CSIS Discussion Paper No. 115

A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources

携帯電話情報を用いた都市における人々の移動特性分析と動的な
人口推定手法に関する研究

ホラノン　ティーラユット

ABSTRACT

Sensing is becoming increasingly mobile and people-centric in a network enabled society. Urban activities are recognized from the footprint of the usages of ubiquitous urban infrastructure. This new paradigm is a combination of urban infrastructures, information and communication technologies and digital networks. The rapid technology development in the area of digital network and telecommunications has a significant impact on our societies, lifestyles and the way we interact with the city. The technological advances and the benefits resulting from the use of these technologies are capable of improving current urban management, and introduce new schemes for urban mobility studies. This dissertation explores a novel practice concerning how mobile devices can be employed as a new urban mobility sensor, location-aware and human in the loop system that enables city wide information collection and analysis.

Despite the growing concern for public health, urban securities and natural disasters, these studies have initially been developed to improve operational efficiency or to plan public space management. Understanding the population dynamics at high spatial and temporal resolutions, especially when incorporated into established population distribution models, will hopefully allow for detailed investigations of populations impacted by natural and manmade disasters and, subsequently, could lead to more appropriate emergency planning and responses, as well as better informed policy decisions.

Another promising application in utilizing digital footprints from the mobile devices is public marketing analysis. The distribution of a population at different points of time in each city space could be an ideal source for marketing executives to decide on a place to activate premium urban advertising opportunities or campaigns that are geographically based with more targeted and efficient marketing. These finsights would allow them to offer more specific ad-funded and LBS services that stimulate uptake and increase their profitability.

In this research we explore and analyze the daily population distribution patterns in the city, using a unique data source from aggregate calling activity over time in a mobile phone network. This study will therefore address the following research questions: Would modification of calling activities be confirmed to correspond to actual population in the area? How do such patterns vary in different parts of the city? And how can human mobility be reconstructed from calling records of mobile phones? The questions are meant to explore whether the population distribution follows a trend of mobile phone call detail records (CDRs) and to investigate using such data for potential develop dynamic population estimation models.

To test the provided research questions, the study had been conducted in several major cities of the world starting from a test base scenario in central Bangkok in 2008. We developed a system call "Mobile Sensing", a web based 2D and 3D GIS, to use as a fundamental tool for mobile phone data visualization and analysis. The initial results depicted and captured time series of interpolated mobile phone call traffic in a grid density surface. This observation leads to the speculation about how one part of the central city is upscale, crowded and how long the area remains busy until people move to another part of the city. The analysis of the mobile phone activities during one day and the mean transformation within a month has been examined to extract each cityscape's communal pattern. In addition, results of the study not only provided a tool for area or zoning analysis but also could be used to specify hidden problems of the particular space over a period of time.

In 2009, millions of mobile phone users in Massachusetts were examined. It was the first time that a huge amount of mobile phone traces were used to analyze the mobility of the city. The numerical algorithms were developed to extract the revisit points of individuals, for instance, home and work locations. We preliminarily analyzed this data by characterizing mobility in a profile-based space (activity-aware map) that describes most probable activity associated with a specific area of space. This, in turn, allowed us to capture the individual daily activity pattern and analyze the correlations among different people's work area's profile. We also investigated how good a correlation exists between the presence of people and the mobile phone activities by inventing "SIM Mobility", a mobile phone traffic simulation system. The product outcome has been validated with

simulated trajectories and MassGIS census data. The results yielded evidence confirming that the mobile phone activities are significantly correlated to the existence of people.

The last part of this dissertation was implemented in the Tokyo metropolitan area. The Tokyo's mobile phone traffic or call detail records (CDRs) were generated from the SIM Mobility simulation system. The calling patterns of the Greater Tokyo region were built from analyzing mobile phone usage surveys, and the simulated trajectories were retrieved from Tokyo Person Trip data. Besides this, we developed new dynamic population estimation methodologies involving two different approaches and investigated the accuracy of results. The first methodology was implemented using a group based approach, and the population weighted factor has been modified by time-dependent OD metrics extracted from the observed mobile phone calling traffic. The second methodology was implemented using an individual based approach utilizing data assimilation process. The estimation results shown greatly accurate prediction and capable of being enhanced the existing algorithm previously developed by The Tokyo Metropolitan Region Transportation Planning Commission. By using a complete scale of this novel data type, the proposed methods would enable real-time reporting of city-wide population estimations and potentially paint a complex and dynamic portrait of the urban dynamic in which users are based.

Using mobile phones as sensing devices to accumulate aggregate "crowdsourced" data for urban analysis is still in the early phases of development. The contributions of this research would pave the way for future extensions to larger and more complex analysis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introductions

1.1 From existing and beyond: Toward the ubiquitous city

Today's cities develop at a rate that outpaces architects' and planners' efforts to shape them. Cities, home to more than half the world's population, can be seen as complex networks of components: citizens, businesses, transport, communications, utilities and other systems. Urban spaces become increasingly interconnected and intelligent in nature. Cities have the chance to accelerate their journey towards sustainable prosperity by making use of new "smart" solutions and management practices.

Focusing on urban mobility, the percentage of the population living in urban areas is expected to rise from 13 % in 1900 to approximately 60 % by 2030, meaning that a large and increasing proportion of the population will limit their daily travel to short / medium distances of less than 100 kilometers, often within entirely the urban environment. (Figure 1.1) Particularly relevant to the urban mobility scenarios, the demand for people transportation will further increase over the next decades, requiring new solutions to optimize human fluidity and energy efficiency in specific situations.

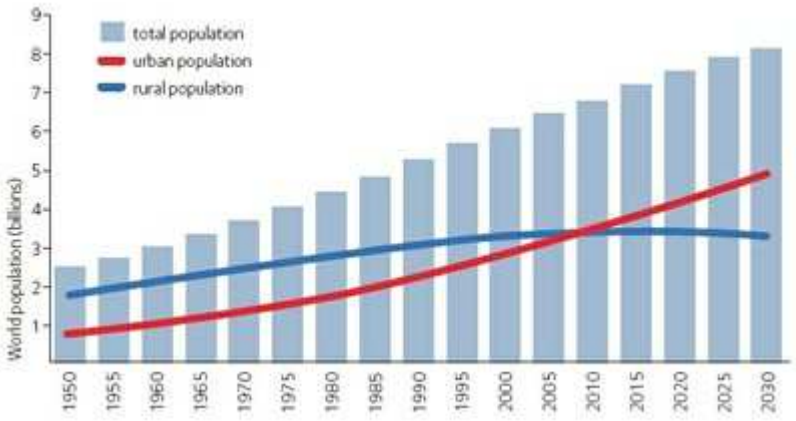


Figure 1.1 Urban and rural population of the world, 1950 - 2030
Source: UN Population Division.

The study of movements of people within a city has had a long history. The population flows in and out of the city are highly dynamic especially in several worlds' mega cities. For instance, Manhattan island in New York City, the island is home to about 1.6 million (2008 - http://en.wikipedia.org/) inhabitants, with roughly double that number during any work day. These highly dynamic situations in busy city centers still pose considerable challenges for state-of-the-art approaches.

The accuracy of estimating the total number of population for an urban area during the day also remains highly questionable and problematic. Many urban planners, sociologists, and even economists have tried various means of estimating populations by using standard multipliers and complex formulas for making inferences from census records. The challenges ahead are significant and complicated and thus will be focused of this research.

1.2 Communications networks as urban fabric

Telecommunication network enables cities to gather more high-quality human mobility data in a timely fashion than ever before. Thinking in term of "scales of networks", mobile communication provides us an ideal solution to create a huge urban fabric in which urban populations simply become part of a network. These networks can be telematic, physical, or even social when people are all engaged.

Implicitly this research underscores the opportunistic of these territories as much as they construct a footing for new intervention practices. In principle, telecommunication infrastructure may be seen as providing a service to people. This means people and infrastructure interact, making new interfaces and ways of representing the existence of any entity that can be seen by the network. The footprints from this interaction clearly become a new source to unambiguously identify people in the real world.

Another reality is that over 4.3 billion people own mobile phones today (March 2010, http://www.gsacom.com/), and the increasing reach of mobile networks creates an unprecedented opportunity for researchers to derive valuable high-level human behavior information in realms of urban planning, mobility and social interactions.

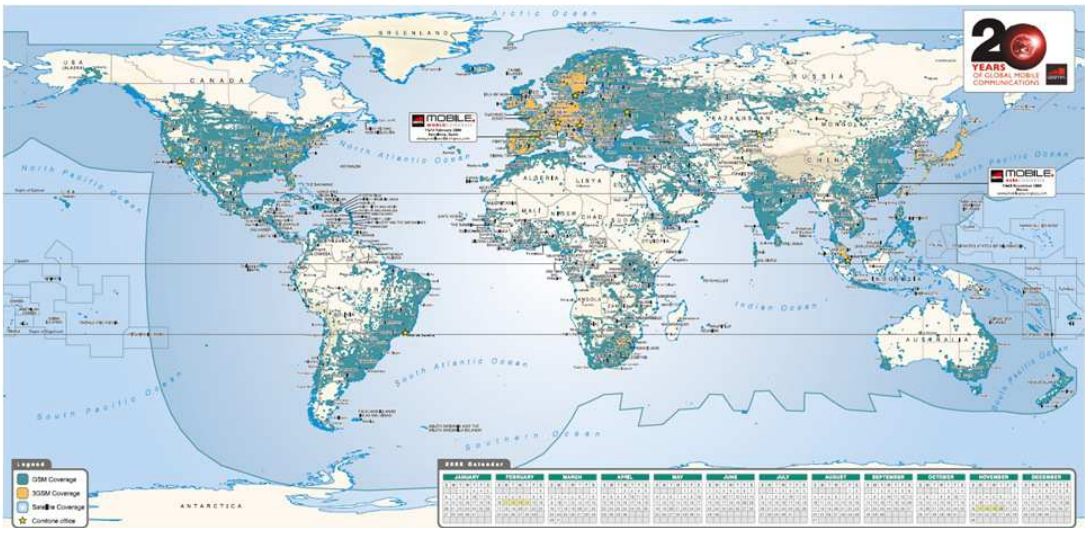


Figure 1.2 World cellular network coverage (2008)
Source: The Global mobile Suppliers Association (GSA).

1.3 Research structure and main findings

This dissertation is organized into seven chapters. The First chapter describes here introduce the backgrounds, problems and new potentials of this research. The remaining sections of this chapter explain the objectives of this research, motivations and research questions.

Chapter 2 briefly describes each of the components of a telecommunication system, definition and terminology and points out the difficulties in examining the mobile phone data for human flow analysis. Also reviews of current and previous works.

Chapter 3 demonstrates the development of research tools, a prototype of mobile sensing platform and web API. The system provides an easy visualizing of mobile phone activity from base station log data. The principle functions of spatial analysis such as Inverse Distance Weighted (IDW), Kriging model, contour interpolation and basic graph analysis had been implemented. A new voronoi based sector interpolation has been introduced in this chapter.

Chapter 4 expresses a use case of analyzing mobile phone traces to identify human daily activity patterns. The article discusses various methodologies and models to extract mobility patterns and spatial profiles from the real mobile traces.

Chapter 5 investigates call-usage patterns mining from real dataset of millions of mobile phone users. A study was conducted to find a minimum behavioral cluster of mobile user groups in inner-city environment. Beside, algorithms for mobile traces reconstruction had been invented to find a common trajectory of the individuals. Last, the correlation of population distribution and mobile phone traffic had been confirmed.

Chapter 6 builds on the concepts introduced in Chapter 5 and focuses on algorithms and models to reconstruct population distribution over time. Tokyo mobile phone usage survey has been conducted to understand calling behavior of user groups in Tokyo. Subsequently, a platform calls "SIM Mobility" was developed for assembling mobile phone call. We applied SIM Mobility to generate the virtual call traffic by utilizing pre-processed Person Trip data and behavioral models. Time-dependent OD matrix method and assimilation process method have been invented to reconstruct human mobility from calling records of mobile phones. The methods contribute significantly to improve the accuracy of population estimation models.

Chapter 7 discusses new findings, future prospect and the possible extension of the proposed algorithm to more complex analysis and concludes the paper.

1.4 Research questions and objectives

Motivations

From the introduction, it is the first time in history that more people live in cities than in the countryside. The accelerated rate of growth of the city today makes urban life steadily more complex than ever. This sudden concentration of population is often raising the problem of traffic congestion, public health and urban securities. Therefore, we need a novel method in order to capture of large-scale quantitative data related to human behavior and their mobility. Figure 1.3 illustrates the conceptual idea and potential use in general.

However, to date, the majority of research in understanding human flow and urban mobility has been done with very limited sample sizes. The estimated population density is mainly meant to census data which is static and refers to night time population. Nevertheless, the vast penetration of mobile phone users and advance development through the innovative use of mobile network are changing our daily lives and also the way to research on the city motion. We now have taken great steps forward to truly become the Ubiquitous city.

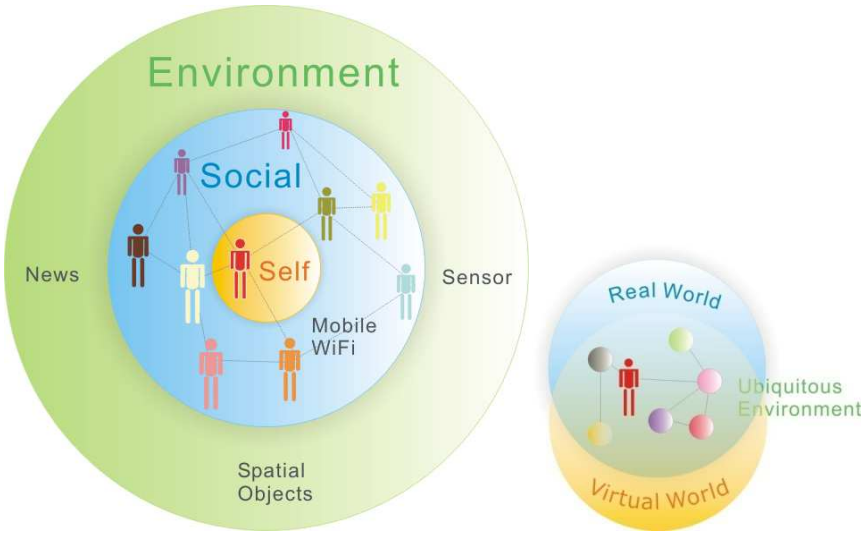


Figure 1.3 A conceptual idea of information sensing

Our hypothesis is that in the urban area where mobile phone penetration rate is high, the mobile device could be represented to the existence of the people and the population density could be determined by mobile phone usage density.

Research questions

In particular, this research examines the extent to which the use of mobile phones traffic data helps to understand human mobility patterns and leads to the development of population dynamic model. The following research questions were developed to examine this aspect.

- How can actual population at a specific time-frame be estimated from calling records of mobile phones?
- How can dynamic people mobility be reconstructed from calling records of mobile phones?

Research objectives

From above research questions, our objective is to develop a methodological framework based on the analysis of mobile phone call detail records (CDRs) data. The research objectives pursued in order to answer the research questions are:

- To implement a standard platform to access, manipulate and visualize mobile phone Call Detail Records (CDRs) and to use this platform to help understand the distribution of calling traffic in the area.
- To confirm the correlation between mobile phone usage and population density.
- To develop a methodology for estimating population density and dynamic mobility of people.

CHAPTER 2

Principle of Mobile Communication and Review of Previous Works

2.1 Evolution of mobile communication: past and future

The first generation of mobile communication was introduced in the late 1970s and this 1G cellular system still transmits only analogue voice information. The development of 2G cellular systems was driven by the need to improve transmission quality, system capacity and coverage. Speech transmission still dominates the airways, but the demand for short message and data transmissions are growing rapidly. 2G cellular systems include Global System for Mobile Communication (GSM), Digital AMPS (D-AMPS), code division multiple access (CDMA) and Personal Digital Communication (PDC). GSM is the most successful family of cellular standards and therefore is the main system in our first study which we will discuss more in the next session. GSM became the dominant technology and is also the basis for one of the two leading 3G technologies UMTS and W-CDMA. The first pre-commercial 3G network was launched by NTT DoCoMo in Japan branded FOMA, in May 2001 on a pre-release of W-CDMA technology. The 3G systems were capable of data transmission at speeds up to 384K bps but can now support downloads in the megabit per second range. Figure 3 illustrates monthly forecasts of Mobile data traffic from 2008 to 2013 and figure 4 shows the laptops and mobile broadband handsets drive traffic growth. The trends help confirm that mobile phone is already part of everyday life with penetration rates rising to 70 per cent and more in many countries. According to a survey amongst more than 1000 Internet leaders, activists and analysts the mobile device will be the primary connection tool to the Internet for most people in the world in 2020 (pew Internet & American life project, December 2008: www. pewinternet.org). In particular, mobile broadband is going to change the way the world lives and to dramatically change the way we interact with the city. This research introduces how mobile phone networks can provide the ubiquitous mobility services in a world where everyone

are connected.

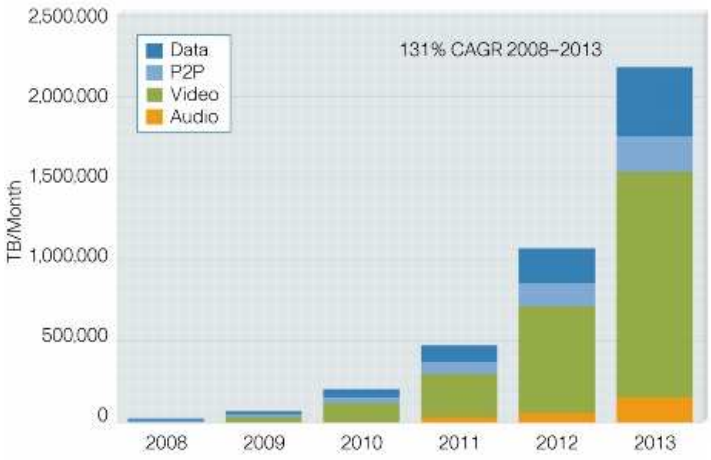


Figure 2.1 Monthly forecasts of Mobile data traffic from 2008 to 2013

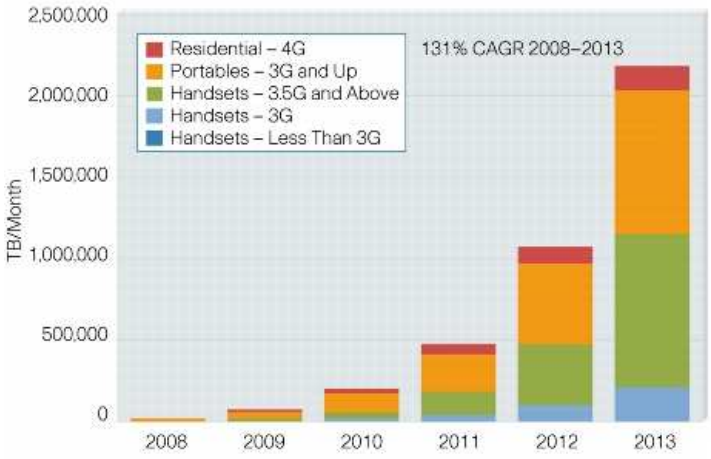


Figure 2.2 Laptops and mobile broadband handsets drive traffic growth from 2008 to 2013

2.2 Definitions and terminology

There are several words and terminologies use in this research paper that are strongly connected with information and communication technology. Some of the most important basic terminologies of GSM system are described in this section.

Global System for Mobile Communication (GSM)

A GSM system is basically designed as a combination of three major subsystems, the Network Subsystem (NSS), the Radio Subsystem (BSS) and the Operation support subsystem (B&CCS & IN). In general, we collect the data from Base Station Controller (BSC) which is a part of Radio Subsystem. When users place a call, a cell phone or mobile station (MS) will establish a communication with the closest Base Transceiver Station (BTS). A coverage area of a single base station indicates the capacity of mobile usage in the area. The initial research, for example in chapter 3, we accumulate the call traffic (Erlang) and other information in each base station with an hour interval. In a very high traffic area, the coverage could be reduced to a hundred-meter radius. However in the suburb, the coverage could be increased, varying from several hundreds to a kilometer. In the case of our study area, in Bangkok, the coverage varies from two to five hundred meters depending on the capacity of the peak usage time.

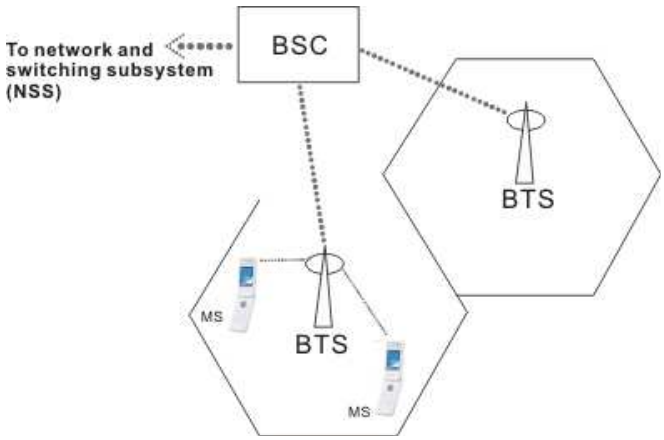


Figure 2.3 Commutation basis of GSM system

Erlang data

An Erlang is defined as a unit of telecommunications traffic measurement. Erlang represents the continuous use of one voice path and it is used to describe the total traffic volume of one hour going through a particular cell antenna. The erlang is a dimensionless unit that is used in telephony as a statistical measure of

offered load. Traffic of one erlang refers to a single resource being in continuous use, or two channels being at fifty percent use each, and so on.

Call Detail Records (CDRs)

For every mobile telephone call there should be a corresponding Call Detail Records (CDRs). It can contain information that the mobile network operator uses for subscriber identification, call charging, services obtained, call routing, Erlang and etc. In this research, we received CDRs data from two mobile operator, namely Advanced Info Service (AIS – Thailand) and Airsage Inc. (Untied State).

Mobile Station (MS)

The mobile station is made up of the subscriber identity module (SIM) and the Mobile Equipment. The SIM card and phone (ME) both have internationally unique identification. The number for the SIM card is called the IMSI (International Mobile Subscriber Identity) and for the phone it is called the IMEI (International Mobile Equipment Identity).

Base Station Transceivers (BTS)

BTSs are always grouped into geographical areas called Location Areas (LA). When a call is established to an MS, the Home Location Register (HLR) will know in which LA the subscriber is. Only the BTSs in that LA will page all the phones in order to find the targeted phone. This saves a lot of signaling resources. When a phone moves from one LA to another, it will request a position update and the HLR will update its database with the new location of the phone.

Base Station Controllers (BSC)

A single BSC (Base Station Controller) controls a number of BTSs and MSs. It handles the radio resource (RR) management of the BTSs such as frequency allocation, timeslot allocation, and handles handovers between cells under its control.

Mobile services Switching Centre (MSC)

The Mobile services Switching Centre coordinates the call setup and termination of MSs in its area and performs the switching between MSs in the local mobile network (PLMN).

Visitor Location Register (VLR)

The Visitor Location Register (VLR) is a temporary database used by an MSC. The VLR will update a subscriber's location when the subscriber moves from one location area (LA) to another in its area. When the subscriber mover from one LA to another under the control of a new VLR, then the old VLR will delete the information of the MS and the new VLR will draw the information from the HLR and also update the HLR with the new location.

Home Location Register (HLR)

The HLR is the central database where a subscriber's profile is held against the IMSI of the SIM card. It contains permanent information such as allowed services, status of supplementary services (such as call diverts) as well as other dynamic information such as the subscriber location.

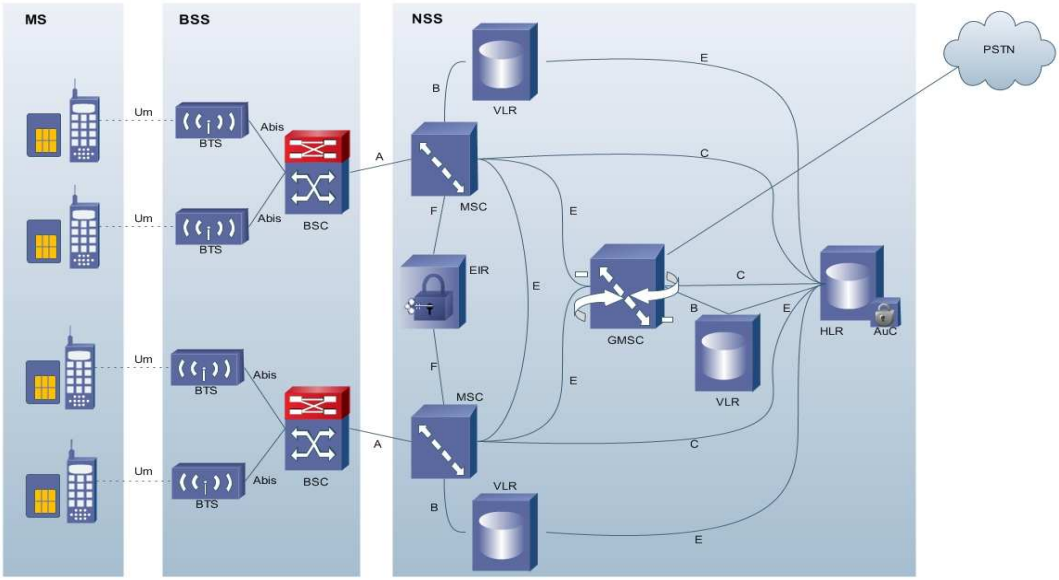


Figure 2.4 Main component of GSM system

2.3 Literature reviews

In recent years, the large deployment of mobile and wireless technologies has provided new means to understand the dynamics of a city. Scientists are discovering that a world buzzing with cell phone calls and text messages has a side incredible benefit: reams of data about who calls whom and about where they are at what time. Researchers are beginning to use that information to answer questions about how people behave, where they travel, and the social networks that connect them.

It is therefore not surprising that considerable research has gone in this new and interesting directions, in particular, developing techniques to bring out and explore large scale human mobility.(Azevedo et al., 2009; Candia et al.,2008; Eagle et al., 2007; Gonzalez et al., 2008)

In reality mining project (Eagle and Pentland, 2007), the project is using data gathered by cell phones to learn more about human behavior and social interactions. The dataset was collected over nine months by monitoring the cell phone usage of 84 participants. The personal reality mining infers human relationships and behavior by applying data-mining algorithms to information collected by cell-phone sensors that can measure location, physical activity, and more. With the aid of some algorithms, they posit, that information could help us identify things to do or new people to meet. It could also make devices easier to use for instance, by automatically determining security settings. More significant, cell phone data could shed light on workplace dynamics and on the well-being of communities. It could even help project the course of disease outbreaks and provide clues about individuals' health.

