博士論文

Dynamic Census: Estimation of demographic structure and spatiotemporal distribution of dynamic living population by analyzing mobile phone call detail records

（ダイナミックセンサス：携帯電話データ分析による
動態人口の属性と時空間分布の推計）

新井　亜弓

Table of Contents

Chapter 1 Introduction

1.1 RAPID SPREAD OF MOBILE DEVICES GENERATING LARGE-SCALE SPATIOTEMPORAL DATA

The pace mobile phones spread globally is exceptional in the history of technology. Since 1978 where the first commercial cellular mobile services established, the subscription rate has increased exponentially. In 2002 the number of subscriptions exceeded one billion, passing the number of fixed-line users, and is reaching the number of the global population [1] In 2013 mobile subscription rate stands at 96% globally where 128% in developed countries and 89% in developing countries [2]. Though the number of subscriptions does not necessarily means the number of mobile phone users, it is evident that the significant part of the global population can be connected to a specific domain, which is the mobile network.

In general, the record of communications through the mobile phone is routinely accumulated for the billing purpose. The data are called call detail records (CDRs), which include time and antenna location of calls and short messaging service (SMS). Because the number of mobile users is huge, CDRs form the large-scale spatio-temporal database. With the sequential information of time and location of individuals, the data allow us to understand the dynamics of human mobility. Besides CDRs, there is the other type of spatio-temporal data deriving from the mobile phone; Global Positioning System logs (GPS). A feature differentiating GPS from CDRs is the recurrence rate of data update. Information obtained through GPS is automatically updated every one to five minutes while CDRs are updated only when the mobile phone is used. Figure 1 compares the feature of CDRs and GPS. Columns (A) and (B) show differences in spatio-temporal resolution between CDRs and GPS, which lead to the resolution of trajectories reconstructed from these data. As described in column (C), CDRs can be collected both through the feature phone and smartphone. On the

other hand, GPS can be collected only through the smartphone because the feature phone is not equipped with GPS. It indicates that the choice of data for understanding human mobility may affect the population under study.

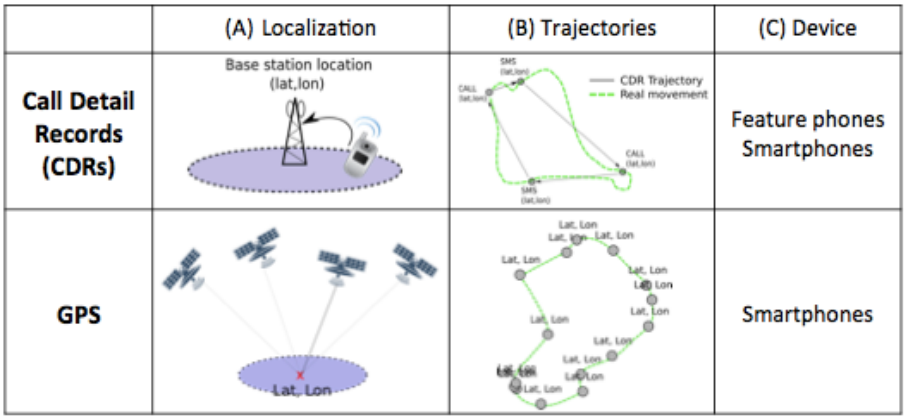


Figure 1. Data localization and trajectories of CDRs and GPS

1.2 UNDERSTANDING HUMAN MOBILITY DYNAMICS AND ITS PRACTICAL APPLICATIONS

In recent years, increasing availability of spatio-temporal data such as CDRs and GPS has advanced the understanding on features and statistical patterns of human mobility. Studies on human mobility often focus on modeling the properties of human mobility patterns. Reference [3] revealed that people routinely visit specific locations by modeling human mobility patterns considering the probability of visiting new places and returning to a previously visited location. Reference [4] found that these repeatedly visited locations are typically home and work places. Reference [5] proposed an algorithm to estimate the home and work location of mobile phone users by examining the time and location distribution of call records. Understanding the human mobility is challenging because of the importance of incorporating the human mobility dynamics into the practical applications to various sectors, e.g. urban

planning, transportation management, public health, etc. Reference [6] developed models to measure the environmental impact of home-to-work commutes by analyzing human mobility patterns through CDRs. They take into account of differences in mobility patterns according to differences in the geographical distribution of home and work places, transportation infrastructure, etc. Reference [7] examined the patterns of interurban movement analyzing volume and distance of human mobility obtained through subway cards. The study provided an approach to modeling flows of urban systems by showing the evolution of polycentric configuration of a city and the dense structure of urban centers. Reference [8] discussed the potential for using CDRs to provide insights to malaria control policy by quantitatively analyzing human mobility. In sum, we can see that the understanding on the human mobility dynamics has been advanced mostly in quantitative manners.

1.3 ISSUES TO BE CONSIDERED TO UTILIZE MOBILE DATA FOR SOCIETAL ISSUES

While the analysis of CDRs and GPS seem to be a prominent approach to understand the human mobility dynamics, it has been increasingly pointed out that there are some concerns for utilizing such data for societal issues. The concerns are primarily summarized into three as followings.

- Sparseness

CDRs and GPS provide a partial view of human mobility, which differs from a full picture of human mobility. The more frequent the data recurrence rate, the more similar the partial view and the full picture. Because CDRs are updated only when the mobile phone is used, the number of records per person per day is limited, e.g. Average number of call records per day for ordinal mobile users in Dhaka is around

3.5. It means we have less than four sets of time and location information on the whereabouts of the mobile user of a day on average.

- Representativeness

CDRs and GPS are the log data from the mobile phone. In other words the data do not include anyone who does not use the mobile phone. It means that the population of the data is only mobile users. Thus, there can be some discrepancy in the trend of human mobility dynamics between actual living populations, which consist of mobile users and non-mobile users, and mobile users, if the characteristics of the two population groups are different.

- Anonymity

Mobile phone log data are generally anonymized to protect the privacy of mobile users. Instead of the attribute information of mobile users such as name, address, age, gender, etc., a random code is assigned to each person, which still allows us to trace the mobility of each person for a given period. Because of this feature, it is difficult to know who are in the data except the fact that they are mobile users.

When we interpret the analysis result of mobile phone data to address societal issues, we need to be careful what is the population under the study. Suppose we use mobile data to analyze how the infectious disease spreads in association with the human mobility. For instance, if the disease is Malaria, we take into account of that small children are particularly vulnerable to Malaria. In addition, we need to note whether the mobile phone data represent the whereabouts of small children. As described previously, an increasing body of study proposes to utilize mobile data to transportation study and urban planning. In fact, mobile phone data can provide the quantitative aspect of human mobility dynamics such as the speed and volume of human mobility. So, the analysis result of the data can be significant inputs. On the other hand, the data have nothing to do with the qualitative aspect of its population. Therefore, the interpretation of analysis results may require additional investigation

when the scope of issues to be addressed matters the quantitative aspect of the population under study.

1.4 STUDY SITE

The study site of this research is Dhaka, Capital of Bangladesh. Bangladesh is composed of six Divisions. Dhaka is part Dhaka Division that consists of South Dhaka city and North Dhaka city. Location of Dhaka Division and Dhaka are shown in Figure 2.

Source:
Left: https://en.m.wikipedia.org/wiki/File:Bangladesh_regions_map.svg (Licensed under the Creative Commons Attribution-Share Alike 4.0 International license and free to share)
Right https://commons.wikimedia.org/wiki/File:Dhaka_Division_districts_map.png (Licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license and free to share)

Figure 2. Areas covered by CDRs

We use CDRs from one of the leading mobile network operators (MNOs) in Bangladesh. Figure 3 shows the area covered by the CDRs. The area is composed of three administrative areas. One is the city of Dhaka, consisting of Dhaka North city and Dhaka South city. These cities are collectively called as Dhaka City Corporation till 2011. Another is part of Dhaka Metropolitan areas outside the city of Dhaka. The other is Municipalities, which are the suburban areas surrounding Dhaka Metropolitan areas. Detailed descriptions of the area are summarized in Table 1.

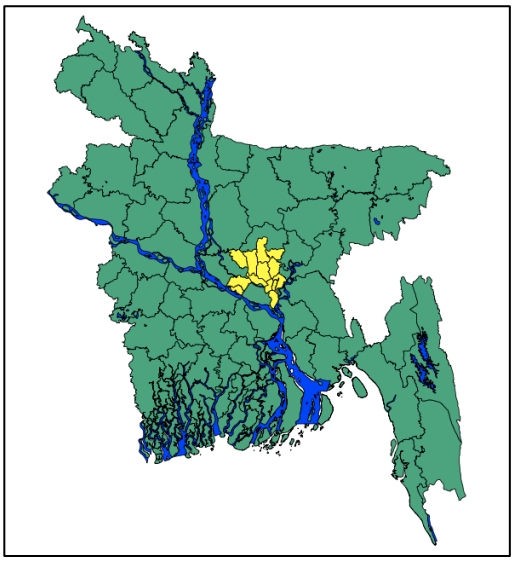


Figure 3. The area covered by CDRs (created by author)

Table 1. Names of Thanas (name of administrative area unit) covered by CDRs

Area type	Name of administrative units covered by each category
Dhaka North city ***Dhaka South city*** ***(Dhaka City Cooperation)***	36 wards 56 wards
Outside of DCC ***(part of Dhaka Metropolitan Area)***	Uttar Khan/Uttara, Dakshinkhan/Uttara, Bashundhora, R/A/Badda, Adarsha Nagar/ Badda, Banashree, R/A/Khilgaon, Simrail and Khorda, Goshpara, Fatulla, Sultanganj/Kamrangirchar, Hasnabad/Keraniganj, Zinjira/Keraniganj, Washpur, Aminbazar/Savar, Birulia/Savar, Ashulia/Savar
Municipality	Narayanganj, Savar, Tongi

Economic background

Bangladesh is one of the poorest countries in the world whose GDP capita is 960 USD. The speed of its economic growth is rapid at an annual rate of 6% GDP since 2011. However, the benefit is not equally distributed in the society and almost quarter of income is held by the 10% of the wealthiest, and thereby huge income disparity still exists [9]. Additionally, the demand for labor force, associated with the economic growth, dose not generate sufficient job opportunities to accommodate the rapid population growth [10]. As shown in Figure 4, Bangladesh's population pyramid is wide at base where median age is 23.5 [11]. It means more than half of the

populations are younger than 24. It also indicates that the economically active population[1] is potentially large.

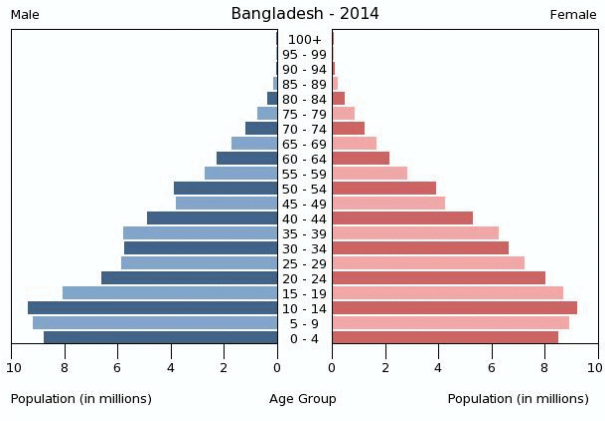


Figure 4. Demographic structure of Bangladesh in 2014 [11]

Urbanization

Despite such insufficient labor demand, economic growth accelerates the speed of rural-to-urban migration in Bangladesh, where people from the rural area have prospects for more income opportunities in the urban area. In fact, urban population in Bangladesh has been increasing annually by 6% since independence [12]. Dhaka is one of largest megacities[2] in the world with the population of 18 million [10]. Table 2 shows the growth of urban population in Bangladesh. It includes six major Divisions in Bangladesh (Dhaka, Chittagong, Sylhet, Khulna, Barisal, and Rajshahi) where the population of Dhaka holds almost 40% of the national urban populations. It implies that the the speed of urban population growth is accelerated particularly in Dhaka. Like other megacities in Asian countries such as Manila and Jakarta, Dhaka has been experiencing rapid urbanization with increasing population inflow from the rural area. Approximately 9% of the Bangladesh population lives in the Dhaka metropolitan area, contributing to 36% of the country's GDP [13].

[1] Population between the ages between15 to 64 is considered to be people who could potentially be economically active [9].

[2] Megacity is defined as the city populated with more than 10 million people.

Table 2. Urbanization in major cities in Bangladesh

Census year	1961				1974			
	Total pop.	Urban pop.	Level of urbanization	Share of urban pop.	Total pop.	Urban pop.	Level of urbanization	Share of urban pop.
Dhaka	15,293	1,073	7.02	40.63	21,316	2,900	13.60	46.23
Chittagong	10,140	569	5.61	21.54	13,876	1,273	9.17	20.29
Sylhet	3,490	71	2.03	2.69	4,759	131	2.75	2.09
Khulna	5,805	311	5.36	11.78	8,768	858	9.79	13.68
Barisal	4,261	119	2.79	4.51	5,427	191	3.52	3.04
Rajshahi	11,850	498	4.20	18.86	17,332	920	5.31	14.67
Bangladesh	50,839	2,641	5.19	100.00	71,478	6,273	8.78	100.00

Census year	1981				1991			
	Total pop.	Urban pop.	Level of urbanization	Share of urban pop.	Total pop.	Urban pop.	Level of urbanization	Share of urban pop.
Dhaka	26,242	5,383	20.51	39.77	33,940	9,620	28.34	43.20
Chittagong	16,940	2,994	17.67	22.12	21,865	4,757	21.76	21.36
Sylhet	5,656	493	8.72	3.64	7,149	755	10.56	3.39
Khulna	10,641	1,737	16.33	12.84	13,244	2,515	18.99	11.29
Barisal	6,510	730	11.22	5.39	7,758	1,001	12.90	4.50
Rajshahi	21,132	2,198	10.40	16.24	27,493	3,799	13.82	17.06
Bangladesh	87,120	13,536	15.54	100.00	111,449	22,447	20.14	100.00

Census year	2001			
	Total pop.	Urban pop.	Level of urbanization	Share of urban pop.
Dhaka	38,987	13,386	34.33	46.80
Chittagong	21,865	5,724	23.73	20.01
Sylhet	7,897	976	12.36	3.41
Khulna	14,605	2,921	20.00	10.21
Barisal	8,154	1,160	14.23	4.06
Rajshahi	30,089	4,438	14.75	15.51
Bangladesh	123,851	28,605	23.10	100.00

Source: [12]. Original source is modified by author.

Although one of major factors attracting the rural population is increasing income opportunities in the urban area, decent jobs are unlikely available to those who are poorly educated or unskilled. So, most of them are engaged in temporal jobs such as construction workers or street vendors in the informal sector. Insufficient income limits their location choice of housing in Dhaka, which is affordable and commutable. Because land prices in residential areas in Dhaka are very high, those without sufficient income inevitably dwell in slum areas whose living environment is extremely poor. Slums are generally formed in riverside and proximity of garbage dumpsites, which mostly lack basic infrastructure such as the supply of clean water, electricity, and reliable sanitation. Most slum residences are temporal structure and

highly dense[3], which leads to slum areas vulnerable to disaster and potentially high risk of epidemics. Despite such bad living conditions, the population in slums is even increasing. Total population increased from 1.5 to 3.4 million and the number of slum communities increased from 3,007 to 4,966 between 1996 and 2005 [14]. Due to the significance of population size and its potential impacts on the entire population in the city, government has been aware of the need of addressing problems of slums. However, successful intervention has been seldom realized yet because no one has a clear picture of slums, i.e. how many people are residing in slums, how they live, where the boundary of slums is, where slums exist, etc. The reason of such ignorance is that populations in slums are generally floating and part of them does not appear in the official statistics of their living areas. This is because they are not registered to the system of their actual living places. Some of them leave their rest of family members and assets in their hometowns if the living in the city is temporal stay for them. Some may not be able to register their living address officially because they illegally occupy public lands. Such nature of slum dwellers adds complications of slums.

One of conventional approaches to resolve such issues, which derive from the urbanization, would be to conduct slum census and mapping. However, it does not enable us to capture the floating nature of slum populations, which potentially impacts on the entire population of the city in case of epidemics and the design of transportation planning. In this context, understanding population dynamics through CDRs is considered to be effective because CDRs can capture the mobility of people regardless of any status as long as the person uses a mobile phone. Considering the size of the portion of slum populations in Dhaka, having a clear picture of human

[3] Population density of slums in Dhaka is 205,020 persons per sq. km whereas that of Bangladesh, which has the highest population density in the world, is 1,004 persons per sq. km. Population density of slums is approximately 200 times greater. Considering that the slum is dominated by single story housings, population density in actual living environment is further [14]. Population density of Japan is almost one-third of Bangladesh.

mobility dynamics through CDRs can benefit policy intervention in various sectors such as public health, transportation, urban planning as well as urbanization.

Mobile phone market and mobile use in Bangladesh

Bangladesh is no exception to other developing countries where access to the mobile infrastructure is much better compared to other basic infrastructures. It is not uncommon that people can use the mobile phone under circumstances where access to clean water, electricity, market, and social services is quite limited [1]. It is most likely because establishing the mobile network is realized just by building an antenna, which let any people to be connected to the mobile network as long as they have a mobile phone. In fact, the mobile phone market in Bangladesh is very vibrant where as many as 124 million active subscribers are using mobile phone service from different operators. Mobile subscription rate is reported to be rapidly increasing, i.e. 55%(2011), 63%(2012), and 74%(2013), by [13]. The real mobile penetration is considered to be less than 44% [15] considering the fact that the number of subscription does not mean the number of mobile users. Table 3 shows the market share of telecommunications operators in Bangladesh. The market is dominated by the three majors; Grameen Phone, Banglalink, and Robi, followed by three other companies. Among the six companies, Grameen Phone and Banglalink are known for the countrywide network coverage where the Robi has limitations in some areas. As the general perception, the mobile phone tariff of Grameen Phone is considered to be relatively expensive and the quality of network is better compared to others.

Table 3. Mobile phone operators' market share in Bangladesh

Operators	Number of active subscribers (million)	Market share
Grameen Phone Ltd. (GP)	52.0	42 %
Banglalink Digital Communications Limited	31.9	26 %
Robi Axiata Limited (Robi)	26.3	21 %
Airtel Bangladesh Limited (Airtel)	8.2	7 %
Pacific Bangladesh Telecom Limited (Citycell)	1.2	1 %
Teletalk Bangladesh Ltd. (Teletalk)	4.0	3 %
Total	123.7	100 %

Source: [16]

In Bangladesh the smartphone is not very common and many people are still using the feature phone. Smartphone penetration rate is reported to be 5%, which holds almost half of the internet penetration rate in Bangladesh [15]. The cost of calling is generally considered be cheaper than texting because free calls are often offered as the bonus of new subscription or part of promotion by the telecommunications company. So, the mobile user tends to prefer calling even though they have a smartphone. It indicates that calling and texting are the common mean to communicate while communications through social networking services using smartphone applications quite are common in the developed country. In this context, analyzing CDRs, which do not include texting and internet communication logs, still allows us to understand the human mobility dynamics of the majority of the mobile users in this study for this moment.

1.5 PROBLEM SETTING AND THE PURPOSE OF THIS STUDY

As discussed, mobile phone log data potentially contain three constraints for utilizing the data to address societal issues; sparseness, anonymity, and representativeness. This thesis aims to overcome these disadvantages by proposing a novel approach to develop a new dataset, Dynamic Census. Dynamic Census is developed by utilizing CDRs of Dhaka from one of the leading telecommunications companies in Bangladesh. Primary features of Dynamic Census are described below.

- Interpolated spatio-temporally

A trajectory directly generated from CDRs for a person is a polygonal chain of a sequential connection of lines, each of which is determine by a pair of time and location of consecutive two call records. The pair of points is potentially neither the time nor location of departure/arrival for a person. One of the primary features of Dynamic Census is that it is interpolated based on the existing road network and the

estimated timing of location changes of the mobile users. It enables us to obtain better spatiotemporal population distribution.

- Labeled with demographic attribute

CDRs area anonymized to protect the privacy of the mobile user. Dynamic Census is labeled with the demographic attribute information, which is estimated through this study. It enables us to specify the mobility of a specific population group under study. For instance, Dynamic Census can facilitate the intervention under a disastrous event by providing the distribution of people, who are vulnerable to disaster such as small children and the elderly.

- Represent actual living population in a given area

The demographic structure of the population in CDRs, which is only the mobile user, is potentially different from the demographic structure of the living population for a given area. It means that there may be a discrepancy in the demographic structure between the two population groups. The demographic structure of Dynamic Census is adjusted to that of the living population. Because of this feature, Dynamic Census enables us to discuss the human mobility dynamics not only of the mobile user but also the living population, presenting in specified areas.

As part of outcomes of this study, we provide the movie of trajectories, which represent the living population and labeled with personal attributes. By using the trajectories, hourly population distributions in 500 sq. grids are provided. From each grid, statistical information on personal-attribute structure and household numbers by income level can be extracted.

Additionally, we provide approaches and analysis results to develop Dynamic Census, and a framework to develop Dynamic Census in other cities. Because the basic data structure of CDRs should be very similar regardless of the telecommunications operators, our approach is applicable to other cities. Also, CDRs exist in any cities wherever the mobile network is available. Considering the mobile

network coverage is reaching almost 100% [1], the impact of this study is significant and global.

As practical applications where Dynamic Census has strong advantages, we suggest following examples:

- Enhance resilience to disasters

Dynamic Census can enhance the resilience to disaster by providing the spatiotemporal population distribution of the living population. For example, Dynamic Census enables us to obtain population estimates at given time and location according to the time of disaster occurrence. It facilitates efficient aid allocation. It is ideal if the spatiotemporal distribution obtained from Dynamic Census is based on the real time CDRs. However, developing such a system and obtaining real-time CDRs for processing seem to be difficult. We suggest that seasonal or monthly estimates from Dynamic Census will be enough as baseline information for local governments. The government can utilize such information for maintaining and strengthening disaster preparedness. Considering that people's behaviors follow some routines to some extent, daily differences can be negligible except for anomaly large-scale events.

- Reduce epidemic risks

Dynamic Census can contribute to the risk reduction in epidemics particularly in urban areas by capturing the dynamic population movement of slum dwellers. Living conditions in slums are generally poor where the risk of the outbreak of disasters is potentially high. At the same time, it is difficult for the government to have enough knowledge on the dynamic population movement of such slum dwellers. Furthermore, they tend not to be included in official statistics and it is highly likely that basic information such as the number of population lacks. Dynamic Census enables the government, practitioners, and researchers to estimate how an infectious disease can

spread according to the location of its outbreak by specifying the movement of high-risk carrier and vulnerable population groups to the disease from Dynamic Census.

Chapter 2 Structure of this thesis

2.1 Structure of this thesis and the process of developing Dynamic Census

The structure of the remaining part of this thesis is as follows. Chapters 3 to 5 address the three bottlenecks of CDRs raised in Section 1: Chapter 3 introduces our approach to overcome the sparseness of CDRs through interpolation. We first investigate the discrepancy of the population presented in CDRs and the living population by specifying the principal population. Principal population groups are determined based on the temporal transition between typical statuses. Then, we interpolate the location of the user for the time band, which does not have call records, by sampling the missing part from a Markov network given the sparse observation and routine distribution. Chapter 4 focuses on the anonymity of CDRs. It explores the relationship between the features of calling behavior and the demographic attributes of mobile users so as to find clues to understanding mobile user attributes of anonymized CDRs through calling behavior. Estimation results of personal attributes are provided. Chapter 5 describes our approach for addressing the population bias in CDRs by estimating the presence of non-mobile users in the living population. First, we propose an approach to estimate the presence of the unobservable population in the households of mobile phone users. Then we demonstrate our approach for estimating the number of the living population based on the distribution of buildings. Chapter 6 provides a process of developing Dynamic Census using analysis results of Chapters 3 to 5, and presents the outcome. Chapter 7 provides the framework to recreate Dynamic Census in other cities. It focuses on key information to be collected from the secondary data to develop Dynamic Census. The significance of key features is explained statistically. A guideline to design a survey for the secondary data collection is provided as well. Chapter 8 includes conclusions

and future prospects. Contribution of this study and what has not been accomplished are described. The flow of entire processes to develop Dynamic Census is summarized in Appendix.

2.2 USING THE HOUSEHOLD AS ONE OF ESTIMATION UNITS

CDRs are the collection of individual call records. Therefore, the smallest unit used for CDR studies is generally an individual and the size of its population is often discussed as the number of populations. In this study, we introduce the household as one of the smallest units for estimating the number of populations. We consider that the use of the household as the unit for estimation is important because the household is one of the smallest decision-making units in the society. Additionally, the unit is widely used for other fields of studies such as public health, economics, and sociology. It means that the outcome of this thesis can be combined with the findings of other fields' studies. In this thesis, we define the household as a person or a group of persons, who 1) occupy a part of or an entire building, 2) usually live together, pool their income, and 3) eat from the same cooking pot. These rules cannot be applied to the household head. For example, the household head can form a household even if s/he is alone and the head can be a member of her/his household even if s/he lives apart from other household members. Regarding the household member, we define as a person, who 1) usually lives in the household regardless of being at home during the survey or temporarily absent, 2) usually born in the household, 3) do not have independent decision-making unit other than this household. We also consider the below people are as household members: 1) A person designated as the household head. 2) Someone who has joined the household through marriage. 3) Servants, lodgers, and laborers currently in the household and will be staying in the household

at least one month. 4) Any children of household members who are staying outside the house and studying in school.

2.3 THE POPULATION DISCUSSED IN EACH CHAPTER

We introduce two types of households for the estimation for this thesis by classifying the living population into two groups as illustrated in Figure 1: *Household A*, which includes mobile users subscribed to *the operator* providing CDRs for this study; and *Household B*, which does not include mobile users of *the operator*. Those who belong to *Household A* consist of the mobile users of *the operator*, and their household members, who do not use the mobile phone of *the operator*. None of those who belong to *Household B* use the mobile phone of *the operator*.

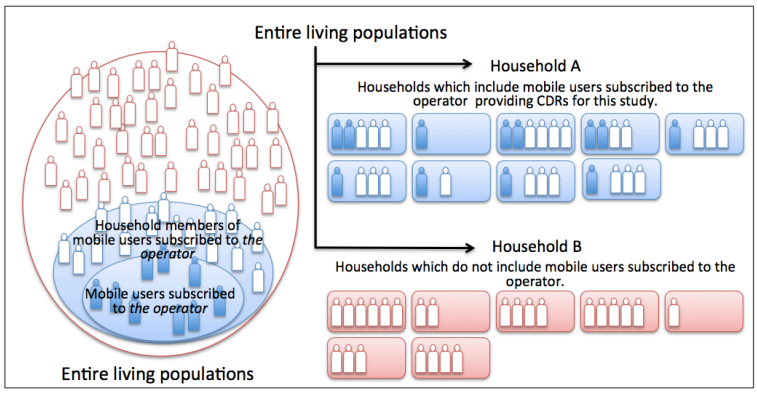


Figure 1. Populations and households discussed in this thesis

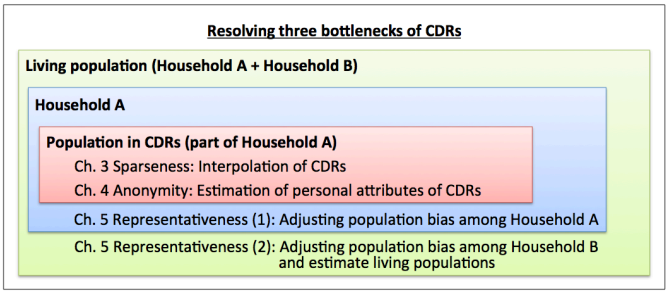


Figure 2. Structure of this thesis

Figure 2 describes three different parts of the living population discussed through Chapters 3, 4, and 5. Chapters 3 and 4 analyze the population in CDRs alone. The red part in Figure 2 indicates the population, of which sparseness and anonymity are resolved. Chapter 5 deals with two different parts of the living population to discuss the representativeness of CDRs and to address the population bias of CDRs. First, we focus on those who belong to *Household A* to adjust the population bias among the household, which includes the mobile users of *the operator*. It is the blue and red part of Figure 2. Then, we discuss those who belong to *Household B*, which is the green part of Figure 2 and finally estimate the living population in the area covered by CDRs.

Chapter 3 Sparseness: Interpolation of CDRs

3.1 BACKGROUND

The large amount of spatio-temporal data collected from pervasive devices has advanced the understanding of human mobility behavior. This understanding enables policy makers and governments to incorporate dynamic aspects of human mobility into public policies and city planning. Based on the assumption that such devices are fairly widespread, the spatio-temporal data can be considered to represent the general population. However, this may not be true, particularly in developing countries, because device ownership biases must be taken into account [1]. Further, Reference [1] suggested that the utilization of such data for societal issues requires knowledge of the types of population that are represented by the data.

The properties of human mobility are represented in a better manner by incorporating periodic modulation of human mobility. References [2,3] contribute to characterizing human mobility by quantifying the interaction between the regularity and randomness in human mobility dynamics. By mining semantically meaningful locations such as home and workplace in anonymized call detail records (CDRs), Reference [4] determine that a few limited locations where people spend most of their time are the means to understanding human mobility and social patterns. A part of human mobility is explained well by taking into account the activity patterns associated with travel because urban travel is considered to be driven by the demand to participate in activities according to Reference [5]. It revealed that travel behavior patterns are explained well by taking account of considering the activity in combination with the socio-demographics. It also proved that socio-demographic characteristics strongly affect the time allocation for activities inside the home and those outside the home. In this context, the extraction of typical behavior patterns

using a few key locations can help in understanding major components of mass populations.

An increasing number of studies investigate the behavior patterns of people by analyzing large-scale spatio-temporal data. Reference [6] describes the hidden structure in human behaviors by analyzing the data collected from 95 academic mobile phone users. The study presents the characteristic behavior of students and staff by extracting the principal components, termed as eigenbehaviors. It analyzes individual eighenbehavior and also describes the community affiliations of populations. While [6] presents partial behavior traits as the principal components, [7] characterize behavior patterns as regular temporal transitions between typical states such as home and workplace, and the Latent Dirichlet Allocation (LDA) topic model is employed to extract location-driven routines. [8] extended the topic model approach to evaluate the similarities and differences in behavior among multiple users by clustering the underlying structure of individual behavior patterns.

In this chapter, we examine the discrepancy between the principal population components of CDRs and those of the living population in Dhaka by comparing typical behavior patterns. First, we profile the typical behavior patterns of the principal population components for mobile users by examining two sets of typical behavior patterns, which are extracted from CDRs and the diary survey data of mobile users. CDRs are inherently sparse; hence, we interpolate CDRs based on the estimated behavior patterns of topics of individual users by employing the LDA topic model. This allows us to transform CDRs, which originally only include information related to time bands with call records, into behavior patterns of mobile users in a continuous manner. Then, we compare the obtained results with the typical behavior patterns of the core components of the general population in Dhaka. Thus, we observe the manner in which principal populations in CDRs differ from those in the living population with regard to behavior patterns.

The contributions of this chapter are as follows:

- The principal population components of mobile users are profiled by comparing behavior patterns extracted from CDRs and those obtained from field survey data. We interpolate sparse CDRs based on the predicted routines and interpret them as sequential activities.
- A novel approach to identifying device domain-specific bias for large-scale spatio-temporal data is proposed. The potential for extending our approach to other areas by using other data is discussed.

The remainder of this chapter is structured as follows: The data used in this chapter is described in Section 3.2. In Section 3.3, we examine the typical behavior patterns of mobile users, which are extracted from CDRs. In Section 3.4, the characteristics of typical mobile users are described and their typical behavior patterns are presented by analyzing the field survey data of mobile users. The diary survey data of the general population is analyzed in Section 3.5. Our conclusions are presented in Section 3.6.

3.2 DATA

In this chapter, wee use CDRs and two sets of field survey data to compare the principal population groups in CDRs and the living population. Sample size, data collection methodologies, and the population represented by the data, are described in this section.

3.2.1 Call Detail Records

We use the Call Detail Records (CDRs) of August 2013 from one of the leading telecommunication companies in Bangladesh (hereinafter referred to as “the company”). The data include the time, antenna location, and duration of calls. We

randomly sampled the call records of 5000 unique IDs. The data comprises the records of all antennas that are located in Greater Dhaka in Bangladesh.

3.2.2 Diary survey data of mobile users

In order to understand the personal attributes and activity of mobile users who use the service of the company, we conducted a diary survey of these users as part of a field survey—the Survey on Patterns of Activity for Comprehensive Explorations of Mobile Phone Users in Dhaka (SPACE), conducted in 2013 and 2014. (Detailed explanations of SPACE are provided in Chapter 4.) Our survey site was Greater Dhaka, which is composed of Dhaka City Cooperation (DCC), surrounding municipal areas (Savar and Karaniganj Upazila), and regions outside of the urban area (OUA, including portions of the Narayanganjt, Gazipur, and Narsingdi Districts). SPACE was conducted from November 27, 2013 through January 4, 2014. We collected information about the schedule of a day in addition to the demographic attributes and main activity for 922 mobile users. The SPACE data do not represent mobile users using the service of the company; these data represent the mobile users corresponding to each income group.

3.2.3 Diary survey data of the living population

We use another set of diary survey data, Person Trip (PT) data, to understand the personal attributes and activity of the general population in Greater Dhaka. The data, surveying the timing, origin-destination, and purpose of trips for a day, in addition to the means of transportation and demographic attributes, were collected by the Japan International Cooperation Agency (JICA) by interviewing 75000 people residing in Greater Dhaka in 2009. This survey was household-based and the sampling methodology was chosen such that it would obtain results that were representative of the general population.

3.3 TYPICAL BEHAVIOR PATTERNS THROUGH THE INTERPOLATION OF CDRS

In this section, we utilize the topic model to discover the latent topics within the CDR dataset. We use Author Topic Model (ATM), which is an extend version of the standard Latent Dirichlet Allocation model (LDA) [9] to comprehensively consider the distribution with respect to day, time, and geographical location.

3.3.1 Methodology

Routine topic model

The users' daily activities can be approximately summarized by a few numbers of basic routines. These routines can be sparsely observed by CDRs and are severely biased by calling preference. A topic model such as the most widely known Author Topic Model (ATM), which is an extension of the Latent Dirichlet Allocation (LDA), has proven to be an effective way to uncover the latent co-occurrence pattern. Inspired by the topic model, we propose our Gibbs-sampling based Routine Topic Model to infer the prototypes of the routines.

