09/20	09/07
11	08/35	10/09	38/75	28/70	23/53	DI	22/44	14/23	04/10	07/16	NM	15/30	10/16	16/13
12	23/58	25/53	33/75	DI	11/17	10/14	06/12	04/09	11/20	11/17	15/35	NM	22/42	20/32
13	05/19	11/08	40/72	26/49	26/48	30/44	29/54	23/33	12/23	11/22	09/15	22/43	NM	11/11
14	11/26	18/23	34/60	19/38	24/36	28/31	28/49	22/28	14/15	10/08	DI	21/33	10/10	NM
15	08/20	11/12	35/65	25/52	20/41	24/41	22/43	17/26	06/10	05/13	05/08	16/31	05/08	09/09

Table 6.2: Google travel time vs. off-peak estimated travel time from CDR

	01	02	03	04	05	06	07	08	09	10	11	12	13
01	NM	19/16	113/112	81/80	-/80	48/55	83/79	63/57	DI	-/40	32/-	-/62	24/21
02	14/18	NM	111/102	81/70	75/70	78/69	67/67	41/43	25/27	34/32	09/11	54/53	09/06
03	107/110	106/98	NM	31/33	DI	36/41	48/56	75/75	90/78	74/70	97/87	60/61	82/79
04	76/74	-/74	28/30	NM	19/25	DI	DI	35/52	-/53	39/39	-/62	DI	51/54
05	-/79	78/75	DI	21/21	NM	11/12	22/24	35/35	45/44	38/35	57/60	27/26	69/61
06	46/46	65/75	33/41	DI	12/13	NM	08/09	25/28	39/39	38/37	54/56	17/17	54/47
07	96/83	68/66	50/60	DI	25/27	09/10	NM	18/19	34/34	33/34	43/47	15/16	63/62
08	DI	43/44	71/71	40/45	31/35	26/28	18/19	NM	13/13	14/15	26/26	10/10	41/39
09	-/31	24/22	79/81	54/55	41/47	37/42	31/36	12/15	NM	06/08	12/13	21/24	24/24
10	35/39	32/28	65/68	37/41	34/39	40/42	32/35	12/15	06/06	NM	16/14	19/20	25/24
11	DI	11/09	96/91	74/75	61/55	49/54	43/48	24/26	12/13	18/19	NM	35/41	17/17
12	-/60	63/47	53/60	DI	20/23	15/19	12/14	10/09	22/22	18/18	36/32	NM	49/47
13	22/23	09/09	87/88	60/63	65/65	52/53	63/64	38/40	25/25	26/25	15/17	46/49	NM
14	30/27	23/24	74/73	46/46	50/48	32/37	50/49	35/30	08/16	08/08	-/16	41/41	10/12
15	38/24	12/11	85/88	61/61	53/56	54/57	51/55	28/34	16/17	17/17	09/11	37/38	08/09

Table 6.3: Morning peak-hour travel time vs. evening peak-hour travel time

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	19/19	35/36	92/88	87/89	86/92	83/78	86/92	82/90	72/75	70/72	61/65	80/88	37/33	54/55	53/52
02	34/37	14/15	91/92	89/94	83/91	84/93	84/88	69/75	47/47	52/53	30/34	74/82	25/27	46/49	27/28
03	88/92	93/95	31/33	46/48	64/71	54/61	64/71	84/90	87/94	81/84	91/93	80/85	84/85	81/85	88/88
04	81/84	87/89	43/44	14/14	39/42	67/71	78/80	71/78	77/84	62/66	83/87	62/66	75/79	63/70	75/78
05	88/93	87/91	65/67	42/41	13/14	31/29	50/50	54/55	67/72	55/54	79/84	34/34	80/83	74/77	72/70
06	82/89	88/92	57/58	67/70	29/32	26/27	36/36	51/53	65/69	60/62	81/85	40/39	78/79	73/77	80/78
07	88/92	81/86	66/68	76/76	44/53	34/36	22/23	39/42	56/60	55/56	65/71	36/38	80/85	75/79	72/70
08	83/88	69/76	81/83	72/73	53/55	51/50	38/40	13/13	23/24	27/26	44/45	18/18	65/72	57/56	51/51
09	68/73	45/50	84/90	75/79	62/71	64/69	52/58	21/24	10/11	17/18	20/21	32/38	40/45	37/37	25/25
10	65/73	47/55	72/75	59/62	48/55	59/61	51/56	24/26	17/18	10/10	33/35	27/29	35/39	21/24	22/22
11	61/64	29/31	88/93	81/84	73/87	74/83	66/72	41/47	21/23	35/35	15/14	54/66	37/37	44/44	24/21
12	83/89	74/84	78/78	63/61	30/34	39/38	36/34	18/18	34/38	29/28	58/62	9/9	69/73	55/58	55/54
13	31/38	25/26	80/85	76/80	73/86	72/84	75/84	63/72	42/45	38/39	37/38	65/74	11/11	21/23	17/15
14	52/54	47/51	80/79	64/63	73/77	75/77	77/81	56/56	40/39	25/22	46/42	58/59	24/23	12/12	22/20
15	43/55	23/29	83/87	74/76	62/75	72/81	65/74	42/53	21/26	21/23	18/23	48/57	14/16	18/21	8/8

