論 文 の 内 容 の 要 旨

論文題目　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）

氏　　名　新井 亜弓

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 2013 mobile subscription rate reached at 96% globally, which evidenced the substantial part of the global population was connected to 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. With the sequential information of time and location of individuals, the data allow us to understand the dynamics of human mobility. In recent years, increasing availability of spatio-temporal data has advanced the understanding on features and statistical patterns of human mobility. Understanding the human mobility is challenging because of the significance of incorporating the human mobility dynamics into the practical applications to various sectors, e.g. urban planning, transportation management, public health, etc. While the analysis of CDRs seem to be prominent as the means of understanding human mobility dynamics, it has been increasingly pointed out that there are constraints for utilizing such data for societal issues. The constraints as summarized as below;

1) Sparseness

CDRs 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.

2) Bias

The population of CDRs is only mobile phone users. Thus, there can be discrepancies in human mobility dynamics between mobile users and actual living populations, which consist of mobile users and non-mobile users, if the characteristics of the two population groups are different. When we interpret the analysis result of CDRs to address societal issues, we need to be aware what is the population under the study.

3) Anonymity

CDRs are generally anonymized to protect the privacy of mobile users. Instead of the attribute information of mobile users, a random code is assigned to each person, which still allows us to trace the mobility of each person for a given period. This feature makes it difficult to know who are in the data except the fact that they are mobile users.

Increasing bodies of study analyzing CDRs have proposed 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 volume of human mobility. On the other hand,

the data alone do not indicate much about their qualitative aspects of the population, and thereby the research on the characteristics of the population in CDRs is limited so far. In this study, a novel approach is proposed to resolve the three constraints of CDRs by proposing a new dataset, Dynamic Census, by utilizing CDRs of Dhaka from one of leading telecommunications companies in Bangladesh (called the operator). The uniqueness of this study is that hidden properties of CDRs are revealed and estimated, analyzing CDRs in combination with field survey data. To understand the population structure and personal attributes, including calling behavior along with their activity, two field surveys (SPACE 2013 and SPACE 2014 − the Survey on Patterns of Activity for Comprehensive Explorations of Mobile Phone Users in Dhaka) were conducted. Basic features of Dynamic Census are described as below;

1) Interpolated spatio-temporally

   The path of the mobile user can be determined by a pair of two consecutive records in CDRs. The pair of points, consisting of time and location information, is potentially neither the time nor location of departure/arrival for a person. In Dynamic Census, blank time bands are interpolated based on the estimated timing of location changes of the mobile users, and then route of their movement is predicted based on the existing road network. It enables us to obtain better spatio-temporal distribution estimates for mobile users.

2) Represent the living population for a given area

   The demographic structure of Dynamic Census is adjusted to that of the living population. It enables us to discuss the human mobility dynamics not only of the mobile user but also the living population, presenting for given areas at given time.

3) Labeled with demographic attribute

   Dynamic Census is labeled with the personal attribute information, which is estimated through this study. It enables us to specify the mobility of specific population groups 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.

The structure of this thesis is as follows. Chapters 2 to 4 propose approaches to resolve the three constraints of CDRs; Chapter 2 investigates the discrepancy of the population presented in CDRs and the living population by specifying the principal population groups of them. In the process of extracting the activity patterns from CDRs, the methodology to interpolate CDRs is provided as well. Chapter 3 describes how these two population groups differ with respect to the personal attribute. In addition, we provide the results of experimental study on estimating the presence of the unobservable population in CDRs. Chapter 4 focuses on the anonymity of CDRs. It explores the relationship between the features of calling behavior and the personal attribute of mobile users. Furthermore, estimation results of personal attributes are provided. Chapter 5 provides a process of developing Dynamic Census using analysis results of Chapters 2 to 4, and presents the outcome. Chapter 6 provides the framework to recreate Dynamic Census in other cities. It focuses on key information to be collected as the secondary data to develop Dynamic Census. A guideline to design a survey for the secondary data collection is provided as well. Chapter 7 includes conclusions and future prospects.