While the promise of reality mining is great, the idea of collecting so much personal information naturally raises many questions about privacy. Some people are nervous about trailing digital bread crumbs behind them. It's crucial that behavior-logging technology not be forced on anyone. But legal statutes are lagging behind data collection abilities which makes it all the more important to begin discussing how the technology will be used.

In Real-time Rome project, the collaboration between Telecom Italia and MIT's

SENSEable City Laboratory in 2007, they explored how researchers might be able to use mobile phone data for an entire metropolitan region to analyze urban dynamics. The Erlang data which is a measure of network bandwidth usage typically collected at the antenna level. The data was collected over four months in late 2006 and covering a region of 47 km^2. The findings suggested that signature analysis can provide an important new way of looking at the city as a holistic, dynamic system. In particular, the mobile phone network lets us develop a real-time representation of those dynamics at the city and city-region scale. This approach can complement traditional collection techniques, which are often outdated by the time they're available to policy makers and the general public (Ratti et al.,2007)

In earlier research on human mobility patterns (Gonzalez et al., 2008), Gonzalez and her team observed the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period. The results had shown a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations. This is describing the recurrence and temporal periodicity inherent to human mobility.

In 2009, Kyunghan Lee studied a mobility model for human walks by proposing a new mobility model called SLAW (Self-similar Least Action Walk) to improve the performance of networking applications. As wireless devices are mostly attached to humans, understanding human mobility patterns leads to an accurate performance prediction of protocols used for these networks (Lee et al., 2009). In contrast of this perspective, understanding the activities occur in the network can lead to a potential predictable presence of human or their mobility patterns in the urban system.

A recent study in Northeastern University (Barabasi et al., 2010), Barabasi used cell phone billing data for 50,000 people in a European country to show that people's travel patterns are extremely predictable. He found that most people travel very little on a daily basis, for instance, 5 to 10 kilometers or so. There were a few individuals, who on a daily basis travel hundreds of kilometers. This finding also suggests that, for the vast majority of the people, there is an average of 93 percent predictability across the user base.

There have been other considerable studies done to understand human mobility and improve the spatial resolution of static population counts through the use of census data (Dobson et al. 2000, Bhaduri et al. 2007). They introduced a project call "LandScan", a population distribution model created by Oak Ridge National Laboratory, seeks to overcome the limitations of aggregated and static population counts by estimating high spatial resolution population locations for both day and night. The project suggested a multi-dimensional dasymetric modeling approach, which has allowed the creation of a very high-resolution population distribution data both over space and time. At a spatial resolution of 3 arc seconds (~90 m), the initial LandScan USA database contains both a nighttime residential as well as a baseline daytime population distribution that incorporates movement of workers and students.

Therefore, any estimate of populations impacted or exposed must consider that census data represent nighttime residential locations and are not accurate for daytime population estimation. By estimating daytime location and disaggregating census blocks for accurate population location, high-resolution population data have already been proven to be vital tools in research and planning for environmental, public health, and disaster concerns (Bhaduri et al. 2007).

2.3.1 The chronological structure and scale of the population dynamic

When we talk about the population dynamic, there are several terminologies to express how to count the population in space. In case of census data or population count within the national level, the population density may describe as household population collected by annual or even longer since if we see the at the national or city scale, the city population may not change rapidly. Population dynamics deals with the way populations are affected by birth and death rates, and by immigration and emigration, and studies topics such as aging populations or population decline. Within this context, population data are constrained both in space and time and do not capture the population dynamics as functions of space and time. In fact, non uniform distribution of human population is quite obvious from simple visual observation of any landscape. From a temporal perspective, the resolution of census information is typically at anywhere between 1 and 10 year cycles and represented as a *nighttime residential* population. Usage of traditional census counts in a daytime

event is irrational. Because of this uncertainty, there is significant potential to misclassify people with respect to their location, These limitations, to a large degree, can be overcome by developing population data with a finer resolution in both space and time at sub-census levels. Geodemographic data at such scales will represent a more realistic non-uniform distribution of population.

Population distribution during the day

Population distribution during the day can be defined as distribution of population in an area during the daytime hours. The question on how to define and integrate daytime movements and collective travel habits into a single measure to produce a better representation of where people are located during an average day is crucial important.

Dynamic day time population or we describe it "ambient population", as it opposes to residential population, it takes into account the movements of individuals through a given area. For instance, no individuals live on an interstate, but they do travel on it from time to time. A resident population would show no one living on the interstate, whereas the ambient population would indicate the presence of individuals based on factors specific to the interstate.

Accurate estimation and representation of dynamic population distribution poses significant challenges and in fact few models have been developed to estimate day time population. The previous known effort to develop an analogous model has been at the Los Alamos National Laboratory (McPherson and Brown 2004; McPherson et al. 2006) that developed nighttime residential and daytime distribution data at a 250 m resolution. Another is LandScan project developed by Oak Ridge National Laboratory. (Bhaduri 2002, 2007) However none of the previous studies provide a methodology in estimating high temporal dynamic population distribution.

2.3.2 Person Trip Survey data and its potential use in this research

Person Trip Survey (PTS), a large scale household interview survey, is conducted to grasp passenger movement or to build travel demand forecasting models. The data has been used by academic researchers for various aspects including urban transportation planning, job accessibility and human flow. (Minamoto et al, 2202;

Kawabata, 2003; Sato et al, 2008) Since Person Trip data provide a details movement of people including mode of transportation, trip purpose and trip time, and because of the difficulty in validating real human mobility and counting the existing number of population in large dimension, thus with this aspect, PTS data would be an ideal source of information for post validation of this research

2.4 Limitations and difficulties in approach

Privacy

Privacy concerns arise when there is a possibility of discovering personal information such as the personal habits, locations, behaviors and lifestyles of individuals inside cities, and to use this information for secondary purposes. Mobile phones are ubiquitous in corporate life, supplying voice, email and browser access to data whenever and wherever information is needed. Although mobile phones are a valued and powerful tool for communication, there are aspects of mobile phones which are problematic, that have arisen due to the power and ubiquity of mobile phones. The ability to not only transmit voice but also to implicitly collect and record geographical locations has created new challenges for researchers in terms of new conflicts between technology Innovation and rights to privacy. This new form of location availability is considered to be very positive, but at the same time creates uneasiness among users. New conflicts and challenges are not only relevant to mobile phones but to other new forms of new communication technology.

Accuracy

One major factor that could impact continued advancement of local analysis on these mobile devices is accuracy of their geolocation estimated position (EP). There are two methods for pinpointing the location of mobile phone users. First is with Global Positioning System (GPS) capability, the mobile phone use signals from satellites to pinpoint location very accurately. But this is not in the normal case since the data sent from the mobile devices normally not include this kind of information.

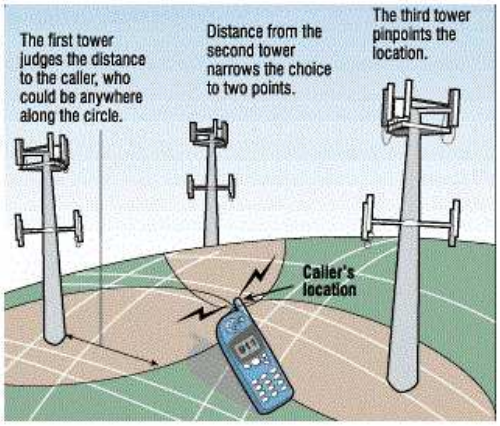


Figure 2.5 Cell Tower Triangulation
Source: http://www.intomobile.com/

The second and less-accurate method but mostly available is often called "Cell Tower Triangulation", referring to how the cell towers which receive a phone's signal may be used to calculate its geophysical location. In most case the accuracy from this method vary from 50-300 meters in dense urban environments.

Availability of data

In almost every country, a new form of researches and studies on mining mobile phones data has limited to the availability of the data from the mobile phone operator. Although the datasets have somehow become available in recent years and have opened the possibility for researchers to study on large-scale urban and social analysis however the support from mobile industry and data availability are still very limited in our experiment.

Data discontinuity

At some point we can consider the estimated mobile phone positions as GPS positions since the logging information include all geographical location and time stamp. Nevertheless, the crucial difference is that there is no continuity of each consecutive point from the Call Detail Record (CDR) since the CDRs are generated only when people use the mobile phones. This dissertation introduces various views to resolve this problem and bring in how CDR could become an effective monitoring probe for real-life application.

CHAPTER 3

Development of Real-Time Mobile Sensing Platform

3.1 Research initiative

In our everyday activities we leave behind footprints from our interaction with the urban environment and its digital infrastructures. This chapter aims to explore a new technique for urban monitoring by mapping of mobile phone usage and its location to represent the urban activities and their evolution through space and time. The massive movement of the people and rise of city congestion could be visualized in near real time on a web-based interface. We implemented mobile sensing platform which interpolate the aggregate mobile sources from the antenna-masts position to predict the population in grid-density surfaces. We then analyze urban patterns at a point of time to illustrate how people experience their city. Furthermore, the urban land-use could be classified from the unique characteristic use of mobile phone in each urban space. We analyzed the activities during one day and examined the mean transformation within a month to extract each cityscape communal pattern. This initial research could derive valuable high-level human behavior information in realms of urban planning, mobility and social interactions.

As we discussed earlier, considerable research has gone into this new and interesting directions. In particular, developing techniques to bring out very large scale urban sensing by leveraging the increasing capabilities found in cell phones (Nicholas D. Lane et al, 2008). Data collected from the cell phones give the foundation for exciting people-centric applications. And the analysis of cell phone use can provide an important new way of looking at the city as a holistic, dynamic system (Reades J. et al, 2007). People-movement maps could lead to

improvements in transport planning, event management and modelling the spread of biological and mobile viruses (Marta C. González et al, 2008, Ratti C. 2007).

Today, maps are not only intelligent, they are ubiquitous and the ubiquity of maps as a form of visual representation produces a new understanding of space. The usage of mobile devices can be treated as a medium for data collection. We developed a system called mobile sensing, an interactive web-based platform for mobile phone and web services data integration. The system analyzes contents and visualizes information in a map interface, which users can access via the internet. This article presents findings in the course of designing and examining how invisible activities can be realized. Obviously this would change the way we see how people interact with their built environment.

For the large scale monitoring of urban space, clusters of Erlang data from mobile base stations are excellent at providing indirect interpretation of spatial patterns of urban life and its temporal dynamic. Erlang data that represent a distribution of call duration in the Global System for Mobile Communication (GSM) network could be performed as aggregate data sources to estimate the population density of a city. This aspect is very useful in the view of public monitoring. It could potentially become a new way to extract or identify invisible problem spots from the complex urbanized areas. Furthermore, mobile sensing is potentially applicable for public marketing analysis. The distribution of a population at different points of time in each city space could be an ideal source to help people decide a place for urban advertising or for opening a shop. In addition, an exploration of mobile sensing data would give the urban planner a better understanding on flowing patterns of people at specific times of a day. If we look at broader contexts, then population transfers from city to city in during special events or holidays could also be determined.

The applications of mobile sensing could be extended to various domains of urban and infrastructure management. Since more than 4 billion people carry mobile phones today, and we are beginning to notice mobile phones' ability to become the ultimate data collection machines.

3.1.1 Study area

The research takes place at Bangkok Metropolitan, Thailand. Bangkok city has a well developed mobile network and has a high degree of mobile usage. We received the support of the Erlang data and base station locations from Advanced Info Service PLC (AIS), a leading mobile operator in Thailand.

Figure 3.1 Study area in central Bangkok. This is a good example for this experiment since the area compose of mix land use including business area and high-rise building zone, Shopping mall, horse racing field and the university. The overlaid voronoi give an idea how base stations and their coverage distribute over the dense city area. The yellow points virtually illustrate the centroid of service area of each antenna range.

3.1.2 Mobile phone log data

In this study, mobile phone log or Call Detail records (CDRs) data had been collected during the period from February to April 2008 that cover a part of the central Bangkok. We collected the data from 50 base transceiver stations and 150 antennas with an hour interval. The data compose of several attributed information, including cell antenna id, base station location, time stamp, handover information, call attempts and erlang data. Figure 3.2 show a plot of Erlang distribution over one month period,

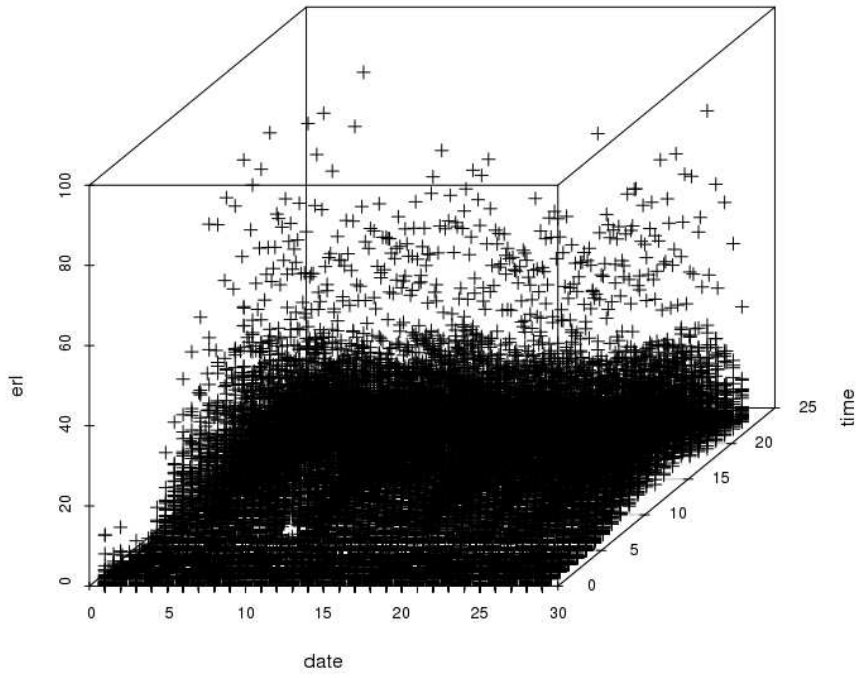


Figure 3.2 Daily Erlang distribution in one month

3.2 System architecture

3.2.1 Mobile sensing platform

The system is intended to visualize the aggregate mobile data sources and perform analysis of temporal human-flow in the city. We therefore developed fundamental tools of spatial exploration and visualization which permitted the data to be obtained, integrated and displayed quickly, easily and flexibly. Since the internet has now made distance virtually disappear and enables us to enjoy a much greater richness of information, we questioned the traditional method of urban monitoring by enhancing the way to instantaneously integrate and obtain large amounts of data via the network.

The key characteristic of the system is that it provides an easy visualizing of spatial data exploration and analysis through the web interfaces. With cell-based probe analysis, it is possible to handle multi-temporal, dynamic information, enhancing the ability to detect and represent changes associated with events and flows.

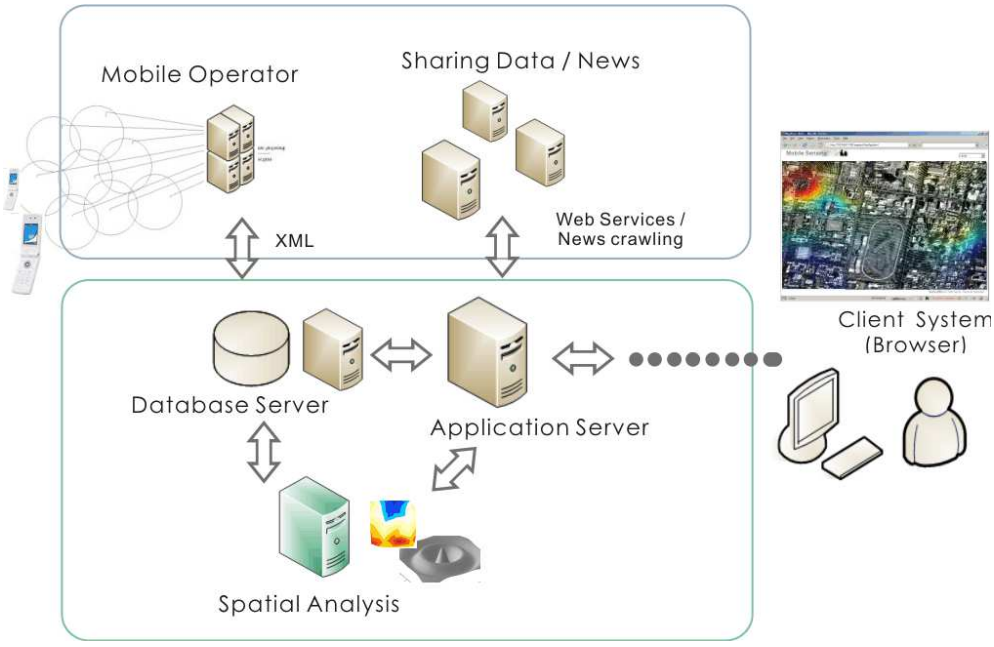


Figure 3.3 System architecture of mobile sensing platform

The mobile sensing platform concept could be described in two different components: a core application along with spatial analysis services that have been implemented in our local environment and the public domain services that acted as data sources for the system (Figure 3.2).

3.2.2 Data Processing and Services API

To maximize compatibility and interoperability, open standard such as XML and web services are utilized for data exchange and sharing. Data analysis is assisted by the open source statistical package R (http://www.r-project.org), integrated into the PostgreSQL database via the PL/R procedural language (Conway, 2003). The current system handles two different services in order to analyze flow of the city, one to obtain statistics service and the other to obtain interpolation service.

Obtain statistics: the statistic service manages two kinds of requests. It can retrieve daily statistics and monthly statistics of the interested area from a request made by a user services interface. This function executes a query request to spatial database servers and returns the processed data back to perform a graph rendering at the browser front end.

Obtain interpolation: the interpolation service provides a method of constructing new data points within the range of a discrete set of known data points. A user service parses a request with relevant parameters including date, time and interpolation method to the spatial analysis server, which query the spatial database server for retrieving mobile points and related information to generate an interpolated image. Finally, the system asynchronously sends the result image back to the originating user service.

3.2.3 Data visualization

The web browser we use as a universal front end, an Ajax mashup, is a hybrid web application which presents a rich UI to update and integrate contents asynchronously from multiple sources. This makes combining data easier, not only spatial data from the host server but also third-party sources from the services available on the internet. The calls can also be made directly to the third-party

sources from the browser or back to the originating server, which acts as a proxy for the third-party content.

We have developed a 2D and 3D visualization prototype system by implementing map APIs from various open source Javascript libraries and Google Earth Plug-in (Figure 3.3). The prototype system could evaluate the grid population density at a point of time and animate them to illustrate the flow patterns of people in the city throughout the day. The dynamic activities in each urban spot could be analyzed as a time signature in a daily and monthly based graph. Various statistics options are implemented to investigate, compare and highlight the hidden intensity of each urban space.

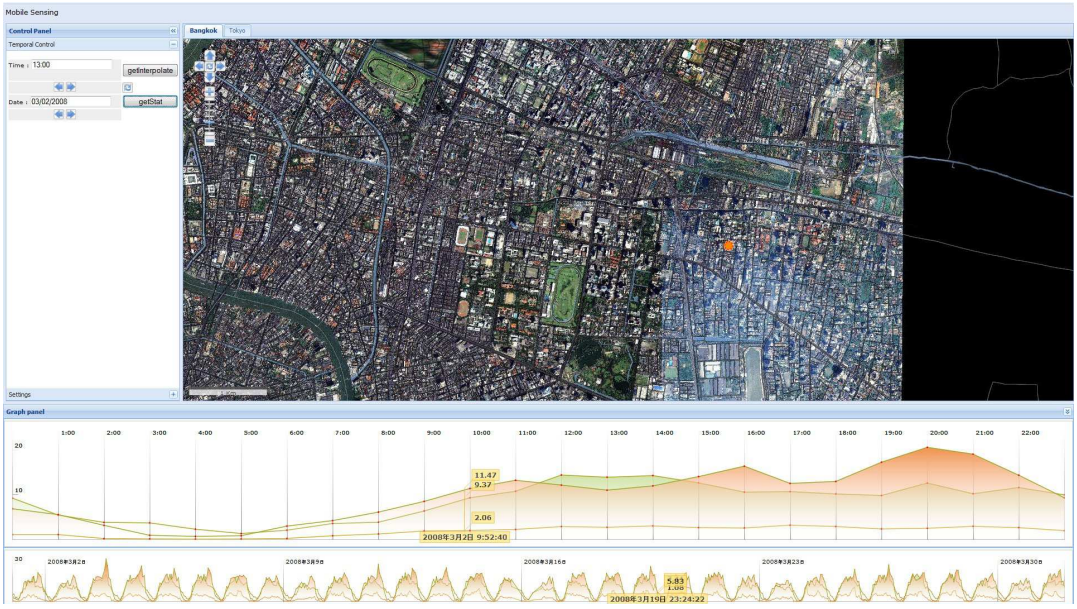


Figure 3.4 A prototype system with temporal control and time based analysis

We applied Google Earth Plug-in that is a subset of the Google Earth 3D graphics rendering engine with KML support for 3D visualization and analysis. At the time we implemented this system, the plug-in has not yet supported two KML's time elements: TimeStamp and TimeSpan which are derived from TimePrimitive. We have to create a separate KML file and display it on a frame by frame basis (Figure 3.4).

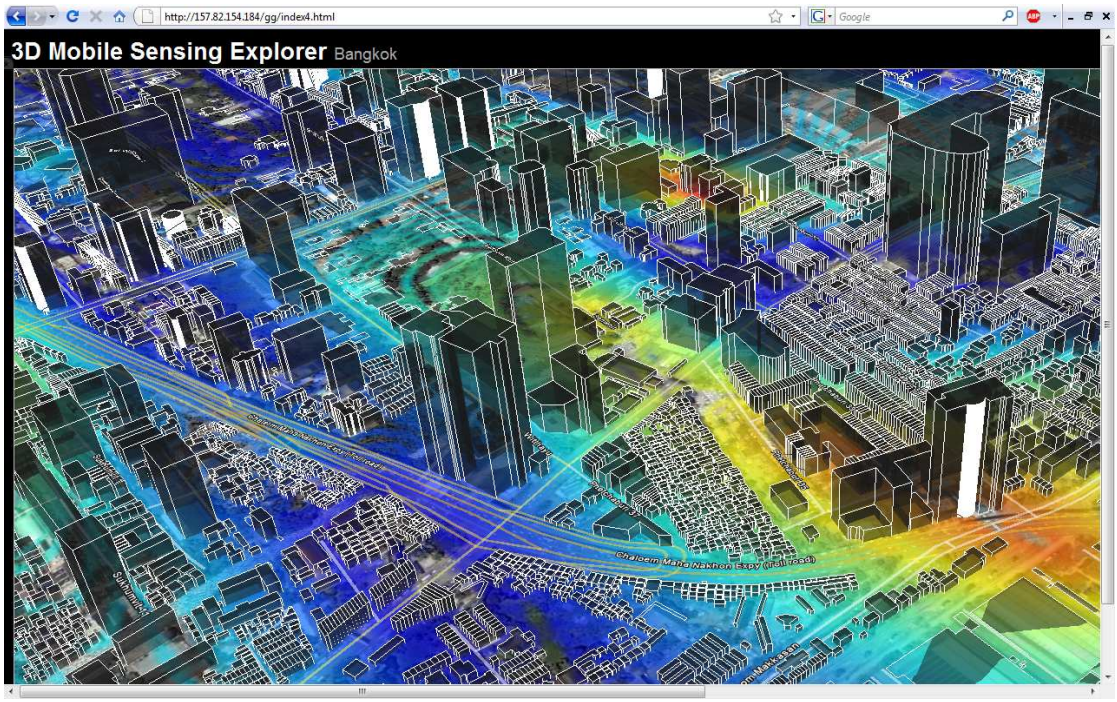


Figure 3.5 Illustrate "Pulse of a City" using Mobile Sensing platform, a browser based prototype 3D system with animated flow patterns

3.2.4 Data manipulation

This section will explain the process by which we first simulated a connection with a mobile operator in our local environment instead of communicating directly to the mobile operator system. The mobile log data during the period from February to April 2008 that cover a part of the central Bangkok area were transformed and inputted into the database. The data established in the database mainly include cell-id, the base station geographic position, update time and Erlang data (Table 1). The Erlang data which is calculated from the call duration are performed as a sample distribution in order to estimate the population density of the whole area.

The study area is covered by 2 Base Station Controllers (BSC) which are composed of 50 base stations and approximately 150 attached cell antennas. The data derived from BSC generally update in hourly intervals. Basically, more than 100,000 of records are fed into the database in one month. We analyze this sample data to see how people experience their city.

Cellid	Latitude	Longitude	Start time	Erlang
APGA1	13.75697	100.5594	2008/03/01 9:00	33.98
APGA2	13.75697	100.5594	2008/03/01 9:00	18.93
APGA3	13.75697	100.5594	2008/03/01 9:00	33.17
ARWG1	13.75138	100.5402	2008/03/01 9:00	20.75
ARWG2	13.75138	100.5402	2008/03/01 9:00	17.93
ARWG3	13.75138	100.5402	2008/03/01 9:00	33.07

Table 3.1 Sample Data from the Base Station Controller (BSC)

3.2.5 Interpolation methods and population density

In order to present population data in a continuous space, we need interpolation techniques to generate a surface from discrete points. There are many interpolation techniques each with their own weaknesses and strengths. For point-based interpolation, the typical examples are conditions based on geostatistical concepts (Kriging), locality (nearest neighbours and finite element methods, IDW, TIN), smoothness and tension (spline), or ad hoc functional forms (polynomials, multi-quadrics). These have been discussed by several researchers including Demers (2000), Mitas et al (1997, 1999) and Pariente (1994). In this paper, we introduced an inverse distance weighted (IDW) and ordinary kriging method to predict population density in our prototype system. The IDW method uses a distance-decay weighting function to determine weights of known centroids for predicting values at positions where observations are not available. The kriging method computes an empirical semivariogram based on observed data to estimate spatial autocorrelation of the interested variable. Based on the spatial autocorrelation evaluated, estimates at unobserved positions can be predicted with minimal kriging error.

As the number and distribution of sample points can greatly influence the accuracy of spatial interpolation, we attempt to increase interpolated resolution by calculating the weight in each cell separately instead of using the mean value. We first generate voronoi tessellation over the discrete set of base station points. Since

the base station is a three-sector cells type, that is, each cell is transmitting in one direction instead of broadcasting all around. (Figure 3.4) We divided each voronoi cell by 120 degrees refer to the antenna installation direction. At last, we re-calculated the new antenna points from the centroid of the segment voronoi (Figure 3.5).