Table 6.4: Travel time from CDR stream comparison between 09 AM and 07 PM hour

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
01	15/19	30/36	76/88	70/89	74/92	78/78	76/92	73/90	60/75	55/72	54/65	65/88	26/33	50/55	47/52
02	31/37	11/15	76/92	68/94	69/91	74/93	69/88	58/75	42/47	35/53	19/34	64/82	18/27	32/49	19/28
03	79/92	66/95	19/33	29/48	53/71	43/61	57/71	67/90	64/94	59/84	74/93	59/85	62/85	59/85	61/88
04	64/84	64/89	33/44	10/14	31/42	54/71	59/80	57/78	62/84	46/66	66/87	48/66	57/79	47/70	55/78
05	79/93	67/91	52/67	28/41	10/14	22/29	42/50	37/55	65/72	46/54	76/84	29/34	58/83	51/77	57/70
06	80/89	78/92	43/58	60/70	19/32	15/27	27/36	45/53	52/69	47/62	64/85	24/39	64/79	64/77	66/78
07	80/92	70/86	60/68	59/76	36/53	27/36	16/23	30/42	45/60	44/56	68/71	23/38	60/85	56/79	46/70
08	65/88	50/76	72/83	53/73	57/55	52/50	24/40	11/13	18/24	23/26	34/45	14/18	52/72	40/56	37/51
09	60/73	36/50	70/90	49/79	56/71	55/69	42/58	20/24	9/11	17/18	17/21	32/38	35/45	32/37	19/25
10	51/73	34/55	59/75	45/62	41/55	53/61	40/56	24/26	15/18	8/10	27/35	23/29	29/39	19/24	16/22
11	53/64	19/31	80/93	60/84	80/87	64/83	62/72	37/47	15/23	31/35	10/14	47/66	36/37	37/44	13/21
12	65/89	56/84	67/78	47/61	25/34	28/38	27/34	14/18	28/38	19/28	49/62	8/9	52/73	43/58	29/54
13	31/38	19/26	70/85	58/80	60/86	63/84	61/84	55/72	35/45	26/39	29/38	54/74	9/11	17/23	13/15
14	46/54	37/51	63/79	47/63	53/77	70/77	61/81	40/56	29/39	16/22	36/42	41/59	19/23	10/12	18/20
15	40/55	17/29	61/87	55/76	55/75	61/81	50/74	44/53	20/26	20/23	15/23	42/57	17/16	20/21	8/8

Table 6.5: Travel time from CDR stream comparison between 02 AM and 07 PM hour

Hour	Direct	Time	Alternative	Time
P-Morn	1 → 3	113.88	1 → 6 → 3	82.10
P-Morn	1 → 7	83.82	1 → 6 → 7	56.47
P-Morn	2 → 4	81.17	2 → 17 → 4	61.42
P-Morn	4 → 15	63.50	4 → 17 → 14 → 15	49.43
P-Eve	1 → 3	112.50	1 → 6 → 3	96.45
P-Eve	1 → 5	80.63	1 → 6 → 5	68.37
P-Eve	2 → 3	102.72	2 → 17 → 4 → 3	89.72
P-Eve	5 → 11	60.17	5 → 10 → 11	49.98

Table 6.6: Example of Measured Shortest Path based on travel time matrix(Peak as P, Morning as Morn and Evening as Eve)

