International Journal of Geo-Information

MDPI

Article

Comparative Perspective of Human Behavior Patterns to Uncover Ownership Bias among Mobile Phone Users

Ayumi Arai [1,*,†], Zipei Fan [2,†], Dunstan Matekenya [3,†] and Ryosuke Shibasaki [4]

1 Earth Observation Data Integration and Fusion Research Initiative, University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan
2 Graduate School of Engineering, University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan; fanzipei@iis.u-tokyo.ac.jp
3 Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; dunstan@mcl.iis.u-tokyo.ac.jp
4 Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan; shiba@csis.u-tokyo.ac.jp
* Correspondence: arai@csis.u-tokyo.ac.jp; Tel.: +81-3-5452-6412; Fax: +81-3-5452-6414
† These authors contributed equally to this work.

Academic Editors: Jamal Jokar Arsanjani, Ming-Hsiang (Ming) Tsou and Wolfgang Kainz
Received: 12 January 2016; Accepted: 30 May 2016; Published: 6 June 2016

Abstract: With the rapid spread of mobile devices, call detail records (CDRs) from mobile phones provide more opportunities to incorporate dynamic aspects of human mobility in addressing societal issues. However, it has been increasingly observed that CDR data are not always representative of the population under study because it only includes device users alone. To understand the discrepancy between the population captured by CDRs and the general population, we profile principal populations of CDRs by analyzing routines based on time spent at key locations and compare these data with those of the general population. We employ a topic model to estimate typical routines of mobile phone users using CDRs as topics. The routines are extracted from field survey data and compared between those of the general population and mobile phone users. We found that there are two main population groups of mobile phone users in Dhaka: males engaged in an income-generating activity at a specific location other than home and females performing household tasks and spending most of their time at home. We determine that CDRs tend to omit students, who form a significant component of the Dhaka population.

Keywords: big data; mobile phone; call detail records (CDRs); demographic structure

1. Introduction

1.1. Background

The large amount of spatiotemporal 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 these devices are widespread, the spatiotemporal data can be considered representative of the general population. However, this may not be true due to data characteristics of call detail records (CDRs). Particularly, device ownership biases must be taken into account in developing countries because ownership varies across different demographic groups [1]. Biases are found in mobile phone user ownership in terms of gender and age group [2]. The utilization of spatiotemporal data for societal issues is suggested to require knowledge of the types of populations

that are represented by the data [1,3]. Without knowing which part of society the data represent, interpretation of the data analysis results may be misleading. In addition, the location distribution observed in CDRs is biased spatially and temporally. Since CDRs are updated only when the mobile phone is used, the records are heavily affected by users' calling behavior [4,5]. This nature of CDRs results in generally sparse data that provide a partial view of actual trajectories [6].

In this study, we examined the discrepancy between the principal population components of CDRs and those of the general population by comparison of typical routines. First, we profile the typical routine of the principal population components for mobile phone users by comparing routines extracted from CDRs and the diary survey data of mobile phone users. CDRs are inherently sparse; hence, we interpolate CDRs based on the estimated routines by employing the topic model. This allowed us to transform CDRs, which originally only include information related to timeslots with call records, into routines of mobile phone users in a continuous manner. Then, the obtained results were compared with the routines of the core components of the general population in Dhaka. Thus, we were able to observe how the principal population in CDRs differs from those in the general population with regard to behavior patterns.

The contributions of our study are as follows:

- The principal population components of mobile phone users are profiled by comparing routines extracted from CDRs and those obtained from field survey data. The sparse CDRs were interpolated based on the predicted routines and interpreted as sequential activities. The ownership bias among mobile phone users is elucidated.
- A novel approach to identifying device domain-specific bias for large-scale spatiotemporal data is proposed. The potential to extend our approach to other areas using other data is discussed.

1.2. Related Work

An increasing number of studies investigate the behavior patterns of people through analysis of large-scale spatiotemporal data. Since the number of mobile phone users is very high, CDRs form large-scale spatiotemporal databases. With the sequential information of time and location of individuals, the data enable us to understand the dynamics of human mobility. Mobile phone data can capture quantitative aspects of human mobility, such as volume and statistical patterns of mobility [7–9]. In recent years, the increasing availability of spatiotemporal data has advanced the research on mobility patterns and their application in sectors, such as transportation [5,10,11], public health [12,13], and urban planning [14]. The properties of human mobility are represented in a better manner by incorporating periodic modulation of human mobility. References [7,9] contributed to the characterization of human mobility by quantifying the interaction between the regularity and randomness in human mobility dynamics. By mining semantically meaningful locations, such as home and the workplace, in anonymized CDRs, Reference [15] determined that a few limited locations where people spend most of their time are the means in understanding human mobility and social patterns. In addition, correlation is found between daily activity pattern and the type of areas, which are considered to be work locations [16]. A part of human mobility is explained well by taking into account daily travel-activity patterns because human mobility is considered to be driven by the demand to participate in activities. Activity in combination with socio-demographics further elucidates human mobility patterns [17]. The socio-demographic characteristics were also proven to significantly affect the time allocation for activities inside the home and those outside. Reference [18] suggested that human activity-travel behavior could be described by the individual spatial behavior, which can be captured by a monthly and seasonal variability in activity. In this context, the extraction of typical behavior patterns using a few key locations can help in understanding major components of mass population.