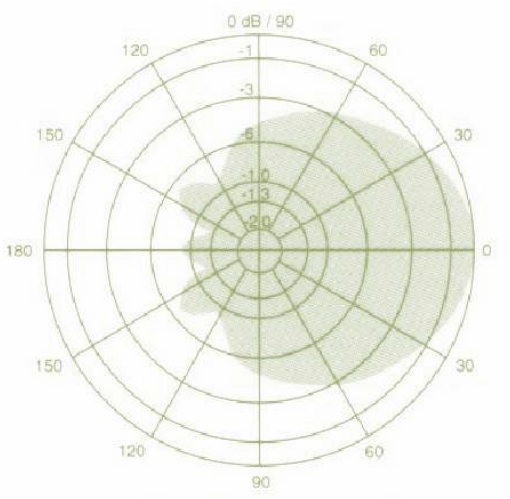


Figure 3.6 Mobile antenna, horizontal plane range
Source: www.matequipement.com

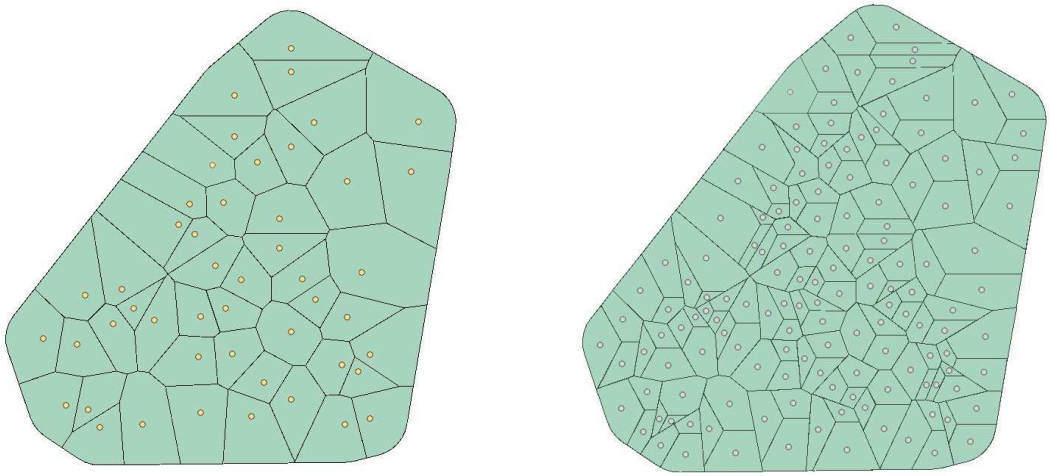


Figure 3.7 Increase interpolated resolution by using voronoi tessellation and the angle of antenna. Left image present original coverage of base station and right image present a sub sampling coverage by this method

Cellid	Latitude	Longitude	Start time	Erlang
APGA1	13.75806	100.56073	2008/03/01 9:00	33.98
APGA2	13.75387	100.56028	2008/03/01 9:00	18.93
APGA3	13.75683	100.55761	2008/03/01 9:00	33.17

Table 3.2 Data modification with voronoi-based segmentation method

3.3 Results and Discussions

We started by presenting our first results on querying the cumulative usage of the mobiles over an hour interval. Histogram and time series statistics could be retrieved from specific locations on a map and displayed in a time-plot based graph. The first two types of graphs, day and month, were produced by implementing a getStatistic webAPI in conjunction with Timeplot, a DHTML-based AJAX widget. The exploration leads to a better understanding of each area's activities during one day as well as the different characteristics between weekdays and weekends. In Figure 5(a), we picked up an office area in central Bangkok, and the time plot shows the trends of increasing activities from early morning up to the peak at noon then decreasing gradually after 5.00 pm. Figure 5(b) and (c) illustrate the overall activity on a monthly basis. We could capture a weekly rhythm of this area which clearly defines high activities on the weekdays and appears to decrease on Saturday and Sunday. Figure 5(c) presents a pulse in April: we can see a week of flat low activities since it was a long holiday in Thailand.

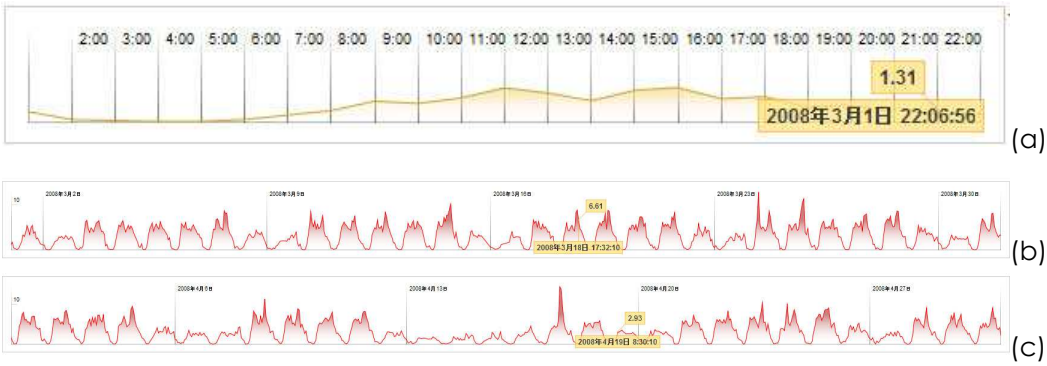


Figure 3.8 Day and month statistics from cumulative mobile usages data

3.3.1 Evolution of urban activities

Another approach which can be employed to explore the data is to generate a surface flow pattern by interpolating the aggregate call traffic. Exploratory analysis of temporal data can give a clear view on how people flow into and out of the city throughout the day. Figure 6 shows the flow patterns in one local area from 6.00 am. until 8.00 pm. This observation leads to the speculation on how one part of the central city is upscale, crowded and how long the area keeps busy until people

move to another part of the city. It is extremely useful for the urban planner to figure out how types of land use, street networks and other urban landscapes could affect the flow and density of the city. Furthermore, results of the study not only provide a new tool for area or zoning analysis but also could be used to specify hidden problems of the particular space over a period of time.

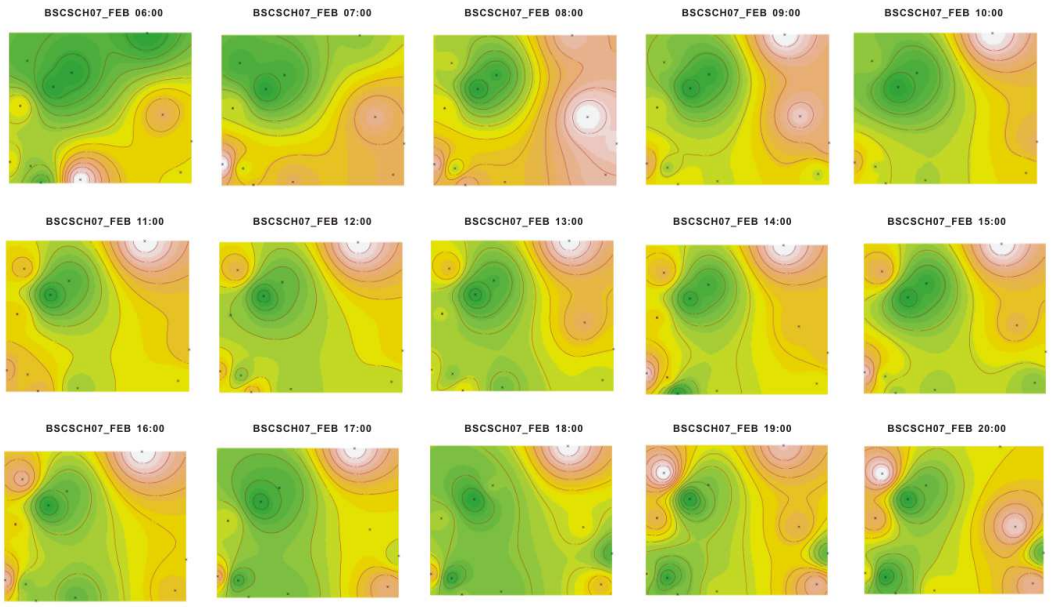


Figure 3.9 The flow pattern from early morning to late evening in central Bangkok

3.3.2 Abnormally detection

In order to clearly highlight extreme values in distribution surfaces, we generate the volume of call density with an overlay contour diagram using a specific color palate.

In Figure 3.9, to illustrate how mobile density could reflect the real world daily activities, we capture Pathumwan area. This is one of the most active spots in Bangkok that has a mixed land use, for example high rise office buildings and a large-scale shopping complex. If we compare the same period of time at 1.00 pm. on Friday and Sunday, the result demonstrates that on Friday the activities are dense at the office area and in contrast, a peak density moves toward the shopping area on Sunday.

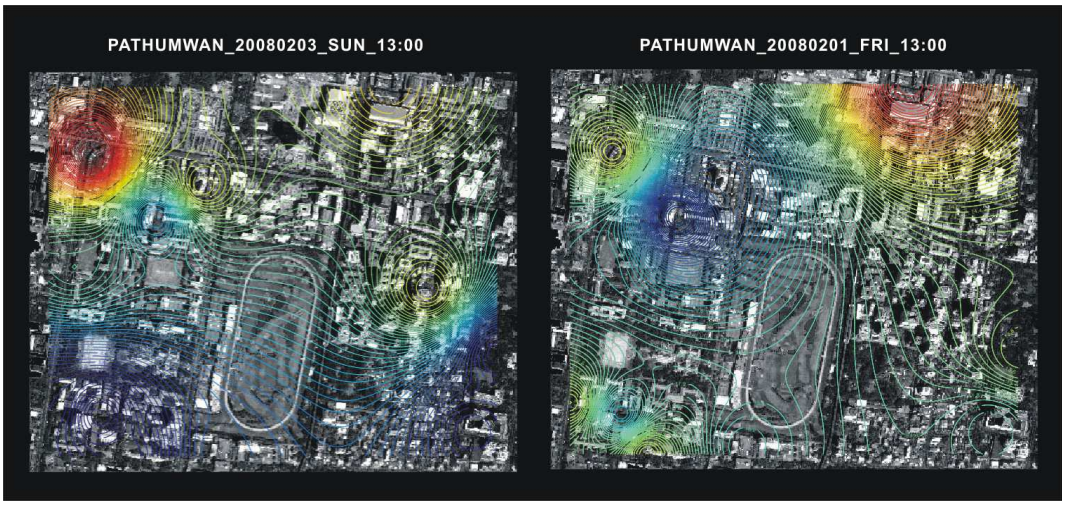


Figure 3.10 Comparison of density contour in Pathumwan area on Friday and Sunday at 1.00 pm.

Obviously enough, this kind of hot spot extraction could be useful to capture some hidden aspects in the urban space. We are planning to collect a longer span of archive data and make a base urban signature in order to implement a real time signature recognition and hot spot extraction.

3.3.2 Activities patterns and Land Use Classification

In general, land use could be roughly determined from the intensity of usage. Low-intensity uses are those that preserve some of the natural features of their sites, involve little construction, and require few public services and facilities, parks and farms, for example. High-intensity uses require major alteration of their sites, a lot of building and a maximum of supporting services and facilities, shopping centers and high-rise apartments, for instance.

The footprints from mobile phone activities could also be presented as a new approach to extract urban land use information from the usage intensity and patterns of use in each area at different timeframes during a day. It is easy to imagine that the activities of people in residential areas and business areas would be different since people regularly go to work or to school in the morning and return home in the evening.

We selected four points from different land use on the map which represent the residential area, the business area, the shopping center and the university respectively. In figure 8 we calculated a mean value with an hour interval in one month. The results clearly describe the unique characteristics of each urban space in term of human activities. The residential area considers increasingly more activities from the evening until late night (Figure 8-1). On the other hand, the business area keeps overcrowding during the day and gradually release after the office hour (Figure 8-2). The Shopping center tended to have a significant increasing of activities after 11 am. after the shops open (Figure 8-3). And the University has less activity in any time of the day because of the data source has been collected during March-April which is the semester break period in Thailand (Figure 8-4).

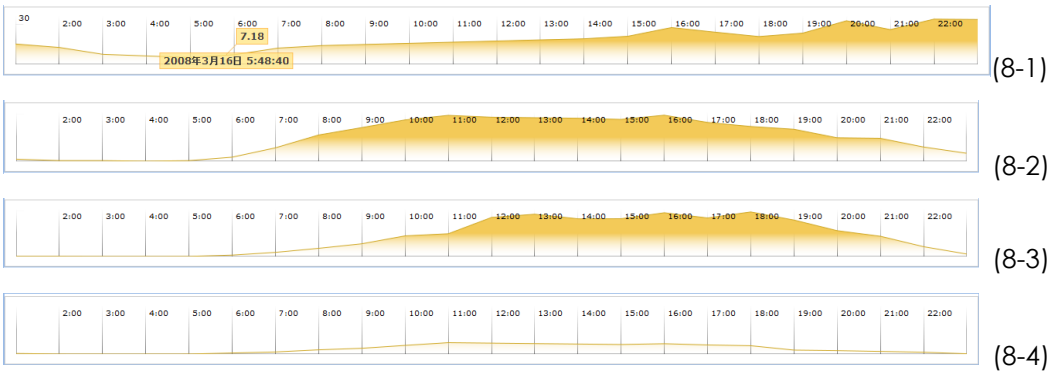


Figure 3.11 The 24 hour graph images illustrates the urban signature in specific land use

Since the study area covers only the center part of Bangkok, it lacks some interesting findings from other kinds of land use such as the night club area, low and high income housing area, national event places and some important transportation nodes such as park and ride. A study on the broader area would leads to a better understanding of each urban land use by evaluating the usage intensity and temporal signature of mobile activities.

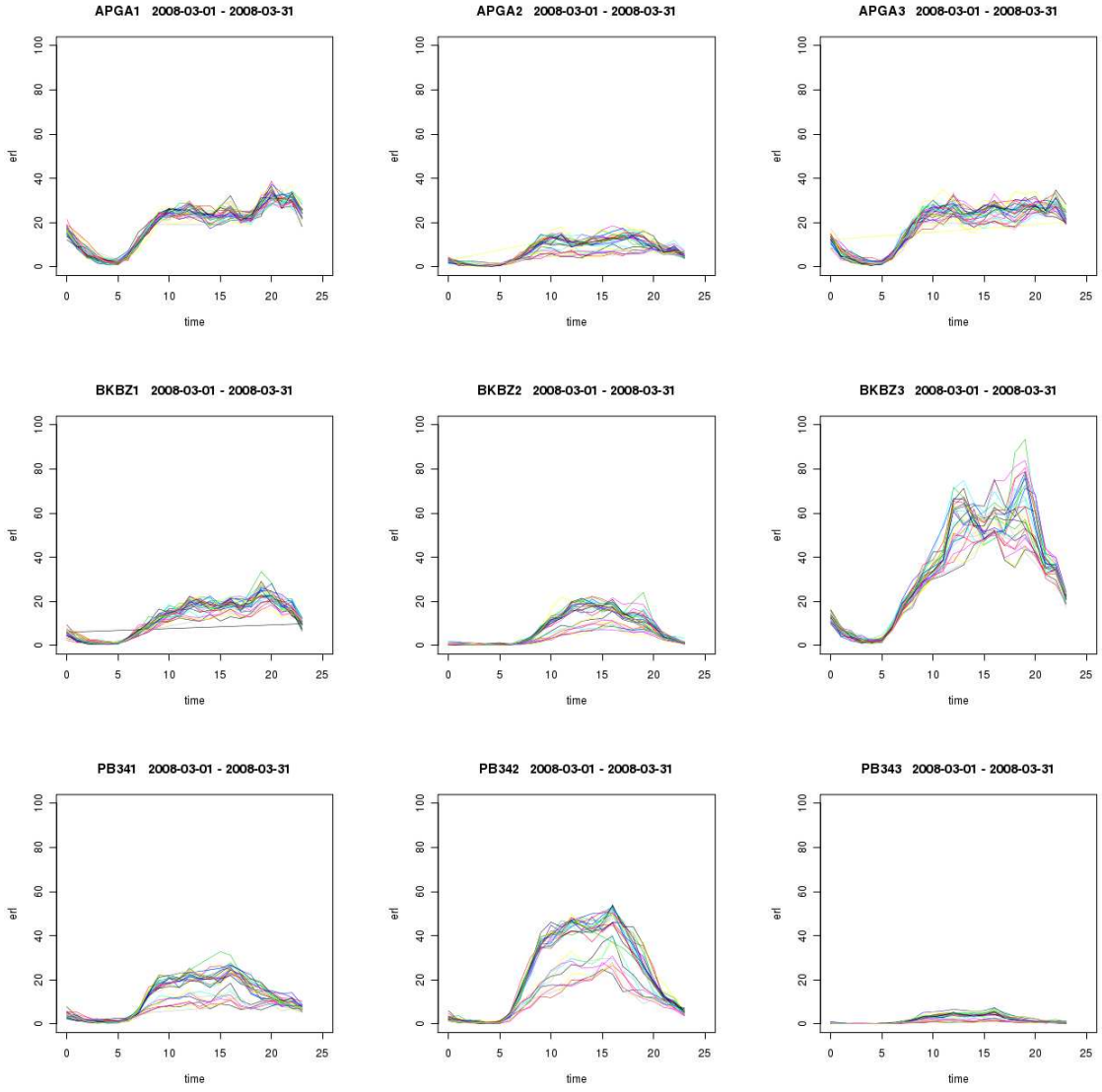


Figure 3.12 Individual mobile phone signature in one day

3.4 Future perspective

In this paper, we have presented the concept design, implemented and illustrated the evolution of city by using the mobile sensing platform. We are planning to extend the study area to other cities as well as include additional sensing data from the public domain services. The initial results of this paper only presented the analysis of urban activities from the mobile phone usages. In the future, we plan to analyze the flow of the city with real time traffic data and urban environmental information from mobile sensor networks. We would like to make the platform as flexible as possible in order to integrate with the public web services. This means the system would not only be used for the data visualization but also data analysis and processing tasks.

Using mobile phones as sensing devices and aggregating "crowdsourced" data for urban analysis is still in the early phases of concept development. Mobile sensing can potential paint a complex and dynamic portrait of the urban environment in which users are based.

Acknowledgements

We would like to thank Advanced Info Services PLC for a support of mobile data and Spatial Dimension Solutions (SDS) for a support of base map and Ikonos satellite images.

Chapter 4

TELEMATIC BEHAVIOR: Identifying Human Daily Activity Pattern Using Mobile Phone Data

4.1 Introduction

Dynamics of human mobility is essential for urban planning and transportation management. Besides geographic space, in this chapter, we characterize mobility in a profile-based space (activity-aware map) that describes most probable activity associated with a specific area of space. This, in turn, allows us to capture the individual daily activity pattern and analyze the correlations among different people's work area's profile. Based on a large mobile phone data of nearly one million records of the users in the central Metro-Boston area, we find a strong correlation in daily activity patterns within the group of people who share a common work area's profile. In addition, within the group itself, the similarity in activity patterns decreases as their work places become apart

This chapter aims to explore to what extent people working in different part of cities and towns has similarity in daily activity patterns by analyzing their individual mobility pattern from the mobile phones usages. For better understanding of the effects of human movement, characterizing human mobility patterns is crucial. For example, without such characterization, the impact of inhabit dynamics in the city cannot be understood. As spatiotemporal and geo-referenced datasets are growing rapidly because of the daily collection of transaction data through database systems, network traffic controllers, sensor networks, and telecommunication data from mobile phones and other location-aware devices, the large availability of these forms of data allows researchers to better characterize human mobility. The additional information of activities associated with human mobility further provides a unique opportunity to better understand the context of human movement, and hence better urban planning and management.

This chapter, we develop the activity-aware map, which provides information about the most probable activity associated with a specific area in the map. With the activity-aware map and an analysis of a large mobile phone data of nearly one million records of location traces, we are able to construct the individual daily activity patterns. This allows us to carry out a correlation analysis of work area's profile and similarity in daily activity patterns.

4.1.1 Related work

A rapidly increasing number of mobile phone users has motivated researchers from various fields to study its social (Turner et al., 2003; Nickerson et al., 2008; Liu et al., 2010) and economic (Kauffman, R.J., 2005; Giray et al., 2009; Li & McQueen, 2008) impact. With the extensive records of mobile phone data such as calling pattern and location of the mobile phone user, analyses have been performed on numerous aspects including behavioral routine (Eagle et al., 2006), social proximity (Clauset & Eagle 2007), call prediction, social closeness (Phithakkitnukoon et al., 2007), and human mobility. (Azevedo et al., 2009; Lee et al., 2009; Candia et al., 2008; Gonzalez et al.,2008)

Understanding dynamics of social networks is beneficial to urban planning, public transport design, traffic engineering, disease outbreaks control, and emergency response management. To study dynamics in human mobility, GPS receiver has been handy for researchers in collecting large real-life traces. Azevedo et al. study pedestrian mobility behavior using GPS traces captured at Quinta da Boa Vista's Park in Rio de Janeiro (Brazil). Movement elements are analyzed from 120-pedestrian collected data. They find that the velocity and acceleration elements follow a normal distribution while the direction angle change and the pause time measure fit better to lognormal distribution. Based on 226 daily GPS traces of 101 subjects, Lee et al. develop a mobility model that captures the effect of human mobility patterns characterized by some fundamental statistical functions. With analytical and empirical evidence, they show that human movement can be expressed using gaps among fractal waypoints. (Rhee et al.,2008) People are more attracted to more popular places.

With a large set of mobile phone data, Candia et al. study spatiotemporal human dynamics as well as social interactions. They investigate the patterns in anomalous events, which can be useful in real-time detection of emergency

situation. At the individual level, they find that the interevent time of consecutive calls can be described by heavy-tailed distribution, which is consistent with the previous reports on other human related activities. Gonzalez et al. examine six-month trajectory of 100,000 mobile phone users and find a high regularity degree in human trajectories contrasting with estimation by Levy flight and random walk models. People tend to return a few frequent locations and follow simple repeated patterns despite the diversity of their travel history. The most recent study in human mobility based on a large mobile phone data by Song et al., whose result is consistent with Gonzalez et al.'s that human mobility is highly predictable. Based on data from 50,000 mobile phone users, they find that predictability in human mobility is independent of distance that each individual regularly travel and show that the predictability is stabled at 93% for all regular traveled distances of more than 10km.

In contrast with other work in human mobility, this work is focusing on human mobility concerning the spatial profile (i.e. type of space or surrounding area such as dinning, shopping, and entertainment) rather than geographical location.

4.2 Methodology

A number of literature have described geographical human mobility pattern concerning movement of people between multiple locations. Here we are interested in characterizing the mobility not by geographic location but its associated spatial profile. This spatial profile-based mobility pattern, in turn, becomes a human activity pattern. In addition, our interest expands to investigation of relationship between this activity pattern and demographic of people. Therefore, in this section, we will describe our methodology used in characterizing space, capturing daily activity pattern, as well as preprocessing our dataset.

4.2.1 Data Preparation

In this part, we use anonymous mobile phone data collected during the period from July 30th, 2009 to September 12th, 2009 by Airsage Inc. of about one million users in the state of Massachusetts, which account for approximately 20% of population, equally spread over space. This includes 130 million anonymous location estimations in (latitude, longitude)-coordinates, which are recorded when the users are engaged in communication via the cellular network. Specifically, the

locations are estimated at the beginning and the end of each voice call placed or received, when a short message is sent or received, and while internet is connected. Note that these location estimations have an average uncertainty of 320 meters and median of 220 meters as reported by Airsage Inc. based on internal and independent tests. For our analysis, we consider the mobile phone data within an area of 33x42 km2, which includes 52 cities (Boston, Cambridge, and others) in the county of Essex, Middlesex, Suffolk, and Norfolk as shown in Fig. 1. The list of the counties and their corresponding area covered (in km2) by this study are shown in Table 4.1.

County	Area Covered (km^2)
Essex	110.30
Middlesex	452.52
Suffolk	154.39
Norfolk	26.12

Table 4.1 List of the counties and their area covered by this study

Within this area in the map, we need to extract mobility traces of each user from the mobile phone data. As the estimation of the user's location is aggregated only when network connection is established, mobility thus can be derived as a temporal sequence of locations. To segment these traces into trajectories so that daily mobility pattern of each individual can be indentifies, we describe here some basic algorithms to extract trajectory and stop.

Let X_k denote a set of sequential traces of user k such that $X_k = \{x_k(1), x_k(2), x_k(3), ...\}$ where $x_k(i)$ is a position i of user k. A trajectory can then be obtained by segmenting X_k with the spatial threshold $\triangle S$. If a distance between adjacent positions is greater than the threshold (distance $(x_k(i), x_k(i+1)) > \triangle S$), then the early position $x_k(i)$ becomes the end position of the last trajectory while the later position $x_k(i+1)$ becomes the starting position of the next trajectory. Once the

trajectories are detected, a stop can be identified as an event during which the user stays in a specific location for a sufficiently long period of time. As each position i contains location and timestamp, i.e. x_k (i) = (lat(i), long(i), t(i)), extraction of a stop depends on time and space. A stop is thus regarded as a sequence of positions { x(j), x(j + 1), x(j + 3), ..., x(j + m) } where the distance between any adjacent positions is less than a spatial threshold Sth i.e., distance(x(j), x(j + 1)) < S_{th} , and time spent within the location is greater than a time threshold T_{th} i.e., t(m) □ t(j) > T_{th} .

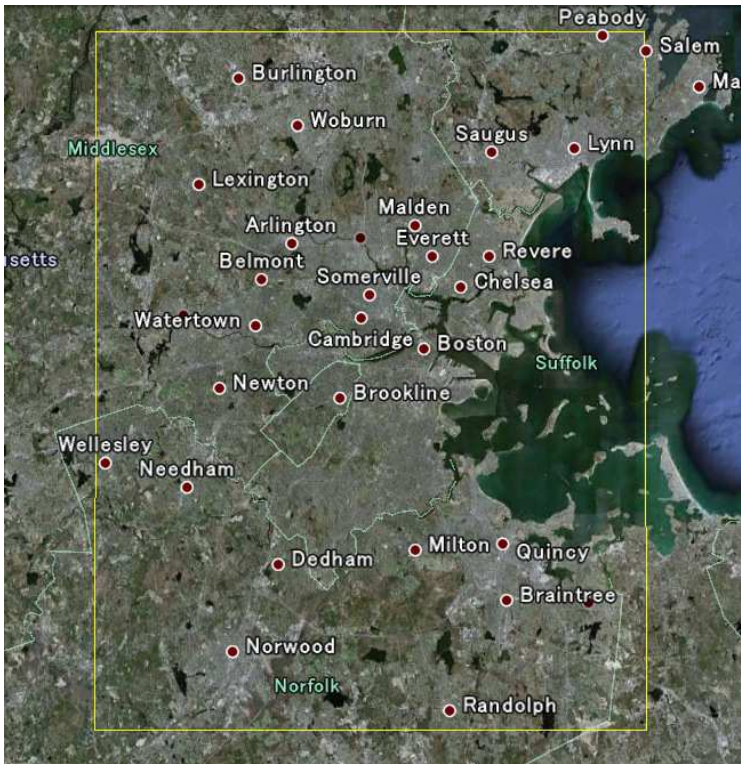


Figure 4.1 Area of study, cropped by yellow line

After stops have been identified, work location of each user is then estimated as a most frequent stop during the day hours. The information about work location allows us to derive the mobility choices of the users, and detect activity patterns throughout the day.