In Chapter 2, the discrepancy in principal population compositions among the living population and mobile users were revealed. Their typical behavior patterns were examined by comparing time spent at significant locations, such as home and work places, between CDRs, SPACE 2013 data, and diary survey data. First, we profiled principal populations of mobile users

by employing the Latent Dirichlet Allocation topic model to extract typical behavior patterns from sparse CDRs. We found two typical behavior patterns: people who spend most of the day engaged in routines activity outside the 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 living population, we found that students form a core component of the living population but are not considered significant among mobile users.

In Chapter 3, the characteristics of the unobservable population, who exist as part of the living population but do not appear in CDRs, were examined. We compared the population structure of mobile users and non-mobile users, which were surveyed through SPACE 2013. In terms of population size, there are roughly 2.4 to 2.8 unobservable people per mobile user identified in CDRs. We found that the majority of the operator's users are males. Additionally, both for males and females, more than 70% of the users are married, and their ages are mostly within the range of late twenties to late fifties. Our findings evidenced that CDRs do not capture specific population groups such as students or children below school enrollment age. Our experimental study revealed that it is possible to estimate the presence of children in CDRs by analyzing calling behavior of mobile phone users. However, the results of the estimation have room for improvement. 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 is not good enough to be applied to large-scale CDRs.

Chapter 4 revealed that calling behavior traits could distinguish gender and occupational types by analyzing relationships between calling behavior and personal attributes, surveyed through SPACE 2013. Analysis results suggest that a higher ratio of calls from home can be a key to distinguishing gender. Females tend to call from home around midday on their primary routine day. In addition, constant frequency and time distribution of calls throughout the week are keys to identifying male users. In addition, the results of estimating gender, presence of children in the household, and personal attribute of mobile users, employing Random Forest, were provided. We used three sets of call records, whose lengths of data acquisition periods vary as two months, three days, and one day, 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 personal attributes. Regarding the estimation of the personal attribute, we found our model is not successful in retrieving users whose population proportion is relatively smaller among the total population, e.g. students. After controlling the population proportion for estimation, the performance was improved slightly but still has room for improvement.

Chapter 5 described the process of developing Dynamic Census using analysis results and methods obtained through Chapters 2 to 4. To obtain the magnification factors to make estimates on the number and the structure of the living population from CDR populations, we used Small-scale Census (SCC) data, which were collected as part of SPACE 2014. SCC

surveyed the population structure and mobile phone ownership of all living populations for a given area in Dhaka. The key of the process is that the type of buildings is used as the proxy of the income level, which was also used for the stratification in the sampling of SPACE 2013 and 2014. This allowed us to estimate the population structure of entire areas of Dhaka based on the distribution of buildings. In addition, we can use income-level wise analysis results, obtained from SPACE data, for interpreting the findings from SCC data. Visualized trajectories by personal attribute and mesh-based hourly population distribution were presented.

Chapter 6 provided a framework to recreate Dynamic Census in other cities by adopting the approach proposed in this study. The framework consists of three aspects, which are closely related to the secondary data collection. We first sorted out key information to be collected based on calling behavior traits, which can distinguish the difference in personal attributes. Additionally, possible alternative data were listed to lower the cost of data collection. Then, we proposed the survey structure to collect the key information through the field survey. Lastly, provided suggestions to determine the sampling method and sample size.

Chapter 7 summarized conclusions. Also, future prospects were provided as the proposal to overcome three limitations of this study. One limitation is the less diverse lifestyle of people in Dhaka where strong social and behavioral norms exist. We expect recreating Dynamic Census in another city, where people's lifestyles are greatly diverse, needs more inputs for estimating the personal attributes of mobile users. Another is the limited number of validation for the model construction. It can be improved by increasing the sample size of validation data, taking account of the decent length of data period to be used for the analysis. The other is the availability of non-aggregated CDRs for recreating Dynamic Census. We expect that opportunities to obtain non-aggregated CDRs will tend to be diminished due to the privacy concern. Therefore, a framework to recreate Dynamic Census using aggregated CDRs will be necessary.