Time	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
Off-P	12.6	12.0	13.4	14.4	12.0	12.7	13.9	12.7	13.0	12.8	12.6	12.9	12.8	12.6	13.4	14.3	13.0
P-Morn	12.2	11.5	12.6	12.8	11.5	12.4	13.8	12.1	11.8	11.7	12.1	12.9	12.1	11.7	11.8	12.0	11.7
P-Eve	11.9	11.2	12.2	12.7	10.8	11.9	13.6	11.6	11.4	11.5	11.6	12.5	12.0	11.5	11.6	12.4	11.6

Table 6.7: Intra Zone Average Travel Speed in km/h(Peak as P, Morning as Morn and Evening as Eve)

Chapter 7

Validation

To validate our results obtained from CDR data, we first collect real travel time data and then compare with our results.

7.1 Collecting travel time data

The travel time data collected while traveling among 17 key locations of Dhaka city's road network over a week has been used in this study. To measure real travel time scenario among these locations, GPS enabled smart-phones are used while traveling through these key locations using different types of vehicles such as cars and buses. These key locations are presented in figure 4.4. Travel time data are collected for both peak hours (from 9 AM to 2 PM and 6 PM to 10 PM) and off-peak hours (2 PM to 6 PM and 10 PM to 9 AM). These key locations and range of peak and off-peak hours are identified from our CDR data, which is described in our methodology section. We also collected travel data between some non-major locations. These travel data include both intra and inter-zone travel.