Analogous to ATM, the CDRs of each cell phone user can be regarded as a document, while the log of each single day is a word within the document. The basic terminology, similar to ATM, can be defined as follows:

1) User topic distribution: $\theta_u \sim Dirichlet(\alpha)$
2) Routine distribution: $\varphi^k \sim Routine\_Dirichlet(\beta)$
3) Topic distribution over day: $\mu_d \sim Dirichlet(\gamma)$
4) Latent topic assignment: $z_{d,u} \sim Multinomial(\theta_u)$
5) Observations: $X_{u,d} \sim Routine\_Multinomial(\varphi^{z_{d,u}})$

The one-day activity of a user is sampled from the routine k and is defined as

$$\varphi^k = \left[\pi_0^k, \pi_1^k, \cdots, \pi_{23}^k\right]$$

where $\pi_i^k, i = 0,1,\cdots,23$ is the probability distribution of the predefined activities (home, work, and other) at the *i*-th hour of the day. Therefore, a routine Dirichlet

distribution applies the Dirichlet distribution separately over all t in φ^k. We infer the latent topic by using Gibbs sampling [10].

Given the one-day-observation $X_{u,d} = \{x^i_{u,d}\}$ of the activities of user u on day d, $x^i_{u,d}$ can be represented by a 4-tuple (u, d, t, a) where t is the time slot and a is the location label, Home/Work/Other. The type of locations is estimated using CDRs from 55 volunteers as training data. We employed Random Forest and clustered the type of locations into the three labels. For the estimation, we first clustered the location of antennas, whose locations are close because the area covered by antennas changes according to the traffic volume and calls from the same location can be recorded as the communication through different neighbor antennas. As a result, we obtain the estimation result with the accuracy of 74%. In the estimation model, the rank in the frequency of calls by location, the proportion of calls by the type of locations, the distribution and timing of calls are found to be significant to improve the estimation accuracy. The statistical significance of the clustering results is difficult to mention because the algorithm of clustering is the collection of clustering trials, in which the best way of clustering is explored and figured out through majority voting without no hypothesis.

Here, we redefine the sampling distribution from routine k:

$$Routine\_Multinomial(k) := \prod_{x \in X_{u,d}} \pi^k_t(a)$$

Algorithm 1 details the key procedure of the routine topic model. Figure 1 shows two examples of routine topics we discovered. Topic 4 describes a typical workday routine for a salaried man (high probability of being at work during the daytime) and a routine for staying at home for an entire day.

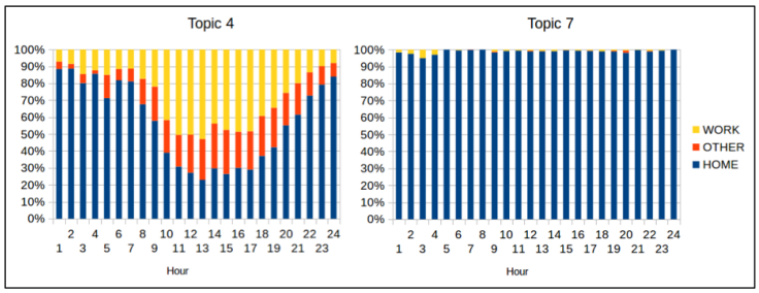


Figure 1. Typical routines for Topic 4 and Topic 7

Algorithm 1

Input: $\{X_{u,d}\}$ CDR dataset; K predefined number of topic number

Output: $\{z_{d,u}\}$ topic assignment; $\{\varphi^k\}$ routine distribution for topic k

(1) $n_{u,k} = 0;\ n_{d,k} = 0;\ n_{k,t,HOME} = 0;\ n_{k,t,WORK} = 0; n_{k,t,OTHER} = 0$

// Random Initialization

(2) For each u, d:

(3) $k = z_{d,u} = RandInt(K)$

(4) $n_{u,k} = n_{u,k} + \#(X_{u,d})$

(5) $n_{d,k} = n_{d,k} + \#(X_{u,d})$

(6) For each $x_{u,d}^i \in X_{u,d}$:

(7) $n_{k,t,a} = n_{k,t,a} + 1$

// Iterative inference

(8) For i=1 to MAX_ITERATION

(9) For each u, d:

(10) $k_{old} = z_{d,u}$

(11) $n_{u,k_{old}} = n_{u,k_{old}} - \#(X_{u,d})$

(12) $n_{d,k_{old}} = n_{d,k_{old}} - \#(X_{u,d})$

(13) For each $x_{u,d}^i \in X_{u,d}$:

(14) $n_{k_{old},t,a} = n_{k_{old},t,a} - 1$

(15) Sample a new topic assignment k from the distribution $p(k) = \frac{n_{u,k}+\alpha}{\sum_k n_{u,k}+K\alpha} \cdot \frac{n_{d,k}+\gamma}{\sum_k n_{d,k}+K\gamma} \cdot \prod_{x \in X_{u,d}} \frac{n_{k,t,a}+\beta}{\sum_a n_{k,t,a}+3\beta}$

(16) $z_{d,u} = k$

(17) $n_{u,k} = n_{u,k} + \#(X_{u,d})$

(18) $n_{d,k} = n_{d,k} + \#(X_{u,d})$

(19) For each $x_{u,d}^i \in X_{u,d}$:

(20) $n_{k,t,a} = n_{k,t,a} + 1$

// Calculate the values to be returned

(21) For each u, k:

(22) $\varphi^k(a) = [\pi_0^k, \pi_1^k, \cdots, \pi_{23}^k] = \left[\frac{n_{k,0,a}+\beta}{\sum_a n_{k,0,a}+3\beta}, \frac{n_{k,1,a}+\beta}{\sum_a n_{k,1,a}+3\beta}, \cdots, \frac{n_{k,23,a}+\beta}{\sum_a n_{k,23,a}+3\beta}\right]$

(23) Return $\{z_{d,u}\}$, $\{\varphi^k\}$

Routine interpolation

Note that CDR data are as sparse as only 2.8 records per day on average. For Dynamic Census, using CDR data directly suffers from a great bias on the user's calling preference. For example, people are more likely to make a call in the evening rather than at midnight. To reduce the bias, we interpolate the CDR data. From our routine topic model, we inferred the latent topic for each user on each day, as well as the routine distribution for each topic. We formulate this problem by sampling the missing part $\setminus X_{u,d}$ from a Markov network given the sparse observation $X_{u,d}$ and routine distribution $\varphi^{z_{u,d}}$:

$$p\left(\setminus X_{u,d} \middle| X_{u,d}, \varphi^{z_{u,d}}\right) \propto \prod_{(x_t, x_{t+1})} f(x_t, x_{t+1}) \prod_{x_t \in \setminus X_{u,d}} g\left(x_t \middle| X_{u,d}, \varphi^{z_{u,d}}\right)$$

f and g are the potentials defined on each edge of Markov network. The first product defines the filtering potentials between each pair of adjacent time slots to smooth the interpolation. The second product defines the factors that the joint distribution of sampling each blank time slot. To sample this joint distribution effectively, we apply a simulated annealing Gibbs sampler to search for the optimal interpolation.

3.3.2 Extracting typical calling behaviors based on spatio-temporal distribution of calls

We apply our extended LDA model to CDRs that are obtained from 5000 sets of records from mobile users. The time pattern that we discovered is shown in Figure 2. The vertical axis represents the number of calls. We extracted three principal topics as the three typical calling patterns of mobile users. As seen in Figure 2, Topic 1 and Topic 3 demonstrate the calling behavior in which people tend to call more during morning hours and during the day. Further, Topic 2 represents the calling behavior of people who tend to call more at night and seldom use mobile phones during the day. The topic is determined by the vocabulary, which is the spatio-temporal distribution of call records in our model, and hence, the patterns in Figure 2 are clustered based on

the pattern of calls in relation to the time distribution of significant locations. As mentioned earlier, locations that are repeatedly visited, such as home and workplace, can explain the behavior patterns of people. Thus, we conclude that, to a certain extent, the patterns extracted by our LDA model can be associated with some significant locations for mobile users.

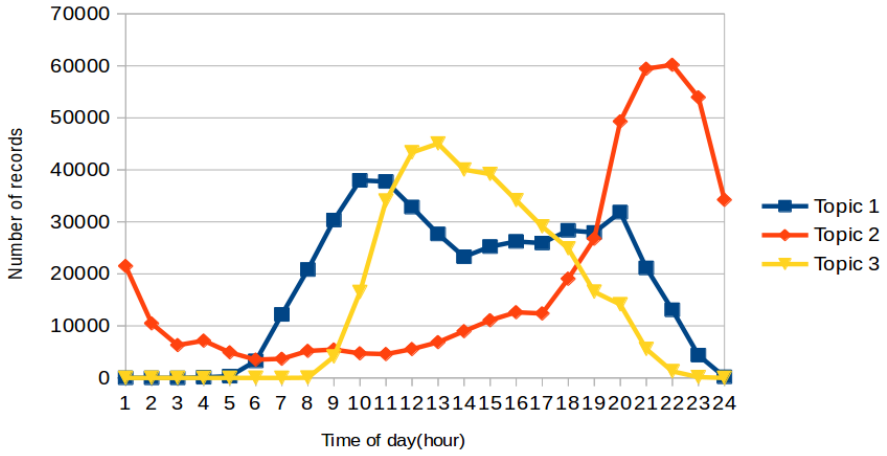


Figure 2. Time patterns of three principal topics extracted from CDRs

3.3.3 Estimating principal behavior patterns of mobile users based on calling behavior

In the previous section, three principal patterns are described based only on call records. Thus, these patterns are generated based on the sparse spatio-temporal matrix of call records, in which only the spatio-temporal slots with call records, while the remaining slots do not include data. We assume that this condition generates a bias that overestimates the spatio-temporal trends of calls. In order to eliminate this bias, we interpolate the blank time intervals of the patterns extracted in Figure 2. It can be proved that an interpolation of the time intervals is equivalent to a normalization of the data points with respect to each time interval. Figure 3 shows the three principal behavior patterns estimated based on the interpolation of the time patterns in Figure 2. From Figure 3, we can observe that after normalization, the

topics varying with time are represented by a percentage of all topics instead of the absolute number of records of each topic. In general, smoothed results are obtained, e.g. No decrease in Topic 2 around midnight (from 11 pm to 5 am), which unveils behavior patterns more objectively than calling behaviors.

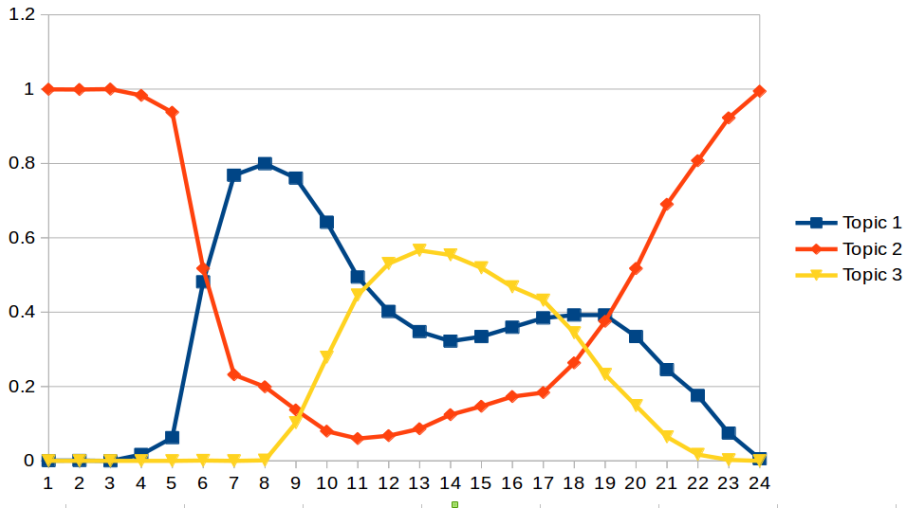


Figure 3. Normalized behavior patterns of the three principal topics

Prior to interpreting the results obtained, we briefly describe the general lifestyle of people in Dhaka. Besides the capital in developed countries, such as New York and Tokyo, most offices, shops, and entertainment venues in Dhaka are closed at night. Thus, a majority of the people residing in Dhaka tend to stay at home or in the vicinity of home late at night. Further, for a majority of Muslim people, who comprise approximately 90% of the population of Dhaka, the day starts with a morning prayer at around sunrise and ends with another prayer at around midnight. Hence, we believe that the spatio-temporal distribution for a majority of the population after midnight is associated with the home location.

Next, we interpret the three behavior patterns seen in Figure 3, extracted from the time patterns of call records. Based on the description above, we assume that the location having the highest probability after midnight is associated with the home

location throughout our interpretation. We begin with Topic 3 because it appears to be a simple pattern. Topic 3 describes the behavior pattern of people who have the lowest overall probability of being outside the home. Among the three patterns, the populations clustered into this pattern are at home starting at the earliest time in the evening. Hence, we assume that this population group is likely to be involved in household activity. Topic 2 exhibits the highest probability after midnight among all the topics, and hence, the probability of being at home for Topic 2 is interpreted with a reasoning that is opposite to the one used for Topic 3. Populations clustered into Topic 2 have the lowest probability of staying at home during the day, and the time at which they return home is the latest among the three topics. Therefore, we assume that this population group is generally engaged in an activity outside home. We interpret that the people belonging to Topic 1 have a very high probability of being at home, using the same approach as used for Topic 3. In general, Topics 1 and 2 are similar because these people generally spend a considerable amount of time outside the home. Topic 1 exhibits two peaks: the first is at approximately 8 am and the other is at 7 pm. Similar to the populations of Topic 2, those of Topic 1 spend a large amount of time outside the home. However, the time at which they return home is earlier than that in the case of Topic 2. For further interpretation, we combine these results with the analysis results of Section 3.4, in which typical behavior patterns of mobile users are described based on field survey data.

3.4 TYPICAL BEHAVIOR PATTERNS OF MOBILE USERS DERIVED FROM SURVEY DATA

In this section, we first obtain the typical behavior patterns of mobile users by analyzing the SPACE data to determine their gender and daily routine activities. Then, we analyze the types of populations that mobile users represent by comparing their behavior patterns obtained in this section and those determined in Section 3.3.

3.4.1 Populations of mobile users

In order to understand the principal populations of mobile users, we examine the SPACE data, which is the diary survey data of 922 mobile users. Table 2 describes the proportion of males and married users among mobile users, classified by income level. The overall proportion of males is greater than that of females. Further, more than 85% of the users are married. Bangladesh has relatively strong social norms of behavior based on gender. For instance, married males are generally engaged in an income-generating activity to support their family. Females tend to stay at home and perform household tasks while taking care of children and other family members. When we take into account the norms and the fact that there is a high proportion of married population, we expect gender-specific behavior patterns to be predominant in the SPACE data.

Table 2. Proportion of males and married users

	High	Middle	Low	Slum
Male user	62%	52%	63%	68%
Married user	85%	86%	87%	86%

Next, we analyze the primary activity of mobile users. Assuming the gender specific trends mentioned above, we classify the type of activity based on gender (male and female) and economic activity (income-earning activity and non-income earning activity). An income- earning activity is any activity that generates monetary income as a return For example, salary earners, part-time workers, self-employed people, etc. are classified as persons engaged in an income- earning activity. The remaining people are engaged in a non-income earning activity, which is any activity that does not generate monetary income as a return. Table 3 shows the distribution of the type of activity by gender. It is evident that more males are engaged in an income-earning activity while a majority of females are engaged in a non-income earning activity, particularly household tasks. This indicates that most of the mobile users subscribed to this company are those who perform responsible roles within the

household, i.e., they have some money at their disposal. The mobile tariff for this company is relatively expensive compared to that for other companies in Bangladesh, and we consider it to be a factor affecting such trends in users.

Table 3. Distribution of main activity by gender

	Income-earning	Non-income earning		
		Household tasks	Education	Other
Male	89%	1%	4%	5%
Female	18%	77%	4%	1%

Table 4 shows the proportion of people who are engaged in a typical activity corresponding to their gender. The values in parentheses represent the proportion of married users. It is observed that trends by gender across all income levels are similar for males and females. Therefore, we can state that, regardless of income levels, there are two types of typical mobile users: the married male engaged in an income-generating activity, and the married female who mainly performs household tasks. Based on the results, these two typical mobile users are termed as Workmale and Housewife.

Table 4. Proportion of people engaged in typical activity, classified by gender and income level

	High	Middle	Low	Slum
Males engaged in an income-generating activity	86% (78%)	87% (81%)	93% (86%)	95% (88%)
Females performing household tasks	75% (78%)	79% (87%)	85% (95%)	71% (78%)

3.4.2 Location of main activity

Considering Workmale and Housewife as typical mobile users, we determine the location of their main activity. It is essential to identify the location of the activity because the behavior patterns extracted from CDRs are described based on the probability distribution of locations. As mentioned in Section 3.3, we cluster the locations recorded in CDRs into three groups: Home, Work, and Other. Table 5 shows the distribution of locations for the main activities of working males and

housewives. In case of working males, 72% of the locations for their main activity are specific locations outside home. This indicates that more than half of typical male users have a specific location outside home for their main activity. In the case of housewives, 94% of the locations for their main activity are home. Owing to the distinctive difference in the location of the main activity for the two principal population groups, it is acceptable to interpret the typical behavior patterns extracted from CDRs based on the analysis results of this section.

Table 5. Distribution of locations for main activity among the principal population in CDRs

Type of location	Workmale	Housewife
Home	7%	94%
Outside the home (in a specific building)	62%	4%
Outside the home (a specific location on the street)	10%	1%
Outside the home (in several buildings)	3%	1%
Outside the home (moving to various locations)	18%	0%

3.4.3 Typical behavior patterns of mobile users

We obtain the location-based behavior patterns of typical mobile users based on the SPACE data. In addition to working males and housewives, we examine the behavior pattern of students, who form the third-largest segment of mobile users although the absolute proportion of this segment is much smaller than the other two. Figures 4(a) and 4(b) show the hourly distribution of the probability of being at Home and Work, respectively. For housewives and students, we considered the primary location outside home as their Work, e.g., the time spent for education for students is considered as Work. We can observe a distinctive trend for housewives: the probability of them being at home is almost 100% throughout the day. Working males and students have relatively similar probability distribution of being at Work, but this probability is much higher for working males. This could be attributed to differences between office hours (e.g., from 9 am to 5 pm) and school hours (e.g., from 8 am to

noon for primary school, and from 10 am to 4 pm for junior high or high school or collage) in Dhaka.

Comparing the analysis result in this section with the description of the three principal behavior patterns described in Section 3.3, we observe that the LDA model can distinguish at least two patterns: (1) Topic 2, which represents the behavior patterns of people who generally spend a majority of their time outside home and return home at a time which is the latest among all the patterns; these people are most likely to be working males, and (2). Topic 3, which represents people who are engaged in an activity related to the home; these people are most probably housewives. With regard to Topic 1, we cannot determine whether it represents students from the given information. However, we believe that the behavior pattern of students can be similar to that of male workers owing to the time duration generally spent outside home. In further studies to distinguish between such similar patterns, we consider taking seasonality, a longer time frame than daily basis estimation, into account because students generally have a term break.

3.5 DISCREPANCY BETWEEN THE POPULATION IN CDRS AND THE LIVING POPULATION

3.5.1 Principal population groups of the living population

In this subsection, we present the background characteristics of all the respondents in the Person Trip Diary Survey in order to get insight into the structure of the general population of Dhaka. Owing to space limitations, we focus on only two key variables: gender and economic activity, as we did in the previous section. In the sample, the number of males was slightly greater than that of females (54% vs. 46%). Table 6 reveals three key population groups based on their activities: respondents engaged in income-generating activity (38%), household tasks (25%), and education (32%). Further, we note that male respondents comprise a majority of the income-

generating group while females are in a majority in the household tasks group. Hence, we focus our analysis on the following key population groups: working males, housewives, and students, and present their typical behavior patterns in Section 2.2.

Table 6. Percentage distribution of survey respondents and their income status by gender

	Income-earning	Non-income earning		
		Hhousehold tasks	Education	Other
Overall	38%	25%	32%	5%
Male	61%	1%	32%	6%
Female	10%	53%	33%	4%

3.5.2 Typical behavior patterns of principal population groups in Dhaka

We report our findings related to typical behavior patterns across the three principal population groups that we obtained from the data, as explained in Section 3.5.1. The results are summarized in Figures 5(a) and 5(b). As we have mentioned in earlier sections, previous research has found that behavior patterns of people can be explained by focusing on important places such as Home and Work. In this regard, we present the distribution of only the key population groups considering (a) Home and (b) Work locations. The results reveal clear and evident patterns. The probability of being at Home is the highest for housewives among the three principal population groups. We find more similarity in the probability of being at Work between working males and students for the general population compared to mobile users. In spite of the minor difference in the result, we can conclude that the behavior patterns extracted from the two sources of data are, in general, similar.

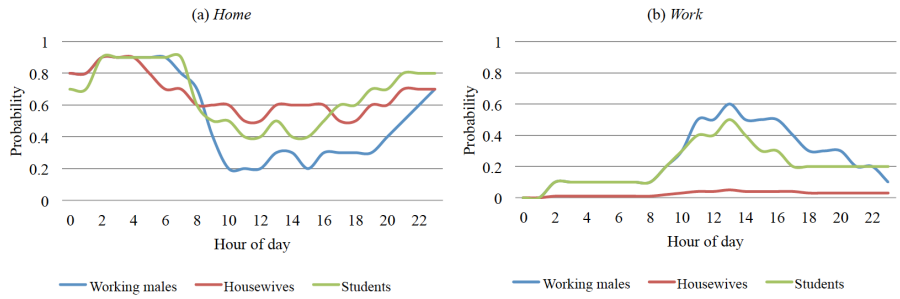


Figure 5. Hourly probability distribution of being at (a) Home and (b) Work for the three principal population groups

3.5.3 Mobile ownership bias

Finally, we discuss the ownership bias of mobile users by comparing the principal population groups of the general population and those of mobile users. Table 7 shows approximate estimates of the proportion of the three principal population groups for the general population and mobile users. The estimate for mobile users that is marked with a + sign indicates a possible minimum estimate because we can calculate the proportion only within each income level, and do not have information about the population distribution across all income levels. It is important to note that, in the general population, there is a sizeable population of students, who receive education as their main activity; however, the corresponding proportion of students among mobile users is very small. Further, the proportion of male workers and housewives is significant in the general population and mobile users. Given the marginal proportion of students among mobile users, we can conclude that CDRs are biased be-cause the data seldom includes students, who comprise a significant proportion of the general population in Dhaka.

Table 7. Proportion of the three principal population groups for two samples

	Workmale	**Housewife**	**Student**
Living population	38%	25%	32%
Mobile users	46%+	26%+	4%+

3.6 SUMMARY

In this chapter, we proposed a novel approach to reveal the discrepancy in principal population compositions between the general population and mobile users by comparing their typical behavior patterns. First, we profiled principal populations of mobile users by employing the LDA topic model to reconstruct typical behavior patterns from sparse CDRs. We found two typical behavior patterns: people who spend most of the day engaged in routines outside home, and those who spend most of their time at home. The results were consistent with the behavior patterns extracted from the diary survey data of mobile users, where we can observe two typical behavior patterns: the male, engaged in income-generating activity outside home during the day; and the female, spending a majority of the time at home, mainly performing household tasks. Comparing the principal populations of mobile users and those of the general population, we found that students form a core component of the general population but are not considered significant among mobile users. The analysis results indicated that CDRs accurately capture the behavior patterns of populations who have a certain amount of money at their disposal in a setting where mobile devices are not yet fairly widespread. Therefore, we suggest that the application of CDRs, targeting the younger generation in particular, needs to take this bias into account because the data do not necessarily represent such populations. We believe that our findings will be useful for the utilization of CDRs in the developing world, which has scarce resources. CDR acquisition does not have additional costs; this data is generally collected for billing purposes by the cellular network company, and is therefore available as long as a mobile network is present.

Although the results are obtained based on a case study, we would like to underscore the fact that this approach can be extended to applications in other disciplines by utilizing other data. In our approach, we use diary survey data to profile the general population; however, obtaining such data is not always easy. For further

studies, we intend to use census data because mobile ownership is recommended as a core topic for the Population and Housing Census by [11]. In fact, a census is recommended every five years, and has been conducted in more than 200 countries in the census round spanning the period from the year 2005 to 2014 [12]. There is considerable potential for understanding domain-specific biases, which can constitute major constraints in the utilization of large-scale domain specific data such as CDRs for addressing various societal issues. However, we would like to mention that the census is conducted based on the registered address and not the actual living location. Further, the official statistics do not generally include certain populations such as squatters in slums and shantytowns, whose populations are substantial and difficult to ignore, particularly in the developing countries In this regard, information collected from the field survey is necessary to a certain extent. In order to generalize our approach, we will consider a trade-off between the requirements of understanding actual real-life situations, which requires data collection from the field, and the costs of data acquisition.

Chapter 4 Anonymity: Estimation of personal attributes of mobile phone users

4.1 Background

Analyses of large-scale mobile phone log data such as GPS logs and call detail records (CDRs) have provided detailed descriptions of human mobility. It is widely agreed that people routinely visit specific locations [1,2], and these are typically home and work places [3]. Various studies have proposed algorithms to estimate the home and work locations of mobile phone users by examining the time and location distribution of call records, which is accomplished by analyzing anonymized CDRs [4,5]. Research based on GPS log location histories measures users' similarities by analyzing the sequential properties of their trajectories, and the hierarchical properties of their location histories. This body of research argues that users with similar location histories also share similar interests and preferences [6]. Although these studies succeed in mining mobility patterns, the outcomes of large-scale data analyses describe the movement of crowds, because the data is anonymized.

In addition to such large-scale datasets, there are other types of data collected through conventional methods. In the field of urban planning and transportation, many empirical studies have attempted to identify the factors that affect human activities and travel patterns, by analyzing data collected through questionnaire surveys. Such data typically consist of socio-demographic attributes, transportation means and origins, destinations, and purposes of movement, all with associated time stamps. Family and social obligations are often presented as significant factors affecting daily travel and activity behavior [7]. This implies that people's activity patterns are constrained by social ties. This observation is consistent with most previous analysis of human mobility patterns from large-scale datasets, where people

routinely visit a limited number of locations [1,8]. Though such conventional approaches may be less efficient in defining human activity and travel patterns in terms of population size and data period length, it can link activity and travel patterns with people's attributes to some extent [9]. Conventional survey data enables common key patterns to be identified from large-scale data, and allows us to understand the hidden properties of anonymized large-scale data. To further investigate the properties of user attributes in anonymized data, it is critical to analyze this large-scale data in combination with such secondary data [10].

In fact, emerging studies are attempting to analyze data derived from mobile phones in combination with secondary data. [11] observed that the ratio of shared phone usage and call type, such as incoming and outbound calls, shows significant differences by gender, on average. In addition, the number of usages during a specified length of time was shown to differ according to income level. This observation is consistent with another research project, which analyzed social connectivity through social networking activities [12]. Utilizing sensor data from volunteer mobile users, [13] proposed prediction models based on user demographic attributes. While these models focus on features derived from smart phones such as acceleration and application usage, they provide features that can be derived from CDRs. For instance, the probability of being at home or work at night is useful for predicting the occupation type. Furthermore, the number of places visited in the evening is useful for predicting marital status.

In this chapter, we provide a novel approach for extracting features from calling behavior, which can be constructed from anonymized CDRs. This technique can reveal distinctive traits that help identify the demographic attributes of mobile phone users. This study is unique, because we focus on extracting lifestyle traits and routines to identify user attributes. We statistically analyze data collected through a field survey that focused on the demographic attributes, calling behavior, and weekly

activity patterns of mobile users. Several key features are empirically derived to analyze user attributes, which are compared with another set of the features generated from CDRs.

Contributions of this chapter are described below:

- Statistical analysis results are provided for calling behavior based on field survey data. We extract calling behavior traits to differentiate gender and occupation types that correspond to the patterns of routine activities. To do so, we introduce the concept of weekly activity patterns, which specifies whether the day's call records correspond to a day where the user is engaged in their primary routine.
- By comparing the analysis results of the field survey data, which include single-day call records of 922 users, and those of two-month CDRs, we describe the advantages and limitations of analyzing call records from the field survey data.
- The results of experimental study on estimating the personal attribute of mobile users are provided. We use three datasets, whose lengths of data period differ, to examine how the daily and weekly routines are important for the estimation.

The remainder of this chapter is organized as follows. Section 4.2 describes the data used for this study. Section 4.3 explains the mobile user lifestyle and calling behavior prototypes that were used for analyzing the survey data. Section 4.4 compares the statistical analysis results of the survey data with those based on the CDRs for 58 volunteers. Section 4.5 discuss the necessity of considering the routine of human behavior through descriptive statistics on calling behavior generated from one-day and two-month CDRs. Section 6 provides the results on the estimation of personal attribute of mobile users. Final section includes conclusions and discussions for further studies.

4.2 DATA

In this chapter, we use two data sources: the aggregated single-day call records of 922 mobile phone users from a field survey, and two-month CDRs for 58 volunteers. A unique characteristic of the survey data is that it includes actual call records from mobile phone users, as single-day records from 922 handsets. The records specify the call's location type and other basic attributes. The data contain neither mobile phone numbers nor any other information explicitly specifying individuals.

4.2.1 Field Survey Data

To understand the hidden properties of CDRs, we conducted a field survey, named the Survey on Patterns of Activity for Comprehensive Explorations of Mobile Phone Users in Dhaka (SPACE). The purpose of SPACE is to understand the calling behavior, characteristics, and lifestyles of mobile phone users. To capture calling behavior along with the diverse characteristics of mobile phone users, we employed two-staged stratified sampling. We first classify all primary sampling units (PSUs) into three groups based on their dominant type of land use: residential, commercial, or industrial. Of these, 15 PSUs are selected, taking population numbers into account. Figure 1 shows the 15 PSUs. Numbers in the map indicate the Ward number. Ten PSUs, which are highlighted in green, are sampled from Dhaka City Corporation (DCC). Three PSUs, which are highlighted in blue, are sampled from Municipalities surrounding DCC. Remaining two PSUs, which are highlighted in orange, are selected from the suburban areas of Dhaka. From each PSU, 18 households are selected for three income groups: high, middle, and low income. This means that we have 270 households for each income group, and therefore 810 households in total. From the 18 households in the low-income group in each PSU, we sample the slum population as part of the low-income group if the ratio of the slum population to the total population in the PSU is more than 25%; otherwise, we do not sample the slum

population for the low-income group. As a result, the low-income group includes 189 households from the slum population.

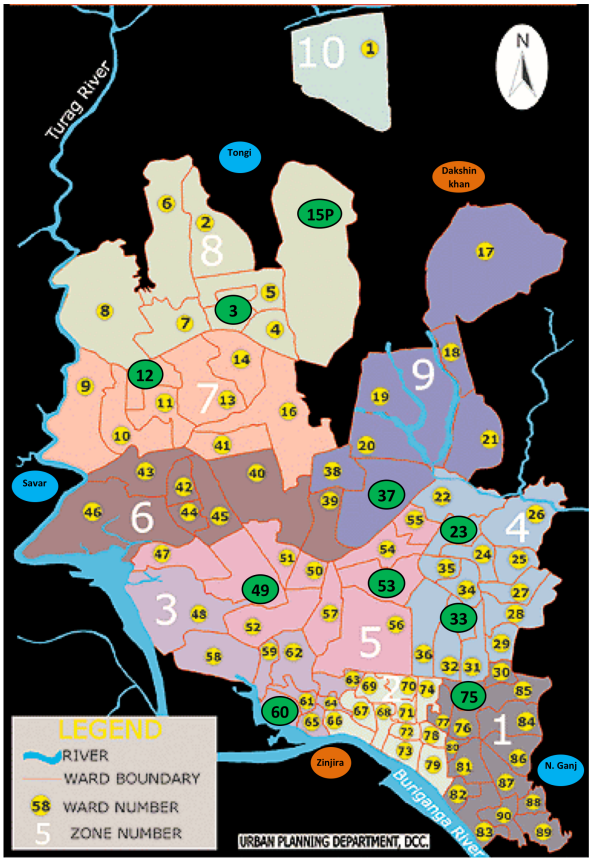


Figure 1. Selected PSUs for the survey

For the following analysis, the 810 households are re-classified into four categories: high, middle, low, and slum income levels. Since there are different numbers of households sampled for each income level, the analysis is based on the ratio and distribution within each income level. Among them, 15 PSUs were selected by considering population distributions. Subsequently, 18 households each were sampled from high, middle, and lower income groups for each PSU. Because it is difficult to obtain household income data for sampling, we set the criteria for the three income groups based on building types, ownership, facility, and durable consumer goods. The survey data include 922 mobile phone users subscribed to the operator.

Table 1. Basic attribute distribution of the mobile phone users in SPACE data

	Male	Female
Users of a specified operator	60%	40%
Roles in the household	**Male**	**Female**
Head of Household	85%	9%
Spouse	0%	76%
Daughter/Son of head/spouse	12%	10%
Others	2%	5%
Age group	**Male**	**Female**
Under 20	3%	5%
20 – 29	16%	28%
30 – 39	31%	33%
40 – 49	26%	24%
50 – 59	16%	7%
60 and more	7%	0%
Primary activity = Occupation type	**Male**	**Female**
Working (not household tasks) = Worker	89%	18%
Household tasks = Household	1%	77%
Schooling = Student	4%	4%
Others = Others	6%	1%
Individual annual income	**Male**	**Female**
0 BDT ~	8%	80%
~ 48,000 BDT (620 USD)	3%	1%
~ 144,000 BDT (1,800USD)	23%	7%
~ 312,000 BDT (4,000 USD)	31%	7%
~ 576,000 BDT (7,400 USD)	14%	2%
~ 996,000 BDT (12,800 USD)	10%	2%
More than 996,000 BDT	10%	1%

Table 1 describes the basic features of mobile phone users for the leading telecommunications operators in the SPACE data. Eighty-five percent of males are categorized as the head of the household, and are generally considered to be financially responsible for the people living with them. Further, 76% of females are categorized as a spouse and 77% of them consider household tasks to be their primary activity. This implies that most female users are married and care for their family members at home. We assume the females are able to allocate some money at their disposal, even if they are not engaged in income-earning activities. Individual income levels among mobile phone users in the SPACE appear to be higher than average for Dhaka, where annual income per capita is 1,350 USD [14]. It partially infers that this operator's mobile phone tariff is relatively high. The primary activity is specified as *Worker, Household, Student,* or *Others*; hereinafter, this is referred to as the occupation type.