Reference [19] described the hidden structure in human behaviors by analyzing the data collected from 95 mobile phone users. The study presents their characteristic behavior by extracting the principal components, referred to as eigenbehaviors. It analyzes individual eighenbehavior and also describes

the community affiliations of populations. While the study presents partial behavior traits as the principal components, Reference [20] characterizes behavior patterns as regular temporal transitions between typical states, such as home and the workplace. The Latent Dirichlet Allocation (LDA) topic model is employed to extract location-driven routines. Reference [21] 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. References [22,23] presented interesting works on the application of the LDA model in large-scale geo-location data to identify latent activity patterns.

While the application of CDRs seems to be a prominent means of addressing societal issues, some problems exist in CDR data. One critical problem is representativeness because analysis results of mobility data may vary according to datasets, which capture different populations [1,24] and different moving processes [25]. For instance, the data of the Oyster card (an electric ticketing system for public transport passengers in Greater London) captures the mobility of transportation users alone [26]. CDRs include mobile phone users alone. Reference [27] found gaps in socio-economic status between mobile phone users and non-users. The application of the analysis results to societal issues may cause problems if a discrepancy exists between the population represented by mobility data and the population under study.

The remainder of this paper is structured as follows: data used in this study is described in Section 2. In Section 3, the characteristics of typical mobile phone users are described and their typical routines are presented by analyzing the field survey data of mobile phone users. In Section 4, we examined the typical routines of mobile phone users, which are extracted from CDRs. The diary survey data of the general population is analyzed in Section 5. Our conclusions are presented in Section 6.

2. Data

2.1. Mobile Phone Data

We use the CDRs of August 2013 from one of the leading mobile network operators in Bangladesh (hereafter referred to as "the MNO"). The data include the time, antenna location, and duration of calls. We randomly sampled the call records of 5000 unique IDs. We did not use the entire dataset because we consider 5000 samples enough to extract the typical routine patterns. The selected samples are evenly distributed geographically in the study site. The data comprises records of all antennas located in Greater Dhaka in Bangladesh.

2.2. Diary Survey Data of Mobile Phone Users

To understand the personal attributes and activity of mobile phone users who use the service of the MNO, 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) [28]. The survey was conducted from November 2013 to January 2014 and covers selected areas of Greater Dhaka. SPACE consists of two parts, namely one-day diary records from mobile phone users and their personal attributes and activity. The former part collects time spent for activities of the day along with call records of the same day. The latter part collects their age, gender, occupation, and major routine activity. We employed two-stage stratified sampling according to land use and household income levels. The areas covered by CDRs are split into 161 administrative areas, which are classified into three groups according to their dominant type of land use: residential, commercial, or industrial. Of these, 15 areas, which consist of 10 residential, two commercial, and three industrial areas, are randomly selected in proportion to their population shares in the total population. From each area, 18 households, each having at least one mobile phone user of the MNO as a household member, were selected from each income group: high, middle, and low. If the slum population in an area is greater than 25%, we sampled that population as part of the low-income group. As a result, we interviewed 922 mobile phone users from 810 households. The SPACE data do not represent mobile phone users using the service of the MNO; these data are considered to represent the mobile phone users corresponding

to each income group. We scheduled the survey on both weekdays and the weekend to reduce bias. In addition, we visited the household according to the availability of household members in the morning, afternoon, and late evening to collect the data from those who work during daytime.