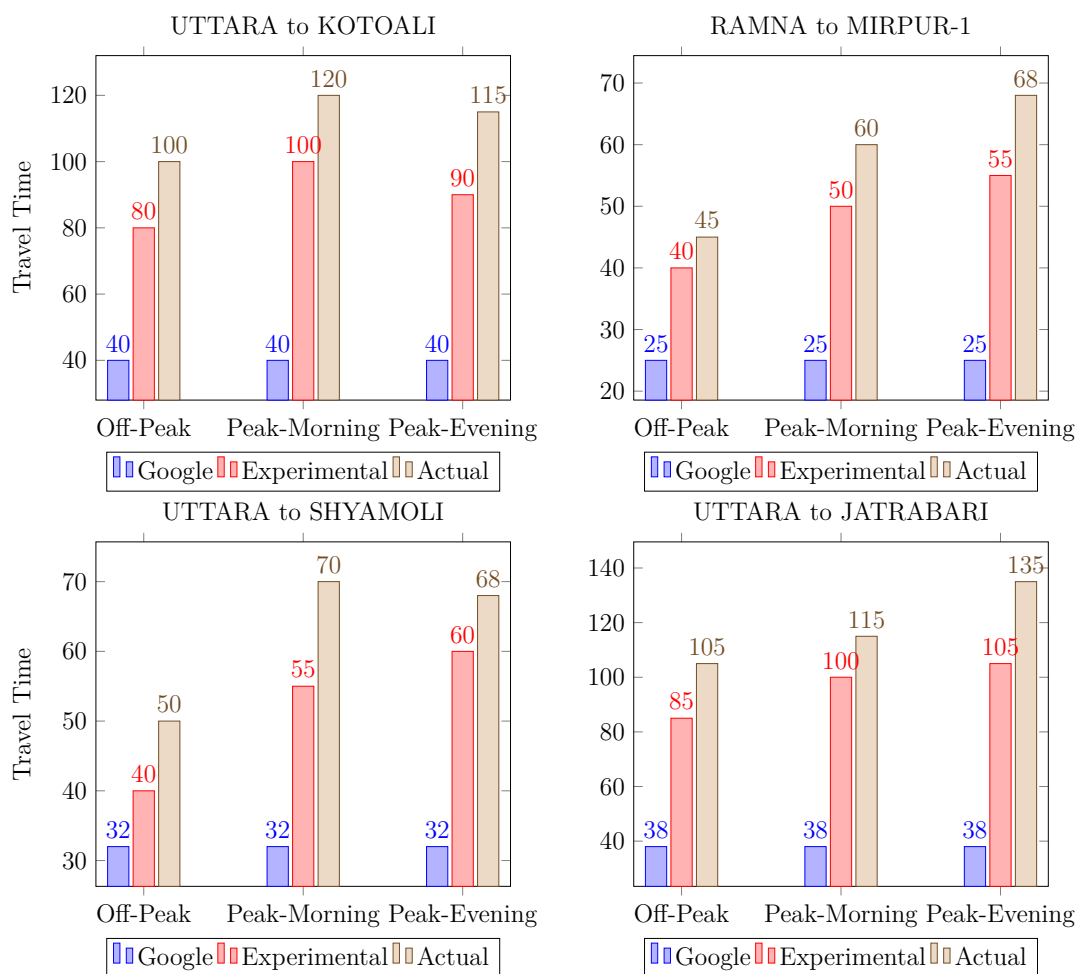


Figure 7.1: Result Comparison of travel time between key locations

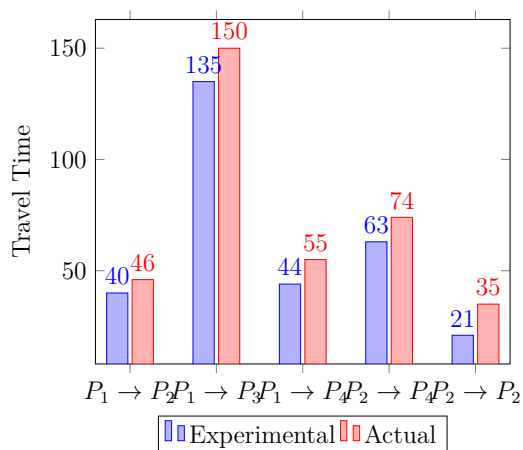


Figure 7.2: Result Comparison of travel time between any two minor locations

7.2 Estimated travel time vs. real travel time

We compare our estimated travel time with both real travel time and Google Map API's generated travel time. Comparison of our experimental results with real travel time data and Google Map's predicted travel time data is shown in Figure 7.1. In this figure, we compared our result in four routes, Uttara (index-03) to Ramna (index-15), Uttara (index-03) to Shyamoli (index-12), Uttara (index-03) to Jatrabari (index-01) and Ramna (index-15) to Mirpur-1 (index-06). We also compared our experimental results with real time data in five different routes between four minor location (arbitrarily taken from Uttara, Mirpur, Jatrabari zone and mark as P_1, P_2, P_3, P_4) shown in Figure 7.2. Experimental results show that our method has less deviation from actual travel time than that of Google Map API generated travel time. Figure 7.1 shows the comparison of actual travel time vs. estimated travel time vs. Google travel time at off-peak, peak-morning and peak-evening hours for a different pair of key locations. We observe that in all cases our estimated travel times are close to the actual travel time for both off-peak and peak hours. However, the Google generated travel time has a huge deviation from the actual travel time.

Result-Source	Off-Peak	Peak-Morn.	Peak-Eve.
Actual Time	40.31	44.67	45.58
Our Approach	35.35	38.55	39.15
Google Map	19.04	19.04	19.04

Table 7.1: The average travel time within 17 key locations

Table 7.1 show the average deviation of our estimated travel time and actual travel time in off-peak, peak-morning, and peak-evening are 12.30%, 13.71%, and 14.11%, respectively. Our estimated travel time between minor location pairs deviate by an average 24%. Note that these slight deviations may come from the recording of CDR data and actual travel time in two different years.

Figure 7.5, 7.3 and 7.4 is three different examples of comparison of travel time between