4.2.2 Spatial Profiling

To model the space, we construct a virtual grid reference by dividing the map into square cells of size 500 meters by 500 meters. Since our interest is in the activities associated with the space, we thus characterize space based on the type of activities expected to be performed within given space. For example, if restaurants were clustered within a particular area, then this area would be associated with eating activity. In this study, we consider four different human activities in which people typically spend time engaging on daily basis. These activities are concerning eating, shopping, entertainment, and recreational. Profiling the map according to these activities requires information about the types of places within each cell. To acquire the information regarding these activities, we search for Points of Interest (POIs) for each cell location. We use pYsearch (Python APIs for Y! search services) version 3.1 for POI search service, and Reverse Geocoding with Geopy (A Geocoding Toolbox for Python) for translating (latitude, longitude) coordinate into a physical address. For each activity category of each cell, we make three search attempts using different keywords. The keywords used for each activity category are listed in Table 4.2. With the limit of 5,000 queries per day restricted by Yahoo, an extensive amount of search time is required inevitably.

Activity	**Keyword used**
Eating	Restaurant, Bakery, Coffee shop
Shopping	Mall, Store, Market
Entertainment	Theater, Bowling, Night club
Recreational	Park, Gym, Fitness

Table 4.2 Considered activities and keywords used for POIs search

Once POI searches are completed, the number of POIs associated with each activity category is recorded for each cell. The raw activity distribution map is then composed of 500x500 m^2 cells where each cell contains distribution of each activity. Each cell C_i contains normalized portion of each activity:

$$C_i = [\alpha_i(1), \alpha_i(2), \alpha_i(3), \alpha_i(4)] \quad (1)$$

where i = 1, 2, 3, ..., N, N is the total number of cells, and normalized portion of each activityα_i (a) in cell i is computed as

$$\alpha_i(a) = \frac{n_{\alpha_i(a)}}{\sum_{i=1}^{N} n_{\alpha_i(a)}} \quad (2)$$

where α_i(a) denotes the number of POIs associated with activity a within the cell i and a = 1, 2, 3, 4 corresponds to eating, shopping, entertainment, and recreational activity, respectively.

Based on our POI search, Figure 4.2 shows a map with the visual grids and POIs found by 12 different keywords (described in Table 4.2) in different colors. To further classify these cells into a more crisp distribution map, we apply k-means algorithm with k=4. The resulting crisp activity distribution map is depicted in Fig. 3 where each cell is classified to one of the four activities according to Bayes theorem:

$$P(a \mid n_{\alpha_i(a)}) = \frac{P(n_{\alpha_i(a)} \mid a)P(a)}{n_{\alpha_i(a)}} \quad (3)$$

The interest here is to find the most probable activity category a for each of the k clusters. Therefore, for each cluster, we find a that maximizes a posteriori (MAP method). So we use Bayes theorem above to compute the posterior

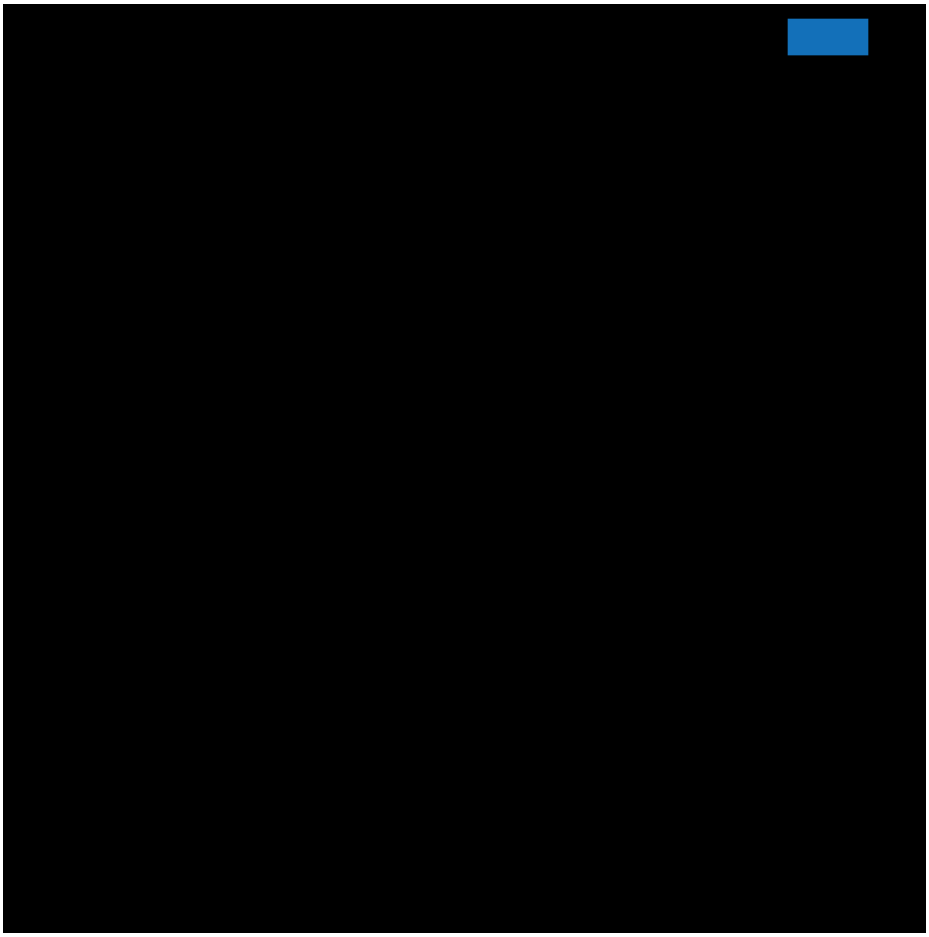


Figure 4.2 POI search results on the map with 500x500 m^2 visual grids.

probability of each activity category as follows:

$$
\begin{aligned}
a_{MAP} &\equiv \arg\max_{a} P(a \mid n_{\alpha_i(a)}) \\
&= \arg\max_{a} \frac{P(n_{\alpha_i(a)} \mid a)P(a)}{n_{\alpha_i(a)}} \\
&= \arg\max_{a} P(n_{\alpha_i(a)} \mid a)P(a).
\end{aligned}
\tag{4}
$$

4.3 Daily Activity Patterns

Generally, people perform different activities throughout the day. A lot of these activities are repeated on daily basis, e.g. eating around 12pm (noon), jogging in the evening, and hence producing recognizable patterns. With our mobile phone data, each user is more likely to engage in an activity during "stop" rather than on the move. Therefore, for each stop, activity is identified according to the crisp activity distribution map.

To infer a daily activity pattern for each user, we divide 24-hour time scale into eight 3-hour segments starting at 5AM as shown in Figure 4.4. So, daily activity pattern is simply a sequence of activities performed by the user during each stop throughout the day. For each user, daily activity patterns are collected over the course of the data collection period. Note that, in this study, we consider only weekdays (Monday, Tuesday, Wednesday, Thursday, Friday) as our speculation is that weekday pattern is different from weekend pattern due to typical work

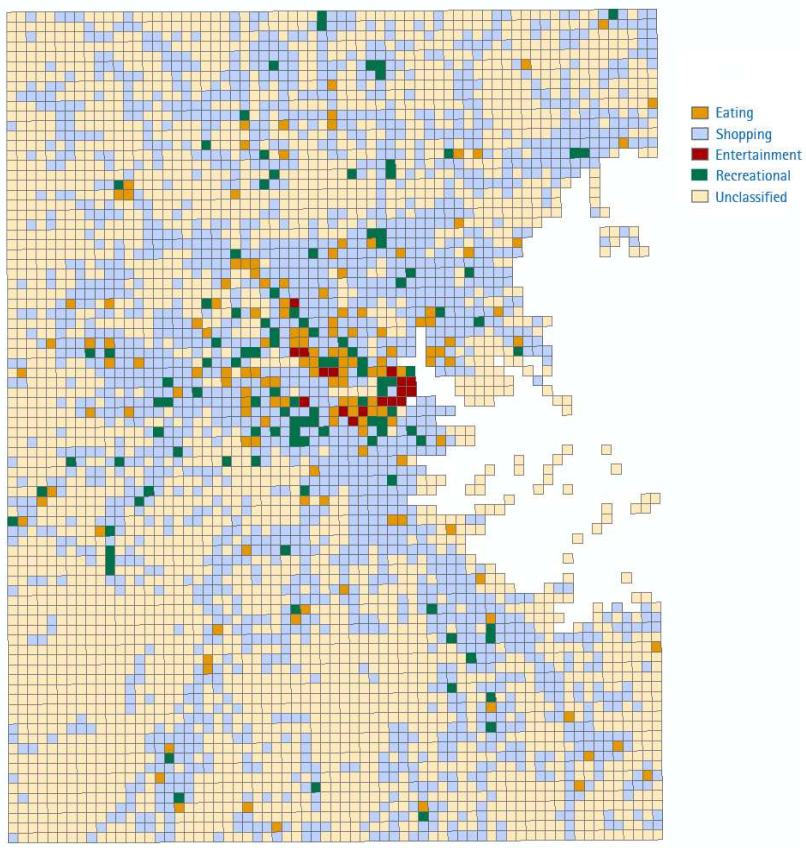


Figure 4.3 Crisp activity distribution map.

schedule and hence different daily activity sequences – this will be addressed and further discussed in our future work.

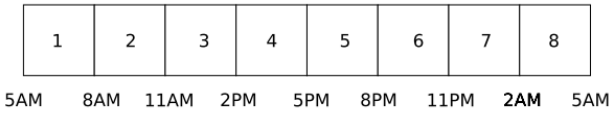


Figure 4.4 The eight 3-hour temporal windows are used to frame the daily activity pattern.

To derive the representative daily activity pattern of each user, we simply assign each segment with the most frequent activity during that time interval over the period of data collection. Precisely, if $\lambda_a^d(t)$ represents the count of activity a on d-th day during time segment t (where t = 1, 2, 3.., 8), then

$$z(t) = \arg\max_a \sum_{d=1}^{M} \lambda_a^d(t) \quad (5)$$

where z(t) is the assigned activity for time segment t and M is the total number of days.

4.4 Work Area's Profile and Similarity in Daily Activity Patterns

The activity map and individual daily activity patterns developed in the previous section allows us to conduct a number of studies that can be useful for better understanding of human behavior in the city. In this present research, we are particularly interested in relationship between people's daily activity patterns and the characteristic of their work area. Do people who work in the same area's category (e.g. eating, shopping, etc.) also have similar daily activity patterns? With the same type of work area, how does distance impact the similarity in their daily activity patterns (e.g. do people who work in an urban shopping area have similar

activity pattern with people who work in a distant shopping area)? In this current study, we are attempting to answer these two questions.

As a first step, we classify the users into four groups based on their work cell's profiles. Each group then consists of a number of different individual daily activity patterns who have a common work cell's profile. To represent each group's activity pattern, we need to find a group signature for further correlation analyses. The representative daily activity pattern or signature of each group can be obtained in a similar fashion with the individual patterns described in the previous section (using Eq. (5)). The derived signatures are shown in Table 4.3.

Group	Group's daily activity pattern
Eating	W–W–W–W–Sho.–Rec.–Rec.–Sho.
Shopping	W–W–W–W–Rec.–Rec.–W–W
Entertainment	Sho.–W–W–W–W–Rec.–Sho.–Sho.
Recreational	W–W–W–W–W–Sho.–Sho.–Sho.

Table 4.3 Signature of each group based on work cell's profile. Note: Eat. = Eating, Sho. = Shopping, Ent. = Entertainment, Rec. = Recreational, W = Work cell.

It can be noticed that there is no Eating element appears in any of other group signatures beside its own group (showing in form of a working activity, W). Our speculation is that it could be caused by first, people normally eat at home (breakfasts) and at work or somewhere nearby workplace (lunches), and second, people are not frequently involved in a phone communication while at eating area. Note also that the patterns are derived from weekday's activities so if weekends-only activities are considered, Eating elements could emerge in the group patterns.

To answer the first question, we need to measure similarity in daily activity patterns among individuals within the same group as well as among other groups. To measure distance (dissimilarity) between two daily activity patterns, we use

Hamming distance, which is normally used to measure distance between two strings of equal length. The distance is essentially the number of positions at which the corresponding symbols are different, which is quite suitable for our case as a series of activities can be considered as symbols. The result of the average Hamming distance within the group is shown in Table 4.4.

Work cell's profile	Average distance
Eating	4.78
Shopping	2.58
Entertainment	4.67
Recreational	3.61

Table 4.4 Average within-group distance.

Using group signatures obtained earlier, we then measure dissimilarity between each group signature and other group's individual patterns. The result of this between-group distance is shown in Table 4.5 in forms of average Hamming distance.

	Eating	Shopping	Entertainment	Recreational
Eating	-	6.53	6.60	6.96
Shopping	4.90	-	4.92	5.05
Entertainment	6.43	6.88	-	7.00
Recreational	5.04	4.81	5.13	-

Table 4.5 Average between-group distance.

As the result of our first investigation, Figure 4.5 illustrates a bar plot intended to make a comparison between within-group and between-group distances where orange bars represent within-group distance while blue bars represent between-group distance. Clearly, it shows that within-group distances are less than between group distances. This implies that people who have a common work cell's profile tend to exhibit more similar daily activity patterns than people who have different work cell's profile.

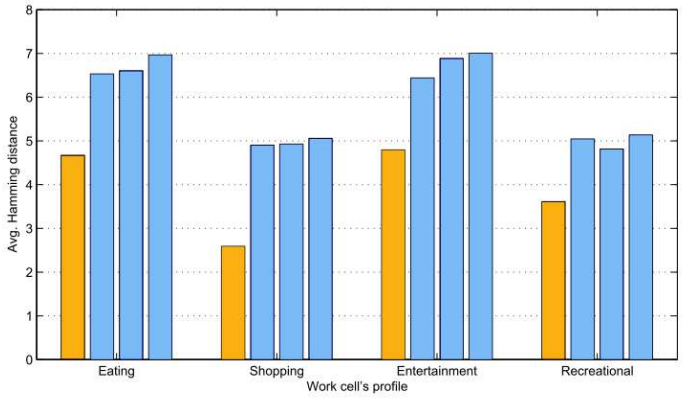


Figure 4.5 When users are grouped together based on their work cell's profiles, within group and between-group distances are illustrated with red and blue bars respectively. This shows higher degree in similarity within the group than between groups.

For the second investigation about the impact of physical distance on the similarity in activity patterns, we decide to proceed by placing a growing spatial window (a circle of an arbitrary radius) onto the map then measure similarity between the users' activity pattern whose work cell located at the center of the window and other users whose work cells are within the vicinity of the spatial window. The similarity is being measured while the radius of the window grows from a small to larger value. The process is repeated for each activity category. This way, we can see the change (if any) in similarity for each work profile as we move away from the center area. Precisely, we choose to grow the spatial window from the center of Boston area with the radius varying from 0.5km to 30km. The result for each

work category is shown in Figure 4.6. We can observe that, overall, the similarity in activity patterns decreases as radius increases, which implies that physical distance has an impact on similarity in daily activity patterns. People whose work area's profile are although the same, their activity patterns tend to deviate more as they work areas become further apart. In summary, we have observed a strong correlation in daily activity patterns within the group of people who share a common work area's profile. Addition, within the group itself, the similarity in activity patterns decreases as the distance between them increases.

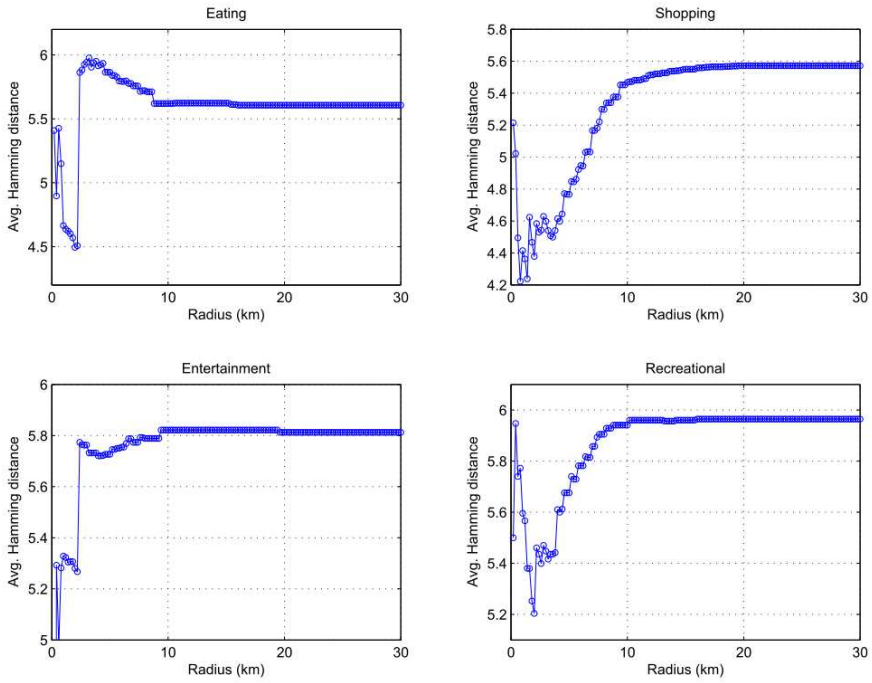


Figure 4.6 Dissimilarity in daily activity patterns is measured by average Hamming distance as the radius varies from 0.5km to 30km for each work cell's profile. The center of the growing radius is near the center of city of Boston. Dissimilarity is between the users whose work cells are within the 0.5km radius and other users covered by growing radius.

4.5 Conclusions

In this chapter, we have developed an activity-aware map that contains most probable activity associated with a specific area in the map based on POIs information. With activity-aware map, we are able to extract individual daily activity patterns from analyzing a large mobile phone data of nearly one million records. Results from our correlation analysis show a strong correlation in daily activity patterns within the group of people who share a common work area's profile. In addition, within the group itself, the similarity in activity patterns decreases as the distance between them increases. This study is the first report of many more to come in using activity-aware map to study inhabitant behavior. So as our future direction, we will continue to investigate on daily activity pattern and its dynamics for better understanding of human dynamics, which in turn benefits urban planning and management.

Limitations of the Study

There are a number of limitations of this study. First and foremost, the lack of continuity of mobility traces due to the fact that the location is estimated from mobile phone data only when connection with a cellular network is made through voice, text, or data communication, which constricts us to a smaller number of users that can be analyzed. Secondly, our POI search is constrained by Yahoo's search limit and capability. Lastly, home and work locations are estimated intuitively according to the data provided. Although ground-truth validation is desired, it would be very difficult to perform due to the privacy issue.

Chapter 5

Calling Pattern and Call Simulation of Crowd Movement

5.1 Introduction

Sensing is going mobile and people-centric in a network enabled society. Urban motilities are recognized from the footprint of the usages of mobile phone. This research introduces how mobile devices can be employed as a new urban sensor and how mobile activities can be represented as dynamic data sources for population estimation. The paper aims to evaluate the presence of People from the Cell Phone Trace Analysis. We demonstrate this assumption by using crowd simulation approach. We first analyze the calling patterns and the mobility patterns of the individual to create a user model. The simulated trajectories are generated from a group of the most active users and therefore are selected in proportion to the number of population in census tract. The random call patterns are assigned to individual agents to formulate the mobile activities in the simulation space. The simulated results express the call volume and the existing of caller location at the point in time. This paper illustrates how the digital footprint from mobile devices can help understand the population distribution of the area in today's excessive dynamic society.

Populations of the cities are intermittently dynamic changing as people commute from place to place, technical speaking across two separate geographic boundaries. The question in estimating population density in high temporal resolutions has reached unprecedented levels, and is therefore the focus of the present study. In this chapter, the usage of mobile devices was treated as a medium for data collection. The fundamental tools have been developed to visualize the mobile phone activities and urban mobility in spatio-temporal space. We calculate the estimated ambient population base on crowd simulation approaches by utilizing a novel data source of mobile phone activities and other

ancillary information. The results formulate aggregate call volume and estimated population density. This paper illustrates how digital footprint from mobile devices can potentially be used as an indicator to estimate the population distribution.

The main purpose of this chapter is to examine the population count at a high temporal resolution. The key problem is how to acquire and estimate the number of population in the whole city space. In this paper, the usage of mobile devices was treated as a medium for data collection. The call distribution in mobile communication network was applied as aggregate data sources to estimate the population density. Previous studies had suggested that erlang measures at particular cells seem to be a decent indicator of actual presence of people. (Horanont and Shibasaki, 2008; Reades et al., 2007) However, erlang data itself could not yield the existing of people who are not using the mobile phone.

It is therefore not surprising that considerable research has gone in this new and interesting direction, in particular, developing techniques to bring out and explore large scale human mobility. Marta C. Gonzalez et al. observed the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period. The results had shown a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations. This is describing the recurrence and temporal periodicity inherent to human mobility.

We often find that data sensed and incorporated in real-time offers additional value to their analysis and simulation models. Live geospatial data provides an immediacy and relevance to applications in domains such as situational awareness, population dynamics, transportation, and evacuation planning. The previous paper (Lieu 2003) described that dynamic approach offers an effective strategy to migrate static application systems to more real-time data-intensive applications which are often the domain of GIS models and users.

Consequently we address the following questions: Would modification of calling activities be confirmed to correspond to actual population in the area? And would land use affect to the consistency between number of call and actual population? We have scheduled the research into two phases. First is to extract calling patterns from a million of mobile phone users collected from July to October 2009 in Massachusetts. Second, we developed a platform to simulate a virtual urban system and call activity to test the population prediction assumption and find how much confidence in estimating the population density by using only anonymous mobile phone traffic based on simulation approach.

5.2 Call patterns analysis

5.2.1 *Mobile phone traffic and population distribution*

In the fourth quarter 2009, there is over 4.12 billion people subscribed to mobile phone communication networks that mean nearly two-third of the world population own the mobile phone today. And the increasing reach of mobile networks creates an unprecedented opportunity to describe the existing of the entire population especially in the big city where the subscriber percentage is high.

In general, we collect the data from Base Station Controller (BSC) which is a part of Radio Subsystem in mobile communication. When people use their mobile phone, their estimated position and time will be logged into BSC database. Our primary assumption remark that there is a possibility to estimate population numbers from the mobile traffic and also we could possible predicts flow and movement of people in the entire city from everyday routine use of the mobile phone. Hence, it is necessary to confirm a correlation between mobile phone CDRs and population density.

The initial exploration has been done by mapping the night time mobile phone user density (figure 5.1) comparing with the census data in tract level. (figure 5.2) The visualization result generally confirms a good match of two dataset.