2.3. Diary Survey Data of the General Population

We utilized 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 include the timing, origin-destination, means of transportation, and purpose of trips for the day, which is a typical day, and the data structure is almost similar with that of the diary survey part of the SPACE data. In addition, demographic attributes, such as age, gender, and occupation, were collected by the Japan International Cooperation Agency (JICA) by interviewing 75,000 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. In the sample, the number of males was slightly greater than that of females (54% *vs.* 46%). Table 1 presents three key population groups based on their activities: respondents engaged in income-generating activity (38%), household tasks (25%), and education (32%). Furthermore, we note that male respondents comprise the majority of the income-generating group while females are in the majority in the household tasks group. Additionally, we found that those who receive education as their main activity constitute almost one-third of the total population in Greater Dhaka. Hence, we focus our analysis on the following key population groups: *working males*, *housewives*, and *students*, and label the rest of the population as *others*.

Table 1. Percentage distribution of survey respondents and their income status by gender.

	Income-Generating Activity	Non-Income-Generating Activity		
		Household Tasks	Education	Other
Overall	38%	25%	32%	5%
Male	61%	1%	32%	6%
Female	10%	53%	33%	4%

3. Typical Behavior Patterns of Mobile Phone Users Derived from SPACE Data

3.1. Population Composition of Mobile Phone Users

To understand the principal populations of mobile phone users, we examined the SPACE data, which is diary survey data of 922 mobile phone users. Table 2 describes the proportion of males and married mobile phone users classified by income level. The overall proportion of males is greater than that of females. In addition, more than 85% of the users are married. Bangladesh has relatively strong social norms of behavior based on gender. Among women, the labor force participation rate in urban areas is 35% while that of men is 80% [29]. Assuming that the sex ratio is almost 1, we can roughly estimate that the population rate of working males is approximately 40% and that of working females is less than 20%. When we take into account the large proportion of the married population among mobile phone users, we expect gender-specific behavior patterns to be predominant in the SPACE data. That is, 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.

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, the primary activity of mobile phone users was analyzed. Assuming the gender-specific trends mentioned above, the type of activity was classified based on gender (male and female) and economic activity (income-generating activity and non-income generating activity). An income-generating activity is any activity that generates monetary income as a return. For example, salary earners, part-time workers, and self-employed people are classified as persons engaged in an income-generating activity. The remaining people are engaged in non-income-generating 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. More males are shown to be engaged in an income-generating activity while the majority of females are engaged in a non-income-generating activity, particularly household tasks. This indicates that most of the mobile phone users subscribed to the MNO are those who perform responsible roles within the household, *i.e.*, they have available money at their disposal. The mobile tariff for this MNO is relatively expensive compared to that for other companies in Bangladesh, and this was considered a factor affecting the user trends.

Table 3. Distribution of main activity by gender.

	Income-Generating	**Non-Income-Generating**		
		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. Trends by gender were observed to be similar for males and females across all income levels. Therefore, we conclude that, regardless of income levels, two types of typical mobile phone users exist: the married male engaged in an income-generating activity, and the married female who mainly performs household tasks. Based on the results, the two typical mobile phone users are termed as *working males* and *housewives*.

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%)

3.2. Location of Main Activity

Considering *working males* and *housewives* as typical mobile phone users, the type of location of their main activity was classified by specifying whether it is home or outside home. The activity of the people is required to be linked to location types because the behavior patterns extracted from CDRs will be described based on the probability distribution of location types in the next section. Table 5 shows the distribution of location types for the main activities of *working males* and *housewives*. In the case of *working males*, 72% of the locations for their main activity are a specific location outside the home. This indicates that more than half of typical male users have a specific location outside the home for their main activity. In the case of *housewives*, 94% of the locations for their main activity are the home. Owing to the distinctive difference in the location of the main activity for the two principal population groups, we found it acceptable to interpret the typical behavior patterns extracted from CDRs based on the analysis results of this section.

Table 5. Distribution of location for main activity.

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

3.3. Typical Behavior Patterns of Mobile Phone Users

In the previous section, we classified the type of location as home and outside home. For each user, a location reported as home is labeled as *Home*. Among outside home locations, a location reported as a workplace is labeled as *Work*. For those who do not have a workplace, they have home and outside home locations only, where the most frequently reported location among the outside home location was selected, and labeled as *Work*. Therefore, every user has locations, which are labeled as *Home* and *Work*. Using these three types of locations, namely, *Home*, *Work*, and *Other*, we obtain the location-based behavior patterns of typical mobile phone users based on the SPACE data. For *housewives* and *students*, we considered the primary location outside home as their *Work*, *i.e.*, the time spent for education for *students* is considered as *Work*. In addition to *working males* and *housewives*, we examined the behavior pattern of *students*, who form the third-largest segment of mobile phone users, although the absolute proportion of this segment is much smaller than the other two. Mobile phone users classified as *students* are mostly college students. Figure 1a,b shows the hourly distribution of the probability of being at *Home* and *Work,* respectively. 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 distributions of being at *Work*, but this probability is much higher for *working males*. This higher probability could be attributed to differences between office hours (e.g., from 9 a.m. to 5 p.m.) and school hours in Dhaka. Weekday and weekend were not differentiated because the type of day for the diary survey, which is one of the datasets used for comparison, was specified just as a typical day for interviewees.