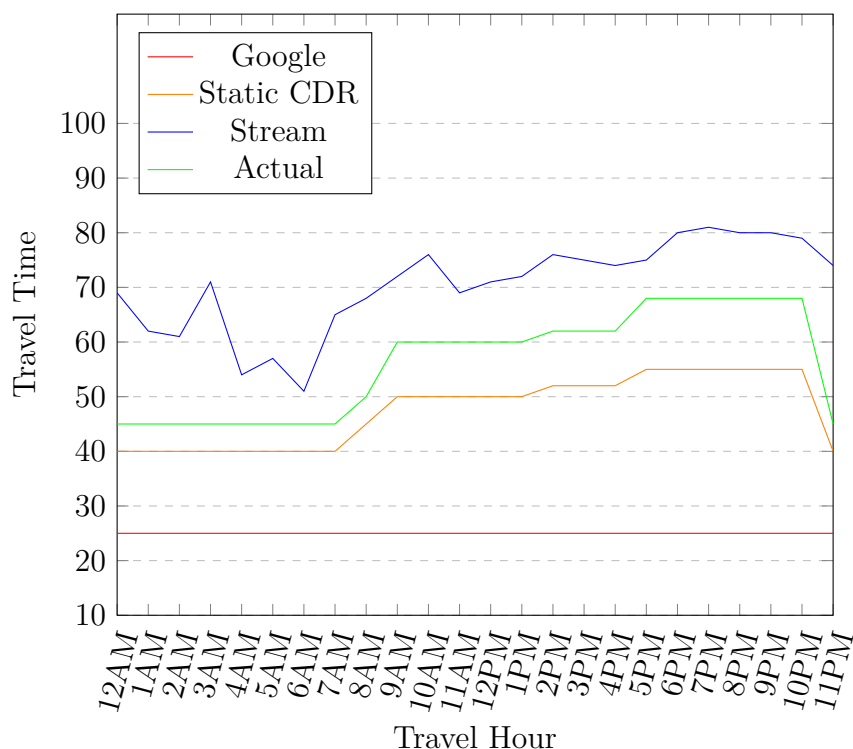


Figure 7.3: Result Comparison of Ramna to Mirpur-1

Google, Actual, Static CDR and Stream CDR. Since our real travel time data is not hourly based hence seems more dissimilar to stream CDR results but exact real travel time data may be similar to results produced from the stream of CDR.

7.3 Summary

In this chapter, we validate the result which is produced from static CDR and stream of CDR by comparing with each other and Google's and real travel time. Our validation indicates the reliable accuracy of our methodology.