4.2.2 Call Detail Records

We used CDRs for 58 mobile users, who receive service from one of leading telecommunications operators in Bangladesh. Their basic attribute information is also collected. It included gender, age, weekly activity patterns, annual income, and call location types such as home and work places. Table 2 describes their basic features. CDRs consist of time, antenna locations, and duration of voice calls from November and December 2013.

Table 2. Basic attribute distribution for 58 mobile users

	Male	**Female**
Users of a specified operator	47%	53%
Roles in the household	**Male**	**Female**
Head of Household	85%	13%
Spouse	0%	81%
Daughter/Son of head/spouse	7%	0%
Others	8%	6%
Age group	**Male**	**Female**
Under 20	11%	23%
20 – 29	19%	35%
30 – 39	33%	23%
40 – 49	15%	13%
50 – 59	15%	6%
60 and more	7%	0%
Primary activity = Occupation type	**Male**	**Female**
Working (not household tasks) = Worker	70%	13%
Household tasks = Household	0%	84%
Schooling = Student	4%	0%
Others = Others	26%	3%
Individual annual income	**Male**	**Female**
0 BDT ~	14%	83%
~ 48,000 BDT (620 USD)	0%	0%
~ 144,000 BDT (1,800USD)	41%	7%
~ 312,000 BDT (4,000 USD)	22%	7%
~ 576,000 BDT (7,400 USD)	15%	3%
~ 996,000 BDT (12,800 USD)	4%	0%
More than 996,000 BDT	4%	0%

4.3 LIFESTYLE AND CALLING BEHAVIOR

To understand activity patterns, it is vital to identify a certain set of locations where people spend the majority of their time. Many studies identify two dominant locations for people as their home and work or school (hereinafter referred to as "Home" and "Primary location outside the home"), which can explain a significant portion of their activity [1]. Identifying the time and location distribution of the users'

whereabouts enabled us to define their activity patterns and portions of their lifestyles. Therefore, in this section we examine the time distribution of calls initiated from Home, Primary location outside the home, or Other by analyzing the calling behavior of SPACE's mobile phone users. Because sequential location histories can infer similarities between people [6], we expect the distributions to vary according to differences in user attributes. We extracted calling behavior prototypes according to gender and occupation type, and examine how their traits differ.

For the analysis provided in this section we introduce the concept of weekly activity patterns, which specifies whether mobile users are engaged in their primary routine on the day the call was made. Routines can be any activities users spend the majority of their time on during the day. The routine that the user follows on the highest number of days during the week is considered as their primary routine.

4.3.1 Behavioral Norms and Diverse Lifestyles

Dhaka is the capital city of Bangladesh; it is one of the emerging economies in South Asia, enjoying an average of 4.8% annual GDP growth per capita [15]. Its strong growth has accelerated the urbanization of Bangladesh, leading to a high concentration of economic and demographic growth in metropolitan areas. Bangladesh currently has very vibrant mobile phone markets, which include 112 million active subscribers in a total population of 155 million [16]. Considering the substantial and increasing subscription rate, the mobile phone is a useful platform for acquiring a general picture of human mobility in Dhaka.

Bangladeshi society has long promoted divisions in social space, and fostered differences in behavioral norms between males and females in various aspects of social, cultural, and religious traditions [17]. As a result, females have often assumed the role of family caregiver [18]. However, attitudes regarding their social roles have been changing, owing to advancements in educational opportunities that have

occurred along with economic development. Consequently, females, particularly those in middle or upper-middle income households, have started to transform their traditional roles significantly [19]. As described in Table 1, mobile users in the SPACE data include a variety of population groups. Therefore, the SPACE data enable us to capture such transformations while traditional behavioral norms continue to exist.

4.3.2 Weekly Activity Patterns

In this subsection, we describe the weekly activity patterns of the 922 mobile users recorded in the SPACE data. Weekly activity patterns are determined based on the number of weekly routines (primary and non-primary), and the number of days on which the primary routine is followed. By classifying the activity according to a pattern, we examine how activity patterns are linked to calling behavior in the following section. We assume that these patterns can also be partially extracted from anonymized CDRs, which allows us to estimate significant locations such as home and work places. Based on this technique, we consider it possible to reconstruct weekly activity patterns from CDRs, to which we can apply this concept.

Table 3 classifies the mobile users into four patterns, according to the following criteria. Pattern (1) captures those who have only one type of routine per week. That is, their weekly activities do not fundamentally change. Fifty-six percent of the users are classified into this pattern. As shown in Table 3, the number of routines is one, and the number of primary routine days is seven. The majority of users classified into this pattern tend to have activities outside of home such as commuting to an office, shopping at a market, or sending their children to school. We assume some of those locations will appear in call records as the Primary outside-home location. The population in Pattern (1) consists of those who are primarily engaged in income-earning activity on a daily basis, and those who primarily perform

household tasks and care for family members at home. All seven days of the week are considered to be their primary routine days, unless they have other routines such as social activities in their community on specific days, or small side businesses that occupy their spare time. In addition, approximately 2% of those in Pattern (1) do not have specific tasks, and therefore remain at home. Most of them are elderly, retired, or young children not enrolled in school. Although they primarily remain at home all day and are not engaged in any specific tasks, all days of the week are counted as their primary routine days because their daily activities do not change.

Table 3. Weekly activity patterns of 922 mobile users in SPACE data

Pattern	Description of people classified into the pattern	Number of routines	Number of days for the primary routine	Proportion (%)
(1)	Those who repeat their primary routine activity every day. Mainly composed of income earners and those who do household tasks.	1	7	56
(2)	Those who repeat their primary routine activity every day, except for Friday. Mainly composed of income earners or are students.	2	6	30
(3)	Those who repeat their primary routine activity every day except for Friday and Saturday. Mostly composed of public sector employees or students.	2	5	9
(4)	Mostly income earners and their activity patterns do not follow Pattern (1), (2), or (3).	1 or more	1 to 6	5

The population in Pattern (2) consists of users who have two routines per week, where six days (except for Friday) are devoted to their primary routine. The majority of users in this pattern are income earners or students, because it is common to have only one non-working or non-school day in Dhaka, which is typically a Friday. As described in Table 3, the number of routines in Pattern (2) is two, and the number of primary routine days is six. Similarly, the population in Pattern (3) primarily consists of users who are income earners or students. However, those who are classified into this pattern devote five days, except for Friday and Saturday, to their primary routine. As listed in Table 3, only a small percentage of users are classified into Pattern (3). Few people in Dhaka have two non-working or non-

schooling days. It is likely that many users in Pattern (3) work in the public sector, because the government sets Friday and Saturday as official holidays. The remaining users are classified into Pattern (4). Most of them are engaged in income earning activity and have both working and non-working days. Their primary routine days do not essentially follow Patterns (1), (2), or (3). For instance, some of them may be engaged in income-earning activity for five or six days per week and their non-working days are neither Friday nor Saturday.

Then, we examine how the four patterns are distributed within major occupation types. Figure 1 shows the distribution of the weekly activity patterns across the occupation types. We note that we do not provide the distribution of the occupation types for weekly activity patterns. This is because we sampled the same number of households from each income group, but the income group distribution among general mobile users was not confirmed. That is, the composition of the occupation type for each weekly activity pattern can be affected by the distribution of the income level, if the population composition differs by income level.

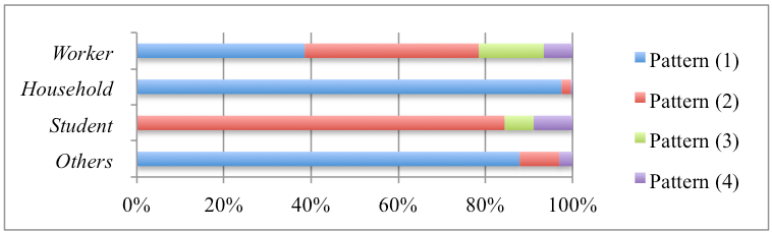


Figure 1. Distribution of weekly activity patterns of mobile users in SPACE data across occupation types

As explained in Table 2, we categorize mobile users into four occupation types: Worker is composed of users whose primary activity is earning an income; Household is composed of users who perform household tasks; Student includes users who are enrolled in school; and Other is composed of users who do not fit into any of the aforementioned categories. As Figure 1 shows, there are certain patterns that are

specific to occupation types. For instance, 40% of Worker is composed of Pattern (1) and an additional 40% is composed of Pattern (2). Household is primarily represented by Pattern (1); similarly, Student is primarily represented by Pattern (3). Assuming that the distribution of call locations can partially reflect the times and locations where people spend the majority of their day, the weekly activity patterns, derived from calling behavior, can be keys to understanding the activity of mobile users.

4.3.3 Calling Behavior by Gender

This subsection compares the calling behavior of males and females by analyzing the call records of the 922 mobile users. First, we examine calling behavior by analyzing call location trends. Figure 2 shows percentages for the number of calls at (a) Home and (b) Primary outside-home location among the total number of calls on the primary routine day. As the figure shows, males and females exhibit distinctly different call location trends. Across all patterns, females tend to call predominantly from Home while males call from both the Home and Primary outside-home location. This indicates that identifying dominant call locations on the primary routine day is important for determining gender. This feature is particularly distinctive among those classified into Pattern (1), where the primary routine is repeated every day.

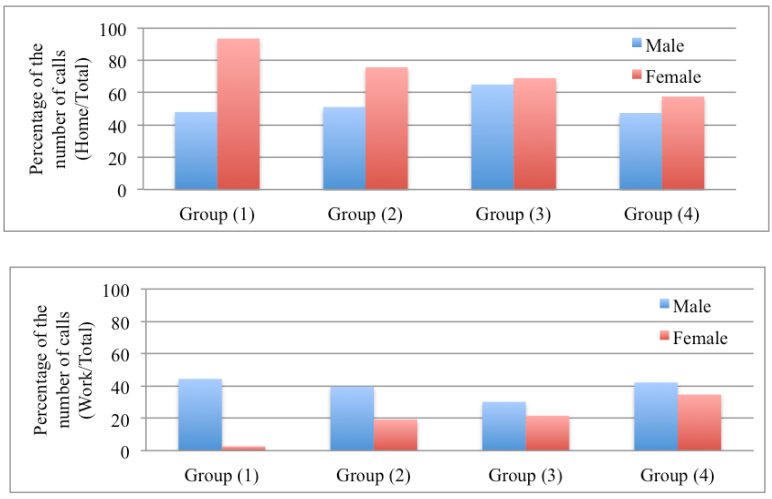


Figure 2. Percentage of calls from (a) Home and (b) Primary location outside the home against total calls on the primary routine day by gender (a) Upper (b) Lower

Similarly, Figure 3 shows percentages of the number of calls from (a) Home and (2) Primary outside-home location compared to the total number of calls on the non-primary routine day. In contrast to primary routine day trends, non-primary routine day trends for males and females are relatively similar. However, the figure shows large differences in the call location distributions of Patterns (2) and (3). This corroborates our assumption that the distribution of call locations, which partially reflects the time spent at the locations, accurately reflects differences in the activity patterns of different population groups. There is no observation for Pattern (1), because those classified into this pattern only have primary routine days. Further, we do not have an observation for Pattern (4) on non-primary routine days in SPACE data. As a result, single-day call records from users classified into Pattern (4) are utilized to represent all of their primary routine days.

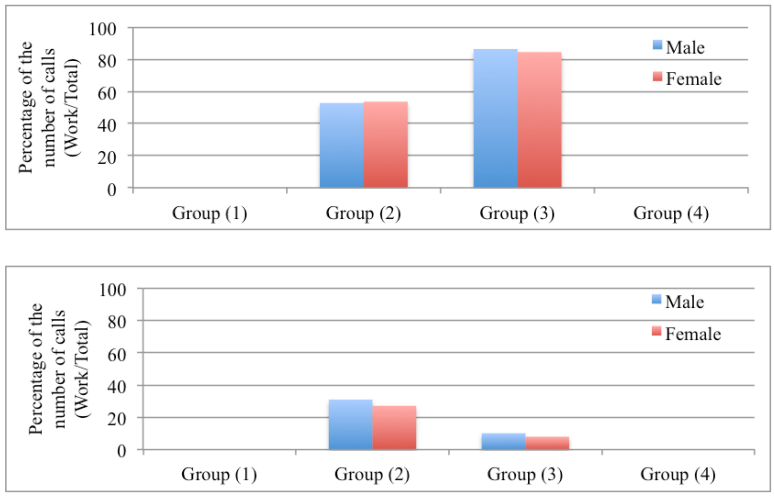
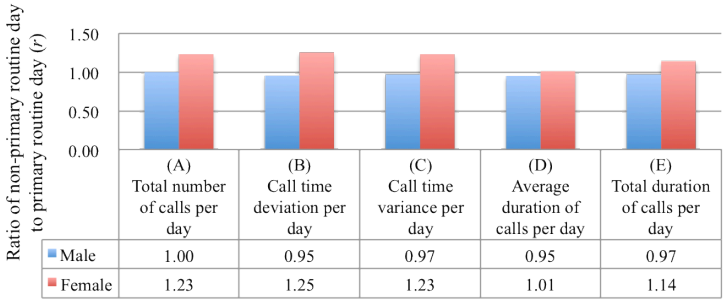


Figure 3. Percentage of calls from (a) Home and (b) Primary location outside the home against total calls on non-primary routine day by gender (a) Upper (b) Lower

We then compared the calling behavior of males and females in terms of time distribution and call length. Using five features, we attempted to capture calling behavior trends based on the frequency, time, and duration of calls for the primary and non-primary routine days. Figure 4 compares the calling behavior of the primary routine day and that of the non-primary routine day by gender. Each ratio r in the table of Figure 4 is obtained by solving the following equation for each feature:

$$r = \frac{\textit{Value of a feature, which is calculated from calling records of } \textbf{\textit{non-primary}} \textit{ routine days}}{\textit{Value of a feature, which is calculated from calling records of } \textbf{\textit{primary}} \textit{ routine days}}$$



	(A) Total number of calls per day	(B) Call time deviation per day	(C) Call time variance per day	(D) Average duration of calls per day	(E) Total duration of calls per day
Male	1.00	0.95	0.97	0.95	0.97
Female	1.23	1.25	1.23	1.01	1.14

Figure 4. Comparison of calling behavior on the primary routine day and non-primary routine day

Features (A), (B), and (C) capture calling behavior based on the call frequency and time of day, while (D) and (E) capture calling behavior based on the duration of calls. For example, the value of feature (A) is the ratio of the number of calls on the primary routine day to the number of calls on the non-primary routine day. The greater the value deviates from unity, the more the users call on the non-primary routine day. The value of (A) for males, which is unity, indicates that the frequency of calls for males did not change between their primary and non-primary routine day. That is, the frequency of calls for males fundamentally did not change according to the type of day. In contrast, the value for females, 1.23, indicates females tended to call more frequently on their non-primary routine days. Based on the trend observed

from feature (A), we conclude that the frequency of calls for males was consistent throughout a week, whereas the call frequency for females fluctuated, in that they tended to make a greater number of calls on the non-primary routine day.

Features (B) and (C) explain trends in the time distribution of calls for the primary and non-primary routine days. Here, the time of calls denotes the time when the user initiated communication with the call's recipient (Missed calls, which are not accepted by the recipient, are not included in CDRs.) Feature (B) is the average deviation of the call times calculated for each user and (C) is the average variance of those calls. The values of both features for males, which are close to unity, imply that calling behavior on the primary routine days and non-primary routine days were similar. That is, on average, if the user tended to call at a specific time on primary routine days, the user exhibited similar calling behavior on non-primary routine days, and vice versa. Conversely, the value of both features for females, 1.25 and 1.23 respectively, significantly deviate from unity, indicating that the call time distribution of the primary routine days differed from that of the non-primary routine days. Similar to the trend in the frequency of calls, the call time distribution trend indicates that calling behavior for males was similar throughout the week; but different for females, where greater variability between the primary routine days and non-primary routine days is observed in terms of call times.

The values for features (D) and (E) show trends in the duration of calls for the primary routine days and non-primary routine days. Feature (D) is the average duration per call per person. For example, the value of feature (D) for males is slightly smaller than unity. This indicates that the average call duration for males was slightly longer on primary routine days compared to non-primary routine days. As for females, average call durations for the primary and non-primary routine days were essentially equal. Contrary to the trends in the frequency and time distribution of calls, the average duration per call for females was similar throughout the week but

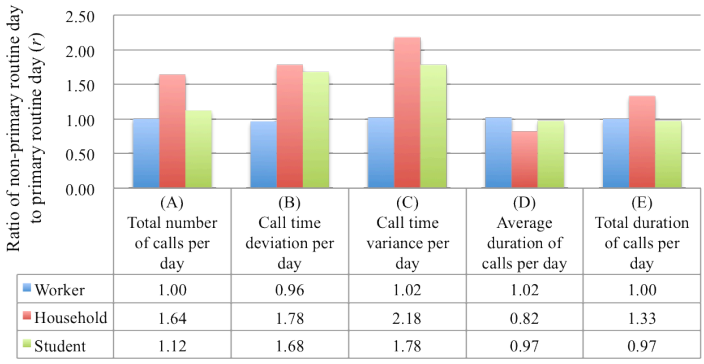
different for males, because males tended to make longer calls on primary routine days. Feature (E) is the average of the total duration of calls per day per person. Similar to feature (D), the value of feature (E) for males is slightly smaller than unity. This implies that the total duration of calls per day for males was slightly longer on primary routine days compared to non-primary routine days. Conversely, the value of (E) for females is greater than unity. This indicates that females tended to spend a greater amount of time on the phone during their non-primary routine days. Thus, regarding the duration of calls, females tended to be on the phone longer during their non-primary routine days, while males tended to be on the phone longer during their primary routine days.

To summarize, we have identified gender-specific traits in calling behavior. By analyzing the five features, we conclude that considering the type of day (primary or non-primary routine day) while examining call records is crucial for extracting gender-wise traits. Assuming that weekly activity patterns can also be reconstructed from the CDRs, we conclude that our findings can be utilized to estimate the gender of the user who produced an anonymized CDR.

4.3.4 Calling Behavior by the Occupation type

Similarly, we compared calling behavior across occupation types. Figure 5 compares the calling behavior recorded on primary routine days and that of non-primary routine days across occupation types. The features used are the same as those in Figure 4. We can observe distinctive differences in calling behavior according to occupation type. For instance, the value of feature (A) for Household, 1.64, indicates that users who were primarily engaged in household tasks tended to call more frequently on their non-primary routine days, while those in the Worker and Student categories appeared to exhibit similar call frequency trends on their primary and non-primary routine days. This indicates that differences in calling frequency between the

primary and non-primary routine days are keys to identifying users who are primarily engaged in household tasks, where greater variability is observed.



	(A) Total number of calls per day	(B) Call time deviation per day	(C) Call time variance per day	(D) Average duration of calls per day	(E) Total duration of calls per day
Worker	1.00	0.96	1.02	1.02	1.00
Household	1.64	1.78	2.18	0.82	1.33
Student	1.12	1.68	1.78	0.97	0.97

Figure 5. Comparison of calling behavior on the primary and non-primary routine days across occupation types

Further, the feature (B) values for Household and Student, 1.78 and 1.68, respectively, and the feature (C) values for Household and Student, 2.18 and 1.78, respectively, largely deviate from unity, while (B) and (C)'s values for Worker are close to unity. This indicates that the call times for Household and Student varied more on non-primary routine days than they did on primary routine days. In contrast, trends for Worker are again similar for the primary and non-primary routine days. That is, Household and Student exhibit more flexibility in the timing of calls on the primary routine days. This indicates that a call time distribution that is consistent throughout the week can be a key to identifying users who are engaged in some type of income-earning activities.

Regarding the duration per call and total duration of calls per day, we observe that users in the Household category have particular trends compared to users in Worker and Student. As shown by features (D) and (E) for Household, those who were primarily engaged in household tasks made shorter but more frequent calls on their primary routine days, whereas the duration per call and total duration of calls per day for Worker and Student did not vary according to the type of day. This indicates

that variability in the call duration and the total duration of calls between primary and non-primary routine days can be a key to identifying users who are engaged in household tasks. Furthermore, we show that combining analysis results for features (A), (B), and (C) enables us to identify users who are primarily engaged in income-earning activities, household tasks, or school, based on calling behavior.

4.4 PROTOTYPES OF CALLING BEHAVIOR

In this section, we narrowed the time window from a weekly basis to an hourly basis, to understand trends in calling behavior across gender and occupation types. As a result, we aggregated the call times by hour in this section. We extracted parts-based representations of calling behavior by conducting vector quantization against the time and location distribution of call records. This enabled us to cluster the data into mutually exclusive prototypes [20].

4.4.1 Method for Extracting Prototypes

We employed non-negative matrix factorization (NMF) for vector quantization. NMF was applied to the call records of the 922 mobile users, where the distribution of call records for a single day is expressed as a 72 x 1 matrix. The first set of 24 elements out of 72 consists of hourly counts of call records for Home. The first element is the total number of calls from Hour 0, occurring between 0:00 and 0:59, and the 24th element denotes Hour 23, occurring between 23:00 and 23:59. The next set of 24 elements is structured similarly to the first set, and it accounts for the number of calls from the Primary outside-home location. In a similar manner, hourly counts of calls for other locations, which are any locations except for home and work/school, are captured by an additional set of 24 elements. As a result, we obtained 922 sets of 72 x 1 column vectors. To employ NMF, we solved the equation below by following an algorithm, which allows only additive combinations [21]:

$$V \approx WH$$

where we obtained non-negative matrix factors W and H given a non-negative matrix V. Given 922 sets of 72 x 1 column vectors, the vectors were placed in the columns of a 72 x 922 matrix V. This matrix was approximately factorized into a 72 x r matrix W, and an r x 922 matrix H. Here we selected r = 3 to analyze differences in the major features of calling behavior. We then defined the cost function that evaluates the quality of approximation for iterative updates of W and H. We calculated the distance between two non-negative matrices and measured the square of the Euclidian distance [22]. As a result of repeated iterations, we obtained an optimal matrix factorization.

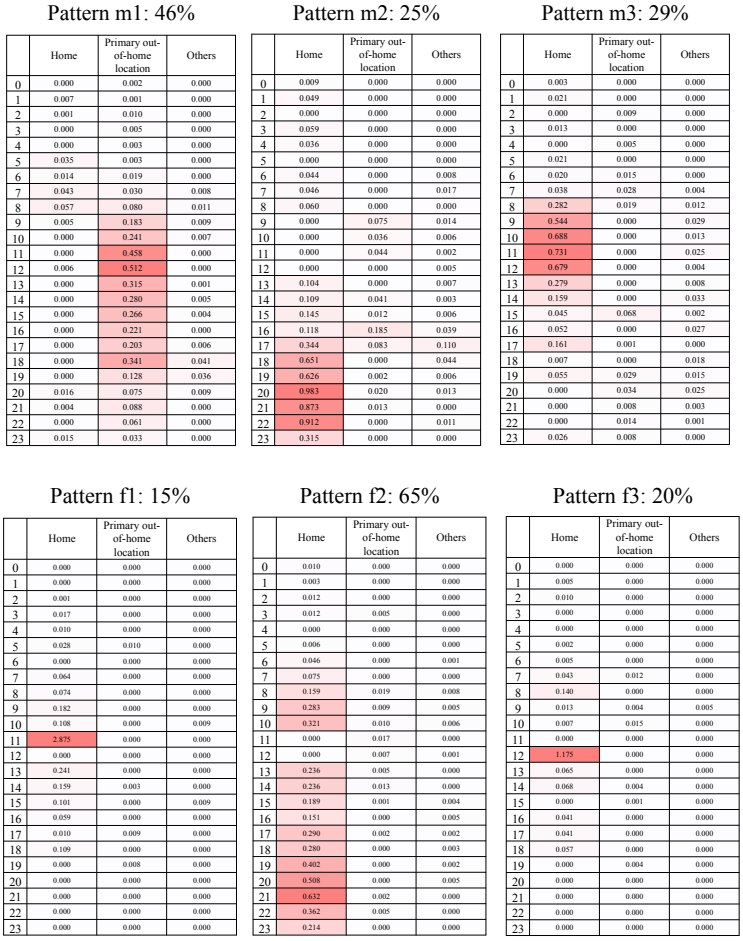
To understand the calling behavior prototypes for males and females, we split the call records obtained from SPACE by gender. Then, we separated the records by primary and non-primary routine days. We assumed that the calling behavior of the primary routine days was different from that of the non-primary routine days, based on the analysis results described in the previous section. We must note that the call records we obtained from SPACE were the aggregation of single-day records for 922 users. That is, the users that generated the records for the primary routine day were not the same as those that generated the non-primary routine day records. To maintain the population composition in terms of income level, we set our survey schedule to randomize the distribution of the days of the week for each income level. Thus, we assume that this condition did not bias our sampling, and did not affect the analysis results.

4.4.2 Prototypes of the Primary Routine Day

Figures 6(a) and 6(b) show three calling behavior prototypes for males and females, whose call records in the SPACE data fall on their primary routine days. Each prototype is expressed in a 24 x 3 matrix, showing time distributions between zero and 23 hours in the rows, and call locations of Home, Primary location outside

the home, and Other in each column from left to right. The intensity of the color indicates the intensity of calls for the specified time band and location. For instance, in each matrix, the cell on the top of the left column represents the intensity of calls for Hour 0 from home. The percentage indicates how significant each component is for explaining the population's call record trends. For example, Pattern m1, whose percentage is 46% and the greatest among the three, is the most dominant pattern for the calling behavior of males on the primary routine day.

As shown in Figure 6, Pattern m1, which is the most dominant pattern for males, shows that males tended to call from Primary outside-home location around midday, peaking in Hour 12. In contrast, the most dominant pattern for females, Pattern f2, shows a peak of Hour 21 at Home. Furthermore, all extracted patterns for females show a higher intensity of calls from Home. This trend follows the one exhibited in Figure 2, which indicates that call locations could be a key to distinguishing the gender of users. Although the trend in call locations partially captures gender differences, frequently recorded hours for the remainder of males, Pattern m2 and m3, are apparently difficult to distinguish from those for females, Patterns f1, f2, and f3. For both males and females, intensive calling hours are Hour 11, Hour 12, Hour 20, and Hour 21 from Home. It is still fair to conclude that the location of calls around midday could be a key to identifying the user's gender. That is, males mostly call from Primary location outside the home and females call from Home around midday on their primary routine days.



Pattern m1: 46%

	Home	Primary out-of-home location	Others
0	0.000	0.002	0.000
1	0.007	0.001	0.000
2	0.001	0.010	0.000
3	0.000	0.005	0.000
4	0.000	0.003	0.000
5	0.035	0.003	0.000
6	0.014	0.019	0.000
7	0.043	0.030	0.008
8	0.057	0.080	0.011
9	0.005	0.183	0.009
10	0.000	0.241	0.007
11	0.000	0.458	0.000
12	0.006	0.512	0.000
13	0.000	0.315	0.001
14	0.000	0.280	0.005
15	0.000	0.266	0.004
16	0.000	0.221	0.000
17	0.000	0.203	0.006
18	0.000	0.341	0.041
19	0.000	0.128	0.036
20	0.016	0.075	0.009
21	0.004	0.088	0.000
22	0.000	0.061	0.000
23	0.015	0.033	0.000

Pattern m2: 25%

	Home	Primary out-of-home location	Others
0	0.009	0.000	0.000
1	0.049	0.000	0.000
2	0.000	0.000	0.000
3	0.059	0.000	0.000
4	0.036	0.000	0.000
5	0.000	0.000	0.000
6	0.044	0.000	0.008
7	0.046	0.000	0.017
8	0.060	0.000	0.000
9	0.000	0.075	0.014
10	0.000	0.036	0.006
11	0.000	0.044	0.002
12	0.000	0.000	0.005
13	0.104	0.000	0.007
14	0.109	0.041	0.003
15	0.145	0.012	0.006
16	0.118	0.185	0.039
17	0.344	0.083	0.110
18	0.651	0.000	0.044
19	0.626	0.002	0.006
20	0.983	0.020	0.013
21	0.873	0.013	0.000
22	0.912	0.000	0.011
23	0.315	0.000	0.000

Pattern m3: 29%

	Home	Primary out-of-home location	Others
0	0.003	0.000	0.000
1	0.021	0.000	0.000
2	0.000	0.009	0.000
3	0.013	0.000	0.000
4	0.000	0.005	0.000
5	0.021	0.000	0.000
6	0.020	0.015	0.000
7	0.038	0.028	0.004
8	0.282	0.019	0.012
9	0.544	0.000	0.029
10	0.688	0.000	0.013
11	0.731	0.000	0.025
12	0.679	0.000	0.004
13	0.279	0.000	0.008
14	0.159	0.000	0.033
15	0.045	0.068	0.002
16	0.052	0.000	0.027
17	0.161	0.001	0.000
18	0.007	0.000	0.018
19	0.055	0.029	0.015
20	0.000	0.034	0.025
21	0.000	0.008	0.003
22	0.000	0.014	0.001
23	0.026	0.008	0.000

Pattern f1: 15%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.001	0.000	0.000
3	0.017	0.000	0.000
4	0.010	0.000	0.000
5	0.028	0.010	0.000
6	0.000	0.000	0.000
7	0.064	0.000	0.000
8	0.074	0.000	0.000
9	0.182	0.000	0.000
10	0.108	0.000	0.009
11	2.875	0.000	0.000
12	0.000	0.000	0.000
13	0.241	0.000	0.000
14	0.159	0.003	0.000
15	0.101	0.000	0.009
16	0.059	0.000	0.000
17	0.010	0.009	0.000
18	0.109	0.000	0.000
19	0.000	0.008	0.000
20	0.000	0.000	0.000
21	0.000	0.000	0.000
22	0.000	0.000	0.000
23	0.000	0.000	0.000

Pattern f2: 65%

	Home	Primary out-of-home location	Others
0	0.010	0.000	0.000
1	0.003	0.000	0.000
2	0.012	0.000	0.000
3	0.012	0.005	0.000
4	0.000	0.000	0.000
5	0.006	0.000	0.000
6	0.046	0.000	0.001
7	0.075	0.000	0.000
8	0.159	0.019	0.008
9	0.283	0.009	0.005
10	0.321	0.010	0.006
11	0.000	0.017	0.000
12	0.000	0.007	0.001
13	0.236	0.005	0.000
14	0.236	0.013	0.000
15	0.189	0.001	0.004
16	0.151	0.000	0.005
17	0.290	0.002	0.002
18	0.280	0.000	0.003
19	0.402	0.000	0.002
20	0.508	0.000	0.005
21	0.632	0.002	0.000
22	0.362	0.005	0.000
23	0.214	0.000	0.000

Pattern f3: 20%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.005	0.000	0.000
2	0.010	0.000	0.000
3	0.000	0.000	0.000
4	0.000	0.000	0.000
5	0.002	0.000	0.000
6	0.005	0.000	0.000
7	0.043	0.012	0.000
8	0.140	0.000	0.000
9	0.013	0.004	0.005
10	0.007	0.015	0.000
11	0.000	0.000	0.000
12	1.175	0.000	0.000
13	0.065	0.000	0.000
14	0.068	0.004	0.000
15	0.000	0.001	0.000
16	0.041	0.000	0.000
17	0.041	0.000	0.000
18	0.057	0.000	0.000
19	0.000	0.004	0.000
20	0.000	0.000	0.000
21	0.000	0.000	0.000
22	0.000	0.000	0.000
23	0.000	0.000	0.000

Figure 6. Time distributions of call locations for (a) male and (b) female on the primary routine day (a) Upper (b) Lower

We then investigated what occupation types and individual income levels represent each prototype described in Figures 6(a) and 6(b). Figures 7(a) and 7(b) describe the distribution of the three patterns illustrated in Figures 6(a) and 6(b) across occupation types Worker, Household, Student, and Others. Interestingly, we can observe that the dominant pattern for males differs according to the occupation type. Regarding Worker, Pattern m1 is the most common pattern, in which users tend to call around midday from Primary outside-home location. Among males, users in the Household and Student categories tended to call at Home around midday as described in Pattern m2. However, the common pattern for females categorized as

Worker, Household, and Student are all the same, Pattern f2, where they tended to call from Home early in the morning and late evening. This indicates that calls from Primary outside-home location around midday can be a key to identifying males who are primarily engaged in income-earning activity.

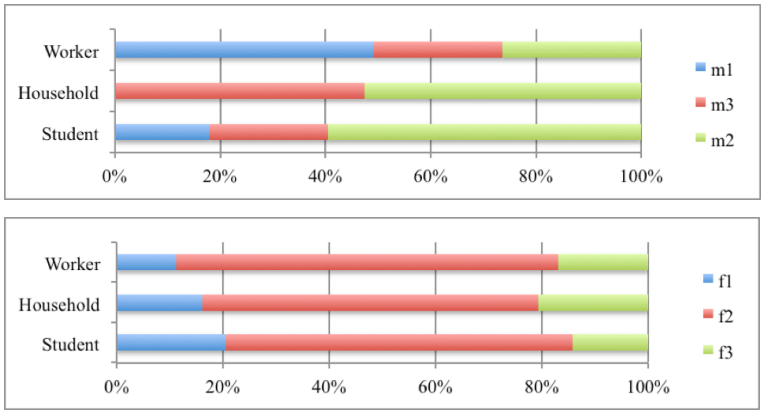


Figure 7. Distribution of occupation types for three principle patterns for (a) female and (b) female on the primary routine day (a) Upper (b) Lower

Figures 8(a) and 8(b) describe the distribution of the calling patterns illustrated in Figures 6(a) and 6(b) for individual income levels. We split call records for males into four groups based on their individual income level. Call records for females were split into three income groups and a separate non-income group, because the majority of the female users did not earn an income. Among males, we observed that their common patterns varied according to their income level. As described in Figure 6(a), the higher income levels contained larger ratios of Pattern m1. Conversely, the lower individual income levels contained larger ratios of Pattern m3. This result is consistent with our field observations where males in lower-income groups tend to be engaged in self-employed jobs whose work locations tend to be their own home. However, no particular trends were observed among females across the income levels. Pattern f2 was the most common calling pattern for all income levels. These trends

imply that calling behavior could be a key to understanding the individual income level for males, but not for females.