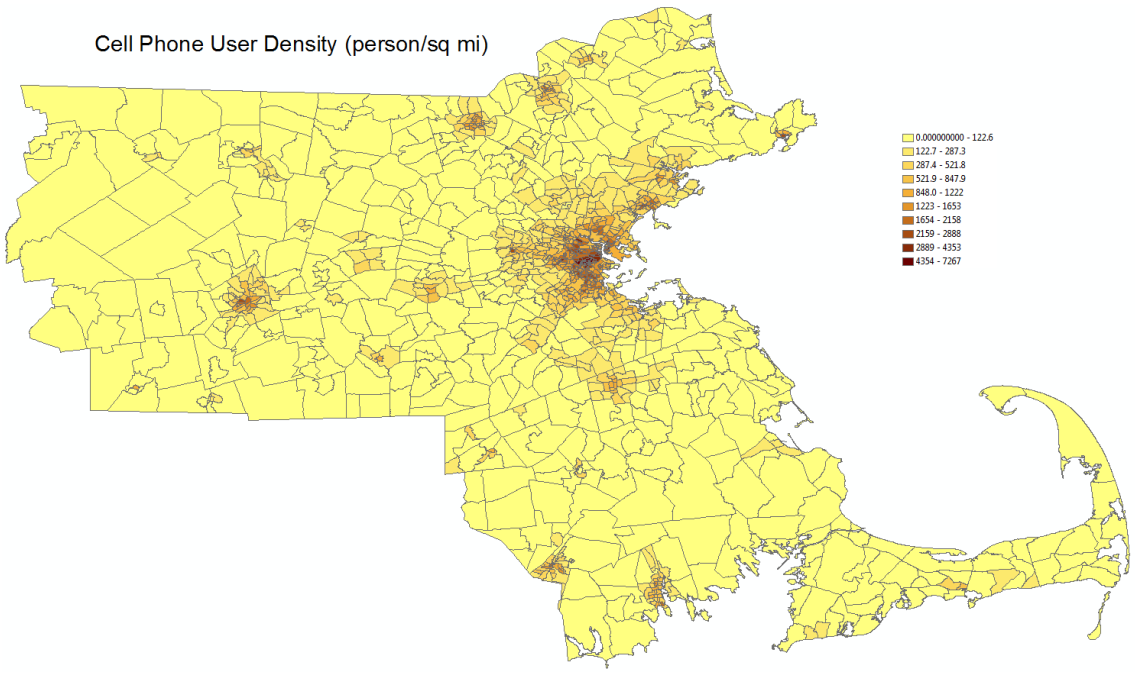


Figure 5.1 Night time cell phone users density, calculated at the census tract level (generally one square mile)

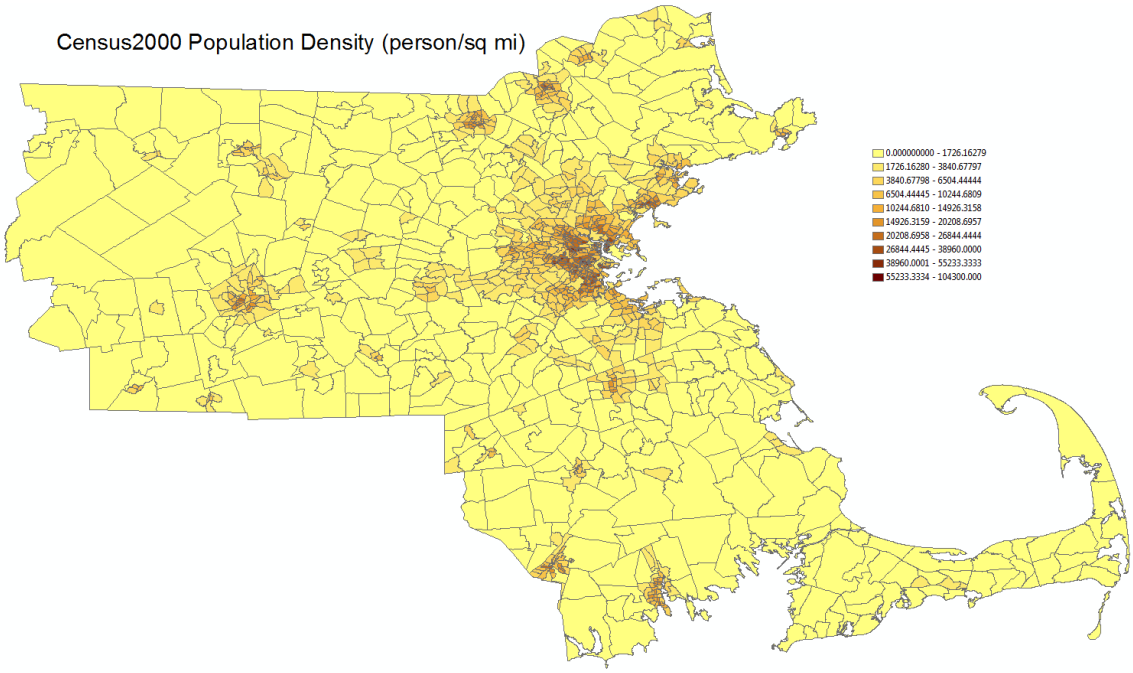
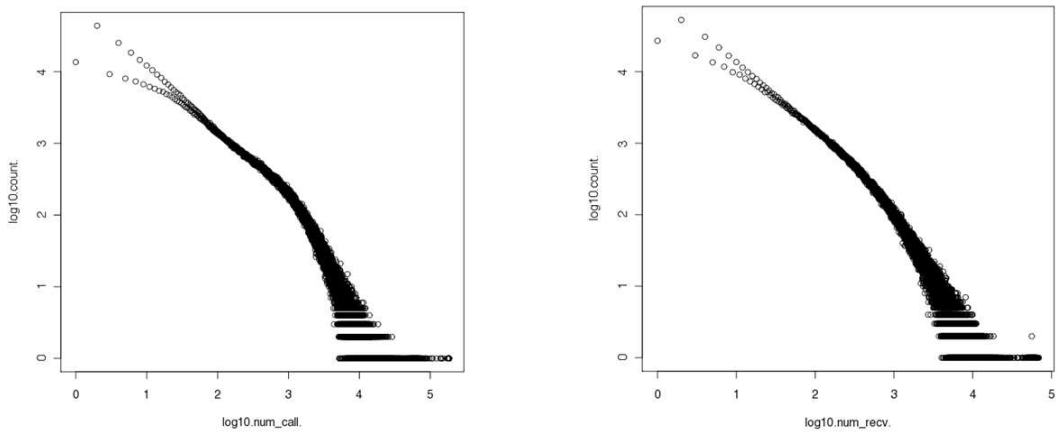


Figure 5.2 Population densities, calculated at the *census tract level* (generally one square mile)

5.2.2 Call patterns extraction methods

The call traffic pattern is an ideal source in designing a calling behavior of each agent in the simulated space. First, by observing the features of the database and then extracting the attributes needed to be mined. Basically Airsage data compose of 3 specific types of cell phone traffic that is while placing a call, receiving a call and using the data connection package. Figure 5.3 illustrate overall distribution of call and incoming call of one million observations in Massachusetts.



(a) (b)

Figure 5.3 Distributions of outgoing call (a) and incoming call (b) in one month. While y axis is population count in log scale and x axis is frequency of call in log scale.

Each single activity comes with the estimated location of the mobile device, method of estimation, approximated distance error and time stamp. Data from the specific group are then summed in an hour interval and examined to assemble a call traffic pattern. The results illustrate how people in Massachusetts behave in using their mobile phone in normal weekdays. We set up the overall process of finding and interpreting patterns from the data by the following steps;

Selection

We developed criteria by first studying a simple conceptual overview of mobile phone activities. We selected a trial by subset a thousand of samples from the

entire users collected in October 2009. The result set were examined separately base on activity group, namely call, receive and data connection activity. The selected procedure had been conducted by plotting the number of mobile activities against the number of people who generate the activity (Figure 5.4). Since we want to make sure that the subset is fully represented to the entire dataset, therefore we have to select them in fair distribution as possible. A straightforward way to cope with this problem is by using binning techniques in which the sample is reduced in resolution to a sufficient degree to ensure that a given peak remains in its bin.

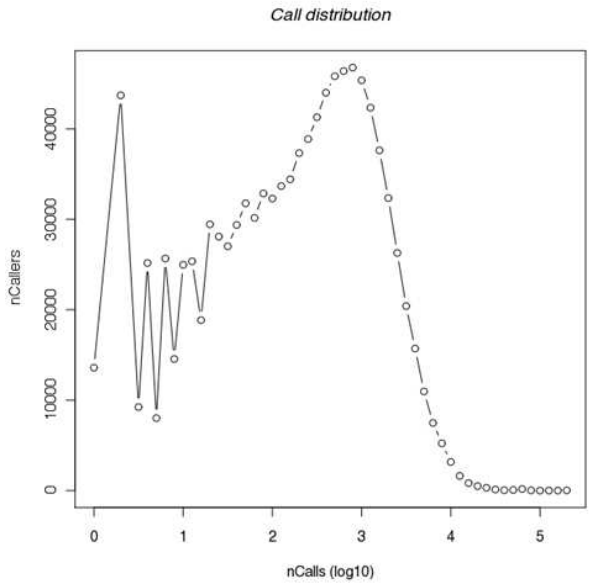


Figure 5.4 Illustrate number of calls and callers

Preprocessing

We aggregate the call activities by an hour and calculate the mean usage of each type in one month. As we focus on the normal working day simulation, the data on weekends and national holidays will be discarded. We then derived the data by formulating the calculation as discussed below.

For each user k, the mean usage for each hour slot t where t = 1,2,3,…,24 can be computed as

$$\bar{x}_k(t) = \frac{1}{N_i}\sum_{i=1}^{N_i} x_k^d(t) \quad (1)$$

Where Ni is the total number of days in experiment and hence d denotes day (d = 1,2,3,...,Ni). Such that mean usage of user k can be expressed as a series of $\bar{x}_k$ (t) for t = 1,2,3,...,24 as

$$X_k = \{\bar{x}_k(1), \bar{x}_k(2), \bar{x}_k(3), ..., \bar{x}_k(24)\} \quad (2)$$

The data which contains a missing value or zero activity in each category was filtered out before process to the next step.

Data mining

Data mining methods are algorithms designed to analyze data or to extract from data patterns into specific categories. In this study we use K-means unsupervised clustering approach. The clustering methods repeatedly run to discover the satisfy patterns in appropriate range of sum of squared errors. Figure 5.5 shows five categories of call clustering produced by *k-means* in heat map and figure 5.6 illustrate individual graph of each cluster.

Interpretation

This section we discuss on the clustering results, three sets of graph indicate the activity patterns of call, receive and data connection are placed at the end of this section. First, we interpret the call activity; figure 5.7 illustrate the clustering of call activity by K-means. We repeat the algorithm to classify it into 4, 5 and 6 clusters and the results demonstrate that the 5 clusters seems to be a decent one. Surprisingly, nearly 60% of people in Massachusetts have very low average call in the weekdays. We further check with the raw data and found out that 15% of population only uses a mobile phone for receiving call and text. Apart from that, 25% of the population has as average call only once a day. The similar results also take place in the case of receiving call and data connection. Since the 60% of the population is large,

we continue investigate this cluster to find out whether there is any difference characteristic in this low activity group. Figure 5.8 reveal the secondary clustering results. The finding yield that more than 50% of this group has very low average call at all time and 12% of population has a possibility to make a call during the day, 11% of the population has peak call in late afternoon and 7% of this group has more call from late afternoon to evening.

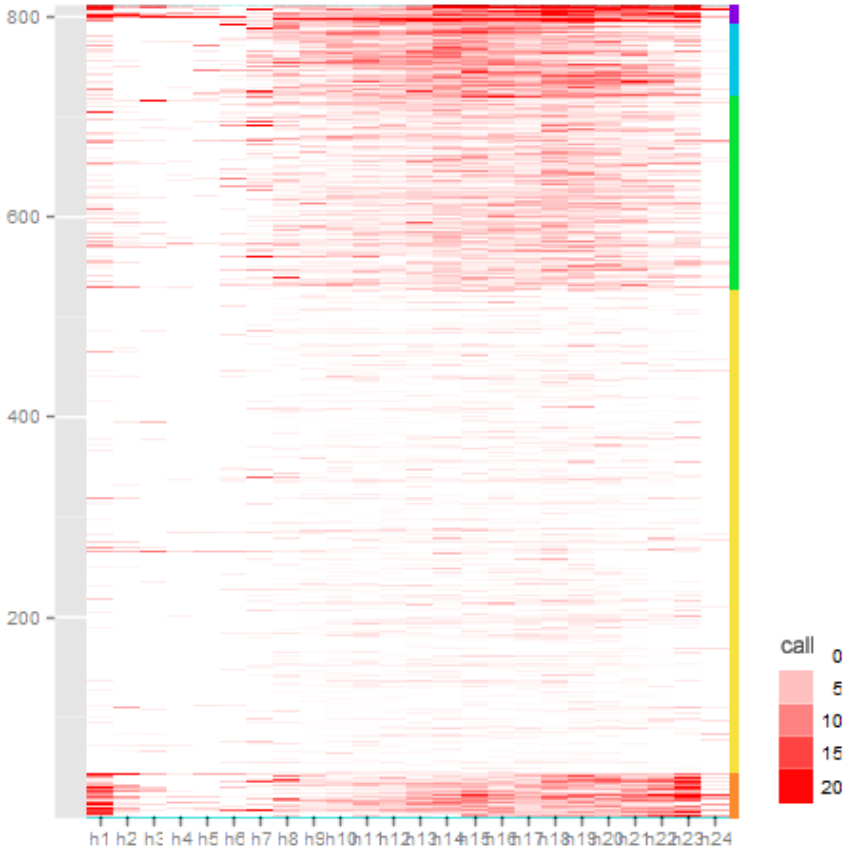


Figure 5.5 The data mining results after clustering using K-Means method

We are focusing on the *calli*ng patterns as it directly describe the user behavior rather than receiving call and data connection in order to generate the user model for the simulation system.

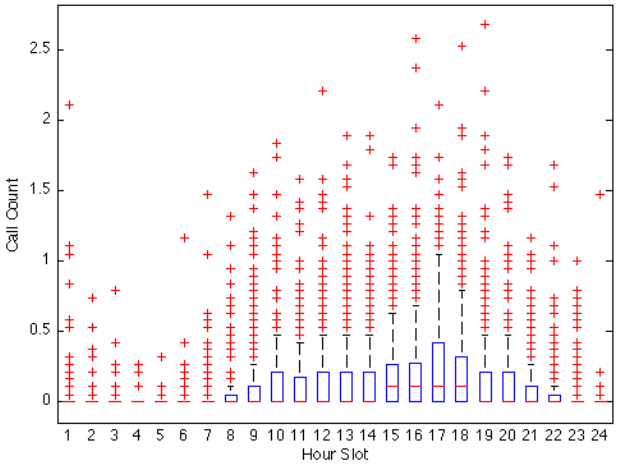


Figure 5.6 (a): 59% of the samples have very less call at all time.

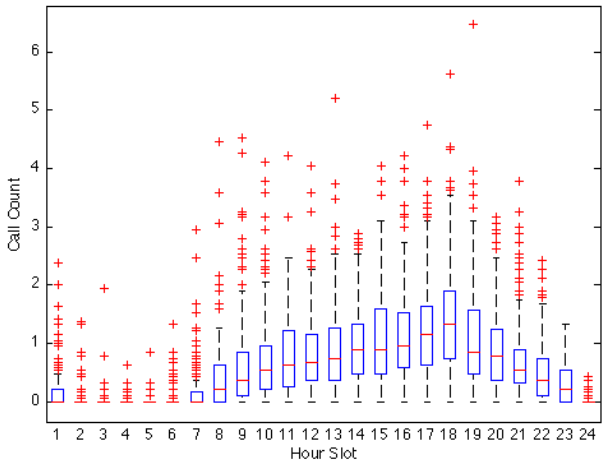


Figure 5.6 (b): 24% of the samples have average one call during the day

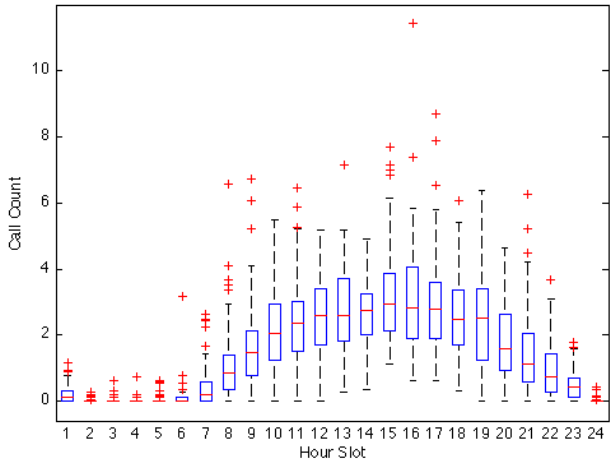


Figure 5.6 (c): 12% of the samples have peak call in the afternoon

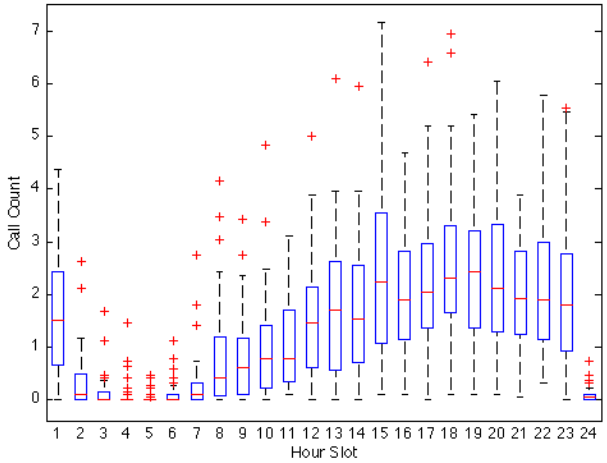


Figure 5.6 (d): 3% of the samples have peak call in the evening

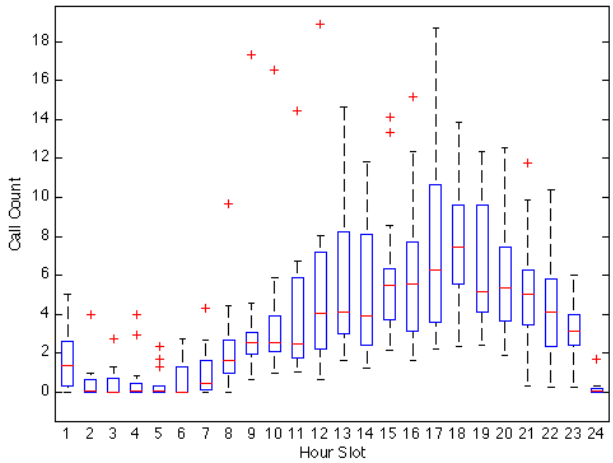


Figure 5.6 (e): 2% of the samples create very high call activity during the day

Figure 5.6 Characteristic of 5 calling patterns in Massachusetts, most of people tends to have less call at all time.

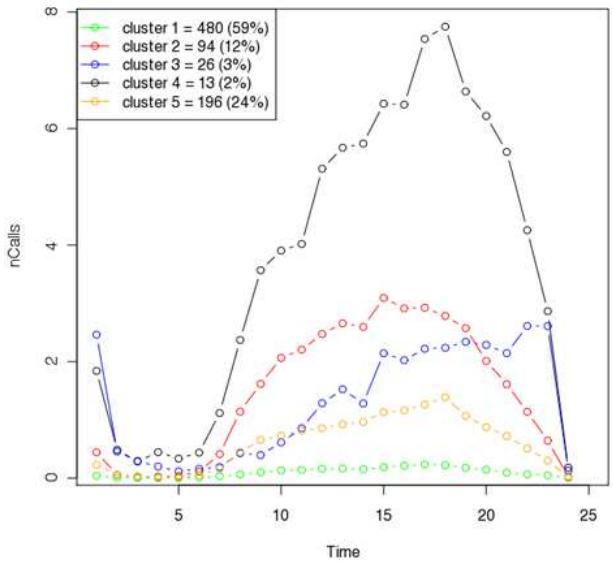


Figure 5.7 Calling patterns with 5 clusters

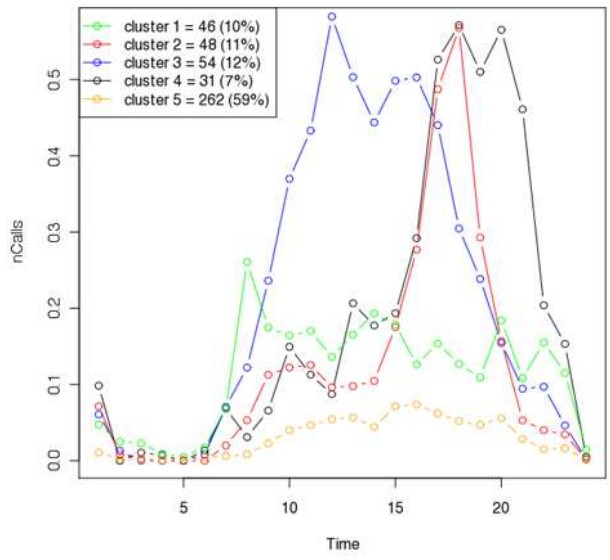


Figure 5.8 Subsequence profile patterns of the main cluster

Time	c1	c2	c3	c4	c5
1	0.041	0.444	2.4631	1.8408	0.2255
2	0.0075	0.0555	0.4581	0.4862	0.0564
3	0.0054	0.0249	0.2962	0.2869	0.0315
4	0.0028	0.0282	0.2004	0.4462	0.015
5	0.0021	0.0362	0.1169	0.3369	0.0131
6	0.0101	0.1139	0.1623	0.4377	0.0498
7	0.0338	0.4133	0.1912	1.1177	0.1658
8	0.0664	1.1429	0.4277	2.3715	0.4496
9	0.1022	1.6195	0.3962	3.5685	0.6561
10	0.1352	2.064	0.6146	3.9062	0.73
11	0.1408	2.206	0.8581	4.0192	0.8158
12	0.1621	2.4746	1.2869	5.3108	0.8597
13	0.1674	2.6587	1.5246	5.6715	0.9244
14	0.1536	2.596	1.2831	5.7423	0.9663
15	0.1892	3.0933	2.1435	6.4269	1.1323
16	0.214	2.9147	2.0246	6.4115	1.1621
17	0.2374	2.9252	2.2212	7.5377	1.2629
18	0.2237	2.7832	2.2373	7.7492	1.3907
19	0.1776	2.5744	2.34	6.6362	1.0685
20	0.1456	2.0115	2.2865	6.2154	0.8763
21	0.095	1.6124	2.1462	5.5985	0.7227
22	0.0657	1.1388	2.6127	4.2554	0.5081
23	0.05	0.6419	2.6096	2.8662	0.3011
24	0.006	0.0268	0.13	0.1823	0.0174

Table 5.1 A *k-means clustering* was applied to the subset samples with k= 5 using the data mining software *weka*

K-Means

Scheme: weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: subset_nordist_1000_avg
Instances: 810
Attributes: 24
Number of iterations: 14
Within cluster sum of squared errors: 65.20075258546107
Missing values globally replaced with mean/mode

Clustered Instances

0 480 (59%) 1 94 (12%) 2 26 (3%) 3 13 (2%) 4 196 (24%)

5.3 The SIM Mobility

In order to test the assumption on measuring the call traffic as a determinant factor to predict actual population numbers, we have developed "SIM Mobility" a 2-dimensional spatial simulation system to simulate movement and virtual call activity of the entire city. The Sim Mobility take the particles or swarm simulation approach which each particle is considered as an individual entity and the interactions between it and the other particles are individually modeled according to their physical movement and their calling patterns. We start testing our assumption by limit the simulation space with in Suffolk County. The simulated trajectories were derived from the generalizing mobile phone location point data of nearly one million records in central Metro-Boston area. (figure 5.9)

Figure 5.9 The simulated trajectories deriving methods. Dots referred to the mobile phone activities of the sample user where the green lines connected consecutive points. The red square defined as stop area of this user.

Please note that there is a limitation on lack of continuity of mobility traces due to the fact that the location is estimated from mobile phone data only when connection with a cellular network is made through either voice, text, or data communication, which constricts us to a smaller number of users that can be analyzed. Besides the user models which perform the calling functions are defined as;

Let {x1, x2, x3,..., xN} be a set of unclassified agents, {y1, y2, y3,..., yN} denote a set of classified agents, and fi ( ) represent the population distribution function of class i where i = 1,2,3,4,5 and N is the total number of agents. The unclassified agents are designated to each class according to the population distribution, that is

$$y_k = f_i(x_k), \qquad (1)$$

Where x_k and y_k are unclassified and classified agent k, respectively.

For each population-based classified agent, the call volume per day is to be assigned corresponding to call distribution function derived from the actual call volume, i.e.

$$g_i = q_i(t) \pm \sigma(t),$$

Where $g_i(t)$ is the call volume per day during t time slot, $q_i(t)$ is the call distribution function, □ denotes the standard deviation of call volume of time slot t, and t = 1,2,3,...,24 represents hourly time slot. The sign ± alternates according to Bernoulli distribution of p = 0.5, such that half of agents are assigned with □ greater than the expected value (mean) and the other half is assigned to □ greater than the expected value.

The mobile phone usage activities are generated by replicating calling functions from random given call patterns. Figure 5.10 show the flow diagram of SIM Mobility system, demonstrate how virtual call traffics were simulated base on the simulated trajectories and user models.

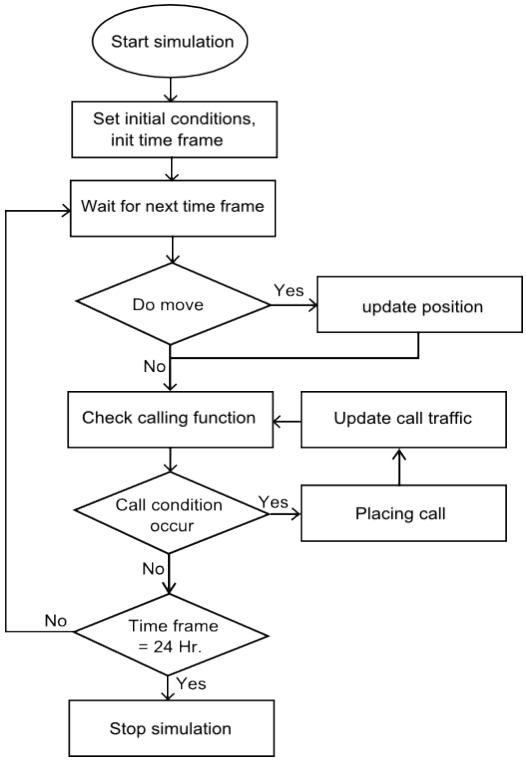


Figure 5.10 SIM Mobility flow diagrams

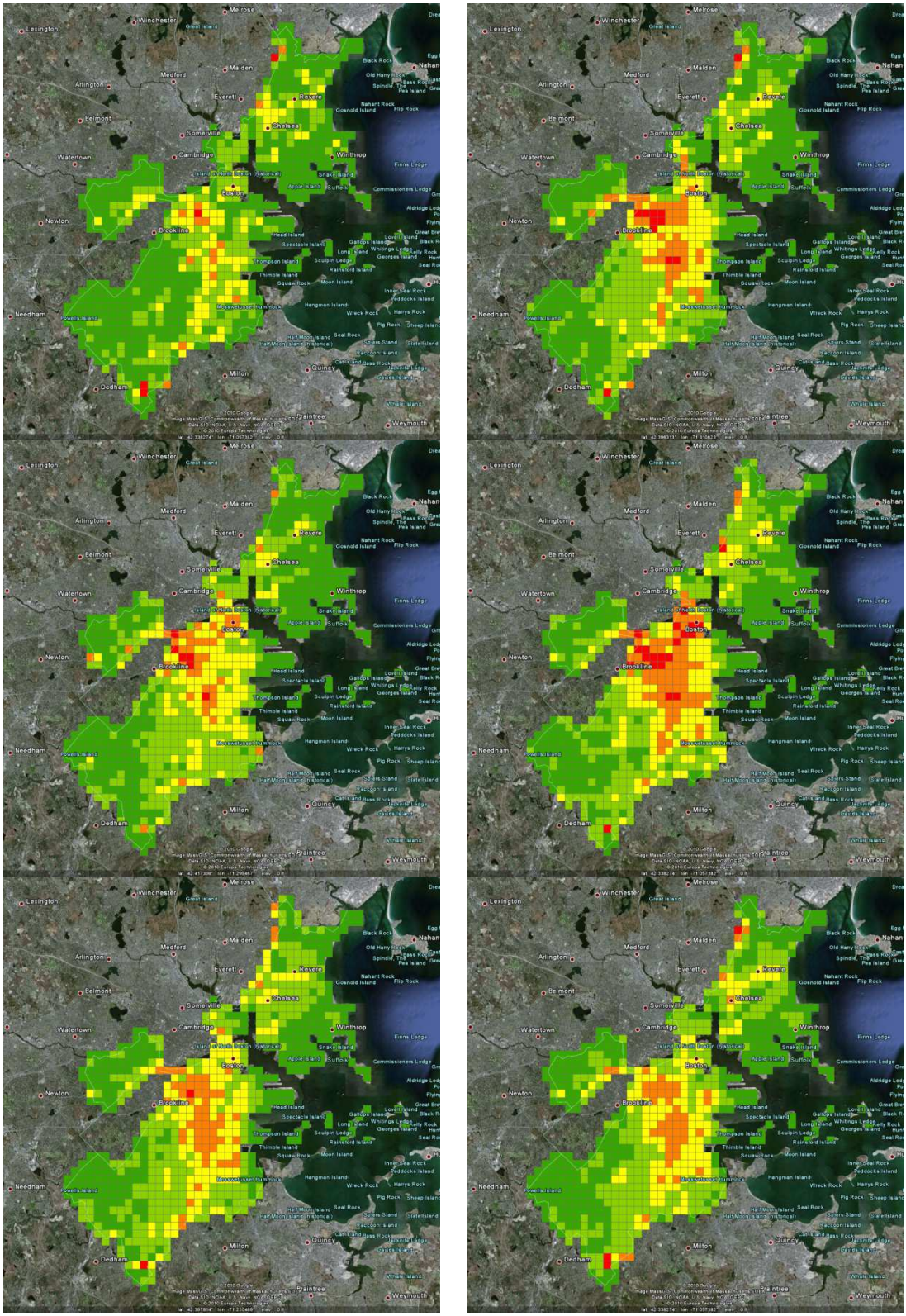


Figure 5.11 Capture of simulated call activities in Metro-Boston (Suffolk County) at 4:00, 8:00, 12:00, 16:00, 20:00 and 24:00 on normal weekday

5.4 Results and discussions

The simulation results provide detail measurements of presence population and number of calls at 1 hour temporal resolution and 500 meters spatial resolution. Figure 5.12 visualize the simulated results of center Boston at 10 am.