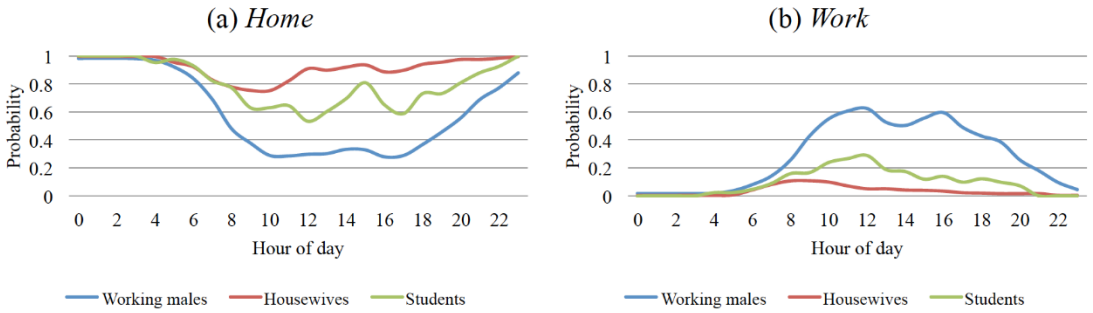

Figure 1. Hourly probability distribution of being at (**a**) home and (**b**) work for three principal population groups.

4. Typical Behavior Patterns Extracted from CDRs

4.1. Methodology

We draw an analogy between discovering a pattern of daily routine from CDRs and discovering a latent topic from documents. The CDRs of each user are considered as a document, and each data point, described by the time stamp and geographical location, is considered as words in the document. The vocabulary in our model describes the temporal distribution and geographical distribution. In this study, we extend the classical LDA model by drawing the time stamp as well as the location from

the latent assignment of topic for each record, and assume that people have similar daily routines but different main locations (e.g., home, workplace). Hence, as shown in Figure 2, we place latent variables of time patterns outside the user plate of the model and latent variables of location inside the user plate. The symbols used in Figure 2 are explained in Table 6. To infer the latent variables, a Gibbs sampling [30] inference is applied as shown in Algorithm 1.

Algorithm 1 Gibbs sampling based behavior pattern discovery

Input: $\{(u,\ d, t, l)\}$ CDR dataset; K predefined number of topic number

Output: $\{z_{u,d,t,l}\}$ topic assignment; $\left\{\phi_u^k\right\}$ topic distribution over location for each user; $\left\{\psi_d^k\right\}$ topic distribution over day; $\left\{\psi_\tau^k\right\}$ topic distribution over time;

$n_{u,k} = 0$; $n_{d,k} = 0$; $n_{\tau,k} = 0$; $n_{u,l,k} = 1$

// Random Initialization

For each record $(u,\ d, \tau, l)$:

$k = z_{u,d,\tau,l} = RandInt\,(K)$

$n_{u,k} = n_{u,k} + 1$

$n_{d,k} = n_{d,k} + 1$

$n_{\tau,k} = n_{\tau,k} + 1$

$n_{u,l,k} = n_{u,l,k} + 1$

// Iterative inference

For i = 1 to MAX_ITERATION

For each $(u,\ d, \tau, l)$:

$k_{old} = z_{u,d,\tau,l}$

$n_{u,k_{old}} = n_{u,k_{old}} - 1$

$n_{d,k_{old}} = n_{d,k_{old}} - 1$

$n_{\tau,k_{old}} = n_{\tau,k_{old}} - 1$

$n_{u,l,k_{old}} = n_{u,l,k_{old}} - 1$

Sample a new topic assignment k from the distribution

$p\,(k) = \frac{n_{u,k}+\alpha}{\sum_k n_{u,k}+K\alpha} \cdot \frac{n_{d,k}+\gamma}{\sum_D n_{d,k}+D\gamma} \cdot \frac{n_{\tau,k}+\gamma}{\sum_T n_{\tau,k}+T\gamma} \cdot \frac{n_{u,l,k}+\beta}{\sum_L n_{u,l,k}+L\beta}$