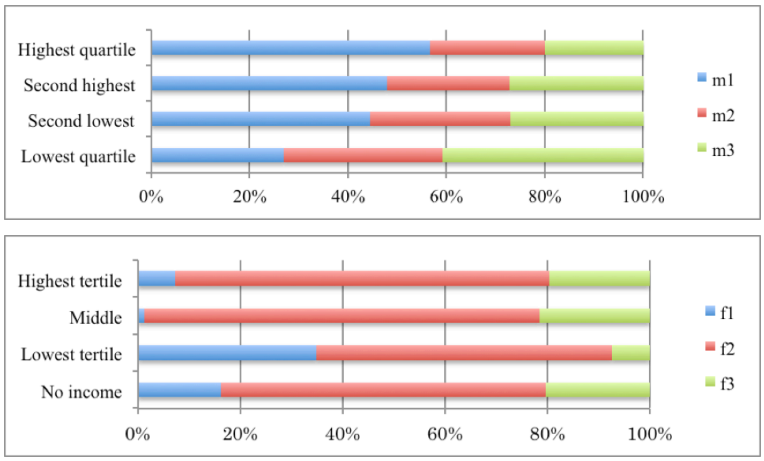


Figure 8. Distribution of individual income levels for three principle patterns for (a) male and (b) female on the primary routine day (a) Upper (b) Lower

It is worth noting that the income level used here is not a household income but individual income, which means we assess how much each person earns annually. Incidentally, we do not observe distinctive differences in calling patterns due to different household income levels when we use the household income level for analysis. We assume this implies that calling behavior strongly reflects the characteristics of individuals, rather than those of households.

4.4.3 Prototypes of the Non-primary Routine Day

Figures 9(a) and 9(b) contain three prototypes composed of time and location distributions of call records. These call records were generated by calls made on non-primary routine days. All patterns for males show that they tended to call from Home on non-primary routine days. By contrast, we observed that some females tended to call from Primary outside-home location on non-primary routine days. As described in Figure 6(b), Home was the most common call location for females on their primary routine days. This is probably because working outside of the home was unlikely to

be the primary routine for the female users. Because females have such opportunities only on limited days (significantly fewer days than males), working outside the home tends to be counted as an activity on the non-primary routine day. This result strongly implies that, among those categorized into Worker, male and female users exhibited opposite call location trends on their primary and non-primary routine days. That is, analyzing the distribution of call locations and the type of day can reveal gender differences among users in the Worker category. We note that the factorization may not be the optimal approach for understanding non-primary routine day trends for our data, because the number of their records is limited fro the type of days.

Pattern m1: 61%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.003	0.007	0.000
4	0.000	0.000	0.000
5	0.000	0.000	0.000
6	0.005	0.000	0.000
7	0.051	0.005	0.000
8	0.174	0.000	0.004
9	0.368	0.063	0.000
10	0.180	0.020	0.031
11	0.000	0.131	0.058
12	0.325	0.041	0.007
13	0.151	0.018	0.000
14	0.424	0.002	0.001
15	0.615	0.011	0.000
16	0.102	0.002	0.032
17	0.912	0.000	0.012
18	1.806	0.002	0.007
19	0.597	0.001	0.000
20	0.982	0.003	0.018
21	1.073	0.000	0.000
22	0.378	0.082	0.000
23	0.081	0.000	0.003

Pattern m2: 16%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.000	0.003	0.000
4	0.000	0.000	0.000
5	1.709	0.000	0.000
6	0.039	0.000	0.000
7	0.000	0.003	0.000
8	0.004	0.000	0.007
9	0.002	0.080	0.000
10	0.002	0.056	0.000
11	0.000	0.051	0.000
12	0.073	0.014	0.000
13	0.010	0.002	0.000
14	0.000	0.004	0.000
15	0.000	0.008	0.000
16	0.008	0.003	0.000
17	0.000	0.000	0.002
18	0.000	0.004	0.003
19	0.001	0.001	0.000
20	0.008	0.001	0.005
21	0.000	0.000	0.000
22	0.003	0.000	0.000
23	0.002	0.000	0.002

Pattern m3: 23%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.036	0.000	0.000
3	0.001	0.000	0.000
4	0.000	0.000	0.000
5	0.000	0.000	0.000
6	0.002	0.000	0.000
7	0.005	0.001	0.000
8	0.060	0.000	0.000
9	0.057	0.000	0.000
10	0.029	0.002	0.000
11	1.109	0.000	0.000
12	0.000	0.000	0.000
13	0.023	0.035	0.000
14	0.159	0.000	0.000
15	0.000	0.000	0.000
16	0.000	0.000	0.000
17	0.000	0.000	0.000
18	0.000	0.000	0.000
19	0.000	0.000	0.000
20	0.000	0.000	0.000
21	0.113	0.000	0.001
22	0.000	0.000	0.022
23	0.000	0.000	0.003

Pattern f1: 36%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.199	0.000	0.000
4	0.000	0.000	0.000
5	0.024	0.000	0.000
6	0.000	0.000	0.000
7	0.078	0.000	0.000
8	0.000	0.000	0.000
9	0.266	0.000	0.000
10	0.053	0.000	0.000
11	0.830	0.000	0.000
12	0.466	0.000	0.000
13	0.000	0.000	0.000
14	0.403	0.000	0.000
15	0.934	0.000	0.000
16	0.603	0.000	0.000
17	0.018	0.000	0.000
18	0.034	0.000	0.000
19	0.165	0.000	0.000
20	0.000	0.000	0.000
21	0.243	0.000	0.000
22	0.000	0.000	0.000
23	0.000	0.000	0.000

Pattern f2: 20%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.000	0.000	0.000
4	0.000	0.000	0.000
5	0.000	0.000	0.000
6	0.000	0.000	0.000
7	0.041	0.000	0.000
8	0.255	0.000	0.000
9	0.013	0.041	0.242
10	0.047	0.000	0.000
11	0.000	0.242	0.000
12	0.003	0.041	0.000
13	0.000	0.242	0.000
14	0.000	0.524	0.000
15	0.000	1.911	0.000
16	0.000	0.450	0.032
17	0.066	0.000	0.000
18	0.000	0.000	0.000
19	0.578	0.000	0.000
20	0.000	0.384	0.000
21	0.000	0.000	0.000
22	0.018	0.000	0.000
23	0.384	0.000	0.000

Pattern f3: 44%

	Home	Primary out-of-home location	Others
0	0.000	0.000	0.000
1	0.000	0.000	0.000
2	0.000	0.000	0.000
3	0.173	0.000	0.000
4	0.000	0.000	0.000
5	0.094	0.000	0.000
6	0.000	0.000	0.000
7	0.084	0.000	0.278
8	0.071	0.000	0.000
9	0.134	0.090	0.000
10	0.000	0.000	0.000
11	0.000	0.000	0.000
12	0.000	0.090	0.000
13	0.150	0.000	0.278
14	0.437	0.232	0.381
15	0.000	0.000	0.103
16	0.000	0.000	0.161
17	0.743	0.000	0.000
18	1.418	0.000	0.103
19	0.054	0.000	0.000
20	1.284	0.000	0.000
21	0.060	0.000	0.000
22	1.290	0.000	0.000
23	0.000	0.000	0.000

Figure 9. Time distributions of call locations for (a) male and (b) female on the non-primary routine day (a) Upper (b) Lower

Figures 10(a) and 10(b) describe the distribution of the calling patterns illustrated in Figures 9(a) and 9(b) for occupation types Worker, Household, and Student. Pattern m1 was the most common pattern for males overall. In contrast, we note that the dominant pattern varied according to the occupation type among females. This indicates that time and location distribution trends on the non-primary routine day varied according to the occupation type among females. Incidentally, as shown in Figure 10(a), we provide no observation for males who are engaged in household tasks on the non-primary routine day in the SPACE data. This reflects the predominant societal norms in Bangladesh, where most males are engaged in income-earning activity while most females perform household tasks.

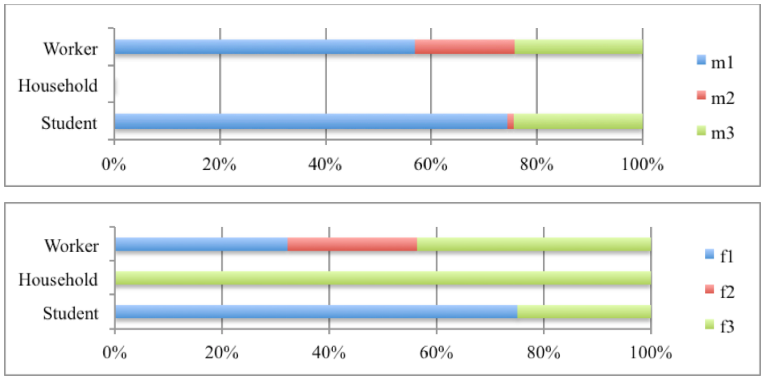


Figure 10. Distribution of occupation types for three principle patterns for (a) male and (b) female on the non-primary routine day (a) Upper (b) Lower

Last, we examined the distribution of calling patterns for individual income levels. The income levels were classified following the same rules used in Figures 8(a) and 8(b). As described in Figure 11(a), the dominant pattern for all income levels among males was Pattern m1, where the intensity of calls was highest at Hour 18. This trend was opposite from that of primary routine days, where time of day and call location distribution could indicate income levels for males. Conversely for females, dominant patterns differed across income levels, as shown in Figure 11(b). Again, the trend was opposite from that of primary routine days. Pattern f3 was common in the middle level, where calls originated on late evenings from Home. However, Pattern f1 was dominant for no-income females where calls tended to be made in the late morning and early afternoon from Home. These trends followed those discussed in Section 3, where males were more flexible on their primary routine days and females were more flexible on their non-primary routine days. Our analysis results indicate that time of day and call location distribution trends are only useful for determining the income levels of female users on non-primary routine days.

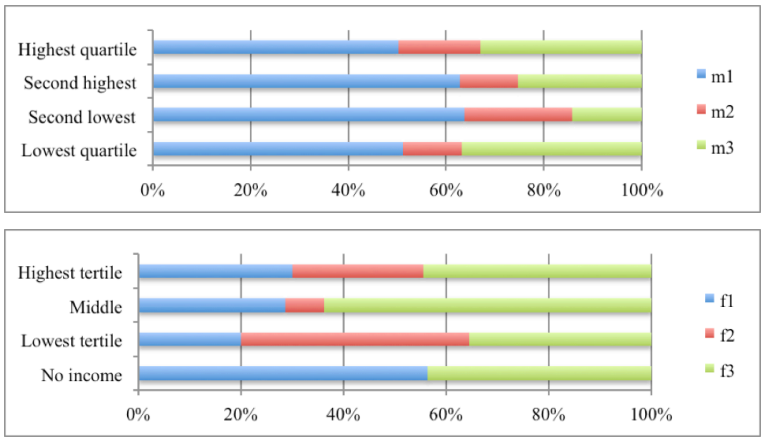


Figure 11. Distribution of individual income levels for three principle patterns for (a) male and (b) female on non-primary routine days (a) Upper (b) Lower

4.5 COMPARING ONE-DAY CALL RECORDS WITH TWO-MONTH CDRS

In this section, we compare basic calling behavior statistics generated from single-day call records from SPACE data and CDRs for 58 volunteers. In the comparison, we employ a set of statistics under two different time frames, referred to as daily and weekly basis time frames. In the daily basis time frame, we generate one set of statistics by calculating averages per day. In the weekly basis time frame, we generate two sets of statistics by separately calculating averages for primary and non-primary routine days. As a result, we can examine the advantages and limitations of capturing calling behavior from survey data.

4.5.1 Calling Behavior by Gender

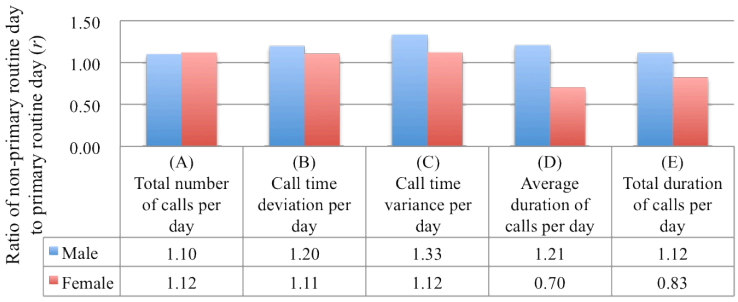
First, we compared SPACE data and CDRs by gender. Table 4(a) and 4(b) present overall calling behavior trends for SPACE data and CDRs in the daily basis time frame. The F/M values represent the ratio of Female (F) to Male (M), which were calculated to compare the trends of (a) SPACE and (b) CDRs. For instance, the F/M value for feature (C) was 0.95 for (a) SPACE and 0.99 for (b) CDRs. That is, both ratios are slightly smaller than unity. This indicates that the feature (C) values for males were slightly higher than those for females for both (a) SPACE and (b) CDRs. By comparing the trends for the remaining features in the two tables, we concluded that their overall trends were similar across the features.

Table 4. Daily basis calling behavior by gender for (a) SPACE data and (b) CDRs (a) Upper (b) Lower

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Male(M)	4.70	2.69	10.65	110.98	458.65
Female(F)	3.85	2.41	10.08	135.79	491.47
F/M	0.82	0.90	0.95	1.22	1.07

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Male(M)	4.07	1.90	8.70	98.98	371.72
Female(F)	3.70	1.85	8.64	113.52	401.11
F/M	0.91	0.97	0.99	1.15	1.08

We then split the CDRs into primary and non-primary routine days. We used the same set of features for *r* that were employed for Figure 4 in Section 4.3. Using CDRs, five features were calculated for males and females in Figure 12. Interestingly, there were two main differences between the trends in the SPACE data and CDRs.



	(A) Total number of calls per day	(B) Call time deviation per day	(C) Call time variance per day	(D) Average duration of calls per day	(E) Total duration of calls per day
Male	1.10	1.20	1.33	1.21	1.12
Female	1.12	1.11	1.12	0.70	0.83

Figure 12. Comparison of calling behavior on the primary and non-primary routine day of CDRs by gender

First, gender differences were more noticeable in terms of call duration, i.e. feature (D) and (E), in the CDRs; conversely, gender differences were more noticeable in terms of the frequency and time distribution of calls, i.e. features (A), (B), and (C), in the SPACE data. Second, differences between the primary and non-primary routine days were more distinctive in the calling behavior of males in the CDRs, while those differences were more significant in the calling behavior of females in the SPACE data. At this time, we are not able to identify the reason we obtained opposite trends when we changed the length of the time frame.

4.5.2 Calling Behavior by Occupational type

Similarly, we compared SPACE data and CDRs by occupational type. Table 5(a) and 5(b) present overall calling behavior trends for SPACE data and CDRs in the daily basis time frame.

Table 5. Daily basis calling behavior by occupation type for (a) SPACE data (b) CDRs (a) Upper (b) Lower

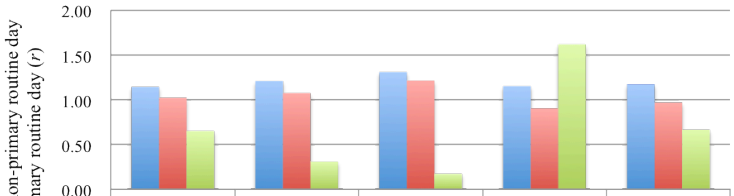
	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Worker(W)	4.73	2.70	10.69	112.34	476.81
Household(H)	3.67	2.40	10.37	137.91	474.40
Student(S)	3.78	2.11	7.61	119.06	383.02
H/W:S/W	0.78 : 0.79	0.89 : 0.78	0.97 : 0.71	1.23 : 1.06	0.99 : 0.80

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Worker(W)	4.12	1.98	9.17	93.65	355.18
Household(H)	3.73	1.84	8.70	126.21	448.38
Student(S)	2.36	1.38	6.22	93.04	201.46
H/W:S/W	0.91 : 0.57	0.93 : 0.70	0.95 : 0.68	1.35 : 0.99	1.26 : 0.57

H/W values represent the ratio of Household (H) to Worker (W) and S/W values represent the ratio of Student (S) to Worker (W), which help compare the trends in the two tables. For example, the ratio of H/W to S/W for feature (C) was 0.97 to 0.71 for (a) SPACE, and 0.95 to 0.68 for (b) CDRs. That is, the ratio of H/W to S/W for (a) SPACE was 1.366, which was calculated dividing 0.97 by 0.71. Similarly, the ratio of H/W to S/W for (b) CDRs was 1.397, which was calculated by dividing 0.95 by 0.68. By comparing the two ratio values 1.366 and 1.397, we observed that the relative trends of three values, W, H, and S, were similar for (a) SPACE and (b) CDRs. By comparing the trends for the remaining features in the two tables, we again concluded that their overall trends were similar across the features, as they were for gender.

We then split the CDRs into primary and non-primary routine day records, as we did for Figure 5 in Section 4.3. Statistics generated under the weekly basis time frame are shown in Figure 13. Apart from the trends observed when analyzing gender

in the previous subsection, we note completely different trends between the CDRs in Figure 5 and the SPACE data in Figure 13. Regarding the different trends derived from the lengthening of the time frame, we will review additional literature and attempt to identify clues for further research in the following subsection.



	(A) Total number of calls per day	(B) Call time deviation per day	(C) Call time variance per day	(D) Average duration of calls per day	(E) Total duration of calls per day
Worker	1.14	1.21	1.31	1.15	1.17
Household	1.02	1.07	1.21	0.90	0.97
Student	0.65	0.31	0.17	1.62	0.66

Figure 13. Comparison of calling behavior on the primary and non-primary routine day for CDRs by occupation type

In sum, our analysis of daily trends showed that similar trends could be observed between the single-day call records, which were gathered from various individuals for primary and non-primary routine days (with the specification of the day type), and the CDRs. This implies that the SPACE field survey data were sufficient to understand the daily traits obtainable from longer-term CDRs. However, when we expanded the time frame from daily to weekly, the single-day records were insufficient to obtain trends that we would expect to extract from longer-term CDRs.

4.5.3 Necessity of Capturing Bounded Nature in Calling Behavior

As described in the previous sections, there are inconsistent trends between SPACE data and CDRs, which appear to be caused by changing time frames. We consider traits that cannot be captured by analyzing collections of single-day call records to be associated with the bounded nature of human activity patterns. This probably results from differences in the sizes of the individual time windows in which

people repeat their routines in different recurrence rates. In fact, different time frames, e.g. 24, 48, and 72 hours, are used to describe the recurrence and temporal periodicity of the bounded nature of human trajectories [5]. Because calling records can trace human mobility patterns to some extent, repeatedly visited locations such as home, work places, and other significant locations are keys to extracting such bounded nature in calling behavior.

By comparing the statistical analysis results from SPACE data and CDRs, we described the advantages and limitations of analyzing calling behavior using collections of single-day call records. We assume that it will be necessary to further examine the routines, which are expressed by returning to a few limited locations on a regular basis within different time frames. One possible method to capture the various time frame lengths would be to expand the field survey's interview time frame from a single-day to three days. However, from a practical standpoint, it may be difficult to obtain meaningful responses by asking respondents to recall activity that occurred during the previous three days. Thus, further investigations are necessary to improve our approach, by developing a method that enables us to capture the bounded nature from calling records.

4.6 ESTIMATION OF PERSONAL ATTRIBUTES

This section describes estimation models and results for the mobile users' personal attributes. Additionally, we revisit the estimation models for the presence of the unobservable and gender, which were discussed in Chapter 3. We previously discussed that collecting call records for a relatively long-term is practically uneasy. In this chapter, we estimate the presence of the unobservable using shorter-term call records, i.e. three-day and one-day call records collected through the field survey.

4.6.1 Data

Considering the necessity of capturing the bounded nature of human behavior, we use thee datasets described in Table 6 for constructing the models. Data (A) is the same dataset used in the previous chapter. Data (B) and (C) are new datasets collected through another field survey, SPACE 2014. The survey was conducted to track the same mobile users, surveyed by SPACE 2013, to collect additional information on calling behavior and lifestyles. Among 922 mobile users, we were able to track 788 mobile users and the attrition rate is 14.5%. The survey collected the call records of consecutive three days, which start from the latest day with any call records for the date of visit for the survey, while SPACE2013 collected only one-day call records. Additionally, SPACE 2014 specifies whether the mobile device replace the latest existing call records if the device is used to make a new call to the same contact. Among 722 mobile users, 359 people (52%) use the device, which does not replace the latest existing call records to the same contact. The remainder includes mobile users, whose devices replace the latest records to the same contact and are not capable of maintaining call records of the surveyed three days. We also found mobile devices with very limited functionality, such as no clock functionality and incapable of recording the duration of each call. Table 7 compare the number of call records to be collected for the consecutive three days for the survey. For Data (B), we extract call records of consecutive three days, which include Friday, to examine differences in calling behavior on the weekday (non Friday) and non-weekday (Friday). It reduces the sample size to 268. We do not classify seven days of a week into the primary routine day and non-primary routine day because the estimation of the type of days increases the uncertainty in estimation results. For Data (C), we extract call records of 291 mobile users whose latest day of the consecutive three days is a weekday.

Table 6. Data used for the model construction

	Data (A)	Data (B)	Data (C)

Length of data period		Two months	Three days	One day
Data source		SPACE2013	SPACE2014	SPACE2014
The type of days	***Non-Friday***	Yes	Yes	Yes
	Friday	Yes	Yes	No
Duration of calls		No	Yes	Yes
Sample size		59	268	291

Table 7. Comparison of the number of call records to be collected from the mobile device for consecutive three days

The number of calls to be collected from the device		**The device can maintain all records of consecutive three days including the duration of each call**	
		Yes (52%)	**No (48%)**
Mean		6.89	4.23
Median		7.05	5.33
Standard Deviation		5	3
Mean by the type of location	***Home***	3.4	2.2
	Work (School)	2.8	1.5
	Other	0.7	0.5

4.6.2 Revisiting the estimation model for the presence of the unobservable population in the household

First, we explain the estimation models for the presence of the unobservable population in the household using Data (B), three-day call records. Data (C) are not used for this because we cannot capture the difference in calling behavior between non-Friday and Friday. As for the definition of the unobservable, we again use the children aged 10 and under for the comparison. Here, we do not use the gender-pooled model, which outperforms the gender-wise models in Chapter 3, because we consider estimating the number of children both from males and females will double-count children in the household. Tables 8(a) and 8(b) describe the estimation results for the presence of children for males and females respectively. The estimation models are slightly better than random guess but are not very successful. Overall, the model for females outperforms the model for males. As shown in Table 9(b), the model for females is inferior to extracting those without the presence of children and tends to retrieve those without the presence of children as those with the presence of children. It means the model tends to overestimate the number of children. Comparing with the estimation models using two-month CDRs (Table 11 in Section 3), the

results of models using three-day call records are overall better with greater values for accuracy. This is likely because we add features, deriving from the duration of calls.

Table 8(a). Estimation results for the presence of children for males

Presence of children	Accuracy	Precision	Recall
Yes	0.571	0.588	0.635
No		0.550	0.500

Table 8(b). Estimation results for the presence of children for females

Presence of children	Accuracy	Precision	Recall
Yes	0.633	0.667	0.778
No		0.568	0.472

Table 9(a). Confusion matrix for the estimation of the presence of children for males

		Condition	
		Yes	***No***
Test outcome	***Yes***	47	33
	No	27	33

Table 9(b). Confusion matrix for the estimation of the presence of children for females

		Condition	
		Yes	***No***
Test outcome	***Yes***	56	28
	No	19	25

Tables 10(a) and 10(b) show significant features used for predicting the presence of children. Two models are overall similar with small difference in the combination and significance rank of features. We found features generated from the duration of calls are useful for estimating the presence of children for both males and females.

Table 10(a). Significant features for predicting the presence of children for males

Rank of significance	Feature	Description

1	***v1***	Average time to start calling in minutes
2	***v2***	Average duration per call
3	***v3***	The ratio of the total duration of calls from Work divided by the total duration of calls from Home
4	***v4***	Total duration of calls
5	***v5***	The latest time to start calling in minutes
6	***v6***	The ratio of the number of calls on Friday divided by the number of calls on non-Fridays
7	***v7***	Total duration of call from Work
8	***v8***	The ratio of the total duration of calls from Other divided by the total duration of calls from Home
9	***v9***	Total duration of calls from Home
10	***v10***	The ratio of the number of calls from Work divided by the number of calls from Home
11	***v11***	Standard deviation of times to start calling
12	***v12***	Standard deviation of hourly probability of making calls
13	***v13***	The proportion of the number of calls between 18:00 to 18:59 among the total number of calls during the three days

Table 10(b). Significant features for predicting the presence of children for females

Rank of significance	**Feature**	**Description**
1	***v1***	Average duration per call
2	***v2***	The ratio of the total duration of calls from Work divided by the total duration of calls from Home
3	***v3***	Total duration of calls
4	***v4***	The ratio of the total duration of calls from Other divided by the total duration of calls from Home
5	***v5***	Average time to start calling in minutes
6	***v6***	The latest time to start calling in minutes
7	***v7***	Standard deviation of times to start calling
8	***v8***	The ratio of the number of calls on Friday divided by the number of calls on non-Fridays
9	***v9***	The proportion of the number of calls between 18:00 to 18:59 among the total number of calls during the three days
10	***v10***	The ratio of the number of calls from Other divided by the number of calls from Home
11	***v11***	The ratio of the number of calls from Work divided by the number of calls from Home

Tables 11(a) and 11(b) show considerable differences in the duration of calls according to the presence of children in the household. Interestingly, both for males and females with the presence of children, (B) Total duration of calls from Work is much longer while (D) Total duration of calls is shorter that those without the presence of children. We consider it reflects the fact that more people are engaged in income-earning activity when they have small children in their households as we discussed in Chapter 3. Additionally, (A) Total duration of calls from Home is almost double for those without the presence of children in the household. Therefore, we emphasize

Table 11(a). Comparison of the duration of calls according to the presence of children within the household among males (duration in second)

Presence of children	Total duration of calls by the type of location			(D) Total duration of calls	(E) Average duration per call
	(A) Home	(B) Work	(C) Other		
Yes	284	541	105	930	86
No	543	383	51	976	127

Table 11(b). Comparison of the duration of calls according to the presence of children within the household among females (duration in second)

Presence of children	Total duration of calls by the type of location			(D) Total duration of calls	(E) Average duration per call
	(A) Home	(B) Work	(C) Other		
Yes	437	66	27	530	98
No	731	27	34	791	167

Besides the estimation of the presence of children itself, there is one issue we need to consider when we want to obtain the estimate on the number of children from the population of CDRs. Counting the number of children, which is calculated from the presence of children in the household both for male and female mobile users, would double-count the number of children in the same household. Recall that we discussed the mobile phone ownership within the household in a typical household in Dhaka in the previous chapter. There is one male mobile user, who is the head of a household, and sometimes his spouse also has a mobile phone. It means that counting the number of children, which is obtained from the estimation model for males, is sufficient. As shown in Tables 8(a) and (b), however, the estimation model for females outperforms the model for males. It is understandable because the behavior of females is considered to be affected by the presence of children in the household.

4.6.3 Estimation of gender

Then, we describe the estimation model for the gender of mobile users using three-day records and one-day call records (Data (B) and (C)). Gender estimation is a necessary step for estimating the presence of children because we use only male mobile users for the estimation. Tables 12(a) and 12(b) show the estimation results,

and Tables 13(b) and 13(c) show confusion matrices for gender estimation, by using three-day call records and one-day records respectively.

Table 12(a). Estimation results for gender using three-day call records

Gender	Accuracy	Precision	Recall
Male	0.806	0.900	0.707
Female		0.741	0.914

Table 12(b). Estimation results for gender using one-day call records

Gender	Accuracy	Precision	Recall
Male	0.869	0.876	0.759
Female		0.753	0.872

Table 13(a). Confusion matrix for the gender estimation using three-day call records

		Condition	
		Male	***Female***
Test outcome	***Male***	99	11
	Female	41	117

Table 13(b). Confusion matrix for the gender estimation using one-day call records

		Condition	
		Male	***Female***
Test outcome	***Male***	120	17
	Female	38	116

Contrary to our expectations, estimation results with one-day call records are slightly better than those of three-day call records. This is probably because three-day call records include the call records of Friday, which may cause some fluctuations in feature. In fact, we initially incorporated some features in the model, which capture difference in calling behavior on non-Friday and Friday. However, these are not effective in the model and thereby dropped in the course of model construction. Furthermore, these results are better than the result of the estimation model using two-month CDRs, Data (A) (Table 9 in Chapter 3), with overall greater values of accuracy, precision, and recall. It indicates that the routineness of daily activity, which is considered to be captured through call records for more than one day, is not essential to extract differences in calling behavior according to gender.

Tables 14(a) and 14(b) shows the features used for predicting gender using three-day call records and one-day call records respectively. The types of call locations, particularly Home and Work, are significant for gender estimation. For instance, four most significant features in models, *v1*, *v2*, *v3*, and *v4*, are related to call records from Home and Work. It indicates that the features, which can capture the partial view of mobile users' activity patterns, are useful. Therefore, the accuracy of the location estimation is very important for gender estimation.

Table 14(a). Significant features for predicting gender using three-day call records

Rank of significance	Feature	Description
1	*v1*	The ratio of the number of calls from Work to the number of calls from Home
2	*v2*	Proportion of the number of calls from Work among the total number of calls
3	*v3*	Total number of calls from Work
4	*v4*	Proportion of the number of calls from Home among the total number of calls
5	*v5*	Average duration per call
6	*v6*	Total number of calls
7	*v7*	Total duration of calls
8	*v8*	Total number of calls from home
9	*v9*	The latest time to start calling in minutes
10	*v10*	Standard deviation of times to start calling
11	*v11*	The ratio of the number of calls from Other to the total number of calls
12	*v12*	Average time to start calling in minutes

Table 14(a). Significant features for predicting gender using one-day call records

Rank of significance	Feature	Description
1	*v1*	Proportion of the number of calls from Home among the total number of calls
2	*v2*	Total number of calls from Home
3	*v3*	The ratio of the number of calls from Work to the number of calls from Home
4	*v4*	Proportion of the number of calls from Work among the total number of calls
5	*v5*	Total number of calls from home
6	*v6*	The ratio of the number of calls from Other to the number of calls from Home
7	*v7*	Total duration of calls
8	*v8*	Average time to start calling in minutes
9	*v9*	The latest time to start calling in minutes
10	*v10*	Average duration per call
11	*v11*	Standard deviation of times to start calling
12	*v12*	Total number of calls

4.6.4 Estimation of personal attributes

Finally, we describe the estimation model for the personal attributes of mobile users. We use four labels for the attribute, Workmale, Housewife, Student, and Other.

The labels are determined based on the principal population groups of the living population and CDRs considering that we the classification result is used for creating Dynamic Census. Workmale is a male who is engaged in income-earning activity. Housewife is a female who are married and whose main routine activity is doing household tasks including caring her family members. We select these labels because they are the two of three principal population groups for the living population of Dhaka according to Chapter 2. Student is a student, who also belongs to the principal population group for the living population. However, they are not the predominant one for the population of CDRs. Table 15 shows the distribution of the four population groups among the total population for the three data. The proportion of Student is very low while the proportions of Workmale and Housewife are high. It indicates that estimation models to predict the personal attribute using these data have following two problems. One is that the algorithm of classification, which tries to lower the error rate of the overall estimation results, put less weight on the estimation error for a small number population group. It means that the estimation model we construct will well predict the Workmale and Housewife, but Student. The other problem is that the small sample size of the training data lowers the dimension of the data. It causes the over-fitting of the model. To resolve the first problem, we propose to control the proportion of Workmale and Housewife by reducing the sample size for the training data. Given the sample size for Student is N, we limit the sample size for Workmale and Housewife to 3×N. For instance, the sample size for Workmale is 30 when that of Student is 10. The samples of Workmale and Housewife, whose size is 3×N, are randomly sampled from the population. The random sampling is done for the sample of each label for ten times if the original population is more than 3×N of the sample size of Student. We compare the results with and without controlling the sample size, and provide best efforts are provided in Tables 16(a), 16(b), and 16(c). Tables 16(a) shows the result without controlling the sample size because it did not

improve the accuracy of estimation results for Student. Tables 16(b) and 16(c) show the result with controlling the sample size. Tables 17(a), 17(b), and 17(c) are the confusion matrices of the results for Tables 16(a), 16(b), and 16(c) respectively. We consider the latter problem is inevitable under this condition.

Table 15. Distribution of the type of personal attributes by dataset

		(A)	(B)	(C)
Length of data period		Two months	Three days	One day
Sample size		55	268	291
Distribution of personal attributes	***Workmale***	18 (33%)	126 (47%)	142 (49%)
	Housewife	23 (42%)	100 (37%)	100 (34%)
	Student	3 (5%)	7 (3%)	9 (3%)
	Other	11 (20%)	35 (13%)	40 (14%)

Table 16(a). Estimation results for the personal attribute using two-month CDRs

Personal attribute	Accuracy	Precision	Recall
Workmale	0.71	0.55	0.61
Housewife	0.69	0.60	0.78
Student	0.93	0	0
Other	0.84	0.75	0.27

Table 16(a). Estimation results for the personal attribute using three-day call records

Personal attribute	Accuracy	Precision	Recall
Workmale	0.77	0.62	0.62
Housewife	0.61	0.57	0.81
Student	0.89	0.33	0.14
Other	0.64	0.38	0.29

Table 16(a). Estimation results for the personal attribute using one-day call records

Personal attribute	Accuracy	Precision	Recall
Workmale	0.79	0.63	0.70
Housewife	0.67	0.47	0.82
Student	0.89	0.40	0.22
Other	0.63	0.20	0.07

Table 17(a). Confusion matrix for the personal attribute estimation using two-month CDRs

		Condition			
		Workmale	***Housewife***	***Student***	***Other***
Test outcome	***Workmale***	11	6	0	1
	Housewife	4	18	1	0
	Student	2	1	0	0

	Other	3	5	0	3

Table 17(b). Confusion matrix for the personal attribute estimation using three-day call records

		Condition			
		Workmale	***Housewife***	***Student***	***Other***
Test outcome	***Workmale***	13	3	1	4
	Housewife	1	17	0	3
	Student	1	2	1	3
	Other	6	8	1	6

Table 17(a). Confusion matrix for the personal attribute estimation using one-day call records

		Condition			
		Workmale	***Housewife***	***Student***	***Other***
Test outcome	***Workmale***	19	5	0	3
	Housewife	1	22	1	3
	Student	3	2	2	2
	Other	7	16	2	2

Overall, the results for Workmale and Housewife outperform random guess (0.25) for three data. Furthermore, features, which are highly significant for the estimation of the personal attribute, are common to those of the gender estimation. It is most likely because of strong social norms for gender-wise activity, i.e. most males in CDRs are engaged in income-earning activity and most females in CDRs are housewives. However, the results for Student and Other are not very successful. We are afraid that it is because the sample size is not large enough to capture the difference in calling behavior of Student from Workmale and Housewife. Initially, we expected that Student might be estimated as Workmale because of their routine activity, i.e. spending specific time at home and school everyday. However, it is not reflected to the calling behavior even for the long-term data such as two-month CDRs. So, further examination is necessary to improve the estimation model for the personal attribute particularly for Student.