We question how much confident the call activity can be correlated with the existing number of people in each coverage at all time. We then simply repeating run the simulation replicate the normal working day in one month and calculated the mean call volume of each grid cell. In addition, to evaluate the degree of consistency, the correlation coefficient was computed to estimate how good the call volume fit to the presence population. Figure 5.13 show calculated R2 in an hour time slot of a complete weekday from Monday to Friday. The result explains that call activity from mobile phone show a positive correlation during the day. More frequent the mobile phone information occurrence a higher correlation we can get.

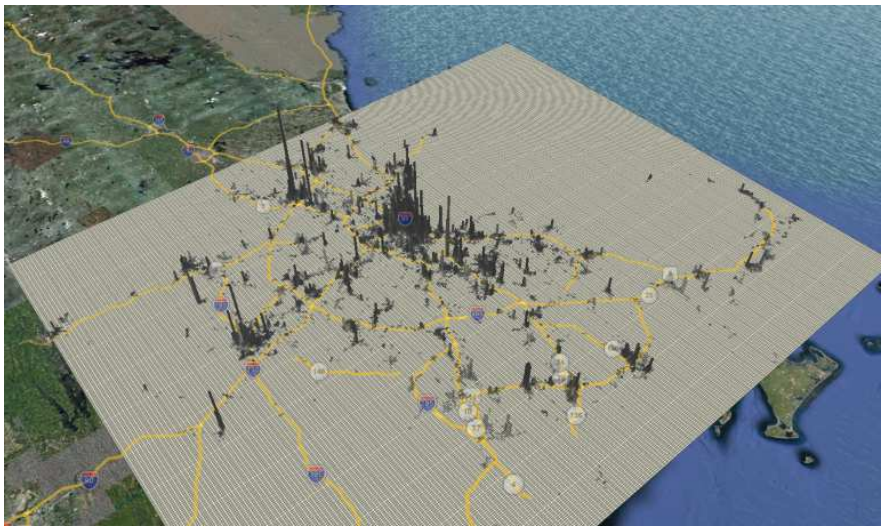


Figure 5.12 Illustrate aggregate call activities of part of greater Boston in 500 by 500 meter grid cell at 10:00 am

A graph show a high correlation during the day with two distinctive peak in the morning and in the evening while the correlation drop significantly at night due to the less mobile phone usage. As previous extensive research had confirmed that

mobile phone call volume has unique characteristic in each specific land use, we further investigate whether different type of land use affect to the consistency between number of call and actual population in the area. MassGIS 2005 land use data was utilized to test our assumption; the land use data was converted into 500 meters grid in order to compare with the simulation results. (figure 5.14)

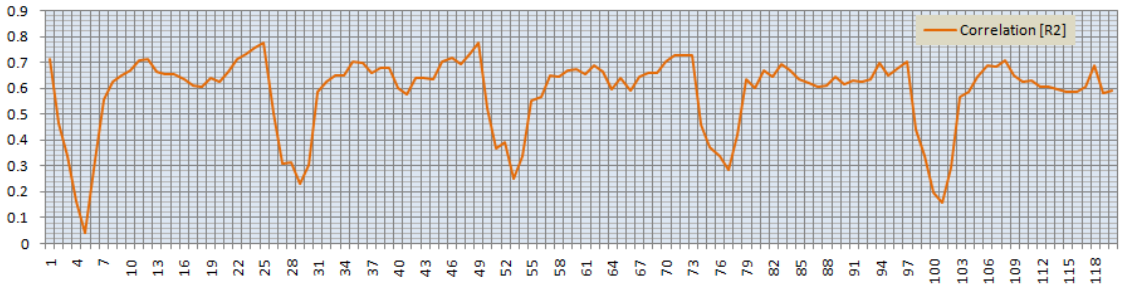


Figure 5.13 Correlation of actual population and call activities in one week period from Monday to Friday

Land Use Code	Land Use Description	Detailed Definition
1	Cropland	Generally tilled land used to grow row crops. Boundaries follow the shape of the fields and include associated buildings (e.g., barns). This category also includes turf farms that grow sod.
2	Pasture	Fields and associated facilities (barns and other outbuildings) used for animal grazing and for the growing of grasses for hay.
3	Forest	Areas where tree canopy covers at least 50% of the land. Both coniferous and deciduous forests belong to this class.
4	Non-Forested Wetland	DEP Wetlands (1:12,000) WETCODEs 4, 7, 8, 12, 23, 18, 20, and 21.
5	Mining	Includes sand and gravel pits, mines and quarries. The boundaries extend to the edges of the site's activities, including on-site machinery, parking lots, roads and buildings.
6	Open Land	Vacant land, idle agriculture, rock outcrops, and barren areas. Vacant land is not maintained for any evident purpose and it does not support large plant growth.
7	Participation Recreation	Facilities used by the public for active recreation. Includes ball fields, tennis courts, basketball courts, athletic tracks, ski areas, playgrounds, and bike paths plus associated parking lots. Primary and secondary school recreational facilities are in this category, but university stadiums and arenas are considered Spectator Recreation. Recreation facilities not open to the public such as those belonging to private residences are mostly labeled with the associated residential land use class not participation recreation. However, some private facilities may also be mapped.
8	Spectator Recreation	University and professional stadiums designed for spectators as well as zoos, amusement parks, drive-in theaters, fairgrounds, race tracks and associated facilities and parking lots.

9	Water-Based Recreation	Swimming pools, water parks, developed freshwater and saltwater sandy beach areas and associated parking lots. Also included are scenic areas overlooking lakes or other water bodies, which may or may not include access to the water (such as a boat launch). Water-based recreation facilities related to universities are in this class. Private pools owned by individual residences are usually included in the Residential category. Marinas are separated into code 29.
10	Multi-Family Residential	Duplexes (usually with two front doors, two entrance pathways, and sometimes two driveways), apartment buildings, condominium complexes, including buildings and maintained lawns. Note: This category was difficult to assess via photo interpretation, particularly in highly urban areas.
11	High Density Residential	Housing on smaller than 1/4 acre lots. See notes below for details on Residential interpretation.
12	Medium Density Residential	Housing on 1/4 - 1/2 acre lots. See notes below for details on Residential interpretation.
13	Low Density Residential	Housing on 1/2 - 1 acre lots. See notes below for details on Residential interpretation.
14	Saltwater Wetland	DEP Wetlands (1:12,000) WETCODEs 11 and 27.
15	Commercial	Malls, shopping centers and larger strip commercial areas, plus neighborhood stores and medical offices (not hospitals). Lawn and garden centers that do not produce or grow the product are also considered commercial.
16	Industrial	Light and heavy industry, including buildings, equipment and parking areas.
17	Transitional	Open areas in the process of being developed from one land use to another (if the future land use is at all uncertain). Formerly identified as "Urban Open".
18	Transportation	Airports (including landing strips, hangars, parking areas and related facilities), railroads and rail stations, and divided highways (related facilities would include rest areas, highway maintenance areas, storage areas, and on/off ramps). Also includes docks, warehouses, and related land-based storage facilities, and terminal freight and storage facilities. Roads and bridges less than 200 feet in width that are the center of two differing land use classes will have the land use classes meet at the center line of the road (i.e., these roads/bridges themselves will not be separated into this class).
19	Waste Disposal	Landfills, dumps, and water and sewage treatment facilities such as pump houses, and associated parking lots. Capped landfills that have been converted to other uses are coded with their present land use.
20	Water	DEP Wetlands (1:12,000) WETCODEs 9 and 22.
23	Cranberry bog	Both active and recently inactive cranberry bogs and the sandy areas adjacent to the bogs that are used in the growing process. Impervious features associated with cranberry bogs such as parking lots and machinery are included. Modified from DEP Wetlands (1:12,000) WETCODE 5.
24	Powerline/Utility	Powerline and other maintained public utility corridors and associated facilities, including power plants and their parking areas.
25	Saltwater Sandy Beach	DEP Wetlands (1:12,000) WETCODEs 1, 2, 3, 6, 10, 13, 17 and 19
26	Golf Course	Includes the greenways, sand traps, water bodies within the course, associated buildings and parking lots. Large forest patches within the course greater than 1 acre are classified as Forest (class 3). Does not include driving ranges or miniature golf courses.

29	Marina	Include parking lots and associated facilities but not docks (in class 18)
31	Urban Public/Institutional	Lands comprising schools, churches, colleges, hospitals, museums, prisons, town halls or court houses, police and fire stations, including parking lots, dormitories, and university housing. Also may include public open green spaces like town commons.
34	Cemetery	Includes the gravestones, monuments, parking lots, road networks and associated buildings.
35	Orchard	Fruit farms and associated facilities.
36	Nursery	Greenhouses and associated buildings as well as any surrounding maintained lawn. Christmas tree (small conifer) farms are also classified as Nurseries.
37	Forested Wetland	DEP Wetlands (1:12,000) WETCODEs 14, 15, 16, 24, 25 and 26.
38	Very Low Density Residential	Housing on > 1 acre lots and very remote, rural housing. See notes below for details on Residential interpretation.
39	Junkyard	Includes the storage of car, metal, machinery and other debris as well as associated buildings as a business.
40	Brushland/Successional	Predominantly (> 25%) shrub cover, and some immature trees not large or dense enough to be classified as forest. It also includes areas that are more permanently shrubby, such as heath areas, wild blueberries or mountain laurel.

Table 5.2 Land use code definitions

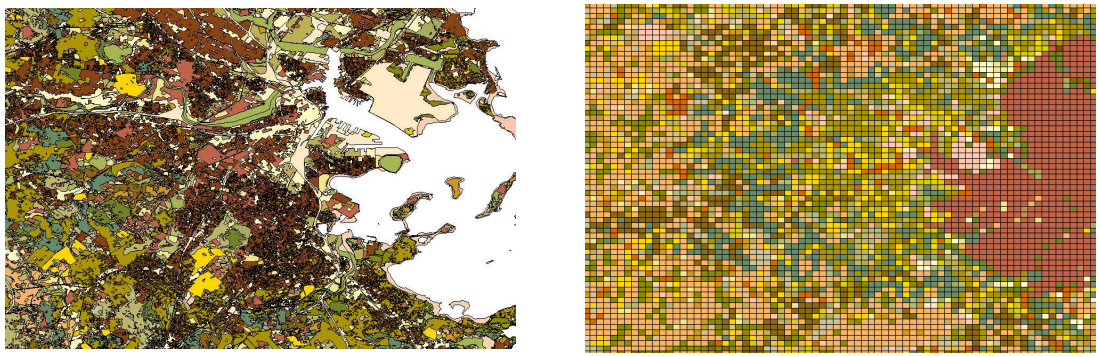


Figure 5.14 Area interpolations of MassGIS land use data, left show the original land use polygon and right show the process data in 500 meters grid

We selected 6 major land use types out of 33 (Table 5.2) in our study area including Multi-Family Residential, High Density Residential, Urban Public/Institutional, Transportation, Commercial and Industrial. Individual land use type show its distinctive characteristic where five of selected samples suggest highly and positively correlate among call activity and presence population during the day except for the industrial area. (Figure 5.15)

Transportation type which covers the major road and public transport giving a

highest correlation with distinguish peak in morning and evening commuting time. Commercial and Urban Public/Institutional are having fairly good translation of calling traffic to population density in the daytime but giving the lowest correlation at night. Besides, we found average correlation in High density residential area during the day; however the nighttime correlations are significantly better than the other land use types. Furthermore, the Industrial areas in Suffolk are having slightly difference interpretation where night time correlations are higher than in the daytime.

In conclusion, this chapter has investigated the opportunity to use mobile phone call detail records (CDRs) to evaluate how mobile phone activity correlate to the actual presence population. The simulation results showed that higher mobile phone activities yield a better correlation between calling records and presence population. Cross-comparison analysis was undertaken against specific land use to describe how similar and different correlations are formed in each land use types. These findings indicate that land use and land cover play a role in shaping the dynamics population estimation within the mobile context. Among the new idea that emerged in this period, this study describes an important scenario in which mobile devices with user-driven contents are fully changing the way of urban mobility research and how to estimate the population dynamics.

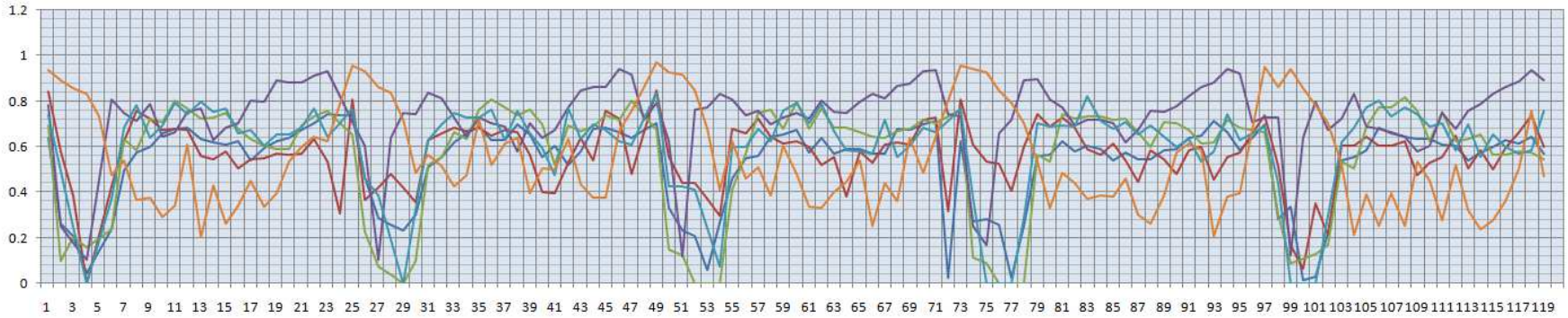


Figure 5.15 Correlation of actual population and call activities by land use type

Chapter 6

PREDICTION OF THE CURRENT: Integration of Mobile Phone calling records and Person Trip Data

6.1 Introduction

The existing crowd flow simulation models fall into two major categories: those based on space syntax theories, and those based on the vehicular micro-simulation models incorporating origin-destination matrix. (Sharma, 2007) The space-syntax based models have a unique strength in their ability to analyze the spatial geometry quickly and generating valuable information about the configuration of space. However it suffers from the exclusion of dynamic effects of flow that are driven by the needs of people to go from A to B. The origin-destination based models are fundamentally micro-simulation models that provide for these dynamic effects. These models provide the valuable information on interaction of agents and densities as a function of time. However these models tend to be complex, both in pre-processing and execution, and reliable information on space effectiveness is only available at the end of a number of simulation runs.

The present work we are focusing on the techniques to effectively predict the probability of human flow from analyzing coarse-grained mobile phone traces. This is different from ordinary GPS or roadside sensor for origin-destination flow estimation in traffic engineering. Unlike GPS and other sensor tracking data, mobile phone log or Call Detail Records (CDRs) data provide uninformed report on locations and time. However, the most beneficial in utilizing CDRs data is a large-scale of analysis can be performed without install additional devices. The individual locations could be determined when the phones are actively activated, namely when the people place or receive calls, sending text messages or using the internet.

In this chapter, we describe the design methodology used to produce statistical mobility model from CDRs data, incorporated with Person Trip Survey (PTS) data to perform an estimation of day time population distribution in the study area. We first analyze calling patterns and mobility patterns of the individual from the mobile phone usage survey to create user behavior models. More detail discussions are placed in the third section of this chapter. The simulated trajectories are generated from PTS data through pre-process of place geocoding and trip mode associated route matching. A brief background of Person Trip and data preparation methods will be discussed in the next section.

Section 4 will provide a methodology framework on how to virtually create mobile phone traffic or Call Detail Records (CDRs) by using simulation. The last part of this chapter describes the mythologies in using these mobile phone call records to estimate dynamic population.

6.2 Tokyo Person Trip Survey: Tokyo as seen by movement of people

The Person Trip Survey (PTS) data has been used in several researches on transportation and planning. It provides valuable information on demographic (age, sex and occupations), travel mode and trip purpose. The main use of PTS data in this research is to generate substantial daily trajectories for the simulation system and for the validation of estimated population results.

The Tokyo Person Trip Survey was conducted in the Tokyo Metropolitan region including Tokyo, Kanagawa, Saitama and Chiba Prefectures and the southern region of Ibaraki Prefecture. This is the arena for the lives and activities of the 34 million people. It is also the world's biggest metropolitan area, playing a pivotal role in the politics, economy and culture of Japan. The movement of people in the Tokyo Metropolitan Region extends widely beyond the metropolitan region, and the provision of suitable transportation to support the diverse activities of these people must be studied as a wide-ranging issue that looks at the whole of the metropolitan region. The Tokyo Metropolitan Region Transportation Planning Commission was established for this purpose. It has been engaged in mutual cooperation and coordination between the prefectures within the Tokyo Metropolitan Region and cities and relevant organizations specified by government ordinance. The Commission's activities began with the first large-scale

transportation fact-finding survey (Person Trip Survey - PTS) in 1968, followed by a second survey in 1978, a third in 1988 and a fourth in 1998, in order to study metropolitan transportation policy from a wide-ranging, comprehensive viewpoint. The results of the surveys have also been widely used as basic data for solving transportation issues in various localities. The 5th Person Trip Survey was conducted in October 2008, and an outline of the results has now been summarized.

Figure 6.1 The coverage area of Person Trip Survey in greater Tokyo region

Survey forms were collected from about 880,000 people who cooperated in this survey. Households are randomly selected and people who are 5 years or more are surveyed. This represents a sample size of approximately 3%. In principle, we are considering PTS data as the moving sampling of a person in simulated environment. This has greatly benefited since the original survey purpose are described as follow;

- "Person trip" refers to the movement of a person
- The Person Trip Survey aims to capture all movements on one day by investigating "what kind of person" moved "when", "for what purpose", "from where", "to where", and "by what modes of transportation".
- The objective of the survey is to obtain an overall grasp and analysis of the

realities of transportation in the metropolitan region on the basis of these survey data, and to explore suitable urban transportation systems for the metropolitan region.

In addition, PTS is basically a person trip diary where the respondents enter the following information for all the trips made in the day.

1. Trip start time
2. Trip origin
3. Modes of travel (e.g. train, bus, private car)
4. Trip purpose (e.g. work, shopping, education)
5. Trip destination
6. Trip end time.

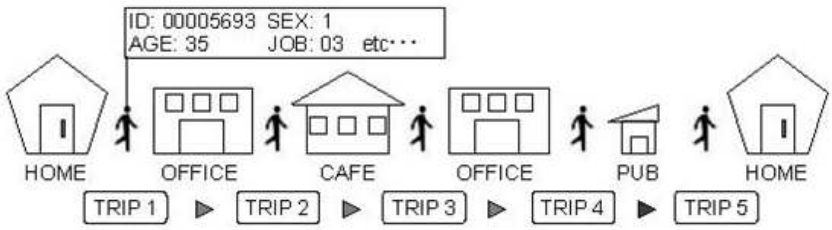


Figure 6.1(b) Person Trip Survey describes overall trips of a person in one day

PTS data also provides detail information of commuting mode including automobile, bus, two-wheeled vehicle, walking and train.

6.2.1 Preprocessing and data structure

Person Trip data were pre-processed before being GIS-mapped. Field address data was first geocode to geographic position. Commuting modes and trip time were used to determine stop points and further for route matching algorithms. Furthermore, in order to work with the SIM Mobility, all traces were interpolated into points with 1 minute interval. These interpolated points represent the existence of people in certain time and hence help confirm the number of populations in the validation process. After that the interpolated points are ported into PostgreSQL, a sophisticated open-source spatial database, which help manage this large

spatio-temporal dataset. Table 6.1 show the preprocess trajectories in tabular format.

pid	pdate	longitude	latitude	sex	age	work	magfac	datum
1020	10:25:00	139.9096	35.72799	1	10	4	37	6
1020	10:26:00	139.9093	35.72864	1	10	4	37	6
1020	10:27:00	139.9081	35.72910	1	10	4	37	6
1020	10:28:00	139.9080	35.72990	1	10	4	37	6
1020	10:29:00	139.9083	35.73092	1	10	4	37	6
1020	10:30:00	139.9087	35.73112	1	10	4	37	6
1020	10:30:00	139.9087	35.73112	1	10	4	37	97

Table 6.1 the preprocess trajectories in tabular format.

6.3 The mobile phone usage survey

In order to simulate individual call activities, the user behavior models have to be developed. Our interest in conducting this survey is to study the behavior of the phone user and therefore create calling patterns of each user groups by using data mining techniques to identify characteristics or patterns found in our survey samples. This section demonstrates the results of this survey to evaluate how people generate their call and data traffics in a normal working day. The results are fairly consistent to the major calling patterns we had extracted from the real CDRs data in Boston in chapter five. We conducted online mobile phone usage surveys which reported quantitative results of calls, messages and internet uses of 1,063 people in the same region that Person Trip surveys were accomplished. The survey samples were selected consistent with the PTS data based on sample's ages and occupations. A summary of the results of the surveys is explained in this section. The original online questionnaire can be obtained from the appendix I and the details analysis can be obtained from the appendix II. The following explain a summary of the general results of the surveys in bulleted lists.

- 100% of the samples have mobile phone.
- 8% of the samples have more than 1 phone.
- 6 % of the samples use smart phone. (ie. iPhone、BlackBerry)

- 66% of the samples use fixed rate service package.
- 45 % of the samples spent 2,000-5,000 yen a month for mobile phone bill and 38% spent 5,000-10,000 yen, 7%spent more than 10,000 and 10% spent less than 2,000.
- Nearly 60% of the samples never turn the mobile phone off, 20% turn it off in specific event such as in the theater, concert hall, etc. And 10% turn it off during sleep, 7% during work and 5% turn the mobile phone off when they commuting on the train.
- 62% of the samples have mobile phone with them most of the time and 30% stay close with mobile phone 24 hours.
- Surprisingly the most frequent use of the mobile phone is email/text service at 37%, following with internet service at 18%. It was reported that only 13% have most frequent use their mobile for calling purpose. Also, they use their mobile to browse internet and email more often than making a call.
- 35% of the samples making a call only few times a week, 26% making call once or twice a day and 25% of the samples are calling with their mobile phone few times a month. Only 12% of the samples reported that they are making call 3-6 times a day and 4% make more frequent call than 6 times a day.
- 63% of the samples have average talk about 1-5 minutes, 19% had reported of 5-15 minutes and 10% have average talk less than 1 min.
- At the work place or school, it was reported that 21% of the samples use their mobile phone at the office only few times a month and 19% never use their mobile phone at work. 15% use their mobile phone at work few times a week, 13% use once or twice a day and 6% use 3-6 times a day.
- During commuting to work or school, 31% report that they never talk on the phone and 19% talk over commuting few times a month. 13% talk few times a week and 9% talk 1-2 time(s) a day. Please note that 25% of the samples need not to travel on a daily basis.
- At home, 35% of the samples make a call with their mobile phone only few times a month, 31% few times a week, 21% once or twice a day and 6% placing call 3-6 times a day. Also 5% they never talking on the phone with their mobile phone at home.
- It was reported that people more frequent talk on the phone at home, when walking on the street, at the shop or department store and at the office.
- Email and internet services have the same usage trend, people seem to use more frequent at home, during commuting with public transportation and when they are at the station.

- 60% of the samples tend to use mobile phone when they are stuck on the road or public transportation.
- 34% of the samples tend to use mobile phone during participating the events, nevertheless 55% tend to not use at this point.
- 45% of the samples tend to use mobile phone in case of accident or disaster occurred and 28% tend to not use.

In conclusion, above summaries help confirm the potential use of analyzing mobile phone CDR as an indicator to describe the population in dense urban environment. The last part of this questionnaire particularly focuses on quantitative use of mobile phone services in a normal working day. We built user behavior models from statistic count of call and email record. Figure 6.2 - 6.6 show the overall results of this survey categorized by area, age, sex, occupation and activity types (call, email and internet) of 1,063 samples in 24 hours.

Since different age groups behave differently and male and female also behave differently. In chapter 4, we demonstrated the use case of using mobile phone CDR as location-aware devices to analyze the activities of each work area's profile. The results explain the correlation of work area's profile and similarity in daily activity patterns.