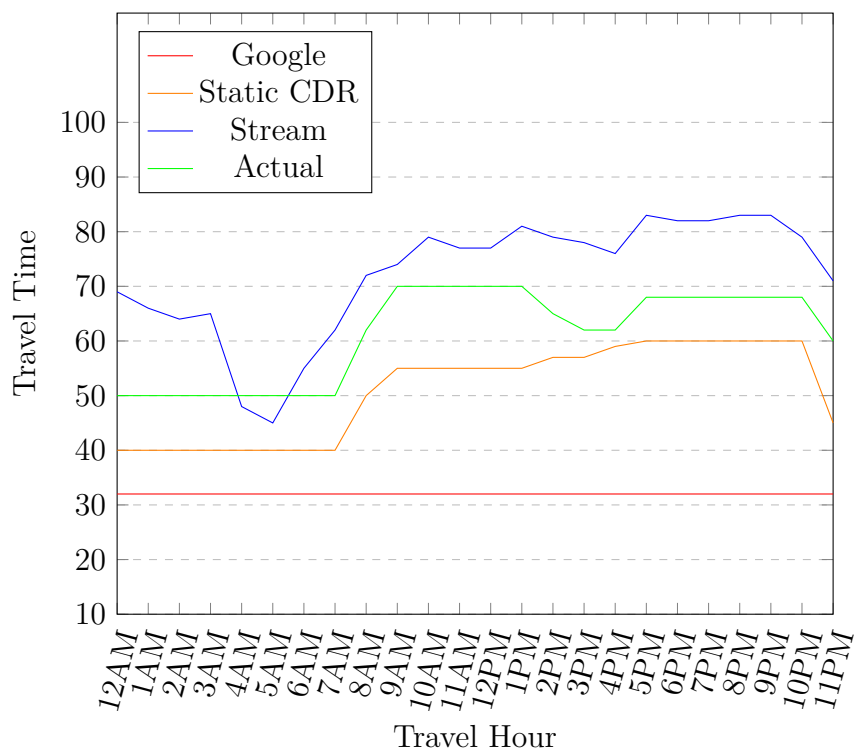


Figure 7.4: Result Comparison of Uttara to Shyamoli

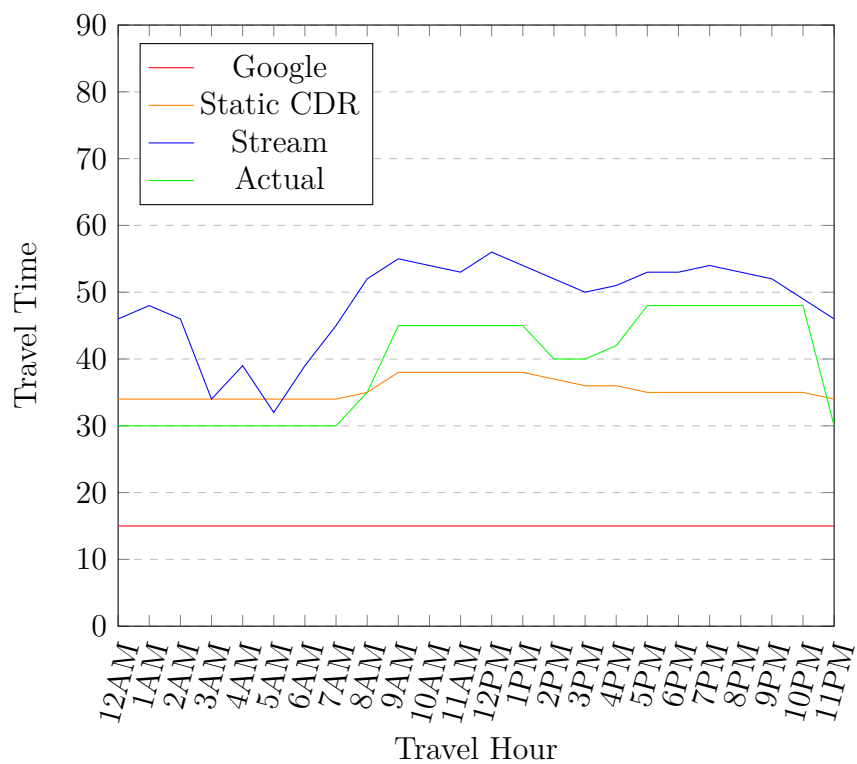


Figure 7.5: Result Comparison of Mirpur-10 to Kawran Bazar

Chapter 8

Conclusion

In this thesis, we have proposed an effective technique that can estimate travel time from mobile phone users call detail records (CDR) accurately. Our approach considers consecutive calls of a user from two different locations and maps this transition to the road network distance in the city to estimate the travel time. Moreover, we also exploit aggregated call detail records of a number of users traveling from a source to destination, to identify peak and off-peak hours, and estimate the travel time accordingly. Our experimental results show that our method can accurately capture the travel times between any pair of key locations and any pair of minor location in the city with a high accuracy. Our estimated travel time only deviates by an average of 13% from the actual travel time, whereas Google generates travel time that deviates by more than 50% from the actual travel time in most of the cases.

List of Publications

1. Md. Mahedi Hasan, Mohammed Eunus Ali, *Estimating Travel Time of Dhaka City from Mobile Phone Call Detail Records*, Proceedings of the Ninth International Conference on Information and Communication Technologies and Development (ICTD 2017), pp. 14, 2017.

References

- [1] Grameenphone ltd. bangladesh. <http://grameenphone.com>. [Online; accessed 20January, 2014].
- [2] Dhaka urban transport network development study, draft final report, 2010.
- [3] Michael Alger, E Wilson, T Gould, R Whittaker, and N Radulovic. Real-time traffic monitoring using mobile phone data. *Online: <http://www.maths-in-industry.org/miis/30> Vodafone Pilotentwicklung GmbH, 2004.*
- [4] Javed Aslam, Sejoon Lim, Xinghao Pan, and Daniela Rus. City-scale traffic estimation from a roving sensor network. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, pages 141–154. ACM, 2012.
- [5] Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, 2007.
- [6] Eric Bouillet, Bei Chen, Chris Cooper, Dominik Dahlem, and Olivier Verscheure. Fusing traffic sensor data for real-time road conditions. In *Proceedings of First International Workshop on Sensing and Big Data Mining*, pages 1–6. ACM, 2013.
- [7] N Caceres, JP Wideberg, and FG Benitez. Deriving origin destination data from a mobile phone network. *IET Intelligent Transport Systems*, 1(1):15–26, 2007.

- [8] Noelia Caceres, Luis M Romero, Francisco G Benitez, and Jose M del Castillo. Traffic flow estimation models using cellular phone data. *Intelligent Transportation Systems, IEEE Transactions on*, 13(3):1430–1441, 2012.
- [9] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.
- [10] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.
- [11] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.
- [12] Corrado De Fabritiis, Roberto Ragona, and Gaetano Valenti. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 197–203. IEEE, 2008.
- [13] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [14] Robert M Groves. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675, 2006.
- [15] Jaroslav J Hajek. Optimal sample size of roadside-interview origin-destination surveys. Technical report, 1977.
- [16] Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.