$z_{u,d,t,l} = k$

$n_{u,k} = n_{u,k} + 1$

$n_{d,k} = n_{d,k} + 1$

$n_{\tau,k} = n_{\tau,k} + 1$

$n_{u,l,k} = n_{u,l,k} + 1$

// Calculate the values to be returned

$\phi_u^k = \frac{n_{u,l,k}+\beta}{\sum_L n_{u,l,k}+L\beta}$, $\psi_d^k = \frac{n_{d,k}+\gamma}{\sum_D n_{d,k}+D\gamma}$, $\psi_\tau^k = \frac{n_{\tau,k}+\gamma}{\sum_T n_{\tau,k}+T\gamma}$

Return $\{z_{u,d,\tau,l}\}$, $\left\{\phi_u^k\right\}$, $\left\{\psi_d^k\right\}$, $\left\{\psi_\tau^k\right\}$

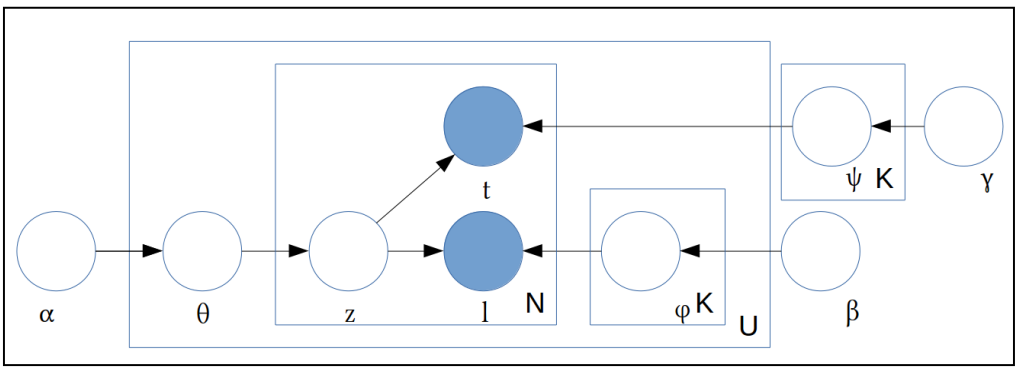

Figure 2. Graph model of our extended LDA model. Shaded and unshaded nodes denote observed and latent variables, respectively.

Table 6. Explanation of symbols used in Figure 2.

Symbol	Explanation
α	Hyper-prior of the user topic multinomial distribution.
β	Hyper-prior of the location multinomial distribution.
γ	Hyper-prior of the temporal multinomial distribution (day and time).
θ	User topic distribution.
ϕ	Geographical location distribution of each topic and each user.
ψ	Temporal distribution for all users with respect to day and time.
U	Number of mobile phone users in the dataset.
N	Number of records of each user in the dataset.
K	Number of latent topics.
z	Latent topic (z = 1, . . . , K).
t	Time stamp of record, represented as (d, τ), where d and τ denote the day and time, respectively.
l	Geographical location of record, described by (latitude, longitude).

4.2. Extracting Typical Spatiotemporal Calling Behaviors Based on Call Records

We applied our extended LDA model to CDRs of 5000 unique IDs from mobile phone users using the algorithm presented at the beginning of this section. The time pattern that we discovered is shown in Figure 3. We extracted three principal topics as the three typical calling patterns of mobile phone users. Figure 3 illustrates the topic proportion at each time and Topics 1 and 3 depict the calling behavior with a dominating high topic proportion during morning hours and during the day, respectively. Moreover, Topic 2 represents the calling behavior of a preference of call at night than at daytime. The topic is determined by the spatial/temporal topic distribution. Therefore, the patterns in Figure 3 are clustered based on the pattern of calls in relation to the pattern of their periodic visit to the same location. As discussed earlier, locations that are repeatedly visited, such as home and the 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 phone users.

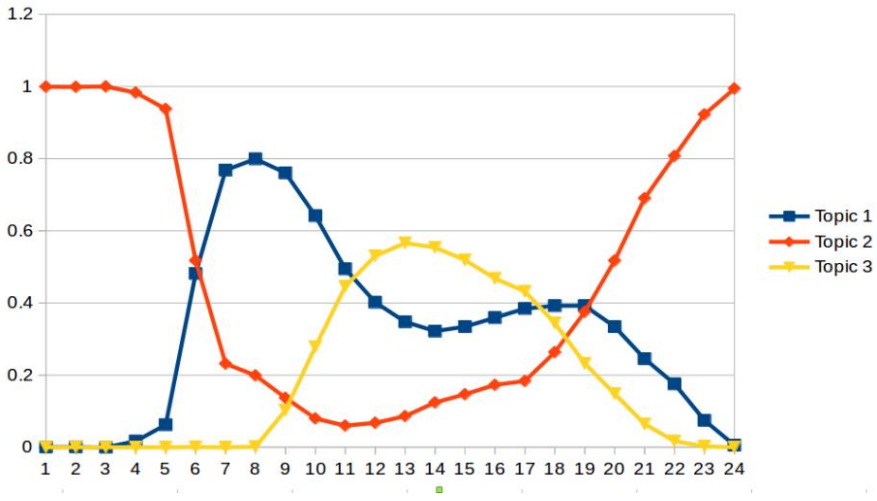

Figure 3. Time patterns of three principal topics extracted from CDRs.