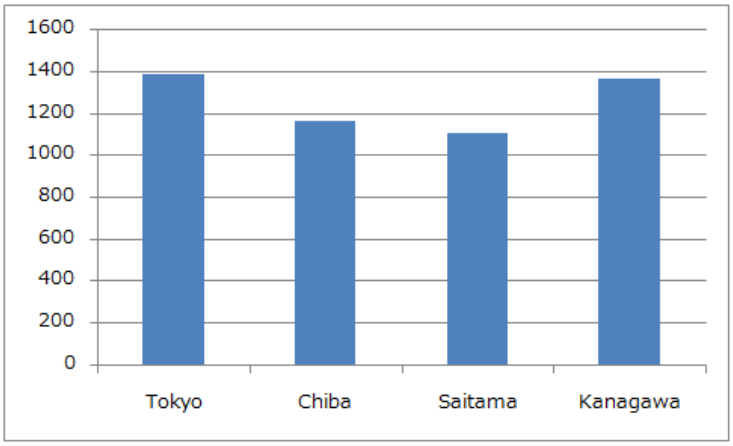


Figure 6.2 Accumulated call activity by area in 24 hours normalized by number of observations

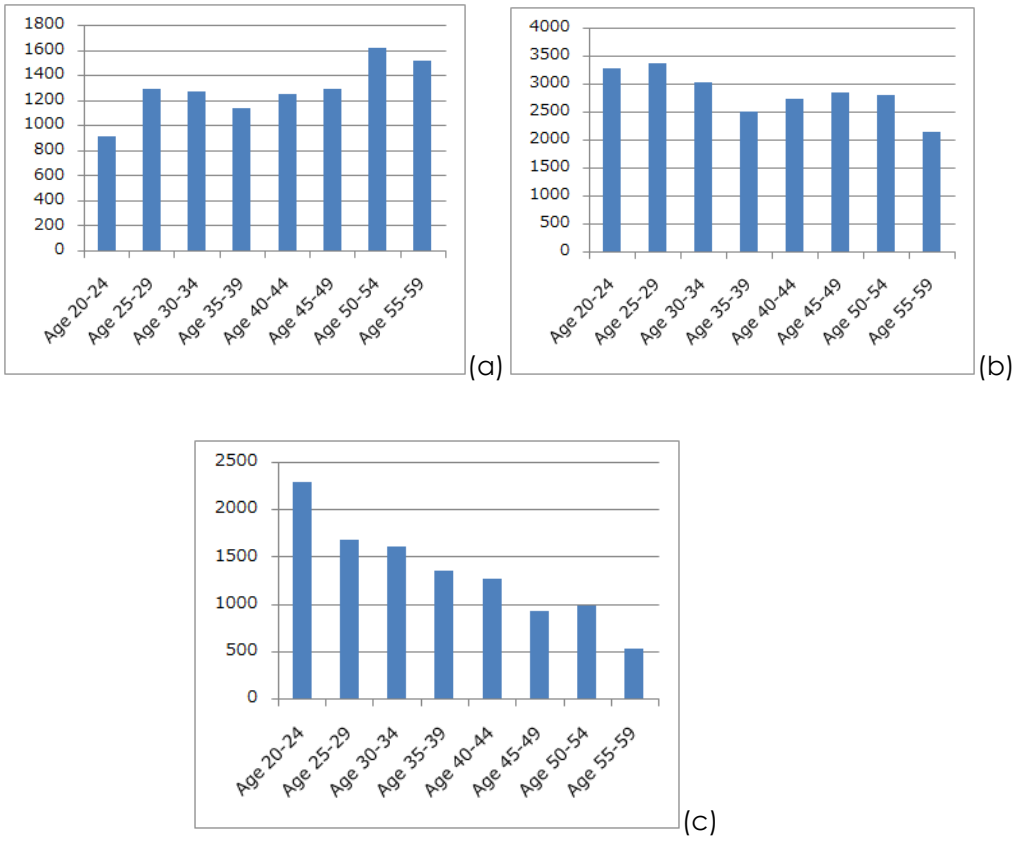
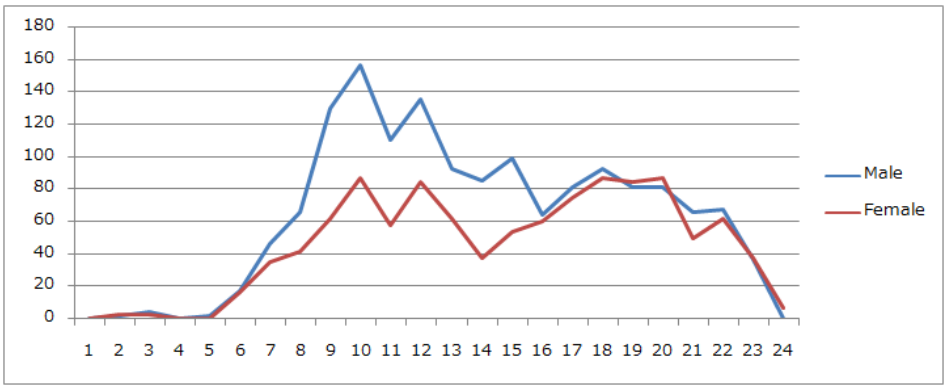
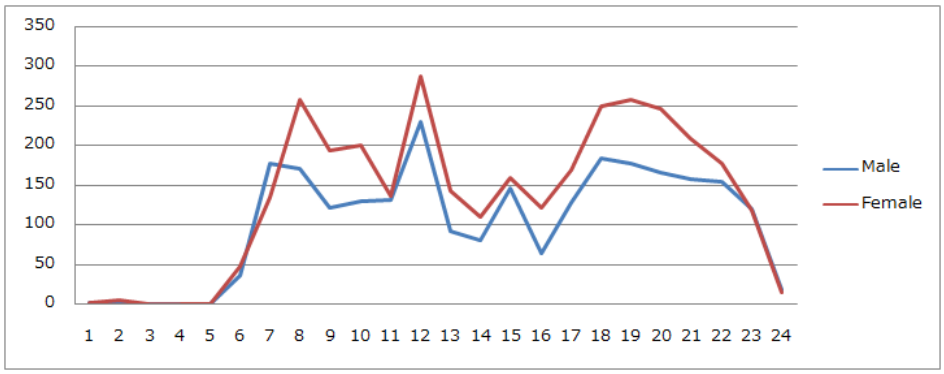


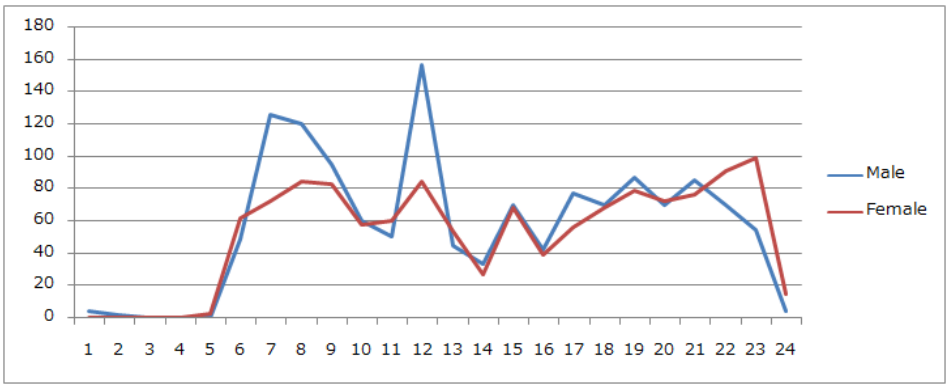
Figure 6.3 Accumulated call(a), email(b) and internet(c) activity by age groups



(a)



(b)



(c)

Figure 6.4 Accumulated call(a), email(b) and internet(c) activity categorized by gender

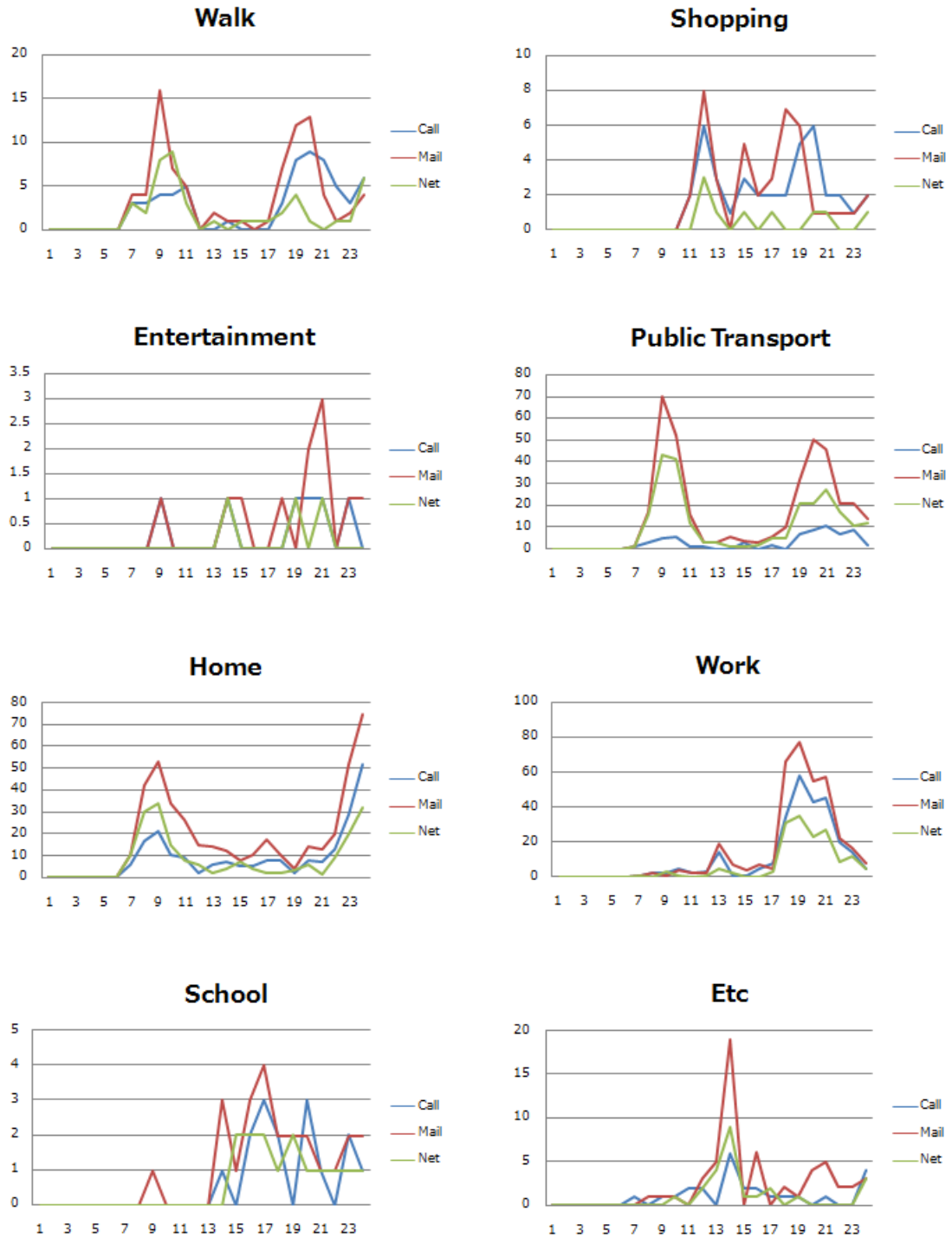


Figure 6.5 Accumulated call, email and internet activity categorized by activity types

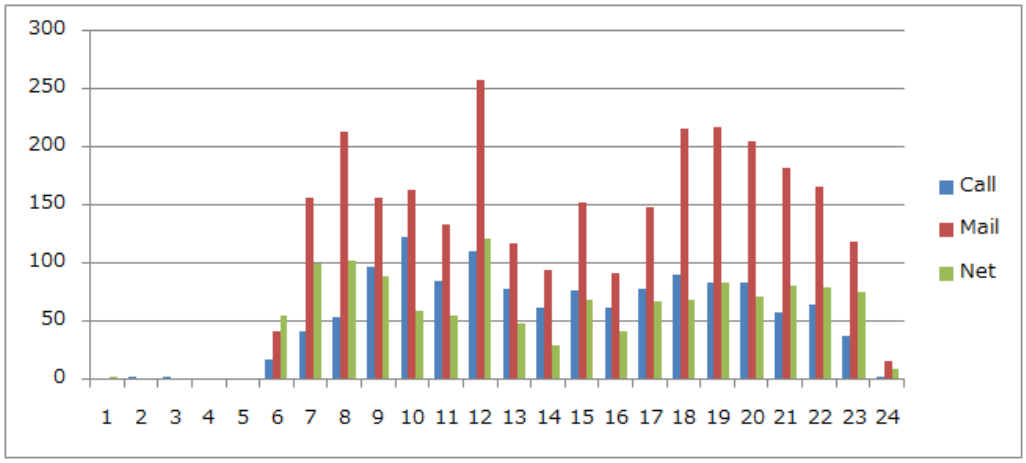


Figure 6.6 Overall accumulated mobile phone activities during one business day

It is clear that the overall mobile phone activities are significant low. From our examination, one issue is that some samples had not replied this question and we judged those to be outliers. Note that, in the experiment we discard simulated internet activity since less than 50% of the samples replied in this part. Figure 6.3 and 6.4 show the accumulated call and email activities of each career groups in 24 hours.

We further calculate an average call and email activities by occupation groups. We consider that occupation groups offer better results in classifying user behavior patterns more than age and sex. The average call and email patterns of each slot t (hour) in any occupation types. Note that the can be computed as

$$\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)$$

Where t = 1,2,3,…,24

and n is the total number of samples exclude the outliers

Table 6.2 present calling patterns of each specific occupation groups in 24 hours.

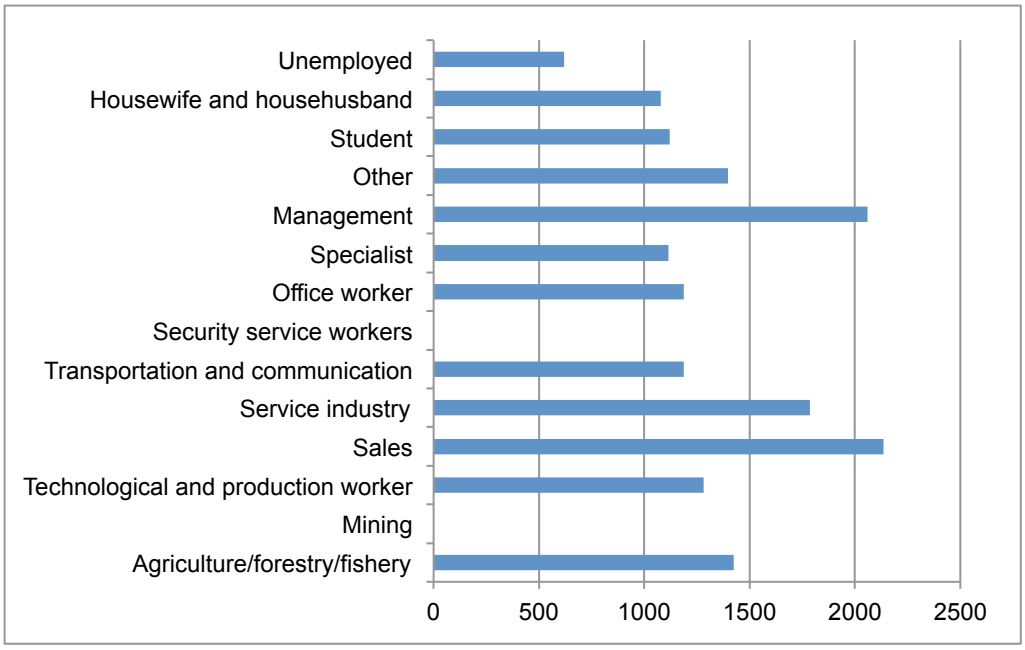


Figure 6.7 Accumulated call activities of each career groups in 24 hours

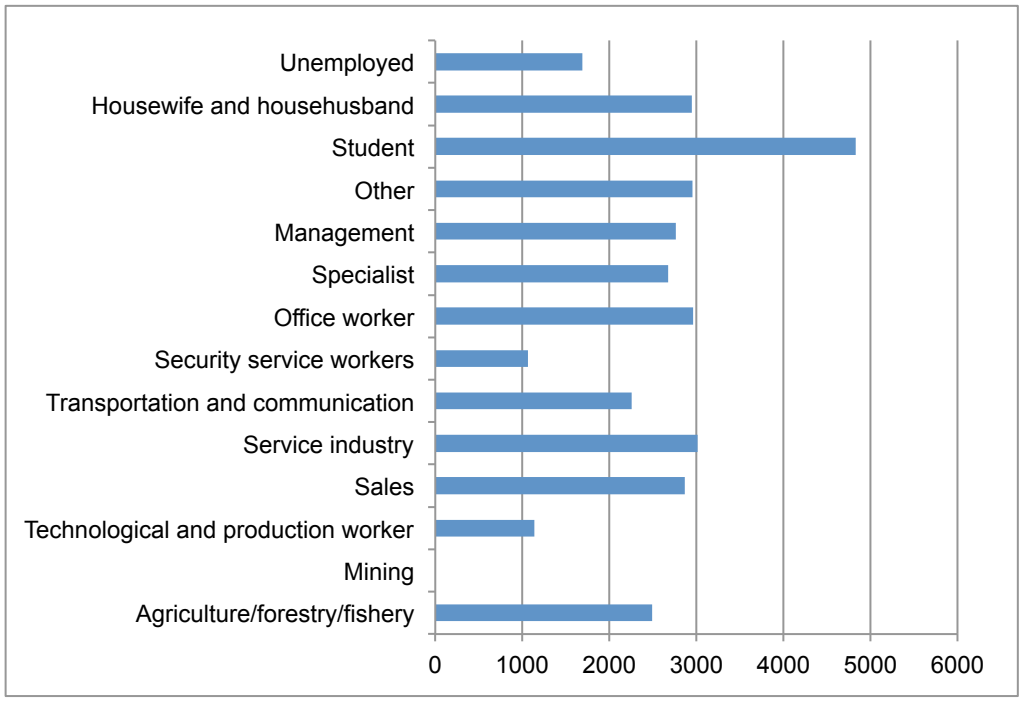


Figure 6.8 Accumulated email activities of each career groups in 24 hours

T	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	3	3	9	12	10	3	3	5	5	3	6	5	1	5	1	1	2	0
2	0	1	3	7	8	6	4	5	5	4	4	3	2	4	4	2	2	1	0
3	1	1	0	0	7	1	3	4	3	6	4	3	3	4	1	0	0	0	0
4	1	1	3	3	4	2	3	4	1	1	3	2	3	4	3	3	3	1	0
5	0	1	1	4	3	4	5	2	1	2	1	2	4	3	3	1	3	2	0
6	1	1	1	7	9	4	6	2	1	4	4	3	3	2	3	3	1	1	0
7	0	2	1	4	5	1	3	3	3	4	2	3	3	3	3	1	2	2	0
8	1	3	2	3	0	0	10	1	2	2	1	1	1	2	2	6	3	5	0
9	1	1	1	3	4	3	2	2	2	3	1	3	3	3	2	2	2	0	0
10	2	0	5	3	0	0	0	3	2	4	2	3	2	1	0	0	1	2	2

Table 6.2 Average Call activity of each occupation type

Type 1: Agriculture 2: Mining 3: Skill worker 4: Sales 5: Service industry 6: Transportation 7:Protective service

8: Business 9: specialist 10: Manager 11: Other 12: Student 13: Housewife 14: unemployed

Type	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22	h23	h24

0	0	0	0	0	0	9	9	0	0	0	9	0	0	0	0	9	9	9	0	0	9	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	12	0	0	6	0	6	0	0	0	3	15	12	0	6	3	3	0	0	0
0	0	0	0	0	5	10	16	7	5	5	12	3	3	7	2	5	13	12	4	5	6	4	2	0
1	0	0	0	0	2	5	9	8	9	9	10	6	5	9	2	5	11	7	7	8	8	5	1	1
0	0	0	0	0	3	4	8	3	1	1	21	4	4	1	4	1	8	10	8	5	4	5	0	0
0	0	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2	8	15	5	3	4	13	4	4	5	3	6	12	10	10	8	9	6	1	0
0	0	0	0	0	2	7	12	8	6	6	18	3	3	4	3	6	7	7	9	8	8	6	1	0
0	0	0	0	0	2	13	10	5	7	6	10	4	2	7	3	7	7	11	12	6	6	4	0	0
0	0	0	0	0	2	10	11	5	9	6	11	4	4	6	10	13	11	9	9	8	6	10	0	0
0	0	0	0	0	2	13	18	11	10	19	15	8	5	8	8	15	11	13	17	21	13	11	1	0
0	0	0	0	0	2	5	7	12	9	4	7	7	6	5	6	4	8	12	6	7	5	4	0	0
0	0	0	0	0	1	6	8	7	11	7	4	8	7	4	4	12	7	6	4	7	6	6	0	0

Table 6.3 Average Mail activity of each occupation type

Type 1: Agriculture 2: Mining 3: Skill worker 4: Sales 5: Service industry 6: Transportation 7:Protective service 8: Business 9: specialist 10: Manager 11: Other 12: Student 13: Housewife 14: unemployed

Note: Pattern probabilities above are maximized by a distribution over 20 days

6.4 Methodology

6.4.1 Data preprocessing and system configuration

The preprocess PTS data contain 1.26 x 10^9 rows in total. Since the data is large, we subset and use the most active area in central Tokyo to test our hypothesis and proposed algorithms. The main Tokyo 23 wards are considered as a simulated environment. (Figure 6.9) We random select 30,000 samples out of 880,000 as observations and use the total PTS data as a ground truth value. The assumption of this experiment is that if the subset samples of 30,000 people could predict the dynamic day time population of 880,000 people then it is feasible to use the same method to predict the real world population from the original PTS data. We finally evaluate how close the sample estimate matches the true value in the whole PTS population.

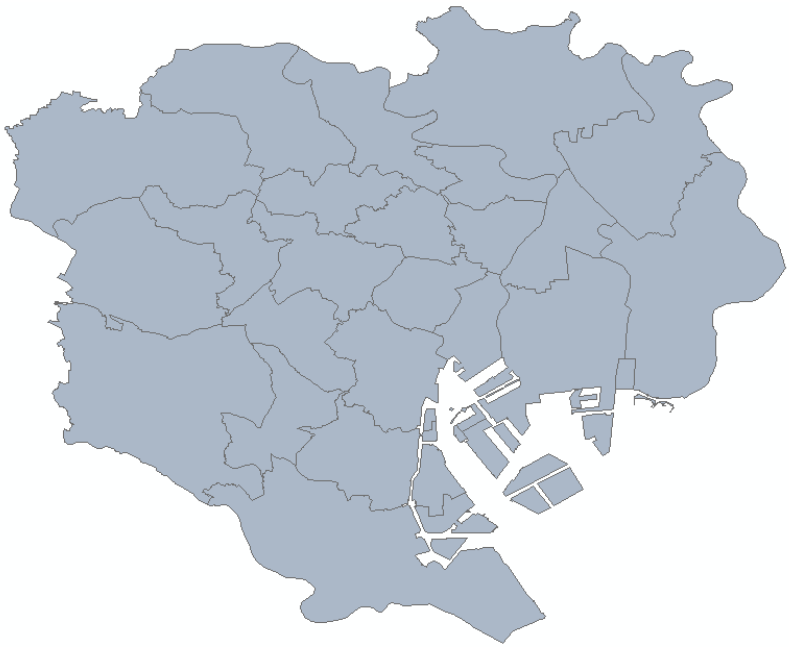


Figure 6.9 Area of study cropped by Tokyo 23 wards.

We have developed the Grid-based mobile phone simulation called "SIM Mobility". The system utilize the preprocess PTS data as simulated trajectories of the individual agents and apply the calling patterns derived from section 6.3 to generate mobile phone activities as user behavior models. Figure 6.10 illustrate the grid based simulation scenario with initial configuration of 250 meters grid as spatial resolution and use 15 minutes interval to accumulate the mobile phone activities as

temporal resolution.

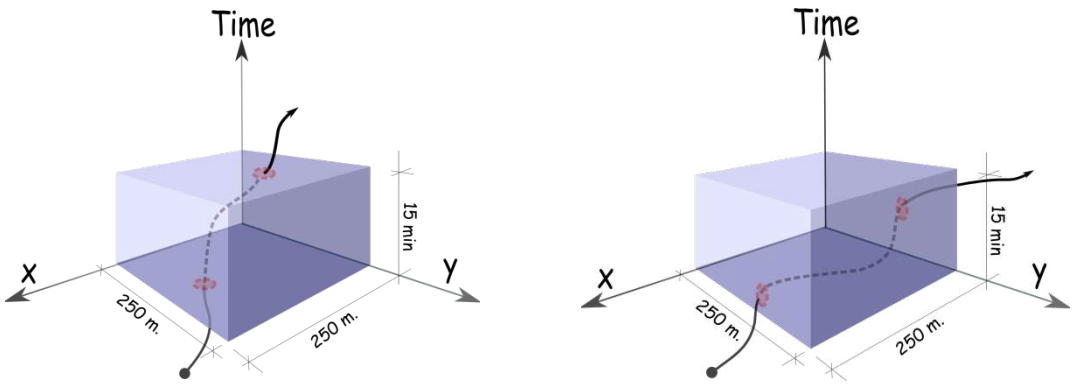


Figure 6.10 Grid dimension of the simulation system

Apart from simulated grid dimension, we configure using 1 kilometer uniform grid to represent mobile phone base station in the simulated environment. In Urban and sub urban the mobile based station can vary from few hundred meters to few kilometers, figure 6.11 giving an example, illustrates the uniform mobile base station and its coverage.

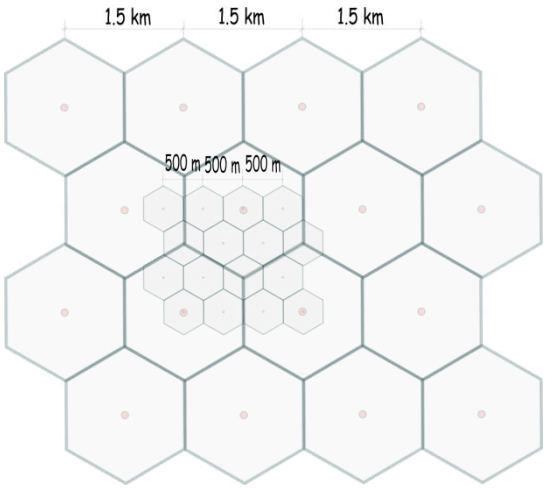


Figure 6.11 Uniform mobile base station, in denser area the coverage may reduce to few hundred meters however the larger grid can be applied for suburb area

First, we use pre-process PTS trajectories and behavior model to simulate mobile phone activities, followed by post-processing methods as illustrated in Figure 6.12.