- [17] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.
- [18] Andreas Janecek, Karin A Hummel, Danilo Valerio, Fabio Ricciato, and Helmut Hlavacs. Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 361–370. ACM, 2012.
- [19] Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, and Helmut Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2551–2572, 2015.
- [20] BS Kerner, C Demir, RG Herrtwich, SL Klenov, H Rehborn, M Aleksic, and A Haug. Traffic state detection with floating car data in road networks. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, pages 44–49. IEEE, 2005.
- [21] Rainer Kujala, Talayeh Aledavood, and Jari Saramäki. Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science*, 5(1):1, 2016.
- [22] Masao Kuwahara and Edward C Sullivan. Estimating origin-destination matrices from roadside survey data. *Transportation Research Part B: Methodological*, 21(3):233–248, 1987.
- [23] Qing Ou, Robert L Bertini, JWC Van Lint, and Serge P Hoogendoorn. A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):747–756, 2011.
- [24] Cuong Pham and Nguyen Thi Thanh Thuy. Real-time traffic activity detection using mobile devices. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, page 64. ACM, 2016.

- [25] Oates R Poushter J. Cell phones in africa: communication lifeline. <http://www.pewglobal.org/files/2015/04/Pew-Research-Center-Africa-Cell-Phone-Report-FINAL-April-15-2015.pdf>, 2015. [Online; accessed 14December 2014].
- [26] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.
- [27] Fabio Ricciato. Traffic monitoring and analysis for the optimization of a 3g network. *IEEE Wireless Communications*, 13(6):42–49, 2006.
- [28] Geoff Rose. Mobile phones as traffic probes: Practices, prospects and issues. *Transport Reviews*, 26(3):275–291, 2006.
- [29] Geoff Rose. Mobile phones as traffic probes: practices, prospects and issues. *Transport Reviews*, 26(3):275–291, 2006.
- [30] Keemin Sohn and Daehyun Kim. Dynamic origin–destination flow estimation using cellular communication system. *IEEE Transactions on Vehicular Technology*, 57(5):2703–2713, 2008.
- [31] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [32] Ashwin Sridharan and Jean Bolot. Location patterns of mobile users: A large-scale study. In *INFOCOM, 2013 Proceedings IEEE*, pages 1007–1015. IEEE, 2013.
- [33] Tamás Tettamanti and István Varga. Mobile phone location area based traffic flow estimation in urban road traffic. *Columbia International Publishing, Advances in Civil and Environmental Engineering*, 1(1):1–15, 2014.

- [34] Wang Tiedong, Fang Tingjian, Han Jianghong, and Wu Jian. Traffic monitoring using floating car data in hefei. In *Intelligence Information Processing and Trusted Computing (IPTC), 2010 International Symposium on*, pages 122–124. IEEE, 2010.
- [35] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic, and Antonio Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 267–279. ACM, 2009.
- [36] Danilo Valerio, Alessandro D Alconzo, Fabio Ricciato, and Werner Wiedermann. Exploiting cellular networks for road traffic estimation: a survey and a research roadmap. In *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pages 1–5. IEEE, 2009.
- [37] Wim Vandenberghe, Erik Vanhauwaert, Sofie Verbrugge, Ingrid Moerman, and Piet Demeester. Feasibility of expanding traffic monitoring systems with floating car data technology. *IET Intelligent Transport Systems*, 6(4):347–354, 2012.
- [38] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.
- [39] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 316–324. ACM, 2011.
- [40] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-driven intelligent transportation systems: A survey. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4):1624–1639, 2011.

Index

k-mean algorithm, 23

t-travel, 17

Actual travel time, 22

APSP, 39

CDR, 7

CDR Stream, 31

Crowdsource, 13

FCD, 12

Google Maps, 13

GPS, 11, 49

Grameenphone Ltd., 4

idle device, 14

Inter *t*-travel, 23

Intra *t*-travel, 23

ITS, 10

Key location, 24

Network distance, 26

OBU, 11

OD Matrix, 14

Off-Peak, 19

Peak, 19

Recent change significance Factor, 36

Sliding window, 33

Static CDR, 17

Threshold, 36

Travel speed, 29

Travel time matrix, 28

Window range, 33