Taking into account the lifestyle of people in Dhaka where most offices, shops, and entertainment venues are closed at night, we assume that the location having the highest probability after midnight is associated with the home location because the majority of the people most likely stay at home or in the vicinity of home late at night. Populations clustered into Topic 3 exhibit the highest probability of being at home for the longest hours after 12 a.m. until 12 p.m. This is similar to the pattern of housewives, which were extracted in the previous section. Populations clustered into Topic 2 show

a clear difference in probabilities of being at home between nighttime and daytime. We assume that this population group is generally engaged in an activity outside the home similar to the pattern of working males, which was also extracted in the previous section. Likewise, the populations of Topics 1 and 2 spend a large amount of time outside the home. Topic 1 exhibits two peaks: the first is at approximately 8 a.m. and the second is at 7 p.m. This is not very similar to the pattern of students extracted in the previous section, which also has narrow peaks in the late morning hours (between 10 a.m. and 12 p.m.) and around 5 p.m. This is probably because populations clustered into this group include not only students but also other populations which cannot be clustered into Topics 2 and 3. Comparing the analysis result with the previous sections, we observe that the LDA model can distinguish two predominant population groups: (1) Topic 2 represents the behavior patterns of people who generally spend the majority of their time outside the home and return home at a time that is the latest among all patterns, and these people are most likely to be *working males*; and (2) Topic 3 represents people who are engaged in an activity related to the home, and these people are most probably *housewives*.

5. Discrepancy between Population in CDRs and the General Population

5.1. Typical Behavior Patterns of Principal Population Groups in Dhaka

In this section, we report our findings related to typical behavior patterns across the three principal population groups obtained from the data. The results are summarized in Figure 4a,b. As discussed 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*. Therefore, we present the distribution of the key population groups considering only (a) *Home* and (b) *Work* locations. The results show 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 phone users. In spite of the minor difference in the result, we conclude that the behavior patterns extracted from the two sources of data are generally similar.

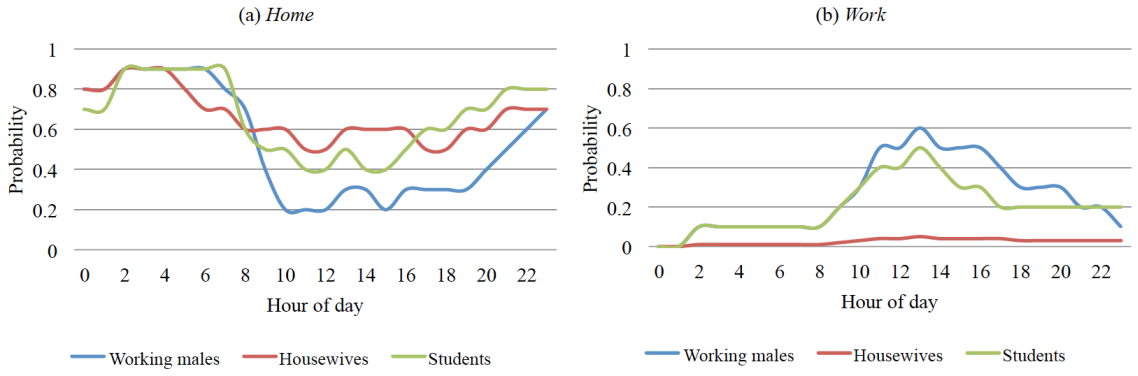

Figure 4. Hourly probability distribution of being at (**a**) home and (**b**) work for the three principal population groups.

5.2. Ownership Bias

Finally, we discuss the ownership bias of mobile phone users by comparing the principal population groups of the general population and those of the mobile phone users. Table 7 shows approximate estimates of the proportion of the three principal population groups for the general population and mobile phone users. The "+" mark in the estimate for mobile phone users indicates a possible minimum estimate. This indicates that proportions were provided as the minimum because these were obtained for each income level but do not have information on the population share.