Table 18(a). Significant features for predicting the personal attribute using two-month CDRs

Rank of significance	**Feature**	**Description**
1	***v1***	The ratio of the number of calls on Friday to the number of calls on non-Friday

2	***v2***	The average time of calls in hour
3	***v3***	The proportion of the number of calls from Home on non-Friday
4	***v4***	The proportion of the number of calls from Home on Friday
5	***v5***	The proportion of the number of calls from home
6	***v6***	Total number of calls
7	***v7***	The number of calls from Home on Friday
8	***v8***	The ratio of the number of calls on Friday to the number of calls on non-Friday
9	***v9***	Standard deviation of the hourly number of calls
10	***v10***	Standard deviation of the hourly probability of making calls
11	***v11***	Total number of calls from Other on Friday

Table 18(b). Significant features for predicting the personal attribute using three-day call records

Rank of significance	Feature	Description
1	***v1***	The ratio of the total duration of calls from Work divided by the total duration of calls from Home
2	***v2***	Total duration of calls
3	***v3***	The ratio of the total duration of calls from Other divided by the total duration of calls from Home
4	***v4***	The latest time to start calling
5	***v5***	Total duration of calls from Work
6	***v6***	Total duration of calls from Home
7	***v7***	The ratio of the number of calls from Work divided by the number of calls from Home
8	***v8***	The proportion of calls from Work to the total number of calls
9	***v9***	The proportion of calls from Home to the total number of calls

Table 18(c). Significant features for predicting the personal attribute using one-day call records

Rank of significance	Feature	Description
1	***v1***	The ratio of the total duration of calls from Work divided by the total duration of calls from Home
2	***v2***	Total duration of calls
3	***v3***	The latest time to start calling
4	***v4***	The ratio of the number of calls from Work divided by the number of calls from Home
5	***v5***	The ratio of the total duration of calls from Other divided by the total duration of calls from Home
6	***v6***	Total duration of calls from Home
7	***v7***	Total duration of calls from Work
8	***v8***	The proportion of calls from Home to the total number of calls
9	***v9***	The proportion of calls from Work to the total number of calls

4.7 SUMMARY

In this chapter we identified calling behavior traits that can distinguish gender and occupation types by comparing the analysis results of SPACE data with those of CDRs. Analysis results suggest that a higher ratio of calls from home can be a key to distinguishing females from males. Females tended to call from home around midday on their primary routine day. In addition, the variability in the frequency and time

distribution of calls from female users was more distinctive, according to the type of day. That is, constant frequency and time distribution of calls throughout the week were keys to identifying male users. Regarding the average duration per call and total duration of calls per day, females tended to use the phone more often on the non-primary routine day, whereas trends for males did not exhibit such variability, according to the type of day. Specifically for males on the primary routine day, the higher the individual income level, the higher the probability that they would initiate calls from their primary outside-home locations around midday. The lower the income level of the user, the higher the probability that they would initiate calls from home in the morning. Conversely, there were no distinctive differences in primary routine day calling behavior for females related to their individual income level. As the basis for this analysis, we introduced the concept of weekly activity patterns, which specifies whether the day of the call corresponds to a day on which the user is engaged in their primary routine. Analysis results suggest that the type of day could be also a key for extracting traits from call records. By comparing the statistical analysis results of SPACE data and CDRs, we also concluded that the current approach is not sufficient for capturing the regularity of individual calling behavior within different time frames.

We also provided experimental results of estimating gender, presence of children in the household, and personal attribute of mobile users, employing Random Forest. For the estimation, we used three datasets whose length of data acquisition periods differs for examining the necessity of capturing the regularity of individual calling behavior within different time frames For the estimation, we found that features, capturing the difference in calling behavior between Friday and non-Friday, do not greatly affect the estimation accuracy of gender and personal attribute. While, these features are important for estimating the presence of children in the household. That is, capturing the regularity of individual calling behavior is important for

estimating the presence of children but gender and personal attribute. Regarding the estimation of the personal attribute, we found our model is not successful in retrieving users whose population proportion is relatively smaller, i.e. students, among the total population. After controlling the population proportion for estimation, the performance was improved slightly and has a lot of room for improvement.

With our work, we exploited the potential of deriving personal attributes from anonymized CDRs. Although experiments were performed with a limited number of CDRs, the techniques developed in this study are capable of extracting gender and occupational traits from large-scale CDRs. Experimental results infer that our approach is capable of constructing the estimation models of gender and personal attributes using anonymized CDRs. However, current models developed with small sample size data have over-fitting problems. So, increasing the sample size for model construction is necessary for more robust models.

Chapter 5 Representativeness: Estimation of the unobservable population in CDRs

5.1 BACKGROUND

With the increasing trend in the growth of global mobile subscriptions, the mobile penetration rate is expected to reach 96% by the end of 2014. The number of subscriptions in the developing world is estimated to represent more than three-quarters of the total, as the speed of growth these areas is almost twice that in the developed world [1]. In some developing countries, it is not uncommon for people to have access to mobile phones even when they do not have bank accounts, electricity, or access to clean water [2]. The mobile phone is arguably one of the most ubiquitous platforms and prominent infrastructures that will allow us to understand and address various issues in the developing world. This is because mobile infrastructure is already more developed than other basic infrastructures and can link large numbers of people with devices to information and technologies distant from them.

Studies of human travel patterns have greatly advanced owing to the rapid diffusion of ubiquitous devices, which generate large-scale spatiotemporal datasets such as GPS logs and call detail records (CDRs). With the capability to trace mass individual trajectories, the majority of these studies focus on quantitating properties of human mobility patterns. The features of repeatedly visited locations are also estimated, taking social norms into consideration. Utilizing the bounded nature of human mobility patterns [3], significant locations such as homes and workplaces can be estimated using spatiotemporal data [4][5]. Apart from the features of trajectories and visited locations, some research attempts to link location histories—derived from GPS logs—to user attributes by assuming that people who have similar location histories share similar interests and preferences. These studies measure user similarity based on the sequence properties of people's trajectories and the hierarchical properties of their location histories [6].

However, the movement patterns described in these studies show the trajectories of crowds because the data are anonymized. Therefore, their major fields of application are concentrated in transportation, where quantitating the volume and speed of mobility can contribute to improving transportation planning and policy interventions. Further, it is increasingly being noted that the population captured by ubiquitous device data is not representative and therefore does not provide an accurate depiction of the general population, because such data can only capture device users. In fact, recent research examines heterogeneous mobile phone owners according to user attributes, such as gender and socio-economic status, comparing mobile users to the general population [7][8]. This allows us to understand that such heterogeneity impacts the estimation results of human mobility analyses using CDRs [9]. This means that the interpretation and analysis of results may be misleading if there is no clear understanding of which parts of society the data represent. This constraint significantly limits the application of data when the population composition of the data matters to the purpose of the application. Although such limitations have been revealed to some extent, thus far, there are no established methods to help us address this issue in the analysis of biased data. Therefore, in this study, we attempt to answer following research questions:

- How dose the population composition of mobile users and non-users differ? How is the difference taken into account when the results of CDR analysis are used to address societal issues?
- Are there any ways to help understand non-mobile users, who are not included in CDRs, through analysis results of CDRs?
- Is it possible to identify the presence of non-mobile users from the calling behavior of mobile users?
- Are there any ways to estimate the number of living population in the area covered by CDRs?

First, we describe the population composition disparities between two groups of people. One is those who are captured by CDRs and the other is those who are unobservable in CDRs but included in the household of the population in CDRs. That is, we discuss the population, which belongs to *Household A*. To do so, we conducted a field survey to collect information on the personal attributes of mobile phone users with service from a telecommunications company in Bangladesh and the members of their households. Based on an analysis of the data, we provide sets of descriptive statistics for the gender, role in the household, and age group of those captured by the CDRs and those who are unobservable in CDRs. Second, we attempt to find clues in the calling behavior of mobile users that indicate the presence of the unobservable population in their household. We suggest that key features of calling behavior enable us to identify hidden properties in the CDRs even when the data are anonymized. Third, we provide the results of an experimental study to identify the presence of the unobservable population and discuss the potential application of our estimation model to large-scale CDRs. Last, we introduce an approach to estimate the number of the living population of the area covered by CDRs. To do so, we try to estimate the number of populations, which belong to *Household B*, sum up with the populations, which belong to *Household A*.

The contributions of our work are described below:

- Descriptive analyses of mobile users and the members of their households, who are not mobile phone users and are therefore unobservable in CDRs, are provided.
- Calling behavior trends according to the presence of those unobservable in CDRs are provided. These traits can serve as clues to understand the unobservable population in CDRs based on the calling behavior of mobile users.
- Approaches to identify the presence of the unobservable in the household of the mobile users under a supervised learning framework using CDRs are provided.

- Approaches to estimate the number of the living population of the area covered by CDRs. To do so, we introduce an approach to estimate the number of households, which do not include any mobile users of *the operator*, by using the distribution of buildings in the area.

The remainder of this chapter includes two main parts, which are differentiated according to the population under the analysis. Section 5.2 explains the data used for this chapter. Sections 5.3 to 5.5 are the first main part. These sections discuss the population, which belongs to the household of the mobile users of *the operator*. Section 5.3 explains the personal attributes and main activity of mobile users and the unobservable in CDRs, who belong to the household of the mobile users of *the operator*. Section 5.4 describes the types of calling behavior that can indicate the presence of those unobservable in the household of the mobile users. Section 5.5 provides the results of our experimental study to identify the presence of the unobservable population from CDRs. In addition, we introduce an approach to estimate the number of populations, which includes the mobile users of *the operator*. Sections 5.6 and 5.7 are the second main part. These sections discuss the population, which belong to the household without any mobile users of *the operator*. In addition, Section 5.7 introduces our approach to estimate the number of the living population in the area covered by CDRs. The final section includes our conclusions.

5.2 DATA

5.2.1 Data used for estimating the number of households which include mobile users of *the operator*

In this chapter, we use three datasets. One is the SPACE data and CDRs from 58 volunteers, which are the same data used in Chapter 1. The SPACE data consist of 3,288 respondents, including 922 mobile users. In addition to one-day call records for mobile, SPACE collected information on the mobile ownership of all household members, including non-mobile users among the surveyed households. As mentioned

before, we employ two-stage stratified sampling based on land use and household income levels, including slum. We consider it crucial to include the slum population in the target population because this population is increasing. In fact, the slum population in Dhaka more than doubled, to an estimated 3.4 million, between 1995 and 2005, and is still increasing, while the total population increased from nine million to 13 million during this 10-year period [16]. Furthermore, this population group is not well captured by official statistics in general. One of the significant aspects of CDRs is that they can link anyone who uses a mobile phone regardless of status or living conditions. This enables us to understand certain populations that are not listed in official records but do exist de facto.

We need to note that the SPACE data do not represent the general population of each income group due to the sampling condition that households without mobile phone users of the operator are rejected from the sample. The ratios of rejection in the high-, middle-, and low-income groups are 23%, 31%, and 28%, respectively. It indicates that the operator is relatively popular among higher-income populations. However, the SPACE data represent the population of households with mobile users of the operator for each income group, which almost fulfills our study's purpose. There were different rates of participation in our survey according to income level. The rates were 30%, 40%, and 48% of households we approached in high-, middle-, and low-income groups, respectively.

We use two months of CDRs for 58 mobile users. In addition to the call record information, we obtained information on household structures, types of activity, and significant locations from the users. Their household incomes skew to higher levels: three-quarters are from households in the higher- and middle-income populations. Forty-seven percent of them are male and 40% are engaged in income-earning activity. Roles within the household are distributed as follows: 47% are household

heads; 43% are the spouses of heads; and the remainder includes the children, parents, and extended family members of heads.

5.2.2 Data used for estimating the number of households which do not include mobile users of *the operator*

In Chapters 3 and 4, we used the field survey data, which were collected through SPACE 2013 and SPACE 2014. The main purpose of these surveys are to understand the calling behavior of mobile users and the population structure of both mobile users and non-mobile users among the households, which include at least one mobile user of the specified telecommunications company. In this chapter, we introduce another field survey data, Small-scale Census (SSC) data, to understand the mobile ownership, the population structure of the actual living population, and the user distribution of all telecommunications companies. SCC was conducted as part of SPACE 2014. In this section, the framework of the SCC and the descriptive statistics on the survey site are described.

Site identification for data collection

Small-scale Census (SSC) was conducted as part of SPACE2014 in December 2014 to obtain the magnification factor for a given area to calculate the distribution of demographic attributes among mobile users and non-mobile users. We use the voronoi area as the smallest unit area of the survey because the location determined through CDRs is based on the cell phone antenna location. Among all voronoi areas in Dhaka, one voronoi area, which overlaps the 15 PSUs of SPACE, was selected as the representative with a fare mix of several types of land use and inclusion of all income levels, employing a five-stage selection process.

In the first stage, voronoi areas in five PSUs (Savar, Narayanganj, Tongi, Zinjira, and Dakhin Khan), which are classified as the municipalities and outside the

urban area, are dropped. This is because of too larger size of voronoi areas, and the homogeneity of income levels and land use. These areas are mostly residential areas of middle or lower income people. Generally, the size of a voronoi is determined to cover similar numbers of populations and the lower population density means larger size of voronoi area. We consider it increases the time and financial costs to survey the same number of respondents for the survey compared to the smaller voronoi area. In the second stage, we dropped voronoi areas in three PSUs within DCC (Motijheel, Rampura, and Kafrul) where the density of antenna is as low as that of the suburban areas. In the third stage, one PSU (Tejgaon Industrial Area) was dropped because large part of the area is used for the industrial activities and majority of people residing in slum areas in an isolated manner. In the fourth stage, voronoi areas of four PSUs (Lalbag, Sutrapur, Ramna, and Dhanmondi) are dropped because of the lack of variety in income levels and land use. In the final stage, we have two candidate PSUs, Pallabi and Mirpur. These two PSUs have similar characteristics and only difference is the income level of the majority of populations. The majority of people in Mirpur are of middle or higher income levels while there is a certain amount of the slum population in Pallabi. Taking account of the significance of capturing the slum population for this survey, we choose Pallabi as our survey site. In Pallabi, there is one voronoi, which includes all income level buildings within the boundary. Therefore, the voronoi is chosen as our survey site. Figure 1 shows the distribution of buildings within the voronoi area, which is surrounded by a hexagon. The voronoi area includes large blank areas, which seem to indicate no buildings. However, approximately 40% of the area is, in fact, occupied by slums. Therefore, we manually mapped the slum areas and visited all households during the survey.

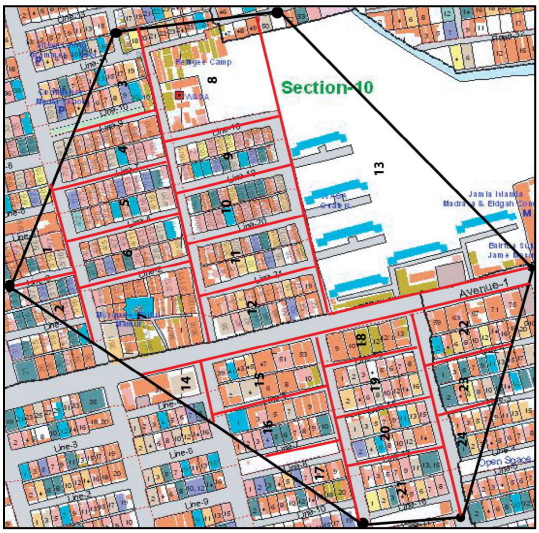


Figure 1. Distribution of buildings in the survey site

Survey structure

SSC consists of two parts; one is census part and the other is building profile part. Below is the description of the each part of the survey;

- Census

Examine the distribution of demographic attributes by income level for a selected area. We surveyed the entire households of the area, which includes mobile phone users of all telecommunication operators and non-mobile phone users. For each household, information on roster, mobile phone usage including the name of telecommunications company to subscribe, and the type and location of main activity was collected.

- Building profile

Investigate the distribution of buildings, including the type of usage for every unit of buildings, by income level. We consider the number of populations obtained through the census part is equivalent to the nighttime population because the home location of the surveyed household member is in this area. Besides, we assume that the daytime population is calculated by summing up a part of household populations from the census part based on the location and type of main activities, and those who

are engaged in commercial activities in the areas during the daytime. So, we additionally surveyed the daily average number of workers and visitors for each unit, including main operation hours if part of a building is used for non-residential purpose. Throughout SCC, the level of income is evaluated based on the type of buildings so that we can estimate the number of population based on the distribution of buildings. The criterion is the same as we used for SPACE 2013 and SPACE 2014.

Survey site

The survey site includes total of 367 buildings, among which one building was not surveyed due to the rejection of the building owner. We surveyed all buildings as long as any part of the building is included in the voronoi area. In addition, we measured the proportion of the coverage by the voronoi for the building when the building is partially included by the voronoi so as to weight the number of populations in the building. Table 1 shows the distribution of building by income level. In the case of slums, the number of buildings cannot be compared with other building in traditional sense because hundreds of households live in shanties under the roof. In general, the low-income building is generally a one-story building, which is non-slum one-story brick-built building with CI sheet/semi-pucca. The middle-income building is a two to six-story building, and the high-income building is a high-rise building, which is seven-story or more. We use seven to separate the middle-income and high-income building because the restriction of building in Dhaka applies special rules to the building, which is seven-story or more, and constructing such buildings requires additional costs. As seen, the number of middle-income buildings holds approximately 60%, followed by the low-income building. According to our criteria of the income level, which is applied for sampling, overall, the middle-income building is predominant in Dhaka considering such types of buildings are commonly observed in Dhaka.

Table 2. Distribution of building in the survey site by income level

Years of built (y)	Slum	Low income	Middle income	High income	Total
y < 15	0	15	67	21	103
15 <= y < 40	0	63	95	10	168
40 <= Y 75	2	47	41	5	95
Total	2	125	203	36	366

From the 366 buildings, we surveyed total of 2,839 households. Tables 3 and 4 describe the distribution of the number of households and household members by income level. As shown in Table 3, the proportion of the middle-income household is the largest, which holds almost a half of the surveyed households. It is probably because the middle-income building is a multi-story building. Second to the middle-income household, the proportion of the slum household holds more than 30% of the total household. Taking account of the area size of slums, the population density in the slum is extremely high.

Table 3. Distribution of the number of households in the survey site by income level

Years of built (y)	Slum	Low income	Middle income	High income	Total
y < 15	0	8	368	223	599
15 <= y < 40	0	112	567	115	794
40 <= Y 75	872	137	372	65	1,446
Total	872	259	1,307	403	2,839

Table 4. Distribution of the number of household members in the survey site by income level

Years of built (y)	Slum	Low income	Middle income	High income	Total
y < 15	0	38	1,428	887	2,353
15 <= y < 40	0	438	2,228	444	3,110
40 <= Y 75	3,575	530	1,716	237	6,058
Total	3,575	1,006	5,372	1,568	11,521

As mentioned, our survey includes the building profile part, which surveys the use of the building and the average number of customers and employees. Table 5 shows the distribution of the number of floor according to the type of use. Interestingly, the case of commercial use in front of the building, counted as 550, is not marginal. It implies that there can be a great discrepancy in the number of populations estimated through the spatio-temporal and static data because these activities are considered to be temporal.

Table 5. Distribution of the number of floor according to the type of use

Type of use	The number of floor									Total
	f0f*	f0	f1	f2	f3	f4	f5	f6	f7	
Residential	0	254	211	169	144	116	86	28	3	1,011
Commercial	550	54	56	6	1	0	0	0	0	667
Manufacturing	56	14	1	1	0	0	0	0	0	72
Other	64	80	21	33	20	34	30	26	1	299

* f0f indicates the front space of the building on the ground[4].

Tables 6 and 7 provide the number of customers and employees per day in the surveyed area. As described in Table 6, the number of people related to commercial or industrial activity, which is almost 20,000 people, is greater than that of residential populations. It indicates great fluctuation in populations according to the time of the interest, and thereby it is crucial to know the population presenting at a given time and location through spatio-temporal data.

Table 6. Total number of customers and employees per day by floor by the type of user

Use of building		Number of people by floor					Total
		f0f	f0	f1	f2	f3	
Commercial	***Customers***	13,811	1,320	1,432	103	1	16,667
	Employees	1,584	188	126	11	1	1,910
Industrial	***Employees***	315	101	10	15	0	441
Total		15,710	1,609	1,568	129	2	19,018

Table 7. Average number of customers and employees by floor by business concern

Use of building		Number of people by floor				
		f0f	f0	f1	f2	f3
Commercial	***Customers***	25	24	26	17	1
	Employees	3	3	2	2	1
Industrial	***Employees***	6	7	10	15	0

5.3 DESCRIPTIVE STATISTICS FOR MOBILE USERS AND THE UNOBSERVABLE POPULATION

In this subsection, we describe the personal attributes of 922 mobile users and the members of their households — that is, of 2,366 non-mobile users — from the

[4] Around the residential areas in Dhaka, it is commonly observed that those who are engaged in informal activity such as the vendor of food, tea, and daily commodity occupy the space in front of the building constructing the temporal structure.

SPACE data. Gender, role within the household, and age group are examined separately for mobile users and non-users. This allows us to show which parts of the population can be captured by CDRs. Descriptive statistics are provided for four income levels throughout this section.

5.3.1 Characteristics of Mobile User Households

Table 1 describes the basic characteristics of households from the SPACE data according to income level. Row (A) shows the average household size, which is almost four for all income levels. This means that there are four persons per household, on average, across all income levels. Row (B) shows the proportion of males, 51% for all income levels. It is close to the 54%, which is the proportion of males, reported from the official statistics of Dhaka [11]. The higher the income level, the more highly educated the users are, as described in row (C). Row (D) shows that the higher the income level, the greater the percentage of mobile users when we disregard the telecommunications provider. Reference [12] reported that the mobile phone subscription in Bangladesh is 69%, but the figure obtained from our field survey is much lower. This is probably because in developing countries, it is common for users to have more than one SIM card so that they can benefit from lower tariffs offered by several operators. Row (F) shows that there is a certain percentage of multiple SIM card holders across all income levels. Interestingly, average number of mobile users per household does not differ according to income level, as described in row (E). Given that the average household size is between 4.0 and 4.1 for all income levels, we can assume that, on average, each person identified on CDRs represents 2.4, 2.6, 2.6, and 2.8 persons from the high-, middle-, low-, and slum income levels, respectively.

Table 1. Basic trends in household attributes by income level

	Income level			
	High	Middle	Low	Slum

(A) Average household size	4.0	4.1	4.1	4.1
(B) Male ratio	51%	51%	51%	51%
(C) Average years of education of users	10.6	8.2	5.6	3.7
(D) Average number of users per household (regardless of provider)	2.6 (66%)	2.3 (57%)	1.9 (46%)	1.7 (43%)
(E) Average number of users per household (who specified the operator as their provider)	1.2 (29%)	1.2 (28%)	1.1 (28%)	1.1 (26%)
(F) Ratio of multi-SIM holders	29%	16%	12%	9%

5.3.2 Attributes of Mobile Users and the Unobservable in CDRs

In this subsection, we examine the personal attributes of mobile users to understand who is represented in the operator's CDRs. Descriptive statistics for the gender, role within the household, and age group are provided by income level. Because the number of households and mobile users differ by income level, we employ proportions to compare the distributions among the four income levels. First, we examine the composition of gender among mobile phone users. Table 2 provides descriptive statistics for the gender of the users. Row (A) shows males are the predominant users overall. Comparing the user percentage among males in row (B) and that among females in row (C), the percentage for males is higher overall for males. Table 3 provides the number of males and females who are assumed to exist but not included in CDRs, given the presence of one male or female in the CDRs, by income level. As discussed previously, we assume that there are roughly 2.4 to 2.8 unobservable persons per male or female in CDRs that we use for this study. For instance, when we find 10 male users in slum areas, they are assumed to represent 27 unobservable persons, 12 males and 15 females.

Table 2. Descriptive statistics: mobile users' gender by income level

	Income level			
	High	**Middle**	**Low**	**Slum**
(A) Ratio of males among mobile users	68%	63%	52%	62%
(B) Ratio of male users to total male population in SPACE data	36%	28%	34%	35%
(C) Ratio of female users to total female population in SPACE data	22%	28%	21%	17%

Table 3. Gender composition of the unobservable per mobile phone user by income level

	Composition of the unobservable by income level							
	High		Middle		Low		Slum	
Per user in CDRs	M	F	M	F	M	F	M	F
Male (M)	1.1	1.3	1.3	1.3	1.2	1.4	1.2	1.5
Female (F)	2.1	0.3	2.3	0.3	2.2	0.4	2.2	0.5

Second, we examine the roles of mobile users within their households. The roles are classified into four categories: household head, spouse, child of the head, and others. People classified as others include the parents of the household head, relatives, servants, or workers living within the household. Figure 1 describes the composition of 100 users in each income level. For each income level with a different number of mobile users, we adjusted the total number of mobile users to 100 to facilitate comparisons of the distributions across the four income levels, e.g. the number of mobile users among the high-income households, who are males and the household heads, is shows as 50 if there are 250 mobile users including 125 household heads for the income level. Each set of 100 users is considered to represent mobile users from the SPACE data in each income level. Majority of males are household heads and majority of females are spouses. This indicates that most users are responsible for household decisions to some extent. Interestingly, the trends are similar across all income levels.

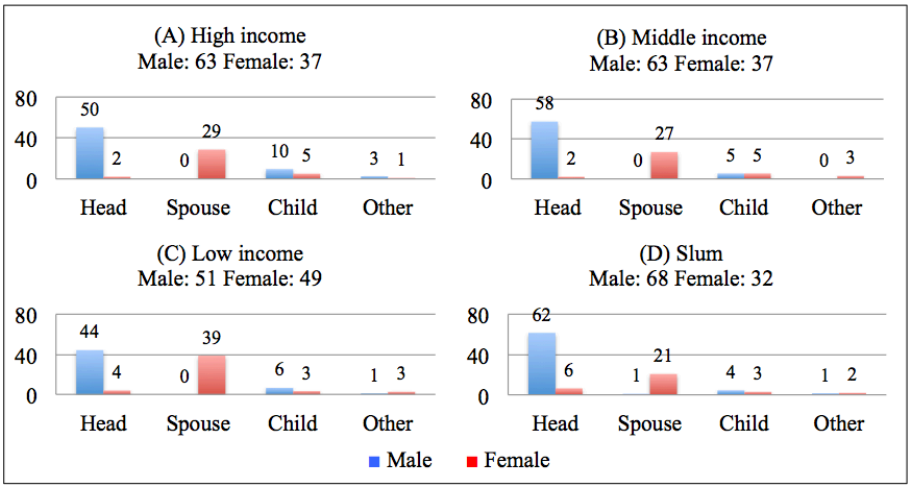


Figure 1. Distribution of roles within the household among 100 users for each income level

Then, we examine the composition of the unobservable by role within the household against the four sets of 100 users. The number of unobservable differs by income level because the proportion of unobservable people to mobile users differs by income level. For instance, if the total populations surveyed for an income level consist of 250 mobile users and 750 non-users, the number of unobservable individuals per 100 users is 300. If another income level consists of 200 mobile users and 400 non-users, the number of unobservable people per 100 users is 200. Figure 2 shows the composition of the unobservable population for each income level per 100 users. As mentioned, the total number of unobservable people varies according to income level because the proportion of non-mobile users to the total population differs according to income level. The greater the number of unobservable in Figure 2, the lower the ratio of mobile users for that income level. Majority of the unobservable are the children of household heads, followed by the spouses of the heads. This trend is consistent across all income levels.

Figure 2. Distribution of roles within the household among the unobservable per 100 users for each income level

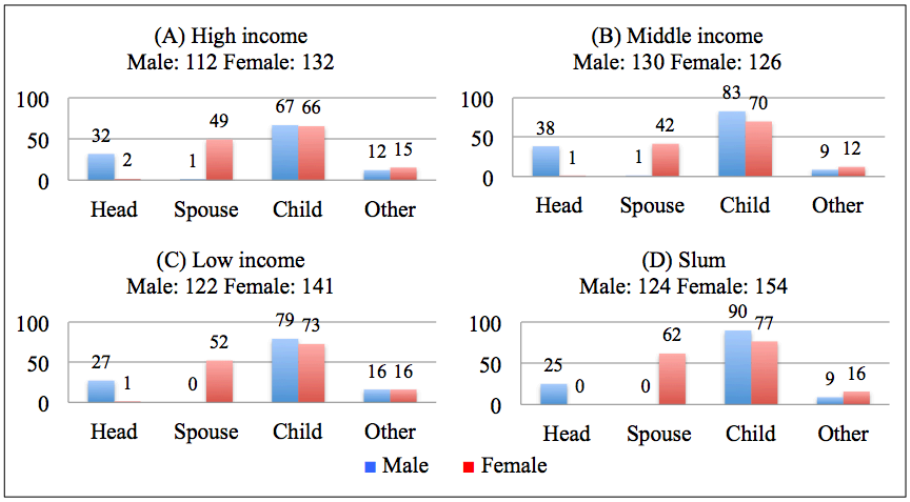


Figure 3. Distribution of type of activity among the unobservable per 100 mobile users for each income level

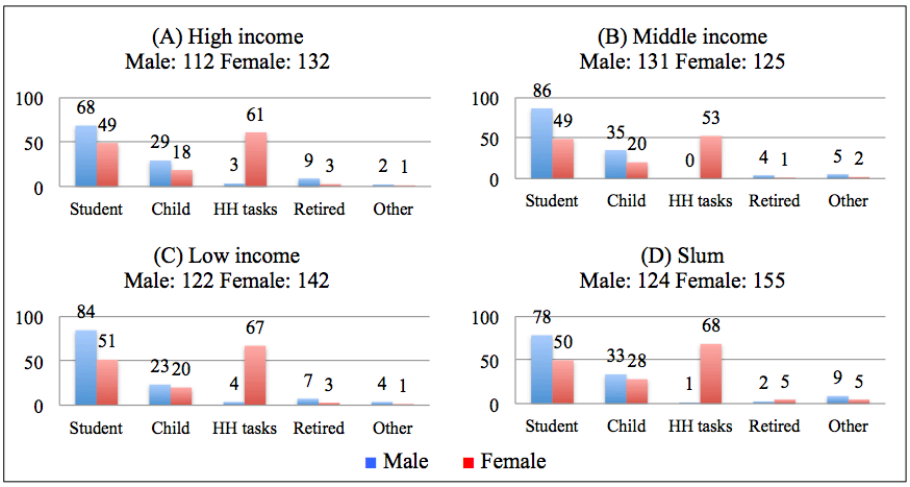


Figure 3 describes the main activity of the unobservable for each income level. "Child" indicates a child below school enrollment age, and "HH tasks" indicates a person who does household tasks. Across all income levels, majority of unobservable males are students and children. This means that most unobservable males are still dependent on their family members. Further, almost half of the unobservable females do household tasks, meaning that they are most likely housewives. The remainder of unobservable females consists of students and children. The results demonstrate that CDRs are not highly capable of capturing the whereabouts of children.

Third, we compare the age group distributions between mobile users and those who do not use mobile phones within the users' households, that is, the unobservable, to extract the population groups that are not captured by CDRs. The age group distribution among users is illustrated in Figure 4 and that of the unobservable is shown in Figure 5. Figure 4 shows that the dominant age group among mobile users is that between 30 and 49. In almost all age groups, there are more male users than female users. The overall trends in the distribution are similar across all income levels.

Figure 4. Distribution of age groups per 100 mobile users for each income level

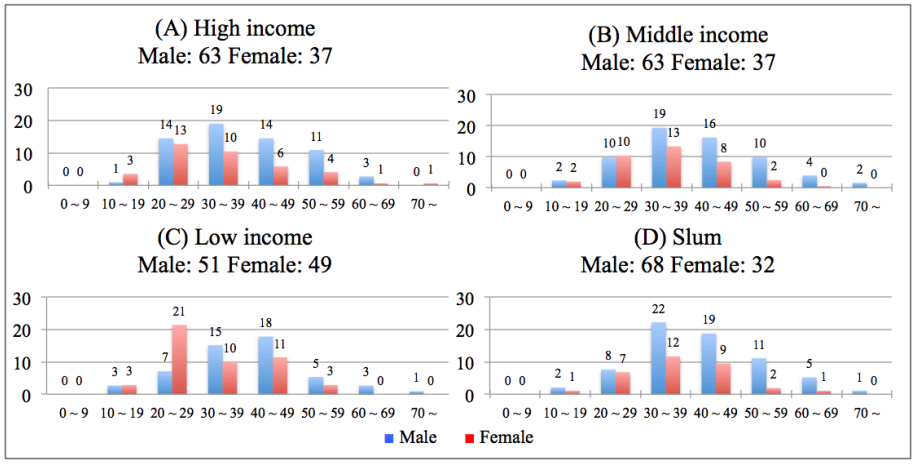


Figure 5 describes the age group distribution among the unobservable per 100 mobile users and by income level. The number of unobservable persons differs according to the income level, because the proportion of users varies. Overall trends are similar across all income levels, while the ratio of younger age groups is greater in lower income levels. Based on the descriptive statistics for the unobservable, we find that vulnerable people, who are often either young or dependent on their family members, such as the elderly, tend not to be included in CDRs. Figure 6 shows the population pyramid of Bangladesh and describes the demographic structure based on gender and age group. The structure is wide at the base with a median age below 25 years. It indicates that half of the population is aged below 25 and is skewed toward the younger age groups. Therefore, the unobservable population is defined as children and students, and tends to comprise a very young population.

Figure 5. Distribution of age groups among the unobservable per 100 mobile users for each income level

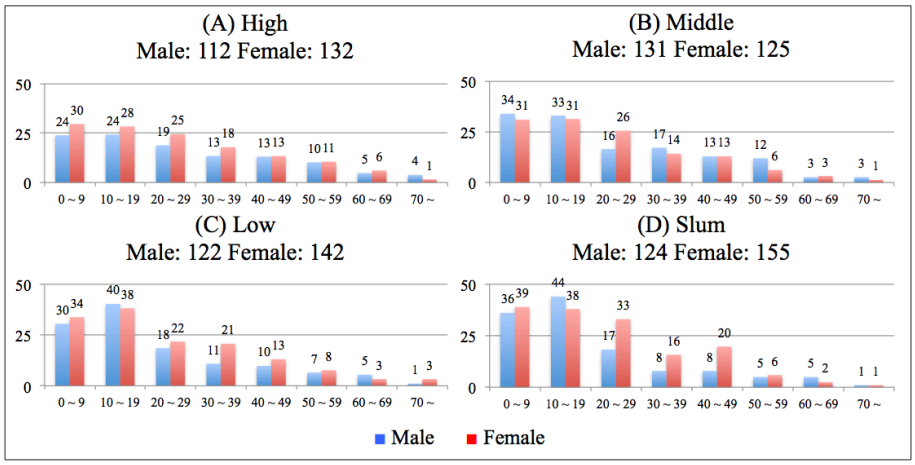
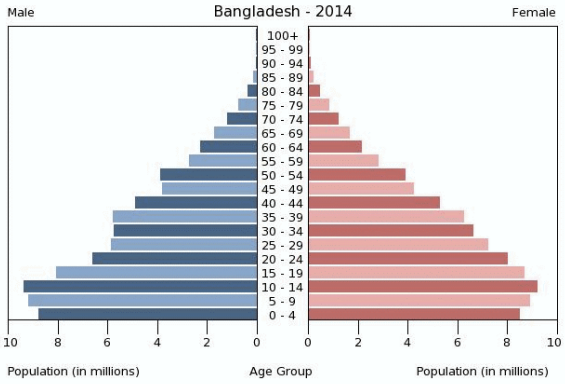


Figure 6. Population pyramid of Bangladesh (source: [13])



Thus, we need to note that the CDRs do not capture the whereabouts of the very young population. Based on the age distribution in Figure 4, it is fair to say that CDRs seldom capture people below the age of 20. These findings describe exactly what we observed in our fieldwork. Typical households in Dhaka tend to comprise nuclear families, which are composed of a household head, a spouse, and some children; it is quite common for household heads to own a mobile phone. Spouses, who are predominantly female because of the very high proportion of male-headed households in Dhaka, sometimes own mobile phones. Most of females do household tasks at home and some females work outside the home. This partially reflects social norms in Muslim societies, where males are supposed to be engaged in income-earning activities outside the home while females stay at home. The remaining

household members seem to have limited opportunities to own mobile phones. They are often the children of the head and the proportion of extended family members, e.g., parents, brothers, sisters, or relatives of the head or spouse, seems to be higher among lower-income levels.

5.3.3 Whereabouts of the Unobservable

We briefly discuss trends in the whereabouts of the younger population among the unobservable population, children and students, which predominantly compose the unobservable population. We found that a young population in Dhaka is primarily engaged in educational activities regardless of their household income levels. Even though many households cannot afford to send their children to public school, many NGO-run schools and educational institutions still provide educational opportunities for them. Except for colleges and universities, school hours are basically only in the mornings or afternoons. Most schools operate from Saturday to Thursday, and Friday is generally a weekend. Some students go to coaching after school hours and others stay at home or around their homes. This means that on weekdays, the majority of students spend half of their days at school and the rest of the day around their homes. It is common to choose schools and coaches that are close to home due to heavy traffic in Dhaka. In other words, we can assume that students generally have similar routines such that their locations and activity patterns are very specific. Therefore, we expect that understanding the whereabouts of children at given times and locations is possible if we can identify the locations of mobile users from households with students.

5.4 CLUES TO FINDING THE UNOBSERVABLE BASED ON MOBILE USERS' CALLING BEHAVIOR

This section explores clues to finding the unobservable in CDRs based on the calling behaviors of mobile users. The results in previous sections show that the majority of the unobservable are students and children below school enrollment age, and hereinafter, we call them children. Hence, we first examine past studies on the link between travel patterns and personal or household attributes, which are related to the presence of children. By doing so, we attempt to identify the features of the activities of people with children in their households. Then, we analyze the calling behavior data obtained from SPACE and two-month CDRs from volunteers to investigate whether the features of the activities extracted from the literature review are reflected in the calling behavior characteristics.

5.4.1 Impacts of the Presence of Children within the Household on Travel-activity Behavior

Conventional trip diary data collected through field surveys have long contributed to enhancing the understanding of human activity and travel patterns for research on urban planning and transportation. Part of human mobility is well explained as travel demand in association with activity patterns, because urban travel is considered to be driven by the demand of people who have the need or desire to participate in activities [14]. In travel demand modeling, travel activities are explicitly combined with the presence of children within households. The presence of children is an important factor affecting the time allocated to outside-the-home and non-work activities [15]. Furthermore, the time allocated to outside-the-home or non-work activities differs among males and females when there are children in the household because children are dependent on their primary caregivers, which are predominantly their mothers. This implies that in terms of locations, calling behavior characteristics may reflect such behavioral differences according to the presence of children [16].

For non-employed people, the presence of children significantly affects travel-activity behavior [17]. Given that CDRs enable us to identify significant locations such as homes and work locations [2], they can provide a partial view of users' time and location distributions. This implies that time allocated to outside-the-home activities, as extracted from calling behavior, can be a clue to finding people who belong to households with children.

5.4.2 Trends in Call Locations among Males and Females

First, we compare trends in call locations between males and females by analyzing one-day call records of 922 mobile users, collected through SPACE. Call locations are classified as home, the primary location outside the home, and other. We underline that this classification only specifies the home as a particular location, and that the primary location outside the home is defined based on the number of call records per location. This definition is based only on frequency because this information can be extracted by counting the number of calls in CDRs. This means that the method applied to the one-day call records from the SPACE data can be applicable to CDRs. We calculate the proportion of the number of calls from home and the proportion of the number of calls from the primary location outside the home for each user as (1) and (2), and take the average by income level. By doing so, we attempt to find differences in call location trends between males and females.

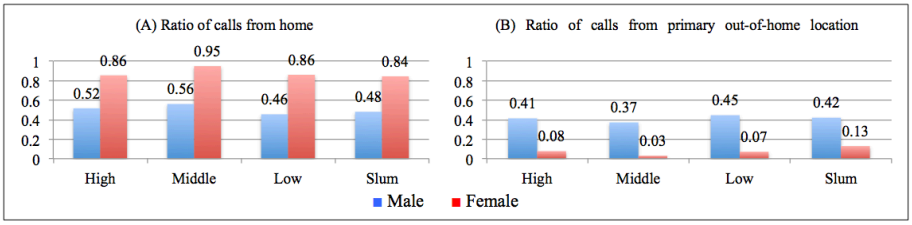
$$\textit{Proportion of the number of calls from home} = \frac{\textit{Number of calls from home per day}}{\textit{Total number of calls per day}} \quad \text{- (1)}$$

$$\textit{Proportion of the number of calls from primary location outside the home} = \frac{\textit{Number of calls from primary location outside the home per day}}{\textit{Total number of calls per day}} \quad \text{- (2)}$$

Figure 7 shows the average proportion of calls per day from (A) home and (B) the primary location outside the home by gender. It shows that females predominantly call from their homes, while males make similar proportion of calls from both

locations. Recall that most female mobile users are spouses of the household head, meaning that they are married, as shown in Figure 1. In Bangladesh, it is still common for married females to stay at home to do household tasks. It is fair to say that the calling behavior described in Figure 6 well reflects trends in time allocation for the home and outside-the-home activity of males and females. This indicates that the proportion of calls from home could be a key to identifying females from anonymized CDRs.

Figure 7. Proportion of calls from (A) home and (B) the primary location outside the home



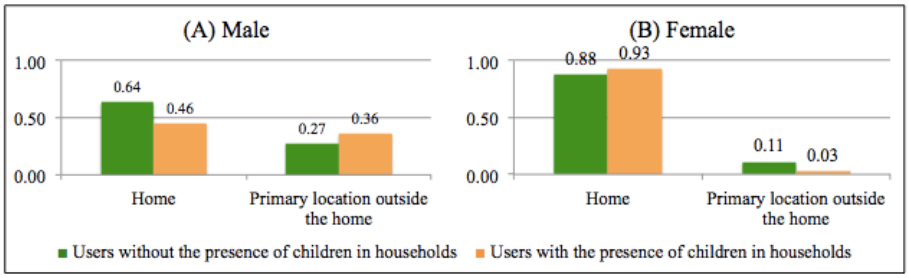
5.4.3 Trends in Call Locations According to the Presence of Children

Second, we examine whether there are any differences in call location trends according to the presence of children in mobile phone users' households. We analyze two-months CDRs for 58 volunteers: 27 males and 31 females. Considering the variety of lifestyles among mobile users, we classify the seven days of the week into two groups: primary routine days and non-primary routine days. We define primary routine days as days in which a mobile user is engaged in her/his primary routine. The primary routine could include any activities on which users spend the majority of their time. For instance, the primary routine for students is to go to school, and that of salary workers is to work at the office. If working days are from Sunday to Thursday, these five days of the week are the primary routine days, making Friday and Saturday the non-primary routine days. In the following, we examine trends in call locations for each type of day according to the presence of children. Then, we analyze trends in call locations on non-primary routine days versus those on primary routine days.

Primary routine days

Figure 8 compares the proportion of calls from home and the primary location outside the home on primary routine days between two groups of users and by gender. (A) shows the trends among males. Those without children in their households tend to make more calls from home. Conversely, those with children tend to make a similar number of calls from home and from the primary location outside the home. Similarly, (B) shows the trends among females. Females tend to call from home regardless of the presence of children in their households. Although we can to some extent see different trends in call locations due to the presence of children, the difference is not sufficient to distinguish between those with and without children based only on call locations.

Figure 8. Proportion of calls by location type on primary routine days among users according to the presence of children in the household

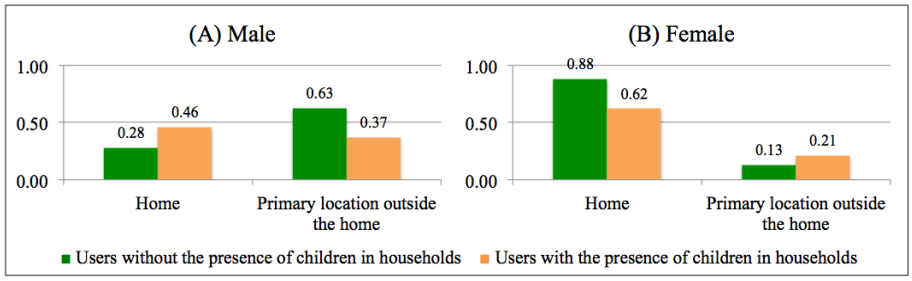


Non-primary routine days

Figure 9 compares the proportion of calls at home and calls at the primary location outside the home on non-primary routine days between the two groups and by gender. Males without children in their households tend to call more from the primary location outside the home. Conversely, those with children tend to make similar proportion of calls at home and at the primary location outside the home. Among males, the trends on non-primary routine days are the inverse of those on primary routine days. As for females, their trends is inverse of those of males but the difference in females is not as significant as that in males. This indicates that males

without children in their households have more flexibility to spend time outside the home. However, this difference appears to be insufficient to identify males with children based on calling behavior alone.

Figure 9. Proportion of calls by location type on non-primary routine days among users according to the presence of children in the household



Non-primary routine days versus primary routine days

Finally, we analyze trends in call locations on non-primary routine days versus those on primary routine days according to the presence of children in users' households. We calculate the ratios, expressed as (3) and (4), and compare traits between males and females.

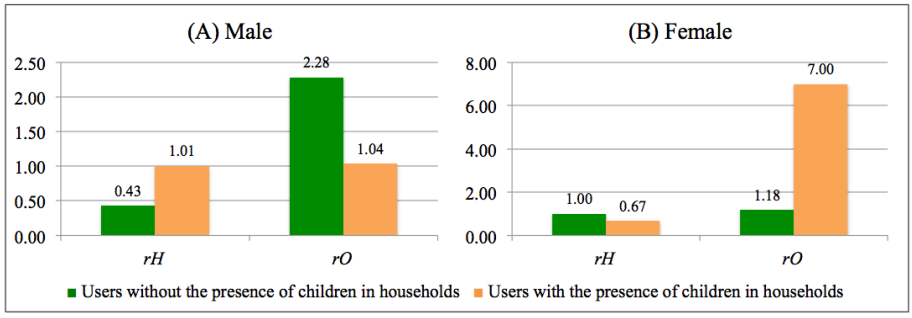
$$r_H = \frac{Value\ of\ equation\ (1)\ on\ non\text{-}primary\ routine\ day}{Value\ of\ equation\ (1)\ on\ primary\ routine\ day} \quad -(3)$$

$$r_O = \frac{Value\ of\ equation\ (2)\ on\ non\text{-}primary\ routine\ day}{Value\ of\ equation\ (2)\ on\ primary\ routine\ day} \quad -(4)$$

Figure 10 compares r_H and r_O according to the presence of children in the household and by gender. (A) shows the values of r_H and r_O among males with children within their households are close to unity. On the other hand, among males without children, r_O is much greater than r_H, which means that they tend to make more calls at the primary outside-home location on non-primary routine days. This indicates that the trends in call locations for males with children are almost unvarying across all days of

the week. In contrast to males, r_H and r_O values among females are close to unity for those without children within their households. That is, males and females exhibit inverse directions of association between the presence of children in the household and trends in call location. The results also suggest that weekly trends in call locations are the key to identifying mobile phone users who have children in their households through their calling behavior.

Figure 10. Trends in call locations (r_H and r_O) according to presence of children in the household and by gender



5.5 IDENTIFYING OF THE PRESENCE OF THE UNOBSERVABLE POULATION IN HOUSEHOLDS

Analysis of the results in the previous section indicates the potential of identifying the presence of children in the household from the calling behavior of mobile users. Additionally, we found that the majority of the unobservable are the very young population. Some studies attempt to predict the demographic attributes and socioeconomic status of mobile users by analyzing sensor data from smartphones [18] as well as calling behavior [19]. The analysis of call records collected through a field survey shows that calling behavior exhibits gender-specific characteristics [20]. In this section, we explore the application of calling behavior characteristics to identify gender and the presence of children among mobile users' households through the analysis of CDRs. This section uses CDRs from 55 out of 58 mobile users. We

drop the data for three people because the number of records is too few considering that the data are also used to construct the estimation model for the gender and presence of children.

5.5.1 Revisit the definition of the unobservable population

To examine user characteristics, calling behavior, and the presence of the unobservable population in households, we discuss the definition of the younger population among the unobservable population, children, for the analysis. We refer to the educational system in Bangladesh where three main educational systems exist. Overall, each of these systems is divided into several levels and summarized as primary education (children aged from six to 10) and secondary education (children aged from 11 to 17). Table 4 shows the distribution of 55 mobile users according to gender and the presence of children in the household. Here, we compare the population distribution under two different thresholds to define the age of children. Comparing the distribution in rows (A) and (B), we find that most mobile users belong to households with children if we set the age threshold as 17. Because the sample size for mobile users without the presence of children in case (B) is insufficient, this section uses 10 as the threshold to define the age of children for further analysis. This setting may limit the coverage of the unobservable population that can be identified by our estimation, but we believe it is still useful. This is because younger children are biologically vulnerable to various environments such as those with diseases, poverty, and disasters [21]. [22] reported that more than 900,000 children under the age of five succumbed to malaria during the year 2000. Therefore, we consider the distribution of such populations of practical use.

Table 4. The number of mobile users according to the presence of children in households

Age of children in the household	Mobile user gender	Number of mobile users (Numbers in parentheses represent the proportion by gender)	
		With the presence	Without the presence

(A) 10 and under	***Male***	16 (64%)	9 (36%)
	Female	16 (53%)	14 (47%)
(B) 17 and under	***Male***	21 (84%)	4 (16%)
	Female	26 (87%)	4 (13%)

5.5.2 Calling behavior by gender

We examine the characteristics and calling behavior of mobile users by gender and the presence of children in households. We investigate the users' characteristics related to activity, which are expected to affect calling behavior. We first compare the characteristics of 55 people and the characteristics of the 922 mobile users of SPACE data described in section 3. Table 5 shows the characteristics and calling behavior of the 55 mobile users. Rows (A) and (B) show that the proportion of male mobile users engaged in income-earning activity is substantially greater than that of females. Additionally, most males are the household head, and most females are the spouses of the head as shown in row (C). This indicates that the majority of the 55 people are working males and housewives. With respect to calling behavior, the proportion of the number of calls from home is greater for females, whereas the proportion of calls from the primary location outside the home is greater for males, shown in rows (E) and (F). We confirm that the overall trends of the 55 users are consistent with the trends observed in the SPACE data in section 3.

Table 5. Characteristics and calling behavior of 55 mobile users

Features		**Male**	**Female**
Sample size		25	30
Characteristics	***(A) Proportion of users who are engaged in income-earning activity***	72%	13%
	(B) Proportion of household heads	88%	13%
	(C) Proportion of household head spouses	0%	73%
Calling behavior	***(D) Total number of calls for four months***	397	399
	(E) Proportion of the number of calls from home to the total number of calls	50%	71%
	(F) Proportion of the number of calls from primary location outside the home to the total calls	18%	7%

3.5.3 Calling behavior according to the presence of children

Table 6 shows the characteristics of mobile users according to the presence of children in the household. Rows (A) and (B) show that the proportion of males engaged in income-earning activity is greater if children are present in the household. Additionally, the majority of males are considered to be the household head. However, the proportion of females engaged in income-earning activity does not differ if children are present in the household. Considering the social norms discussed previously, it is not surprising that more males are engaged in income-earning activity to raise children while females stay at home if the household has children as household members. Row (D) reveals that the mobile user belonging to the household with children is younger by approximately 10 years on average. This result shows the generational difference in household members because of the presence of children. We expect that this feature can help to explain the age differences among mobile users in future studies.

Table 6. Characteristics of mobile users according to the presence of children in households

Characteristics of mobile users	Mobile user gender	Presence of children	
		Yes	No
(A) Proportion of users who are engaged in income-earning activity	***Male***	86%	44%
	Female	12%	14%
(B) Proportion of household heads	***Male***	94%	78%
	Female	6%	21%
(C) Proportion of household head spouses	***Male***	0%	0%
	Female	88%	57%
(D) Age	***Male***	45	54
	Female	32	41

Table 7 shows the calling behavior of mobile users according to the presence of children by gender. Calling behavior is extracted from three aspects: (A) the frequency of calls, (B) the distribution of the timing of calls, and (C) weekly patterns. Feature (A) is calculated as the total number of calls during the four months. The trends in the frequency of calls for males and females are inverse depending on to the presence of children in the household. Feature (B) is the standard deviation of the

average probability of calls per hour for four months. For each person a set of 24 values, each of which is the probability of calls per hour, is calculated to obtain the standard deviation of the values. The greater the value, the more variation in the timing of calls. Interestingly, the trends for males and females are again inverse depending on the presence of children in the household. For the weekly pattern shown in row (C), we calculate the ratio of r_O' to r_H', which is obtained from the equations (3)' and (4)'.

$$r_H = \frac{Value\ of\ equation\ (1)\ on\ non\text{-}primary\ routine\ day}{Value\ of\ equation\ (1)\ on\ primary\ routine\ day} \quad -(3)'$$

$$r_O = \frac{Value\ of\ equation\ (2)\ on\ non\text{-}primary\ routine\ day}{Value\ of\ equation\ (2)\ on\ primary\ routine\ day} \quad -(4)'$$

These equations are the modified versions of equations *(3)* and *(4)* where the type of day is originally classified as either the primary routine day or the non-primary routine day. In *(3)'* and *(4)'*, non-Friday and Friday are used instead of the primary and non-primary routine day to relax the condition when examining the trends of different population groups. This is because identifying the primary and non-primary routine days of anonymized CDRs requires additional analysis and generates uncertainty. Row (C) shows a difference according to the presence of children for both genders. This result indicates that the power to explain the difference remains after relaxing the condition of classifying according to the type of day. (Friday and Saturday in Bangladesh are similar to Saturday and Sunday in many western countries. Government offices are officially closed on these days. In general, Friday is a holiday if the institution has only one holiday per week according to our field observations.) However, the trend for both genders is in the same direction.

Table 7. Calling behavior according to the presence of children in the household

Calling behavior	**Mobile user gender**	**Presence of children**	
		Yes	***No***

(A) Total number of call records	***Male***	497	220
	Female	398	400
(B) Standard deviation of the hourly probability of calls	***Male***	9.94	4.68
	Female	10.47	10.15
(C) Ratio of r_O' to r_H'	***Male***	0.23	7.82
	Female	0.14	7.32

3.5.4 Identifying gender and the presence of the unobservable population

Using the feature extracted in the previous section, we propose estimation models to identify mobile user gender and the presence of children in the household using CDRs from 55 mobile users. We employ Random Forest, a powerful ensemble method based on the collection of tree classifiers using the R package "randomForest" [23]. We generate features from three different aspects of calling behavior; frequency, timing, and weekly patterns as was the case when analyzing calling behavior. For the model feature selection, we first calculate the correlation among all feature combinations and remove highly correlated features with an absolute correlation of 0.75 or higher. Then, we drop features with redundant Gini importance. The Gini importance is the mean Gini gain produced by a feature over all trees. The greater the value is, the more significant the feature in the estimation model is. Tables 8 and 9 show the features used for the gender prediction and Random Forest results for gender. Table 8 shows that the three most significant features, *v1*, *v2*, and *v3* are generated by specifying the day of the week, non-Friday and Friday, for call records. This implies that the relaxed condition still finds some weekly trends. Another three features, *v1*, *v3*, and *v6*, are related to call records from users' significant locations such as home and the primary location outside the home. This indicates that accurate identification of significant locations from CDRs is critical for gender identification. Table 9 shows the prediction results for gender. The accuracy is superior to random guess, which is 0.5.

Table 8. Significant features used for the prediction of gender

Rank of importance	Feature	Description
1	*v1*	Proportion of the number of calls from home on Friday
2	*v2*	Ratio of the number of calls on Friday to the number of calls on non-Friday
3	*v3*	Proportion of the number of calls from the primary location outside the home on non-Friday
4	*v4*	Total number of calls
5	*v5*	Time (hourly basis in 24 hours) at which the user called most in the pm
6	*v6*	1 if the most frequently called location for the user is home

Table 9. Random Forest results of gender

Class	Accuracy	Precision	Recall
Male	0.709	0.696	0.640
Female		0.719	0.767

Table 10 shows the features used for identifying the presence of children in the household when all samples are pooled. We do not provide the rank of importance for the prediction by splitting the population by gender because the result for the pooled case outperforms the split cases. Table 11 shows the Random Forest results for the presence of children in the household; the features capturing weekly patterns, *v1* and *v5*, are significant. Additionally, the features extracting the timing or distribution of calls within a day, *v2* and *v3*, are significant. The accuracy for the pooled case is superior to random guess but exhibits difficulty in capturing the traits of mobile users without the presence of children in the household.

Table 10. Significant features used for the prediction of the presence of children in the household

Rank of importance	Feature	Description
1	*v1*	Ratio of the number of calls on Friday and that of non-Friday
2	*v2*	Standard deviation of hourly probability of calls
3	*v3*	Time (hourly basis in 24 hours) at which the user called most in the pm
4	*v4*	Total number of calls
5	*v5*	Ratio of r_O' to r_H'

Table 11. Estimation results of the presence of children in the household

Gender	Class	Accuracy	Precision	Recall
Male	***Yes***	0.571	0.667	0.615
	No		0.444	0.500
Female	***Yes***	0.533	0.556	0.625
	No		0.500	0.429
Pooled	***Yes***	0.655	0.700	0.719
	No		0.590	0.565

5.6 DEMOGRAPHIC STRUCTURE OF FOUR POPULATION GROUPS

In this section, we compare the demographic structure of four population groups to investigate differences according to the telecommunications company to be subscribed and mobile phone ownership. Analyses throughout this section are done by income level to see whether any distinctive differences in trend exist according to the income level:

(a) The users of *the operator* whose CDRs are used for this thesis. They are the mobile users belonging to *Household A*.

(b) The user of other telecommunications operators except for *the operator.* All of them are mobile users but do not appear in CDRs due to the difference of the telecommunications companies to subscribe.

(c) Non-mobile users.

(d) The unobservable population of CDRs. Population group (d) includes anyone who does not use mobile phone of *the operator* specified by (a). So, the definition of the unobservable population is slightly different from what was used in Sections 3.3 and 3.4.

Figure 2 shows the distribution of gender among the four population groups. As seen from (a) and (b), predominant mobile users are commonly males regardless of the telecommunications operator to be subscribed. The lower the income level, the more predominant the male user. As a result, females hold the greater proportion of non-mobile users as shown in (c). Figure 3 shows the proportion of the number of populations represented by each population group. The greater the proportion, the better the representation of a population group. As seen from (a), (b), and (c), the male better represents the population of mobile users regardless of telecommunications operators overall. Interestingly, as show in (a), for both male and

female users of *the operator*, the higher the income level, the better the representation of the population group. On the other hand, as seen from (b), not specific difference in the proportion of the representation due to the income level among male users, while the higher the income the better representation of a population group among females.

Figure 2. Distribution of gender among four population groups

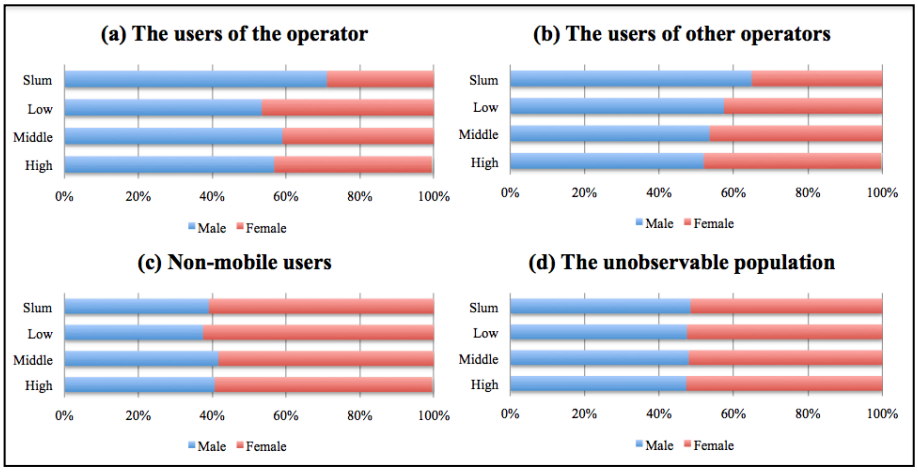
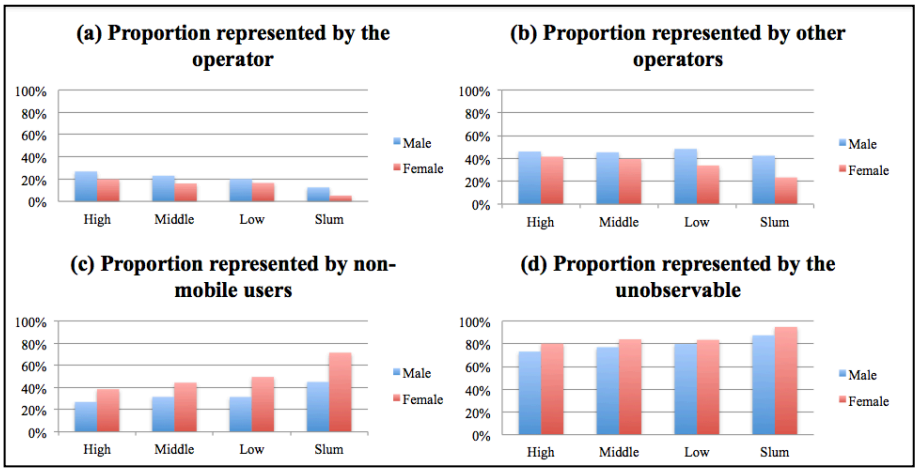


Figure 3. Proportion of the number of populations represented by four population groups by gender



Based on the observation from Figures 2 and 3, we can summarize that there is a certain level of gender gap in access to the mobile phone in Dhaka. The trend is more significant for *the operator* compared to the rest of operators. It is probably because the mobile phone tariff of the operator is relatively expensive than others, and the

male generally have more degree of freedom in the use of the money at their own disposal considering the social context in Bangladesh. As shown in Figure 3(b), among female mobile users subscribed to other operators, the proportions among the higher income is not very different from those of males.

Figure 4 provides the distribution of main activity among the four population groups. As seen from (a) and (b), overall, the working male is the predominant mobile user followed by the housewife regardless of telecommunications operators. Interestingly, more than 90% of the male seem to have the mobile phone if they have jobs according to (c). Majority of the non-mobile users are the student and other, but the proportion of the student is greater in higher income groups. It is perhaps because higher-income people tend to have longer years of education enrollment.

Figure 4. Distribution of main activity among four population groups

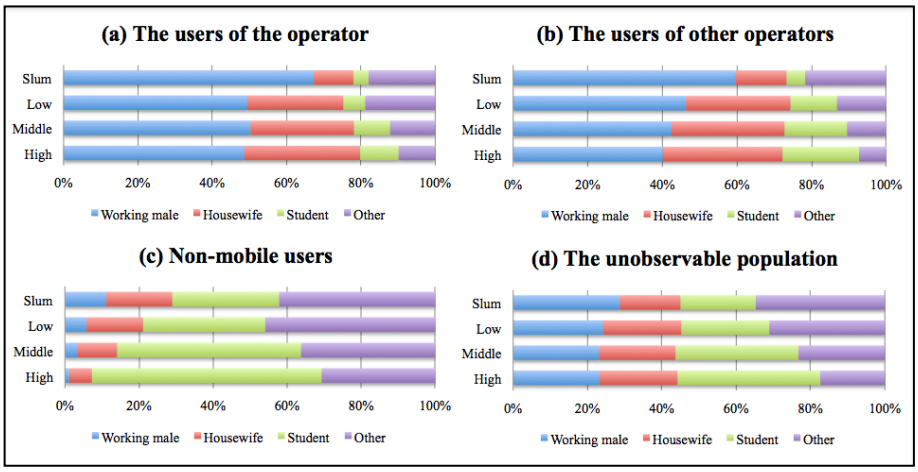


Figure 5 shows the proportion of the population represented by the four population groups. As seen from (c), the non-mobile user better represent the student and other across all income levels. As shown from (a) and (b), the trend of the proportion of the representation between the operator and other operators is overall similar despite that the proportion of the representation for the working male mobile user is almost the same across all income level for the user of other telecommunications operators. We can summarize that the activity of the mobile user explained by CDRs does not

significantly differ due to the difference in the telecommunications operator. Figure 6 compares the two population groups, one is the mobile user and the other is the sample population. As seen from (b), the lower the income level is, the greater the proportion of the non-mobile user is. Interestingly, the proportion held by Company B, C, and other is indifferent across all income level. It indicates that we can get approximate estimates for the number of the mobile users of other telecommunications operators and non-mobile users given the number of the mobile users of the telecommunication operator, which can be calculated by the CDRs used in this study.

Figure 5. Proportion of the number of population represented by four population groups

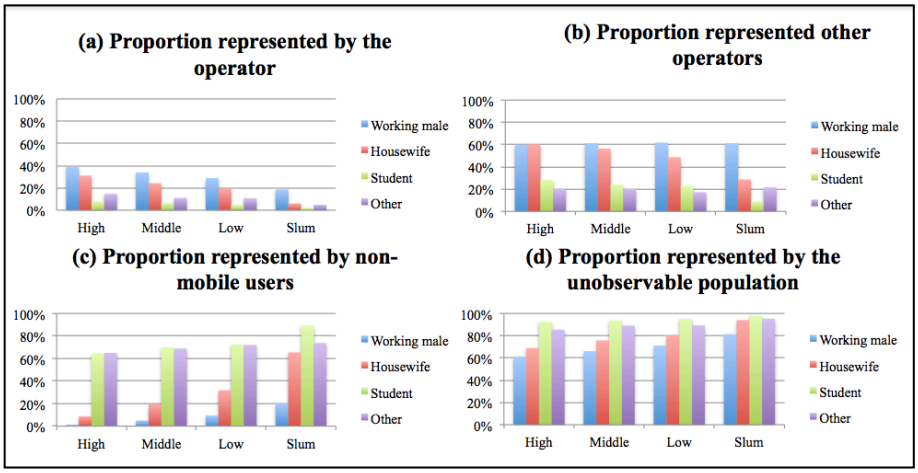
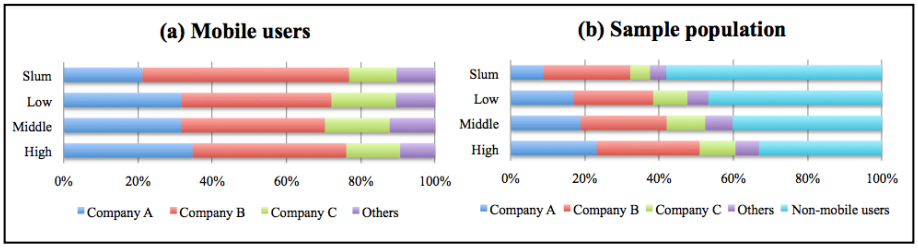


Figure 6. Distribution of telecommunications operators among two population groups



To sum, we find that overall population structure in terms of gender and main activity do not differ according to the telecommunications company to be subscribed. However, we are not able to examine differences in behavioral patterns among them

because all information we have on the activity of those who do not use mobile phones of *the operator* is only the type of activity. Therefore, further data collection from those population groups is necessary to investigate the behavior difference among the people who are classified to the same type of main activity but belong to different population groups.

5.7 ESTIMATION OF THE NUMBER OF THE ENTIRE LIVING POPULATION

We describe the process for estimating the number of the entire living population, and their demographic structure, from CDRs. For the area unit for calculation throughout this process, we use the Voronoi area, which is determined by the distribution of antennas. The area under this study is divided into 1,362 Voronoi areas. The sizes of Voronoi areas vary according to the population density because the antennas are distributed so as to optimize the traffic volume per antenna. It means that the higher population density, the smaller Voronoi area size. The area sizes range between 0.002 sq. km and 192 sq. km. We use the household as the smallest unit for estimating the number of populations by taking account of the household member structure. We consider that the use of the household as the smallest unit for estimation is important. This is because the unit is one of the smallest units of decision-making in the society and the unit is often used as an analysis unit for other studies such as economics. To use two different frameworks for estimation, we classified the living population into two groups as illustrated in Figure 2: *Household A*, which includes mobile users subscribed to *the operator* providing CDRs for this study; and *Household B*, which does not include mobile users of *the operator*.