For instance, the proportions of *housewives* among mobile phone users for high, middle, low, and slum levels are 30%, 29%, 27%, and 26%, respectively. The overall proportion of *housewives* among mobile phone users can be at least 26%. In the general population, a sizeable population of *students* exists, with education as their main activity. However, the corresponding proportion of *students* among mobile phone users is very small. Furthermore, the proportion of *male workers* and *housewives* is significant in the general population and among mobile phone users. Considering that the population pyramid of Bangladesh is wide at the base [31] and *students* generally belong to relatively younger population groups, we conclude that CDRs are biased because the data seldom include students, who comprise a significant proportion of the general population in Dhaka.

Table 7. Proportion of the three principal population groups.

	Male Workers	***Housewives***	***Students***
General population	38%	25%	32%
Mobile phone users	46%+	26%+	4%+

6. Conclusions

In this study, we proposed a novel approach to elucidate the discrepancy in principal population compositions between the general population and mobile phone users by comparison of their typical behavior patterns. We profiled principal populations of mobile phone users through SPACE and diary survey data of mobile phone users. We found that working males and housewives are two dominant population components of mobile phone users of the MNO. We also succeeded in extracting their behavioral patterns from CDRs by employing a topic model. Analysis results presented two typical behavior patterns, namely people who spend most of the day engaged in routines outside the home, and those who spend most of their time at home. These findings were consistent with the behavior patterns extracted from the diary survey data of mobile phone users, where we observed two typical behavior patterns: the male engaged in income-generating activity outside the home during the day and the female who spends the majority of the time at home, mainly performing household tasks.

Comparing the principal populations of mobile phone 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 phone users. The analysis results indicated that CDRs capture the behavior patterns of working males and housewives. Therefore, we suggest that the application of CDRs, targeting the younger generation in particular, takes this bias into account because the data do not necessarily represent this population group. We believe that our findings will be useful for the utilization of CDRs in developing countries, which have limited resources. CDR acquisition does not have additional costs. The data are generally collected for billing purposes by MNOs and are therefore available as long as a mobile network is present. Our study shows that the potential for understanding domain-specific biases, which can constitute major constraints in the utilization of large-scale domain-specific data such as CDRs, exists through the analysis of CDRs in combination with secondary data. However, the activity patterns of people in this case are not very complex because people's lifestyles are affected by strong social norms. Applications of this study in other areas would require more analysis on the relationship between the features of the principal population groups and their daily routines. Additionally, conducting a large-scale field survey is expensive. 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 Reference [32]. A census is recommended every five years and has been conducted in more than 200 countries in the census round spanning the period from 2005 to 2014 [33]. We consider utilizing such data to lower the cost of data acquisition for application in other study areas.