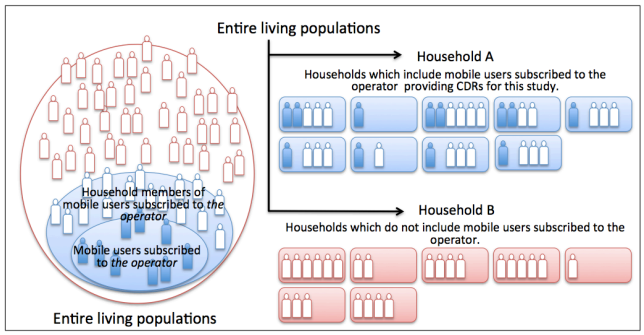


Figure 2. Composition of the entire living population

The number of the living population obtained here is considered to be the proxy of the nighttime population, assuming that most people spend their time at home during the night. We use SSC data to obtain descriptive statistics necessary to obtain the number and structure of the living population from CDRs of *the operator* and building map data. On the map, residential buildings are classified into four groups: high, middle, low, and slum buildings based on the height and material of buildings. The criteria for the classification of buildings for the map data and SSC are almost the same. Therefore, we can relate the category used for the map with that of the SSC data.

5.7.1 Population belonging to Household A and its structure

N_A (the number of members in *Household A*) is the sum of N_{AM} (the number of mobile users of *the operator*) and N_{AN} (the number of mobile users of *the operator*). N_{AM} can be obtained by calculating the number of unique IDs in CDRs of *the operator*. We assume that people generally do not purchase more than one SIM card from one telecommunications operator, although owning multiple SIM cards is not uncommon. Using the method proposed by [4], we can obtain the number of mobile users with and without the presence of children in the household (N_{AM_w} and N_{AN_wo}, respectively), where N_{AM} is the sum of N_{AM_w} and N_{AM_wo}.

We calculate N_{AN}, the number of non-mobile users belonging to Household A, based on the estimation result of the presence of children in the household of the mobile user. The distribution of personal attributes for non-mobile users in *Household A* according to the presence of children aged below 10 is provided in Tables 6(a) and 6(b).

Table X. Distribution of personal attributes among non-mobile users among members of *Household A*

(a) When mobile user belong to a household that has children

Average household size	**Distribution of personal attributes among non-mobile users**			
	Workmale	Housewife	Student	Other
4.2 $=HH_w$	13% $=p_{AN\_w\_w}$	17% $=p_{AN\_w\_h}$	37% $=p_{AN\_w\_s}$	32% $=p_{AN\_w\_o}$

(b) When the mobile user belongs to a household without the presence of children

Average household size	**Distribution of personal attributes among non-mobile users**			
	Workmale	Housewife	Student	Other
3.6 $=HH_{wo}$	21% $=p_{AN\_wo\_w}$	21% $=p_{AN\_wo\_h}$	42% $=p_{AN\_wo\_s}$	17% $=p_{AN\_wo\_o}$

As discussed in 4.2, we can see that the non-mobile users are predominantly of the Student category. Considering that the average household size is 4.1, there are roughly two children even if the household does not have children under aged 10. The number of the population belonging to *Household A* is given as equation *(1)*:

$$N_A = N_{AM\_w} \times HH_w + N_{AM\_wo} \times HH_{wo} \quad -(1)$$

To calculate the number of the population by personal attributes among non-mobile users, we use the proportions given in Tables 6(a) and 6(b). For example, the number of Student among the non-mobile users, N_{AM_s}, is given as equation (2).

$$N_{AM\_s} = N_{AM\_w} \times (HH_w - 1) \times p_{AN\_w\_s}/100 \\ N_{AM\_wo} \times (HH_{wo} - 1) \times p_{AN\_wo\_s}/100 \quad - (2)$$

5.7.2 Population belonging to Household B and its structure

Populations classified as the members of *Household B* consist of mobile users of other operators and non-mobile users. However, we do not have any clues about them from the CDRs. Thus, we first estimate N_B (the number of populations belonging to *Household B*) and examine the population structure using SSC data and building map data. Using the map data, we calculate the number of buildings for the i^{th} Voronoi. We obtain the numbers of buildings in the high, middle, low, and slum levels as ${N_b}^H{}_i$, ${N_b}^M{}_i$, ${N_b}^L{}_i$, and ${N_b}^S{}_i$ respectively. Throughout this section, ^{C} denotes ^{H}, ^{M}, ^{L}, and ^{S}, which indicate the high, middle, low, and slum income levels, respectively.

Table 7 shows the average number of *Household B* per building by income level. We calculate the number of populations by income level using the number of buildings by income level. The low-income building is generally a one-story building, which is a non-slum, one-story brick building with cloth-inserted (CI) sheet/semi-pucca. The middle-income building is a two-to-six-story building, and the high-income building is a high-rise building that is at least seven stories high. We use seven as a threshold to separate the middle-income and high-income buildings because restrictions on the construction of buildings in Dhaka applies special rules to buildings that are at least seven stories high, and constructing such buildings requires additional costs. The number of buildings in the slum cannot be compared with other types of buildings in the traditional sense because in a slum, hundreds of households live in small spaces such as a box under a large roof. A slum is generally formed as a colony of such units, and we consider the number of colonies to be the number of buildings in the slum. Through SSC, we surveyed two colonies as a slum in the selected Voronoi area and interviewed 872 households. From Table 7, the average number of *Household B* per building for each income level is obtained as ${F_B}^H$, ${F_B}^M$, ${F_B}^L$, and ${F_B}^S$.

Table 7. Average number of Household B per building by income level for surveyed Voronoi (ith Voronoi)

Income level	Number of *Household B*	Number of buildings	Average number of *Household B* per building
High	201	36= $N_b^H{}_i$	5.6 =F_B^H
Middle	686	181=$N_b^H{}_i$	3.8 =F_B^M
Low	143	91=$N_b^H{}_i$	1.6 =F_B^L
Slum	614	2=$N_b^H{}_i$	307.0 =F_B^S

Table 8 lists the proportions of *Household A* and *Household B* by income level. The higher the income level, the greater the proportion of *Household A*. This means that higher-income people tend to subscribe to *the operator* and, in fact, we observed that the tariff of *the operator* is relatively expensive compared with other operators. The proportion of *Household B* by income level is obtained as p_B^H, p_B^M, p_B^L, and p_B^S.

Table 8. Proportions of Household A and Household B by income level

	High	Middle	Low	Slum
Household A	50%	48%	44%	29%
Household B	50% (=p_B^H)	52%(=p_B^M)	56%(=p_B^L)	71%(=p_B^S)

Equation *(3)* gives the number of populations belonging to *Household B* in the i^{th} Voronoi by income level, and thus the number of populations belonging to *Household B* is given as equation *(4)*.

$$N_{B\,i}^C = N_{bi}^C \times F_B^C \times p_B^C/100 \text{ - (3)}$$

$$N_B = \sum_{k=1}^{n} \left\{ N_{Bk}^H + N_{Bk}^M + N_{Bk}^L + N_{Bk}^S \right\} \text{-(4)}$$

To calculate the number of populations by personal attributes among the populations belonging to *Household B*, we use the proportions given in Table 9, which shows the distribution of personal attributes among people who belong to *Household B*.

Table 9. Distribution of for personal attribute groups among the population belonging to Household B

Income level	Distribution of personal attributes			
	Workmale	Housewife	Student	Other
High	21%	17%	24%=$p^H{}_{Bs}$	38%

Middle	30%	23%	29%=p^{M}_{Bs}	18%
Low	32%	20%	22%=p^{L}_{Bs}	26%
Slum	32%	15%	18%=p^{S}_{Bs}	35%

For example, N_{B_s} (the number of Student among the population belonging to *Household B*) is given as equation *(5)*:

$$N_{B\_s}=\sum_{k=1}^{n}\left\{\begin{array}{l} N^{H}_{Bk}\times p^{H}_{Bs}/100 \;+\; N^{M}_{Bk}\times p^{M}_{Bs}/100 \\ +\, N^{L}_{Bk}\times p^{L}_{Bs}/100 \;+\; N^{S}_{Bk}\times p^{S}_{Bs}/100 \end{array}\right\} \text{-(5)}$$

5.7.3 Number of the entire population

Based on the previously mentioned methods, N (the number of the entire population) is given as equation *(6)*, where N_A and N_B were determined in equations *(1)* and *(4)*.

$$N = N_A + N_B - (6)$$

We note that, because of space limitations, we did not provide all calculations to obtain the number of populations by personal attributes. This can be calculated using equations *(2)* and *(5)*, which are provided as examples in 5.7.2 and 5.7.3.

5.8 SUMMARY

CDRs are receiving increased attention due to their ability to capture the mobility patterns of large-scale populations. Research on a variety of topics such as transportation, urban planning, disaster management, and public health is flourishing, utilizing the movement of people to address societal issues. However, the application of such data may be misleading if the population groups captured by CDRs are not relevant to the issue at hand. In fact, discrepancies between the populations captured by CDRs and the actual living population are noted in several studies. In this study, we demonstrated gaps between those who are captured by CDRs and those who are unobservable. We found that CDRs do not fully capture the movements of entire population groups in Dhaka. In terms of population size, there are roughly 2.4 to 2.8

unobservable people per mobile user identified in CDRs. Considering that for three income levels out of four, males represent more than 60% of mobile users; we can say that the majority of the operator's users are males. In addition, more than 70% of the users are married, and their ages are mostly within the range of late twenties to late fifties. Our findings show that CDRs do not capture specific population groups such as students or children below school enrollment age. This implies that the application of CDRs needs to take such biases into account.

We also provide clues to identifying households with children from the calling behavior of mobile users. We first examine trends in call locations between males and females, and then compare their calling behaviors on primary routine days and non-primary routine days according to the presence of children. The results show that male users with children in their households exhibit consistent trends with regard to calling locations regardless of the type of day. Interestingly, female users exhibit the inverse trend. Although we demonstrated trends in call locations according to gender and the presence of children within the household, we discussed little on the whereabouts of the children. We will further examine time and location distribution of the unobservable including children in future studies.

Our experimental study revealed that it is possible to estimate the presence of the unobservable population in CDRs by analyzing calling behavior of mobile phone users. However, the results of our study can be improved. Limited sample size of validation data may have caused over-fitting of our models, which implies that the dimension of our data is too sparse and the model cannot be applied to large-scale CDRs. Throughout this paper, we focused on capturing the differences in routine activity by classifying the type of days into two groups, primary and non-primary or non-Friday and Friday, and analyzing four-month CDRs. As an alternative approach, we plan to use SPACE data, which are the one-day call records of 922 mobile phone users with personal attribute information to extract features and construct estimation

models. This may require further studies because the data has limited information on calling behavior because of the length of days. However, it can provide increase the sample size for developing the estimation models.

Finally, we would like to emphasize that our findings imply that there is considerable potential to utilize CDRs to address issues related to part of the vulnerable population in the developing world. In many capitals of developing countries, urbanization is a common issue that accompanies rapid economic growth. This is because rural people migrate to urban areas expecting to find more income-earning opportunities in cities. Most of these people, who are poorly educated and lack assets, have to reside in vulnerable areas where the risks of disasters and infectious disease are potentially high. To improve such conditions, it is crucial to understand how many and what kinds of people reside in which places. However, official statistics cannot provide such information because most of these people are not registered in the places in which they are living but rather in their hometowns. Estimates on the spatiotemporal distribution of the population in CDRs represent useful data for policy intervention. We highlight that identifying the approximate distribution of the unobservable population in CDRs from the analysis of mobile phone calling behavior is a valuable component of and application for CDRs. We believe that this study can shed light on such areas by providing clues to understanding the whereabouts of the vulnerable, including the unobservable population from CDRs.

Chapter 6 Development of Dynamic Census

6.1 RATIONALE OF ESTIMATING THE NUMBER OF THE LIVING POPULATION

The availability of good data is crucial for effective policies [1]. According to [2], the quality of data is defined as "*Data of high quality if they are fit for their intended uses in operations, decision making, and planning.*" However, populations residing in informal settlements, which are in need of assistance for better being and generally accepted as overcrowded, underserviced, and dilapidated settlements [3], are potentially underrepresented in many national sample surveys. It means that there are discrepancies between the numbers reported and situation on the ground particularly for slums, which remain a general feature of the urban areas in most regions in Africa, Central and Latin America, and Asia [4]. Reference [5] reported that one third of urban populations in developing world reside in slums. In addition, slums are a significant economic force and it is approximately 60% of employment in the informal sector of the urban population. Not only its size but also potential risks in slums are also a reason why government needs to address issues in slums. The risk of pandemic is potentially high due to poor sanitation and limited access to clean water. So is the risk of secondary disasters because dwellings in slums are generally temporal and easy to collapse. Therefore, issues related to slums are becoming un-negligible in many counties. However, the speed of slum upgrade is generally slow. One of the reasons is the difficulty to see the entire picture of slums. This is because most of slum populations are floating and it is difficult to capture based on official statistics. Sometimes they even do not have citizenship because they just temporarily migrate from the rural area leaving their assets and family members in their hometown seeking better job opportunities in the urban area. To address these issues we propose the novel approach to estimate the distribution of demographic attributes

of entire population on the ground analyzing CDRs from a specific telecommunication company.

6.2 SUMMARY OF THE PROCESS TO DEVELOP DYNAMIC CENSUS

In this chapter, we summarize the process of developing Dynamic Census. Dynamic Census is a dataset that is developed from CDRs by addressing three bottlenecks of CDRs for utilizing CDRs for societal issues: sparseness, anonymity, and population bias.

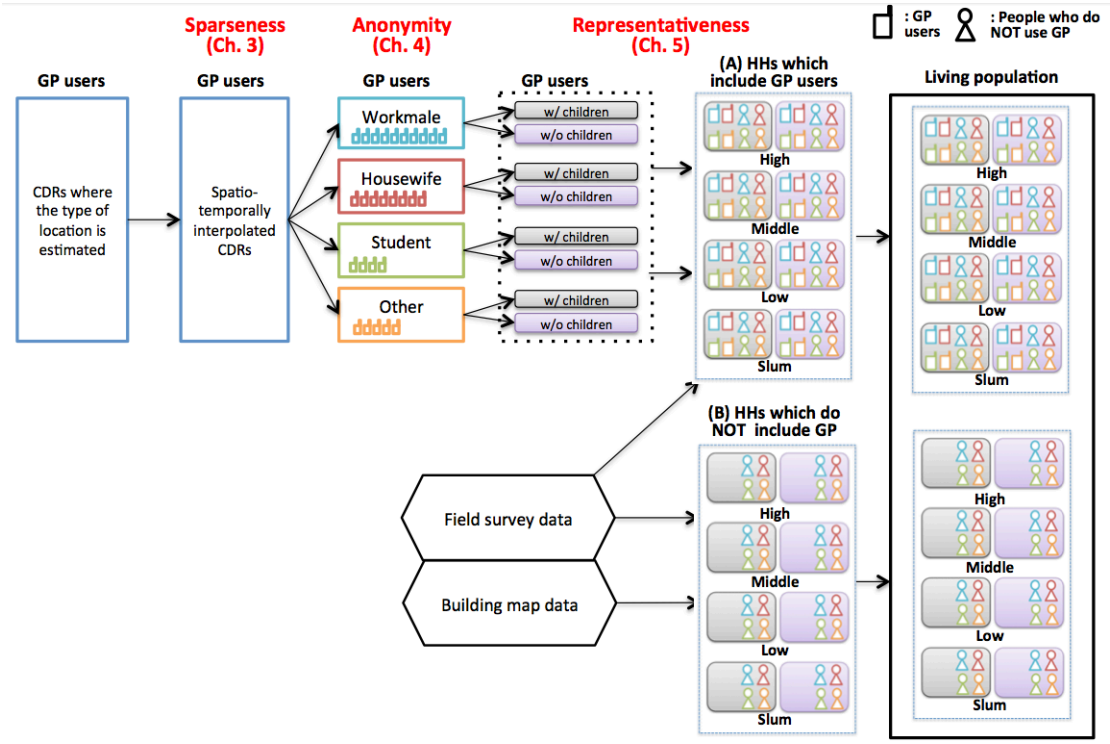


Figure 1. Stages of the process to develop Dynamic Census and its visualization

Figure 1 summarizes the three-stage processes proposed in this study: lowering the sparseness of CDRs (Chapter 3), estimating the personal attributes of mobile users (Chapter 4), and adjusting the population bias in CDRs to represent the living population (Chapter 5). The population discussed in Chapters 3 and 4 is the

population in CDRs. The population covered by Chapter 5.2 is the mobile users, subscribed to *the operator*, and their household members (*Household A)*. Chapter 5.3 discusses the entire living population. It covers the population of the households, which do not include the mobile users, subscribed to *the operator* (*Household B*).

The entire process consists of four parts and is described below. Detailed explanations on the analysis methods, which are used in this section, are omitted because these were already explained previously.

1) Obtain CDRs with the result of interpolation, personal-attribute estimation, and the type of household estimation by:
   (a) Identifying the type of locations for the hourly time bands without call records by the interpolation, employing the estimation model described in Chapter 3.
   (b) Identifying the personal attributes of mobile users, employing the estimation models described in Chapter 4.
   (c) Identifying the type of households to which mobile users belong and estimating the number and personal-attribute structure of the unobservable population who are in *Household A* and *Household B*.
2) Obtain the trajectories of the living population in the area under study by;
   (a) Splitting the processed CDR based on the voronoi area of home locations. Based on the number of the unobservable considering their personal attributes and income levels, the number of population of CDRs is weighted so that the total number of populations represent the living population;
   (b) Splitting the building map data of the study site into voronoi units, whose centroids are determined based on the locations of mobile phone antennas.
3) Disaggregate the trajectories of Dynamic Census by;
   (a) Disaggregating the spatial resolution of the home location, which is the centroid of each voronoi area, to the location of building basis, which are

originally determined by the process described in 1) (a). The number of populations to be allocated for each building is determined based on the process described in Section 5.5; and

(b) Disaggregating the spatiotemporal resolution of the trajectory between all pairs of two points, which is originally hourly basis, is disaggregated into every five-minute basis by the routing method described in Section5.5. The trajectory for each pair of two points, which is originally the Euclidian distance, is converted to the non-liner trajectories based on the road network.

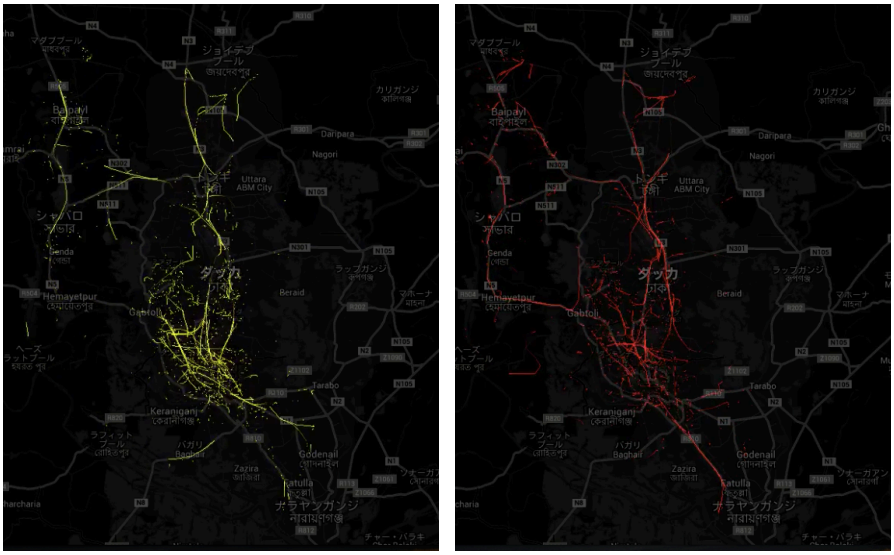
4) Aggregate the population distribution to 500 m^2-grid resolution by;

(a) Splitting the area under study into 500 m^2 grids. Hourly changes in population distribution along with the personal attribute for each grid is computed.

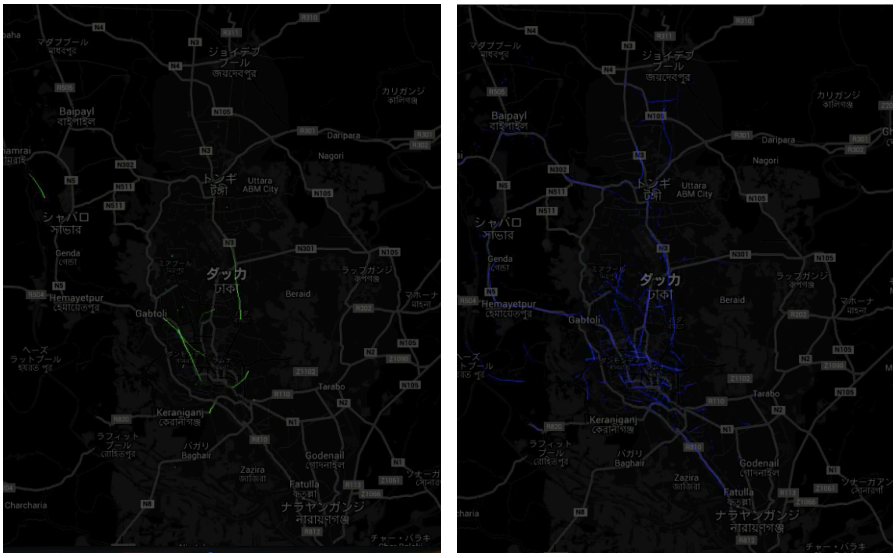
6.3 VISUALIZATION

6.3.1 Reconstructing the trajectories by routing

To reconstruct the trajectories for Dynamic Census from CDRs, we use processed CDRs, in which sparseness and anonymity are already addressed using the methods described in Sections 3, 4, and 5. First, we disaggregate the spatial resolution of processed CDRs, which was originally based on the centroid of the Voronoi area, to the location of the building basis within the Voronoi. Next, we transform the liner path of mobile users in CDRs, which was originally the line segment connecting two antennas, to nonlinear trajectories using a road network. As a result, we obtain nonlinear trajectories for the processed CDRs. Then we use the trajectory generated from CDRs and adjust the distribution of personal attributes to the structure of the living population by adopting the methodology proposed in Section 5 for each Voronoi area. Figures 4(a), 4(b), 4(c), and 4(d) show screen captures of the trajectory movies of a weekday for Workmale, Housewife, Student, and Other.



Figures 7(a) and (b). Trajectories of Workmale (left) and Housewife (right) at around 5pm on a weekday



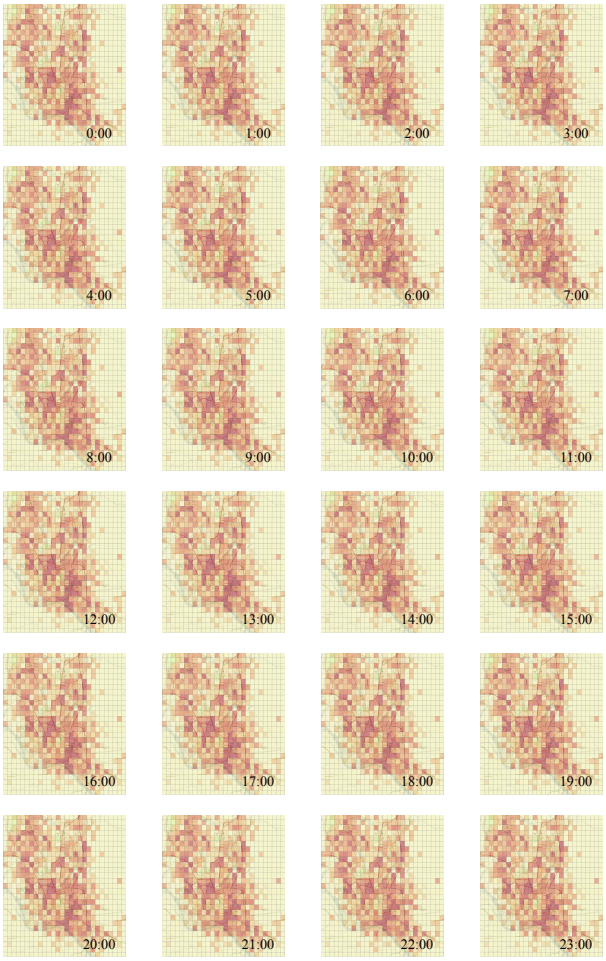
Figures 7(c) and (d). Trajectories of Student (left) and Other (right) at around 5pm on a weekday

6.3.2 Calculating the grid-based spatiotemporal population

Showing trajectories in movies is one of useful ways to understand the dynamics of population movement intuitively. However, such trajectories themselves

are not very useful for policy makers and practitioners when they want to combine with secondary information for policy intervention. We consider the -based map to be user-friendlier than the trajectories because the gridded map can be easily overlaid with other thematic maps. Using the trajectories generated for visualization in 6.3.1, we compute the gridded spatiotemporal population. We calculate the hourly population distribution at the 500-m^2-grid level. The rectangle, which encloses our study site, has a size of 32.5 km by 66.5 km. We use the 500-m^2 meter gridded map to divide the rectangle and obtain 8,645 grids. In addition, statistical information such as personal-attribute structure distribution per Voronoi on hourly basis can be extracted. This can facilitate policymakers and researchers to consider the human mobility dynamics of the living population when designing policies.

Figure 5 shows the hourly population distribution of Workmale on a weekday in 500-m^2grids. We chose Workmale as the most highly mobile population group among the four. Contrary to our expectations, the hourly population distribution of Workmale does not change dramatically between daytime and nighttime. A possible reason is that the distance of people's mobility is limited owing to persistent traffic jams in Dhaka. In fact, the ratio of the daytime population to the nighttime population is 0.89, where the daytime population is 6.5 million and nighttime population is 7.3 million [9]. This is a 10% change in population number.



Figures 8. Hourly population distribution of Workmale on a weekday

Chapter 7 Framework to recreate Dynamic Census in other cities

To recreate Dynamic Census in other cities by adopting the approach proposed in this study, it is evident that secondary data play crucial roles. In this chapter, a framework to collect secondary data is provided from three aspects; key information to be collected, survey structure for the data collection, and sampling.

7.1 KEYS TO RECREATING DYNAMIC CENSUS

In Chapter 1 three problems of CDRs, which are the sparseness, bias, and anonymity are raised, to develop Dynamic Census. So, the framework to recreate Dynamic Census is described in the context where these three problems are addressed. Table 1 summarizes the key information and suggested approaches for addressing the problems. It provides the general framework of the secondary data collection and detailed descriptions on information items are provided in Table 2. As discussed in Chapter 1, CDRs provide partial view of human mobility, which can be explained by the spatiotemporal distribution of repeatedly visited locations such as home, work place, and school, i.e. significant locations. It means it is important to extract the relationship between the partial view of people's activity, which can be obtained from call records, and the people's activity along with their demographic attributes, obtained from their schedule and characteristics, through the analysis of survey data. As described in row (B), some information needs to be collected both from mobile users and non-mobile users for addressing the bias problem while the population to be sampled for rows (A) and (C) is the mobile phone user alone. One of keys in the secondary data collection to recreating Dynamic Census is to determine an indicator, which allows us to calculate the magnification factor for obtaining the entire living population of the study site as explained in (c) of row (B). In this thesis, we used the number of buildings as the indicator. This is because we found we can

use it as the proxy of the income level where access to the mobile phone varies to some extent, e.g. the higher the income level, the greater the proportion of mobile phone users. In addition, we assume their mobile usage varies according to the income level, e.g. the higher the income level is, the more amount of money they spend on the mobile phone. In this thesis, the distribution of buildings was extracted from the building data, which are assumed to be generated from aerial photos and include slum settlements. We tried not to use the map, which are published by the official institution, for this study because the slum settlement is normally recorded as vacant land. It is highly recommended to find information, which can capture the actual condition on the ground without some filters or bias.

Table 1. Key information and suggested approach to recreate Dynamic Census

Problems	Suggested approach	Key information to be collected
(A) Sparseness	(a) Analyze the relationship of spatiotemporal distribution between the people's significant locations and call records.	• Call records from the handset along with the user's schedule including time, locations, and activities.
(B) Bias	(a) Analyze the distribution of the demographic attribute and mobile phone usage among the living population.	• Demographic structure including activity and mobile phone usage of living population.
	(b) Examine market share of mobile phone companies (depends on the CDRs used to recreate Dynamic Census).	• Distribution of telecommunications companies subscribed among mobile users • Proportion of non-mobile users among the living population.
	(c) Identify a key indicator to calculate the magnification factor to estimate the population structure of the entire living population based on the structure of the living population, sampled through (2) and (3).	• Information to estimate the distribution of entire living population of the target area, in which Dynamic Census is to be recreated.
(C) Anonymity	(a) Analyze the relationship between calling behavior and demographic attributes.	• Call records from the handset along with the users demographic attribute and schedule. • Weekly routine and routine activity per user.

Table 2 provides the detailed description on the information item to be collected. The items are organized by the population group to be sampled. Living population in Table 1 consists of the mobile user and non-mobile user. So, if the population to be sampled includes both of them, the survey is the household basis.

Because most of items are common to both of them, it is suggested to conduct the household basis survey for the secondary data collection instead of interviewing mobile phone users alone. In the table, some optional items are added considering these items are necessary under some specific setting. During the fieldwork, we encountered three issues due to the popularity of the feature phone in our study site, which are most likely the case in many developing countries for now. First, an existing call record is replaced with a new record when a call is made to the same number. To cope with this issue, we asked whether the interviewee made any calls listed in the record, and ask approximate time if the interviewee found the old record is replaced. Second, we found that recording the actual time at the interview and time shown in the interviewee's handset is critical to collect the correct time of calls. During the pre-test of our field survey, it was often observed that the clock of the feature phone is not correctly set. It is common especially among the lower-income people because they do not know how to set it. Apart from the smartphone, time setting of some feature phones is reset when it is switched off or SIM card is removed from the handset. (It is not uncommon that people have more than one SIM card and occasionally choose it to minimize the cost of calls.) Third, we found some feature phones, which are very cheap and whose functionality is very limited, store only a few records of calls. These kinds of issues need be noted when Dynamic Census is recreated under a setting where the income level of large part of the society is not very high.

Table 2. Information items to be collected as the secondary data

Population to be sampled		**Information to be collected**	
Mobile user	**Non-mobile user**	**Categories**	**Items**
√	√	Personal attribute	• Gender • Age • Routine activity (An activity in which the person is primarily engaged

		(Optional)	• Type of activity (Specify whether it is income-earning or non-income earning activity)
√		Call records (Optional)	• Day of a week, time, duration, and the type of locations of calls • Specification of the callers (Phone sharing among family members is common in some developing countries) • Actual time at the interview and time shown in the interviewee's handset
√		Schedule	• Starting time and ending time of activity • Type of the location of the activity
√	√	Mobile ownership	• Whether the person has own mobile phone
√		Mobile usage (Optional)	• Names of the telecommunications operator (Owning more than one SIM cards is common in some countries) • Specification of the primary and secondary SIM card based on the frequency of usage
√	√	Housing	• Stories of building (Considered as the income level in the case of this thesis) • Number of housing units per floor of the building

It is ideal if we can always conduct surveys to get all necessary information from the field. However, conducting the interview survey is generally costly. Table 3 provides the list of data, which can be alternatives to the key information suggested in Table 2. We tried to list alternative data from public domain data or data which are collected by the public sector. We expect that such data can be used for recreating Dynamic Census if the purpose of usage is for increasing social welfare. It needs to be noted that Origin-Destination survey data and Population and Housing Census data are generally published in the aggregated form. It means that the unit or level of aggregation affect the process to relate the information derived from the census data and other datasets used to recreate the Dynamic Census.

Table 3. Possible alternative data to the secondary data described in Table 2

Key information suggested in Table 2	**Alternative**	**Notes**
Schedule and personal attributes	Origin-Destination survey data	• Some surveys exclude people who are part of the living population but do not affect the volume of human mobility, e.g. infant.
Mobile ownership	Population and Housing Census data	• Information of the mobile phone ownership is recommended as a core topic for the Population and Housing Census by (United Nations, 2008). • The location of home used in the survey is based on the registered address. Thus, there can be a gap between the population used in the survey and living population for a given area.
Housing (Partial)	Satellite image or arterial photos	• Distribution of buildings can be estimated analyzing the data. However, the estimation of the population distribution will be challenging.

7.2 SURVEY STRUCTURE

To collect the secondary data, we suggest the structure of the survey described in Table 4. At the first stage, it is highly recommended to conduct a baseline survey to understand the local context of mobile phone usage and to collect information to design a calling behavior survey and small-scale census survey. We underline the importance of finding a key indicator to calculate the magnification factor, which is used to estimate the population structure of the entire living population. It is because the population structure of regions within the area covered by the entire living population most likely differs from that of sampled area. Furthermore, it is crucial to determine the indicator for planning surveys of the second and third stages. In the case of Dhaka, the population structure is defined by the type of buildings including the number of stories, which is the key indicate for this study. It means that we obtained the distribution of four income-level population groups for the study site by using the type of buildings. As mentioned, it is considered as the proxy of the household income level. We also used the type of buildings as on of the stratifiers for the sampling for the second stage.

Table 4. Structure of the survey

Stage of the survey	Purpose of the survey	Population to be sampled
(1) Baseline survey	• To collect information to design (2) and (3), e.g. - Information for sampling - Usage of the mobile phone in the local context • To identify a key indicator to calculate the magnification factor to estimate the population structure of the entire living population based on the structure of the living population, sampled through (2) and (3)	N/A
(2) Calling behavior survey	• To collect information on calling behavior and household structure of mobile phone users, e.g. - Personal attribute - Call records - Schedule - Mobile ownership • Mobile usage	• Mobile phone users of the telecommunication operator, whose CDRs are used for the study • Household member of the mobile users specified for (2)
(3) Small-scale census survey	• To collect information on the structure of the living population for a given area, which is used to estimate that of the living population, e.g. - Personal attribute - Mobile ownership - Building profile	• Entire living population of the study site, which include both any mobile users and non-mobile users

At the second stage, the survey focuses on collecting information on the personal attribute and calling behavior of mobile users. We suggest surveying mobile users of the telecommunications company whose CDRs are used for recreating Dynamic census. This is because data collected through this survey are used to generate features to identify the personal attribute of mobile users analyzing CDRs. To minimize the budget for the survey, conducting the calling behavior survey for all companies increases the cost because collecting call records from handsets is the most time consuming part for this stage. Furthermore, given that budget is limited, sample size for the mobile users subscribed to the specific company will be reduced. Rather that that, we suggest surveying non-mobile users, who are the rest of the household members of the mobile user, if the budget allows. Such information can be used to identify the unobservable population, who are not included in CDRs. Also, it can be background information for planning the small-scale census survey. We would note that this part could be omitted from this stage because information on the personal attribute and mobile phone ownership is collected from the sample of entire living population in the third stage.

At the third stage, the population to be sampled is the entire living population. The survey of this stage is a census survey, which collects information on the personal attribute and mobile phone ownership from each member of households residing in a given area. In addition, information on the key indicator needs to be collected from all households so that the data collected through this survey is linked to the indicator, which determines the population structure for the sample. In the case of Dhaka, we surveyed the building profile for all buildings included in the survey site. We needed to know the number of households and household members, residing in each of four different types of buildings, because the number and spatial distribution of the entire living population in Dhaka are to be calculated based on the number of buildings with

the specification of the type of buildings, including the number of stories of the building.

7.3 SAMPLING

Sampling is an important component for collecting a subset of individuals who are the population under study. In this section, we discuss sampling from two aspects, which are considered to affect the survey design; one is the choice of the sampling method and the other is the sample size. Suggestions on the sampling method and sample size are provided for the calling behavior survey and small-scale census survey.

Regarding the sampling method for the calling behavior survey, we recommend employing the multi-stage stratified sampling. Populations in cities, for which recreating Dynamic Census is considered, are mostly diverse, and stratification is necessary to capture the homogeneity of the population structure. As one of stratifiers, the key indicator, which is determined in the baseline survey and is used for obtaining the magnification factor to estimate the number of entire living population, is used to divide the sample into strata. In the case of Dhaka, two-stage stratified sampling is employed using the type of buildings and that of land use for dividing the sample. It is ideal to realize the proportional allocation for the size of the sample in each stratum. In the case of Dhaka, the survey is household basis because one of stratifiers used for the stratification is the type of buildings, which is the proxy of the household income level. The unit of the survey depends on the design of the survey.

It is difficult to suggest the exact sample size necessary for the calling behavior survey for recreating Dynamic Census in other cities not knowing what sampling method is used in what context. As examples, we provide the sample size and effect size of selected key features, which are used to identify gender for the

mobile phone user analyzing CDRs previously. It means that we examine whether the key feature explains the difference between males and females with certain statistically significance. If the observed difference is small, we investigate whether it is just because of sample size, and thereby sample size increment helps extract the difference. We consider that the demonstration of a process to obtain expected sample size and effect size for the key features can be a useful reference for determining the sample size for the calling behavior survey to recreate Dynamic Census in other cities.

Here, we briefly define the term, the sample size and effect size (ES), used in this section. Sample size is the number of observations of a sample, which are selected from a population under the study. We make inferences about a population by analyzing the sample. ES is a quantitative measure to evaluate the magnitude of effect or association between two or more variables [1]. There are various ES indices according to the purpose of the statistical tests. For instance, ES index for t tests of independent means of the phenomena *A* and *B* for the non-directional case is provided as below;

$$d = \frac{|m_A - m_B|}{\sigma} \quad - \quad (1)$$

where m_A and m_B are the population mean of groups *A* and *B*, σ_A and σ_B are the standard deviation of groups *A* and *B*, and $\sigma = \sqrt{(\sigma_A^2 + \sigma_B^2)/2}$ is the standard deviation of groups *A* and *B*. The ES index can evaluate the effect size of the mean difference by expressing score distance in units of variability, i.e. *d* represents how the means differ by two standard deviations. It is quite common to use null-hypothesis significance testing as a tool for examining the data. However, limitations of the null-hypothesis significant testing has long been pointed out due to three reasons; it is sensitive to sample size, it has inability to accept the null hypothesis, and null-hypothesis significance testing cannot determine the practical significance of

statistical relationships [2,3,4,5]. On the other hand, the effect size is mostly considered to be resistant to sample size influence, and thus provide a truer measure of the magnitude of effect between variables [6]. It is recommended to use the effect size in addition to null-hypothesis significance testing [7]. Because we intend to provide references for determining the sample size to recreate Dynamic Census, we consider it is better to use key features, whose value differences can be statistically evaluated using indicators mentioned above. Here, we select two features extracted from calling behavior, which are considered to be useful to see gender differences.

First, we use t test for two independent means to investigate whether observed differences in calling behavior have statistical significance. We consider it important that such features at least have a certain statistical significance because we cannot proceed to the statistical power analysis and sample size planning if these features do not have the significance. The arithmetic means is one of the most frequently used measures for comparing the location of two groups of samples. Formally, it is applied under the assumption where the population sampled are normally distributed and of equal variance. However, departing form the assumption generally has negligible effects on the validity of both Type I and Type II error calculations. This is particularly true for non-directional test and sample size increase about 20 or 30 cases [8].

Table 4 describes the t test results for non-directional cases. We use Welch's t test because the standard deviation of two samples, males and female, are not equal.

Table 4. t-test results

Feature	Mean		Standard deviation	t-value	p-value
(A) Proportion of the number of calls from home to the total number of calls	***Male***	0.4962	0.2817	-2.8726	0.0059
	Female	0.7116	0.2708		
(B) Proportion of the number of calls from the primary location outside the home to the total calls	***Male***	0.1342	0.0450	2.6618	0.0126
	Female	0.0159	0.0718		

Based on the results and equation *(1)*, we obtain the ES index *d* for features (A) and (B) as 0.7796 and 1.9744, respectively. To interpret the obtained *d*s, three values, U_1, U_2, and U_3, are provided in Table 5. These values are defined as measures of overlap (U) associated with *d* with the assumption that the populations compared are normal and with equal variability. U_1 is the percentage of non-overlap among the distributions of two populations. U_2 is the percentage in the population with the greater mean exceeds the same percentage in the population with the smaller mean. U_3 is the percentage of the population with smaller mean, which the upper half of the case of the population with greater mean exceeds. In the case of feature (B), we find d=2.0 where U_1 equals to 81.1%. It means that there is 81.1% of combined area not shared by males and females. In this case, the highest 84.1% of males exceeds the lowest 84.1% of females, thus U_2=84.1%. The upper half of males exceeds 97.7% of the 97.7% of females, so that U_3=97.7%. Cohen (1988) provided operational definitions for the value of ES index *d* as small size effect (d=0.2), medium size effect (d=0.5), and large size effect (d=0.8). Therefore, we can say that the ES for features (A) and (B) is large.

Table 5. Equivalent of *d*

d	**U_1**	**U_2**	**U_3**
0.8	47.4%	65.5%	78.8%
2.0	81.1%	84.1%	97.7%

Source: [2]. Original table is modified by the author.

In addition to the ES, we introduce the concept of statistical power that is defined as "the probability that it will lead to the rejection of the null hypothesis, i.e. the probability that it will result in the conclusion that the phenomenon exists" (Cohen, 1988). The power depends upon three parameters: the significance criterion, sample size, and ES. We can use the power table when the three parameters are specified. It means we can determine the sample size *n* with the specification of a certain ES, significance criterion α, and the amount of power the investigator desires. Table 6

provides the power of non-directional t test of two populations at α=0.05. If we anticipate to have the power=0.8 with the ES=0.5 (medium), the sample size of each group is 64 and thereby 128 samples in total is necessary.

Table 6. Power of non-directional t test of two populations at α=0.05

Sample size	Power		
	ES=0.2 (Small)	ES=0.5 (Medium)	ES=0.8 (Large)
10	0.07	0.18	0.39
20	0.09	0.33	0.69
30	0.12	0.47	0.86
40	0.14	0.60	0.94
50	0.17	0.70	0.98
52	0.17	0.71	0.98
54	0.18	0.73	0.98
56	0.18	0.74	0.99
58	0.19	0.76	0.99
60	0.19	0.77	0.99
64	0.20	0.80	0.99
68	0.21	0.82	*
72	0.22	0.85	*
76	0.23	0.86	*
80	0.24	0.88	*
84	0.25	0.90	*
88	0.26	0.91	*
92	0.27	0.92	*
96	0.28	0.93	*
100	0.29	0.94	*

* Power values with this point are greater than 0.995.
Source: [8]. Original table is modified by the author.

As the specification of desired power as power=0.8 is suggested as a convention by [8]. Therefore, we can summarize the necessary sample size *N* for small, medium, and large ES at power=0.8 for a=0.01, 0.05, and 0.10 for mean difference in Table 7. We note that the numbers in the table is the sample size *N* for each group. Therefore, to obtain the total sample size, the number should be multiplied doubled.

Table 7. Sample size for small, medium, and large ES at power=0.8 for α=0.01, 0.05, and 0.10

α								
0.01			0.05			0.10		
Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
586	95	38	393	64	26	310	50	20

Source: [2]. Original table is modified by the author.

As for the sample size of the small-scale census survey should be larger than that of mobile phone users surveyed for the calling behavior survey because the number of entire living population is much greater. In the case of Dhaka, the sample size of the living population in the small-scale census is around 11,500 and that of the mobile users in the calling behavior survey is around 900, which is approximately 8% of the 11,500. According to the survey results, described in Figure 6 in section 5.3, the proportion of the mobile user of the telecommunication company under study is between 9% to 20%+ depending on the income level.

7.4 SUMMARY

In this chapter, a framework to collect secondary data for recreating Dynamic Census in other cities was provided. We suggested how the survey for data collection is to be designed based on the concept that the secondary data are used to resolve three problems of CDRs, such as sparseness, bias, and anonymity. It is essential that the secondary data include information on calling behavior, which can relate CDRs and the data collected from the field. As for sampling, sampling methods and reference information, which is considered to be useful for determining the sample size, are described. We underline that the framework provided in this chapter is merely a framework and thereby careful investigation on local context during the preparation stage is vital. Additionally, we tried to provide the list of alternative data for the secondary data, which are in public domain or can be accessible for recreating Dynamic Census. However, the data on the list are limited for now. Further investigation is necessary to identify what data sources can provide what variables necessary for the study.

Chapter 8 Conclusions and future prospects

8.1 CONTRIBUTIONS OF THIS STUDY

This thesis proposed a novel approach to overcome three constraints of CDRs for utilizing the data for societal issues, namely, sparseness, bias, and anonymity of the data. By addressing the constraints, a new dataset is developed and named as Dynamic Census. Dynamic Census is the data of large-scale human mobility where the mobility is provided as the trajectory; the population represents the actual living population for a given area; and the personal attribute of individuals in the data are identified. In addition to that, a framework to recreate Dynamic Census in other cities is provided. Contributions of this study are summarized as below;

3) Provided the method to develop Dynamic Census by;
    - Identifying the population bias of CDRs and the presence of the unobservable population in CDRs, by analyzing CDRs in combination with the field survey data;
    - Interpolating sparse CDRs with respect to the spatiotemporal aspect. The location type for the time band without the call records of the mobile phone user is determined employing the Latent Dirichlet allocation model. For visualization, trajectories are estimated using road network data of Dhaka; and
    - Determining the personal attribute of mobile phone users through analysis of calling behavior of mobile users using CDRs in combination with the survey data. Estimation model employing Random Forest is developed.
4) Provided the framework to recreate Dynamic Census in other cities by;

- Explaining key information to be collected as the secondary data for identifying the unobservable population and estimating the personal attributes of mobile phone users;
- Recommending the structure of the field survey and detailed information items to be collected. A list of alternative datasets for part of the secondary data is suggested; and
- Providing reference information for determining the sample size for the secondary data collection. Instead of the exact sample size, the examples of the sample size for identifying traits in calling behavior are provided based on key features used in the estimation in this study.

5) Shed light on the population, which tends not to be included in general statistics due to systematic reasons by;

- Capturing actual living population in a given area under study. As the examples of the population, seasonal migrants and slum dwellers in the urban area are expected to be captured by Dynamic Census. The size of these populations is increasing due to urbanization associated with rapid economic growth in many developing countries. Because the impact of such populations to the city is not negligible, improving the understanding on their conditions is crucial for effective policy intervention.
  - Seasonal migrants: Those who temporarily reside in the urban area to get temporal jobs during the idling season for agricultural production, i.e. dry season. They are not included in the registration system in living locations in the urban areas because their registrations remain in their hometown with their family members and assets.
  - Slum dwellers: Those who illegally occupy public land will not be included in some official registration systems because they occupy the land, which is of public and is not registered as residential land. Slums

are formed in vulnerable locations such as riversides and coastal areas, and thereby potential risks of infectious disease and disaster are quite high with poor sanitation and infrastructure. Seasonal migrants are sometimes part of this population.

These contributions help policy makers and researchers to realize better social welfare in various sectors, including urban planning, transportation, disaster management, and public health, by providing more accurate estimates on the spatiotemporal distribution of the actual living population under study. Though approaches provided in this study are based on a case study of one city, it is possible to apply these approaches to CDRs in other cities for recreating Dynamic Census. This is because CDRs are routinely accumulated wherever the mobile network is available and the key components of data are principally common. Additionally, CDRs can be collected regardless of the type of mobile phones, both feature phones and smartphones.

I also underline the importance of conventional official statistics, and Dynamic census cannot simply replace such data for this moment, given the nature of such conventional data, e.g. Household and Population Census data. Such data can provide the viewpoint of the long-term static transition in the population dynamics. So, both data are rather mutually complementary where Dynamic Census is superior in revealing individual human mobility dynamics. However, conducting the census survey is costly because the data collection process for entire information items is still corresponding to individual interviewees, who submit their information by themselves. It implies that there is the potential that Dynamic Census can contribute to reducing the cost of census survey by replacing part of Census data because part of components of both data are common.

8.2 FUTURE PROSPECTS

Lastly, I propose future prospects by suggesting how limitations of this study can be addressed.

1) Comparative studies are necessary for evaluating how accurately the proposed framework can make estimates according to different social/cultural background of study sites
   - Dynamic Census was developed through a case study of Dhaka. Frameworks of developing Dynamic Census were provided as the generalization of the methodology. However, it is not mentioned that how accurately the proposed framework can make estimates according to different social/cultural background of study sites because the demonstration is done through a single case study. Considering the strong social norm for the role in the family and the Muslim society, the trend of personal attributes, such as gender and the role in the household, observed in Dhaka population is considered to be relatively simple. To evaluate how the proposed method can make estimates, comparative studies are necessary.
   - Under a setting where the lifestyles of people are diverse, it is necessary to extract more features from calling behavior to capture the traits of people's activity. To specify what kind of additional information is necessary, it is essential to specify which part of the proposed method can be used as the framework of the analysis to resolve three constraints of CDRs. In this thesis, listed alternative secondary data were limited. By doing so, we expect that we can provide more alternative information items.

2) Improve estimation results

The estimation results of the presence of the unobservable population, the personal attributes of mobile phone users, and the number of the living population can be improved. We provide four possible approaches as below:

(a) Collect/use validation data, which require less costs of data collection

When Dynamic Census is recreated in a study site where feature phones are common, we need to collect validation data through field survey. Under this circumstance, collecting larger number of samples is difficult because collection procedure is manual. Therefore, one of possible ways to improve the estimation results would be constructing the mode using CDRs of shorter terms, which are easier to collect compare to the longer-term data. If Dynamic Census is recreated in a study site where smartphones are common, we can utilize mobile apps to collect validation data, which allow us to communicate with mobile users directly. We expect the spread of smartphones can contribute to reducing the cost of validation data collection and thereby improvement of estimation results.

We need to note that there will be certain population groups who do not use smartphones even if smartphones become common. Thus, it is necessary to examine approaches to collect validation data for these people even if the portion is marginal, e.g. combination of surveys through mobile apps and field survey.

(b) Extracting features, which explain calling behavior traits, with less information losses:

CDRs originally have spatiotemporal dimensions (multi-dimensional) but it was converted to mono-dimensional features in this study. We labeled the type of locations and computed statistical values for calling time distributions, which is considered to have caused significant information losses.

(c) Developing methods to capture the distribution and type of buildings from imagery data:

Current population estimates heavily rely on the data quality of building map data. We emphasize that using imagery data is vital because it capture the situations of the ground directly and do not filter out anything.

(d) Examine the discrepancy and similarity in activity patterns of people in CDRs and the unobservable, who are classified to the same personal attribute:

In this study, we used the mobility patterns of people in CDRs for reconstructing those of unobservable people. We assumed that activity patterns of the people in CDRs and the unobservable were similar if people in two population groups were classified to the same personal attribute. This is partly because we surveyed the schedule of mobile users alone due to the limitation of budget and time. To examine whether there is a discrepancy in the activity patterns between people in CDRs and the unobservable, it is necessary to survey the schedule of the unobservable and to compare it with that of people in CDRs. Regarding small children, whose presence in the household was estimated, it is difficult to extract their behavioral patterns from CDRs because they are rarely included in CDRs. Most of them are classified as “Other” (one of personal-attribute labels), in which the elderly are also classified. Because most of small children and the elderly generally spend majority of time at home, we consider our estimation results depict their whereabouts to some extent. However, some small children go to the kinder garden and some educational institutions, which are difficult to capture using CDRs. Therefore, field survey on such populations are vital to improve estimation of activity patterns of these population groups.

(e) Employ other machine learning methods:

In this study Random Forest is used for clustering. However, it is better to employ several other methods such as Support Vector Machine and Kernel

method and compare the accuracy of estimated results. These methods are also common approaches for clustering, which require defining features prior to running the model. Recently, Deep Learning, one of machine learning methods, has been increasingly utilized in the field of image processing. It dose not require predefining the input features and thereby can reduce the risk of excluding important features in the model. So far, we cannot find the examples of application of Deep Learning to spatiotemporal data analysis but it would be possible to utilize existing algorithm for spatiotemporal data. Deep Learning requires the large volume of data for modeling and the huge capacity of data processing environment. Therefore, it may take time to utilize existing algorithms but worth-trying for improving estimation results of Dynamic Census.

3) Develop validation methods

- Dynamic Census is a dataset that can depict the living population, part of which are not generally included in official statistics such as census data. Because of this, it is difficult to find data, which can validate the outcomes of Dynamic Census though the validation is very important when the data are used for improving social well-beings. Therefore, it is vital to develop a method to validate overall trends of Dynamic Census.
- One of validation methods we can propose for this moment is partial validation. If we additionally collect data from other Voronoi areas in Dhaka, which will be surveyed just as we did through Small-scale Census, we can validate part of Dynamic Census. However, we need to consider a time difference in data acquisition between CDRs and validation data, considering that mobility of people in Dhaka is high. For example, in our panel field survey, we can trace 75% of households which were originally surveyed one year before.

4) Develop data aggregation methods by which information losses for creating Dynamic Census are minimized

- Dynamic Census was developed using anonymized CDRs, which allowed the analysis on the patterns of individual calling behavior and mobility. It was very fortunate for us that we could access to raw and non-aggregated CDR for this study but it will become increasingly rare. This is because the provision of non-aggregated CDRs tends to be perceived as sensitive issues and thereby mobile network operators prefer providing processed data to raw data. There are a couple of processing methods with less information losses such as mixing noise and data aggregation. The former one is with less information loss because the data are still on individual basis. However, we found practitioners were generally more comfortable with aggregated data than the data with noise. This is because releasing data on individual basis may raise people's privacy concern even if the data are properly processed. Therefore, data aggregation methods, which still allow us to recreate Dynamic Census is critical as future studies. We expect opportunities to recreate Dynamic Census can be significantly increased if the framework to recreate Dynamic Census is developed by using aggregated CDRs.

Despite the limitations described above, I believe that this study can provide more opportunities for policy makers and practitioners to incorporate the dynamics of human mobility into policy making and project design for better social well beings using CDRs. It will not take long to observe much higher penetration rates of smartphones in developing counties. It means that sparseness of data used to recreate Dynamic Census is gradually addressed and the partial view of human activity depicted through mobile phone log data is closer to actual activity. On the other hand, representativeness of data will last because it is almost impossible to collect

uniformly from entire population through a specific device. There may be a potential that allocation of universal identification codes specific to individuals may address such a problem. In this regard, integration of data collected through indoor environment can improve the accuracy of Dynamic Census and the framework will be still contribute to understanding the dynamics of the living population through data obtained from mobile devices.

References

Chapter 1

[1] World Bank. 2012. 2012 Information and communications for development: maximizing mobile. Washington, DC: World Bank.

[2] International Telecommunication Union. The world in 2013: ICT facts and figures. Retrieved 23 April, 20105, from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf

[3] Song, C., Koren, T. K., Wang, P, and Barabási, A. L. 2010. Modelling the scaling properties of human mobility. Nat. Phys. 6, 10, 818-823.

[4] González, M. C., Hidalgo, C. A., and Barabási, A. L. 2008. Understanding individual human mobility patterns. Nature. 453, 7196, 779-782.

[5] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M. J., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In Pervasive Computing, 133-151. Springer Berlin Heidelberg.

[6] Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012, June). Human mobility modeling at metropolitan scales. In Proceedings of the 10th international conference on Mobile systems, applications, and services (pp. 239-252). ACM.

[7] Roth, C., Kang, S. M., Batty, M., & Barthélemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. PloS one,6(1), e15923.

[8] Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E., & Snow, R. W. (2013). Mobile phones and malaria: modeling human and parasite travel. Travel medicine and infectious disease, 11(1), 15-22.

[9] World Bank. 2014. World Development Indicators 2014. Washington, DC: World Bank.

[10] Murad, S. M. W. 2009. The trends of labor market in Bangladesh and its determinants. MPRA Paper No. 32408.

[11] Central Intelligence Agency. 2013. The World Factbook 2013-14. Washington, DC: Central Intelligence Agency.

[12] Zuman, A. K. M. H., Alam, K. M. T., and Islam, M. J. 2010. Urbanization in Bangladesh: Present Status and Policy Implications. ASA University Review, 4 (2), 1-16.

[13] World Bank. In Bangladesh, the alternative to urbanization is urbanization. Retrieved 23 April, 2015, from http://blogs.worldbank.org/endpovertyinsouthasia/bangladesh-alternative-urbanization-urbanization

[14] Centre for Urban Studies, National Institute of Population Research and Training, and MEASURE Evaluation. 2006. Slum of urban Bangladesh: mapping and census. Dhaka, Bangladesh and Chapel Hill, USA.

[15] Grameenphone. Overcoming barriers to internet usage in Bangladesh. Retrieved 28 April, 2015, from http://www.telenor.com/wp-content/uploads/2014/06/04-Grameenphone-IFA-presentation-FINAL.pdf

[16] Bangladesh Telecommunication Regulatory Commission. Mobile phone subscribers in Bangladesh March 2015. Retrieved 28 April, 2015, from http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-march-2015

Chapter 3

[1] Wesolowski, A., Eagle, Nathan., Abdisalan M. N., Snow, R. W. and Buckee, C. O. 2013. The impact of biases in mobile phone owner-ship on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81).

[2] Song, C., Qu, Z., Blumm, N., and Barabási, AL. 2010. Limits of predictability in human mobility. *Science*, 327(5968): 1018-1021.

[3] González, M. C., Hidalgo, C. A., and Barabási, AL. 2008. Understanding individual human mobility patterns. *Nature*, 453(7196): 779-782.

[4] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In Pervasive Computing, 133-151. Springer: Berlin Heidelberg.

[5] Lu, X., and Pas, E. I. Socio-demographics, activity participation and travel behavior. 1999. *Transportation Research Part A: Policy and Practice*, 33(1), 1-18.

[6] Eagle, N., and Pentland, A. S. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7), 1057-1066.

[7] Farrahi, K., and Gatica-Perez, D. 2011. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1), 3.

[8] Zheng, J., and M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. 2012. In the *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 153-162. ACM.

[9] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research,* 3, 993-1022.

[10] Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences,* 101(Suppl. 1), 5228–5235.

[11] United Nations. 2008. Principles and recommendations for population and housing censuses: revision 2. Statistical papers Series M, No.67/Rev.2. New York.

[12] United Nations. 2010. World population and housing census programme. Retrieved 9 February, 2015, from http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/censusclockmore.htm.

Chapter 4

[1] Song, C., Koren, T. K., Wang, P, and Barabási, A. L. 2010. Modelling the scaling properties of human mobility. *Nat. Phys*. 6, 10, 818-823.

[2] Song, C., Qu, Z., Blumm, N., and Barabási, A. L. 2010. Limits of predictability in human mobility. *Science*, 327.

[3] González, M. C., Hidalgo, C. A., and Barabási, A. L. 2008. Understanding individual human mobility patterns. *Nature*. 453, 7196, 779-782.

[4] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M. J., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In Pervasive Computing, 133-151. Springer Berlin Heidelberg.

[5] Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, 239-252. ACM.

[6] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W. Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 298-307. ACM. New York, NY.

[7] Mo, K., Tan, Ben., Zhong, Erheng., and Yang, Q. 2012. Report of task 3: your phone understands you. Paper presented at Nokia mobile data challenge 2012 workshop, Newcastle, UK, 18-19 June 2012.

[8] Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. 2012. Understanding mobility in a social petri dish. *Scientific reports*, 2.

[9] Pas, E. I. 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. Environ. Plann. A, 16, 5, 571-581.

[10] Lu, X., Bengtsson, L., and Holme, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, 109, 29, 11576-11581.

[11] Blumenstock, J., and Eagle, N. 2010. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, 6. ACM.

[12] Aarthi, S., Bharanidharan, S., Saravanan, M., and Anand, V. 2011. Predicting customer demographics in a mobile social network. In *Proceedings of 2011 IEEE International Conference on Advances in Social Networks Analysis and Mining*, 553-554. IEEE.

[13] Lu, X., and Pas, E. I. 1999. Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice*, 33(1), 1-18.

[14] Financial Express. Where Dhaka stands today. Retrieved 30 May, 2014, from http://www.thefinancialexpress-bd.com/old/index.php?ref=MjBfMDJfMjZfMTNfMV85Ml8xNjEyOTE=

[15] World Bank. 2013. World development indicators. 2013. Washington, D. C., World Bank.

[16] Bangladesh Telecommunication Regulatory Commission. Mobile phone subscribers in Bangladesh 2014 April. Retrieved 23 May, 2014, from http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-april-2014

[17] Sultana, N. 2004. Polygamy and Divorce in Rural Bangladesh. *Empowerment*, 11, 75-96.

[18] White, S. 1992. Arguing with the Crocodile: Gender and Class in Bangladesh. London: Zed Books Ltd.

[19] World Bank. 2012. Information and communications for development 2012: maximizing mobile. Washington, D. C., World Bank.

[20] Lee, D. D. and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*. 401, 788-791.

[21] Lee, D. D. and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing 13. Cambridge, MIT Press.

[22] Paatero, P. and Tapper, U. 1997. Least squares formulation of robust non-negative factor analysis. Chemometr. Intell. Lab. 37, 23-35.

Chapter 5

[1] International Telecommunication Union. The world in 2014: ICT facts and figures. Retrieved 7 June, 2014, from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf

[2] World Bank. 2012. Information and communications for development 2012: maximizing mobile. Washington, D. C., World Bank.

[3] M. González, C. A. Hidalgo, A. L. Barabási. 2008. Understanding individual human mobility patterns. Nature, 453, 7196 (2008) 779-782.

[4] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. J. Martonosi, J. Rowland, A. Varshavsky. Identifying important places in people's lives from cellular network data. in: K. Lyons, J. Hightower, E. M. Huang (Eds.), Pervasive Computing, Springer Berlin Heidelberg, 2011, pp. 133-151.

[5] C. Song, T. K. Koren, P. Wang, P, A. L. Barabási. Modelling the scaling properties of human mobility. Nature Physics, 6, 10 (2010) 818-823.

[6] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W. Y. Ma. 2008. Mining user similarity based on location history. In Proceedings of The 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. November, 5-7. New York, USA. ACM.

[7] J. Blumenstock, N. Eagle, Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of The 4th ACM/IEEE International Conference on Information and Communication Technologies and Development. December 13-15, 2010. New York, USA. ACM. doi:10.1145/2369220.2369225

[8] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, C. O. Buckee. Heterogeneous mobile phone ownership and usage patterns in Kenya. PLoS One, 7, 4 (2012) e35319.

[9] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, C. O. Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society Interface, 10, 81 (2013) 20120986.

[10] Stimson Center. 2009. Dhaka: developing resilience. Retrieved 23 September, 2014, from http://www.stimson.org/images/uploads/case_study_abridged_dhaka.pdf

[11] Bangladesh Bureau of Statistics, Community Report Dhaka Zila: Population and Housing Census 2011. Dhaka, Ministry of Planning Bangladesh Bureau of Statistics, 2012.

[12] Bangladesh Telecommunication Regulatory Commission, Mobile phone subscribers in Bangladesh 2014 April. Retrieved 23 May, 2014, from http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-april-2014

[13] Central Intelligence Agency, The World Factbook 2013-2014. Washington, DC, 2013.

[14] X. Lu, E. I. Pas. 1998. Socio-demographics, activity participation and travel behavior. Transportation Research Part A, 33 (1998) 1-18.

[15] Y. Lee, M. Hickman, S. Washington. Household type and structure, time-use pattern, and trip-chaining behavior. Transportation Research Part A, 41 (2007) 1004-1020.

[16] D. M. Scott, P. S. Kanaroglou. An activity-episode generation model that captures interaction between household heads: development and empirical analysis. Transportation Research Part B, 36 (2002) 875-896.

[17] E. I. Pas, 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. Environment and Planning A, 16, 5 (1984) 571-581.

[18] S. Brdar, D. Culibrk, V. Crnojevic. Demographic attributes prediction on the real-world mobile data. In Proceedings of Mobile Data Challenge by Nokia Workshop, in Conjunction with International Conference on Pervasive Computing, 2012. Newcastle, UK.

[19] J. Blumenstock, Calling for better measurement: estimating an individual's wealth and well-being from mobile phone transaction records. In Proceedings of The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Workshop on Data Science for Social Good. August 24-27, 2014. New York, USA. ACM.

[20] A. Arai, A. Witayangkurn, H. Kanasugi, T. Horanont, X. Shao, R. Shibasaki, Understanding user attributes from calling behavior: exploring call detail records through field observations. In Proceedings of The 12th International Conference on Advances in Mobile Computing and Multimedia. December 8-10, 2014. Kaohsiung, Taiwan. ACM. doi: 10.1145/2684103.2684107

[21] I. Bates, C. Fenton, J. Gruber, D. Lalloo, A. M. Lara, S. B. Squire, A. Theobald, R. Thomson, R. Tolhurst, Vulnerability to malaria, tuberculosis, and HIV/AIDS infection and disease. Part 1: determinants operating at individual and household level. The Lancet infectious diseases. 4 (2004) 267-277.

[22] World Health Organization. Report on infectious diseases: scaling up the response to infectious diseases, Geneva, 2002.

[23] A. Liaw, M. Wiener. Classification and regression by randomForest. R News 2(3), 2002, 18-22.

Chapter 6

[1] Agarwal, S. 2011. The state of urban health in India; comparing the poorest quartile to the rest of the urban population in selected states and cities. *Environment and Urbanization*, 23(1), 13–28.

[2] Defeo, J., and Juran, J. M. 2010. Juran's Quality Handbook: The Complete Guide to Performance Excellence 6th Edition. McGraw Hill Professional.

[3] United Nations. 2011. UN Data Glossary. Retrieved April 2, 2015, from http://data.un.org/Glossary.aspx

[4] Kit, Oleksandr, Lüdeke, M., and Reckien, D. 2013. Defining the bull's eye: satellite imagery-assisted slum population assessment in Hyderabad, India. *Urban geography* 34(3), 413-424.

[5] United Nations Human Settlements Programme. 2013. State of the world's cities 2012/2013: Prosperity of cities. Routledge.

[6] Shaw, R., and Mallick, F. H. 2013. Disaster Risk Reduction Approaches in Bangladesh. A. Islam Eds. Springer.

Chapter 7

[1] Snyder, P., and Lawson, S. 1993. Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, 61, 334 – 349.

[2] Cohen, J. 1992. A power primer. *Psychological Bulletin*, 112, 155–159.

[3] Cohen, J. 1994. The Earth is round (p<.05). *American Psychologist*, 49, 997–1003.

[4] Loftus, G. 1996. Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161–171.

[5] Osborne, J. 2008. Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology*, 28, 151–160.

[6] Ferguson, C. J. 2009. An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532.

[7] Wilkinson, L. and APA Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, 54(8), 594–604.

[8] Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences: Second Edition. Academic Press, Inc: New York.

Appendix

Acknowledgements

I would like to extend the deepest appreciation to my committee chair Professor Ryosuke Shibasaki, who has the attitude and the substance of a genius: he continuously and convincingly conveyed a spirit of adventure. Without his guidance and persistent help, this dissertation would not have been possible.

I wish to express my sincere thanks to the committee members of my Ph.D. dissertation, Professor Atsushi Deguchi, Professor Yukio Sadahiro, Professor Masahide Horita, and Associate Professor Yoshihide Sekimoto, who provided valuable comments and suggestions to improve the thesis greatly.

I would like to thank Associate Professor Xiaowei Shao, who provided valuable guidance for the machine learning and academic paper writing. I would like to thank Dr. Apichon Witayangkurn, Mr. Hiroshi Kanasugi, Mr. Fan Zipei, and Dr. Teerayut Horanont, who are my research collaborators as well as advisors. You greatly helped me develop research ideas to academic achievements. I have learnt much through discussion. I would also like to thank Dr. Toshikazu Nakamura, who helped me learn programming, for your kind and patient instruction.

I take this opportunity to express gratitude to all secretaries of Shibasaki & Sekimoto laboratory, who supported my research activity in many ways. Without help of Ms. Reiko Honma and Ms. Kumiko Akieda, my research activities extended outside the laboratory would not have been completed successfully, in particular.

I would like to appreciate the mobile network operator and volunteers who provided data for the research. I also thank to GRENE-ei, funded by Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), for the funding of part of my research.

I also wish to express my gratitude to one and all of the laboratory members, who directly or indirectly have lent their hands.

I thank my husband, Shunsuke, for his love, continuous encouragement, and support throughout the venture. I would have never been able to be here without you.