Acknowledgments: We thank the mobile network operator, the Japan International Cooperation Agency, and the volunteers who provided the data for this research. Part of this work was supported by GRENE-ei, funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Author Contributions: Authorship has been included and strictly limited to researchers who have substantially contributed to the reported work. Ayumi Arai designed and conducted the field survey and analyzed the diary survey data of mobile phone users. Zipei Fan designed algorithms for clustering populations in CDRs. Dunstan Matekenya analyzed the diary survey data of the general population. Ryosuke Shibasaki administered the field survey and CDR data acquisition. All authors discussed the results and their implications.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wesolowski, A.; Eagle, N.; Noor, A.M.; Snow, R.W.; Buckee, C.O. The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface* **2013**, *10*. [CrossRef] [PubMed]
2. Frias-Martinez, V.; Virseda, J. On the relationship between socio-economic factors and cell phone usage. In Proceedings of the 5th International Conference on Information and Communication Technologies and Development, Atlanta, GA, USA, 12–15 March 2012; pp. 76–84.
3. Blumenstock, J.; Eagle, N. Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th International Conference on Information and Communication Technologies and Development, London, UK, 13–16 December 2010.
4. Kang, C.; Sobolevsky, S.; Liu, Y.; Ratti, C. Exploring human movements in Singapore: A comparative analysis based on mobile phone and taxicab usages. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, IL, USA, 11–14 August 2013.
5. Demissie, M.G.; de Almeida Correia, G.H.; Bento, C. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. *Transp. Res. Part C: Emerg. Technol.* **2013**, *32*, 76–88. [CrossRef]
6. Candia, J.; González, M.C.; Wang, P.; Schoenharl, T.; Madey, G.; Barabási, A.L. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A: Math. Theor.* **2008**, *41*. [CrossRef]
7. González, M.C.; Hidalgo, C.A.; Barabási, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
8. Barabási, A.L. Scale-free networks: A decade and beyond. *Science* **2009**, *325*, 412–413. [CrossRef] [PubMed]
9. Song, C.; Koren, T.K.; Wang, P.; Barabási, A.L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818–823. [CrossRef]
10. Iqbal, M.S.; Choudhury, C.F.; Wang, P.; González, M.C. Development of origin–destination matrices using mobile phone call data. *Transp. Res. Part C: Emerg. Technol.* **2014**, *40*, 63–74. [CrossRef]
11. Wang, H.; Calabrese, F.; Di Lorenzo, G.; Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In Proceedings of 13th International IEEE Conference on Intelligent Transportation Systems, Madeira Island, Portugal, 19–22 September 2010.
12. Balcan, D.; Colizza, V.; Gonçalves, B.; Hu, H.; Ramasco, J.J.; Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 21484–21489. [CrossRef] [PubMed]
13. Buckee, C.O.; Wesolowski, A.; Eagle, N.N.; Hansen, E.; Snow, R.W. Mobile phones and malaria: Modeling human and parasite travel. *Travel Med. Infect. Dis.* **2013**, *11*, 15–22. [CrossRef] [PubMed]
14. Becker, R.A.; Caceres, R.; Hanson, K.; Loh, J.M.; Urbanek, S.; Varshavsky, A.; Volinsky, C. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Comput.* **2011**, *10*, 18–26. [CrossRef]
15. Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.J.; Rowland, J.; Varshavsky, A. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*; Lyons, K., Hightower, J., Huang, E.M., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2011; pp. 133–151.
16. Phithakkitnukoon, S.; Horanont, T.; Di Lorenzo, G.; Shibasaki, R.; Ratti, C. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*; Salah, A.A., Ruiz-del-Solar, J., Meriçli, C., Oudeyer, P.-Y., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2010; pp. 14–25.
17. Järv, O.; Ahas, R.; Witlox, F. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transp. Res. Part C: Emerg. Technol.* **2014**, *38*, 122–135. [CrossRef]

18. Lu, X.; Pas, E.I. Socio-demographics, activity participation and travel behavior. *Transp. Res. Part A* **1998**, *33*, 1–18. [CrossRef]
19. Eagle, N.; Pentland, A.S. Identifying structure in routine. *Behav. Ecol. Sociobiol.* **2009**, *63*, 1057–1066. [CrossRef]
20. Farrahi, K.; Gatica-Perez, D. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*. [CrossRef]
21. Zeng, J.; Ni, L.M. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012.
22. Hasan, S.; Ukkusuri, S.V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C: Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]
23. Hasan, S.; Ukkusuri, S.V. Location contexts of user check-ins to model urban geo life-style patterns. *PLoS ONE* **2015**. [CrossRef] [PubMed]
24. Arai, A.; Witayangkurn, A.; Horanont, T.; Shao, X.; Shibasaki, R. Understanding the unobservable population in call detail records through analysis of mobile phone user calling behavior: A case study of Greater Dhaka in Bangladesh. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications, St. Louis, MO, USA, 23–27 March 2015; pp. 207–214.
25. Szell, M.; Sinatra, R.; Petri, G.; Thurner, S.; Latora, V. Understanding mobility in a social petri dish. *Sci. Rep.* **2012**, *2*, 1–6. [CrossRef] [PubMed]
26. Roth, C.; Kang, S.M.; Batty, M.; Barthélemy, M. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE* **2011**, *6*, e15923. [CrossRef] [PubMed]
27. Rice, R.E.; Katz, J.E. Comparing internet and mobile phone usage: Digital divides of usage, adoption, and dropouts. *Telecommun. Policy* **2003**, *27*, 597–623. [CrossRef]
28. Arai, A.; Witayangkurn, A.; Kanasugi, H.; Horanont, T.; Shao, X.; Shibasaki, R. 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, Kaohsiung, Taiwan, 8–10 December 2014; pp. 95–104.
29. Rahman, R.I.; Islam, R. *Female Labour Force Participation in Bangladesh: Trends, Drivers and Barriers*; ILO: Geneva, Switzerland, 2013.
30. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, 5228–5235. [CrossRef] [PubMed]
31. United Nations. World Population and Housing Census Programme. 2010. Available online: http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/censusclockmore.htm (accessed on 9 February 2015).
32. United Nations. *Principles and Recommendations for Population and Housing Censuses: Revision 2*; Statistical Papers Series M No. 67/Rev.2; United Nations: New York, NY, USA, 2008.
33. United Nations. *World Population Aging: 1950–2050*; United Nations: New York, NY, USA, 2001.