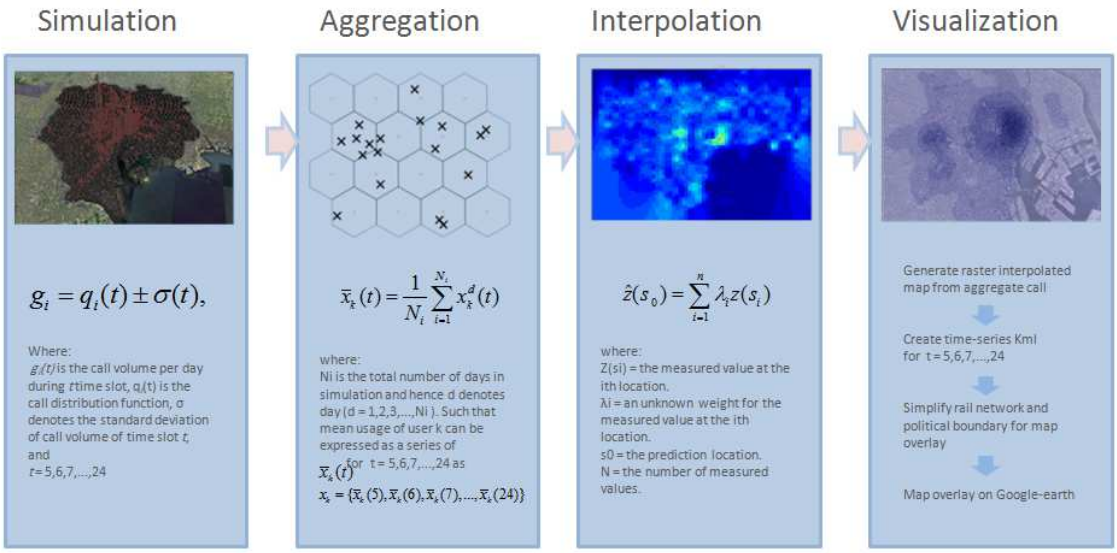


Figure 6.12 Structure and procedural steps of simulation system

6.4.2 Simulation of call traffic

We use discrete event simulation in order to analyze the distribution of mobile phone activities. Sim Mobility assigns a unique numerical ID when each agent is created. The call behavior patterns are assigned to each individual agent based on the occupation attribute data. To generate the call activity, a uniform random probability is used to determine whether a call will be occurred regarding the present underlying distribution of each calling patterns. A uniform random number from 0 to 20, UR[0, 20], is first obtained. The call occurs if the result number is lying within the distribution function displayed in table 6.2 and 6.3. A generated call timestamp would be the solution of t for R(T, t) = UR[0, 60]. Besides, the calling probability is computed based on spatial location and event; specifically no call will be made when the transportation mode is motorbike or bicycle. The same approach is repeated for every agent for 24 hours. In this experiment, we continued simulate the weekday activities for one month, which is 20 times running task had been carried out in our preliminary test.

Let {x1, x2, x3,..., x_N} be a set of PT agents, {y1, y2, y3,..., y_N} denote a set of classified agents(base on occupation), and $f_i(.)$ represent the population distribution function of occupation class i where i = 1,2,3,…,15 and N is the total number of agents.

$$y_k = f_i(x_k),$$

Where x_k and y_k are PT and classified agent k, respectively.

$$g_i(t) = q_i(t) \pm \sigma(t),$$

where gi(t) is the call volume per day during t time slot, qi(t) is the call distribution function, □ denotes the standard deviation of call volume of time slot t, and t = 5,6,7,...,24

Visualization of the simulated call information can play a big role in our understanding of urban mobility and human flow. Preliminary simulation results are parsed to interpolated module developed in chapter 3 to the grid interpolation images. Figure 6.13, 6.14, 6.15 display a capture of mobile phone usages in Tokyo's 23 words in one day.

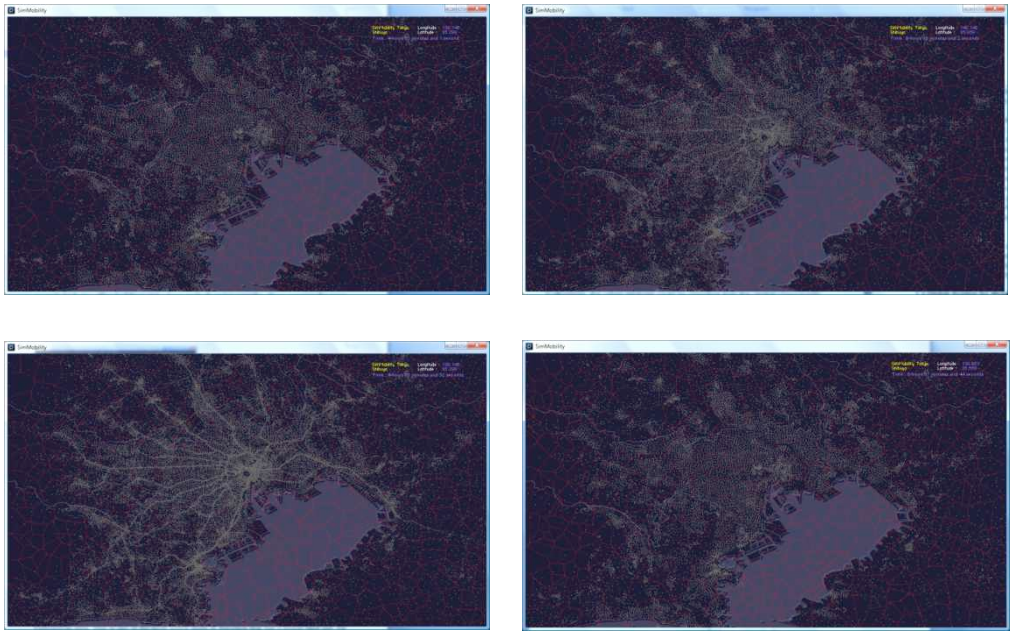


Figure 6.13 A SIM Mobility prototype system illustrates Tokyo activity at 05:00 am.(upper left), 12:00 am.(upper right), 6:00 pm.(lower left) and 12 pm.(lower right)

Figure 6.14 Illustrate call activities in simulated grid config at 250 x 250 meters

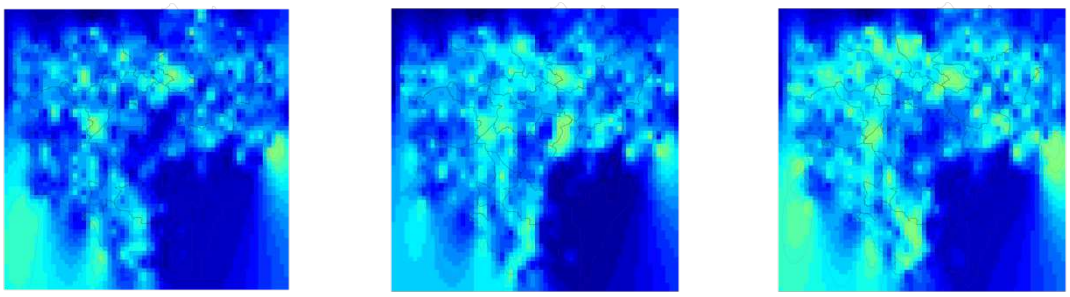


Figure 6.15 (a) Interpolation of mobile phone activities at morning rush hours (6-8 am.)

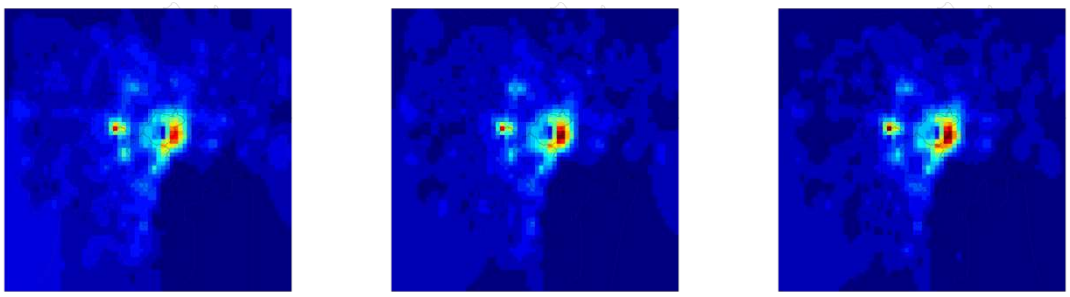


Figure 6.15 (b) Interpolation of mobile phone activities at morning office hours (10-12 am.)

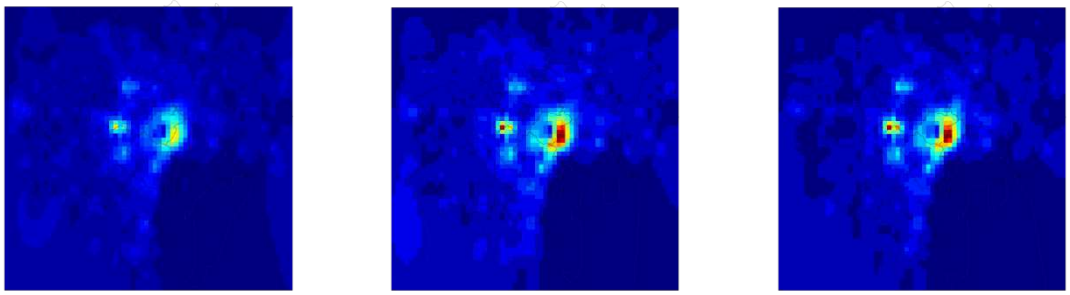


Figure 6.15 (c) Interpolation of mobile phone activities at afternoon office hours (1-3 pm.)

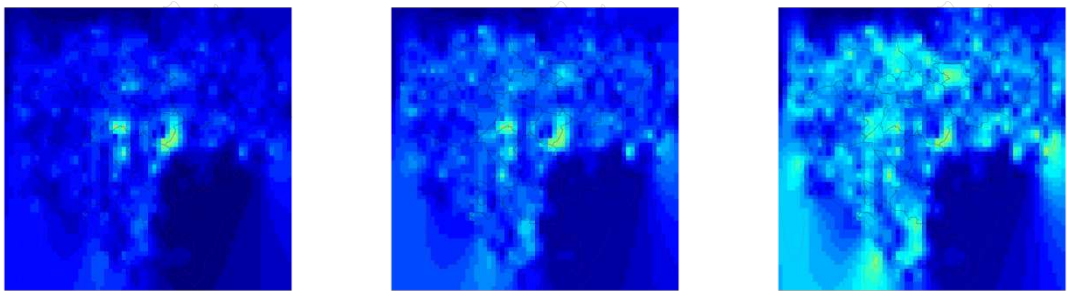


Figure 6.15 (d) Interpolation of mobile phone activities at evening commuting time (6-8 pm.)

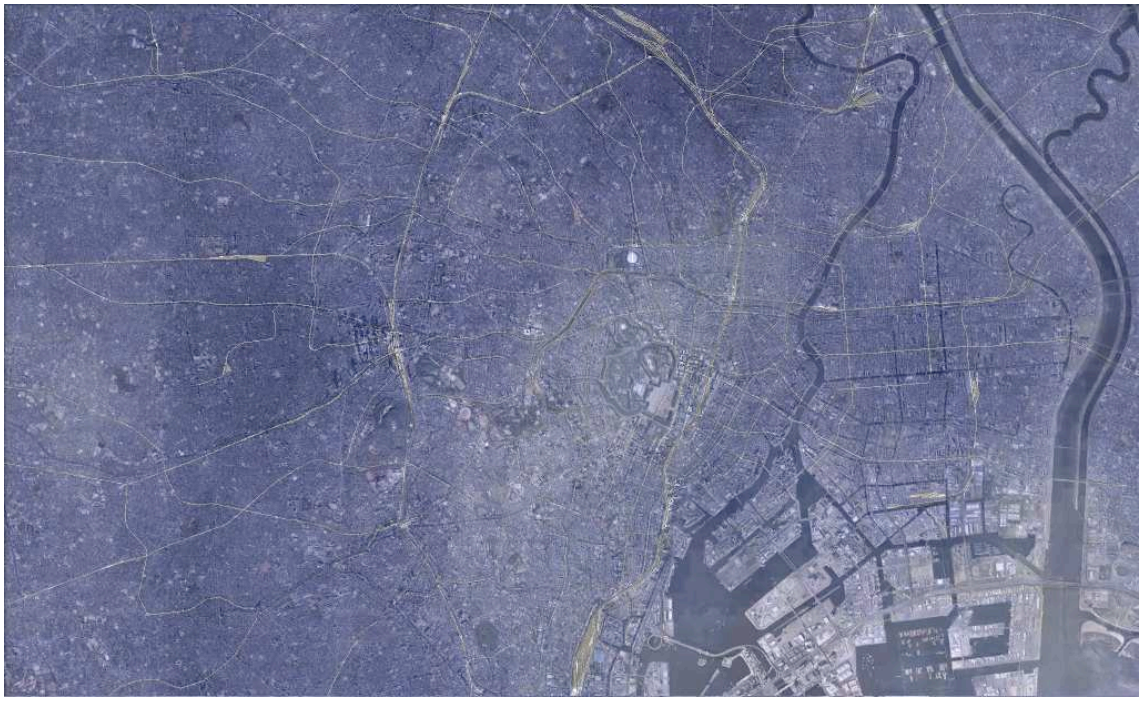


Figure 6.16 Visualization of mobile phone activities at morning commuting time overlay on map with train network

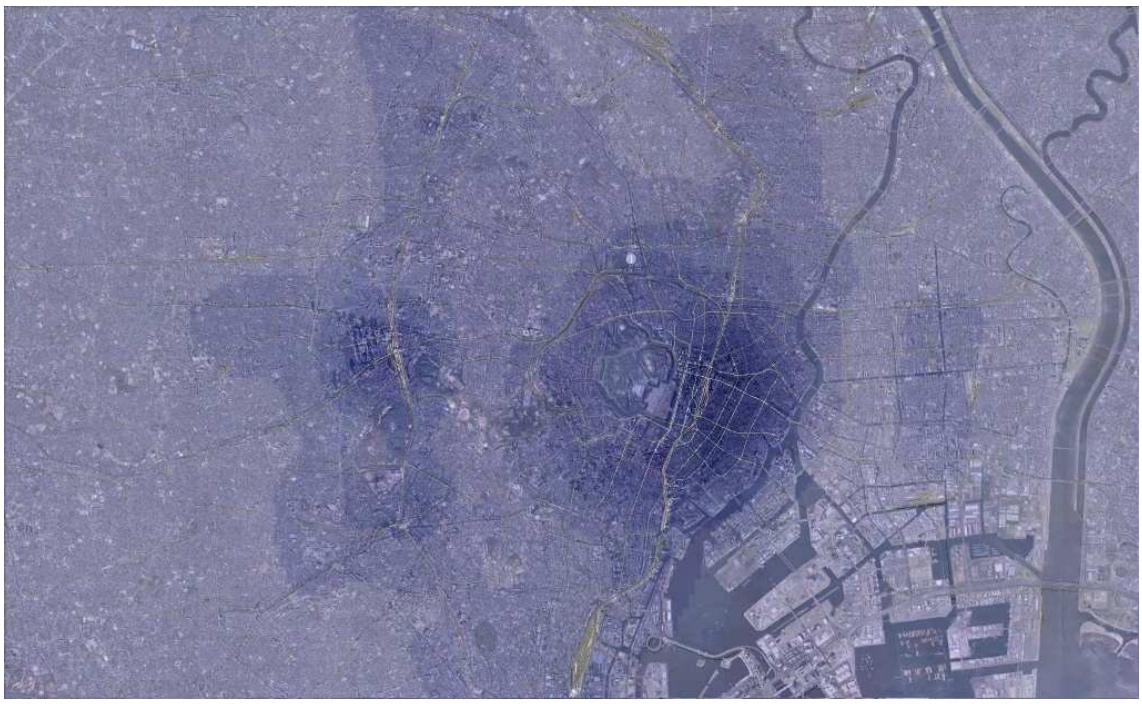


Figure 6.17 Visualization of mobile phone activities at 10 am. on normal business day overlay on map with train network

6.5 Time-dependent dynamic OD weight modification

The theoretical framework of the time-dependence based Origin destination (O-D) matrix analysis is discussed in this section. In general, O-D matrices are not observable unless the vast majority if people can be tracked on a continuous basis, which become reality with the increasing penetration of mobile phones. The fundamental approach is to utilize mobile phone traffic as real time flow indicator to modify magnification weight factor in order to reconstruct the population density in the area. In the previous section, we discussed how to simulate mobile phone activities from the pre-process PTS trajectories and call behavior model, the followed section will explain the post-processing methods, how to generate the O-D matrix from mobile phone CDRs and how to utilize generated O-D matrices for estimating population density.

The original Person Trip Survey (PTS) data come with a magnification factor which is a ratio used to reconstruct real number of population in the area. (Table 6.1) However, range of estimated error for the total population becomes large when person move cross the zone in the day time. Table 6.4 present the estimated population with PT magnification factor and the true value. Note that, the numbers of population are calculated from counting the person who exists in specific grid cell and in specific time frame. In the following case, we use Japan mesh level 3 (1 x 1 kilometer) grid cell with time interval of 15 minutes.

code	True Value	PT magnify
53393514	1188	2320
53393515	865	920
53393516	1437	1956
53393517	927	1665
53393518	2340	4109
53393519	832	1263
53393520	85	241
53393521	582	622

53393522	1080	1606
53393523	1809	3669
53393524	1749	3888

Table 6.4 Real population and estimated population at one square kilometer grid and 15 minutes time frame.

Where code is the 1x1 kilometer grid id from Japan mesh level 3, True value is the estimated population count from the entire 880,000 people of original PT data, PT magnify is the estimated population count from the subset 3 % of PT data and multiply by its magnification factor.

Since the mobile phone activities are related to the actions of people and thereby to the overall dynamics of cities, how they function and evolve over time, meaning that it represented stochastic mobility patterns of the entire population in the city. (In contrast to PTS data that only give mobility patterns of the samples). In this case, our assumption is that by combining the mobility patterns from the mobile phone, we could improve the accuracy of estimated population at the certain time.

In order to understand the flow of the city, the first question is to reconstruct the individual trace from the CDRs data. However, the frequency of simulated daily call activities is less we then simply demonstrate this methodology by applying it to estimate day time population. The optimal strategy for generating flow in this context is to evaluate home and work locations in order to create trip O-D matrix in the study area. A likely location of home place is estimated from a simple model with two criteria measurements. First is the frequency of call and another is the regularity of revisit to the observed places. In this experiment, we define the resolution of estimated home-work locations by using 1 x 1 kilometer grid and use time frame of 5 - 8 am. in the morning and 10 – 12 pm. at night to calculate home locations. The time between 10 am to 4 am is considered to calculate work locations. The computation is defined as;

Home location = $\max \sum_{i=1}^{n} H_{ij}(t) * n^{-1}$ (1)

Where:

H_{ij} = Number of call at base station location

t = Expected time at home (5,6,7,22,23)

n = Number of simulation days (20 days)

Work location = $\max \sum_{i=1}^{n} W_{ij}(t) * n^{-1}$ (2)

Where:

W_{ij} = Number of call at base station location

t = Expected time at work (10,11,12,13,14,15)

n = Number of Simulation days (20 days)

Figure 6.18 show home and work location from above methodology link with straight line connecting home-work nodes. In our framework, to evaluate dynamic population estimation, the Origin-Destination (O-D) matrices provide a crucial input as a city-wide mobility indicator. Since a dynamic O-D matrix describes the aggregate demand for trip interchange between specific zones of the network over a series of time intervals. We further aggregate the process home–work nodes from grid-level to larger geographic zone, for instance, Tokyo wards area in our case.

Figure 6.18 Home and work location of samples in Tokyo 23 wards

Consider a network composed of G nodes, H and W be the amount of origins and destinations from and to which samples are allocated and P be the amount of O-D pairs traversed by flows. E_h denote the number of samples departing from origin h to destination w during estimation interval $T \in \tau$ and contributing to the flows traversing links during count interval $\tau \in t$, where T is the study period that typically refers to the business hours period. Then, the dynamic O-D matrix estimation at an interval T can be expressed as:

The resulting O-D matrices are used as input to a Dynamic Time-dependent O-D matrix population estimator for mapping the estimated zone mobility into a set of path and link flows. The first set of origins and destinations defined based on the estimated home-work locations is illustrated by figure 6.18

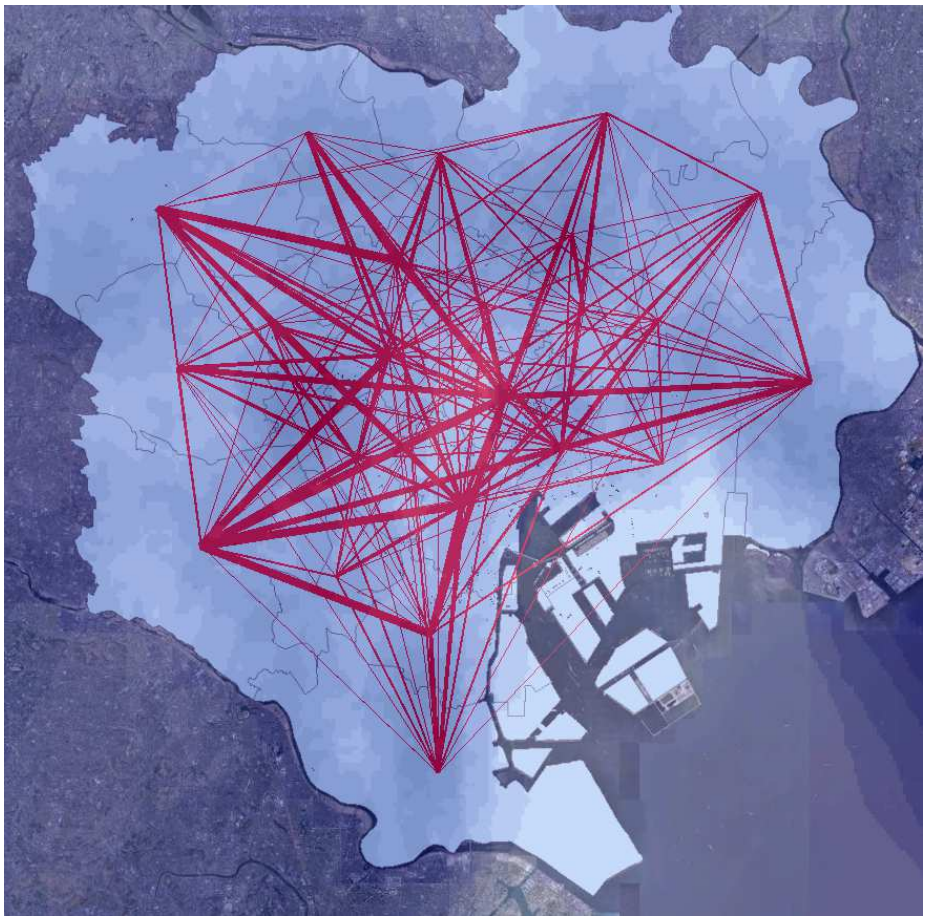


Figure 6.19 A day time O-D mapping graph of Tokyo 23 Words

The above O-D mapping graph is considered as time-dependent relationship of travel demand and link volumes between zones, as defined by the fractions of each extracted home-work pairs, which constitute the assignment matrix.

6.5.1 Population reconstruction

The fundamental concept of population reconstruction was explained by magnification weight factor in the Person Trip Survey. We therefore are proposing an approach to calibrate this method by using O-D Matrix Adjustment. The O-D matrix adjustment gives logic on the use of simplified rules to represent the movement and interaction between zones. Then modification of magnification weight is accomplished by weighted mean function. A sample O-D matrix that provided in Table 6.5 and 6.6 describe link flow from and to Chiyoda-ku, one of ward area in Tokyo.

h_jcode	w_jcode	outflow
13101	13113	2
13101	13109	2
13101	13108	2
13101	13101	20
13101	13103	4
13101	13106	1
13101	13104	1
13101	13102	2

Table 6.5 O-D matrices explain link flow from Chiyoda-ku

Where

h_jcode = home administrative area

W_jcode = work place administrative area

h_jcode	w_jcode	inflow
13104	13101	2
13105	13101	5
13106	13101	3
13107	13101	2
13108	13101	2
13109	13101	6
13110	13101	2
13111	13101	3
13112	13101	8
13113	13101	1
13114	13101	3
13115	13101	6
13116	13101	2
13117	13101	4
13118	13101	2
13119	13101	6

13120	13101	5
13121	13101	5
13122	13101	3
13123	13101	6

Table 6.6 O-D matrices explain link flow to Chiyoda-ku

Where
h_jcode = home administrative area
W_jcode = work place administrative area

6.5.2 O-D weighted modification function

The new approach using O-D weighted modification function is based on estimated weight of interchange population group in the area. This caused to reduce the estimated error by changing the weights of each individual in the specific zone through the existing mobility. To reduce complexity, we demonstrate this algorithm by using a single area; the input–output pairs are listed in table 6.5 and 6.6

From table 6.5, the specific zone (jcode -13101) is composed of 14 outflow links and 20 nodes still remain in this area. We first calculate weighted mean of each group in this zone. Note that, group refers to the occupation groups and weight refers to the magnification weight specify in Person Trip Survey data. The weighted mean can be computed by:

$$\bar{X} = \frac{\sum_{x=i}^{n} w_i \cdot x_i}{\sum_{x=i}^{n} w_i}$$

Where

w_i represent the bounds of the partial sample. In this case, bound refer to smaller zones used in computing home and work location, partial sample mean outflow numbers of specific occupation group in the same zone.

The process was repeated for every zone in the study area (23 wards). The results give values of gradient weighted mean of each occupation group by zone.

We next calculate the inflow weighted mean of destination zone, refer to table 6.6, we use the same formulation as for the case described in previous step, that is;

$$\bar{X} = \frac{\sum_{x=i}^{n} w_i \cdot x_i}{\sum_{x=i}^{n} w_i}$$

Where:

w_i represent the bounds of the partial sample. In this case, bound refer to smaller zones used in computing home and work location, partial sample mean inflow numbers of specific occupation group in the same zone.

For more informative, the formula can be rewrite as;

$$\bar{a} = \frac{w_1 a_1 + w_2 a_2 + w_3 a_3 + ... + w_n a_a}{w_1 + w_2 + w_3 + ... + w_n}$$

Where

w = inflow numbers from each of origin zones

a = weighted mean of an occupation group of origin zones

$\bar{a}$ = weighted mean of an occupation group of destination zone

To reconstruct population of each zone we use the result gradient weighted mean from the previous step. In general, we can proceed multiple gradient weighted mean by applying specific time-dependent O-D matrices. For instance, the same process could derive the gradient weighted mean at time t+1 if the O-D matrices at time t are identified. In this example, we apply the results weighted mean to estimate day time population since we merely process day time average OD matrix (work place). The population reconstruction can be computed as;

$$P = \sum_{j=1}^{k} \sum_{i=1}^{n} \bar{X}_j * x_i$$

Where

P = estimated population

k = number of groups in the area (occupation)

n = number of members in group

$\bar{X}_i$ = gradient weighted mean of group j

6.5.3 Experimental results and conclusions

The comparison between the estimated population distribution and the real distribution allows the assessment of the quality of this novel approach. The resulting population distribution matches relatively well with the real distribution. This section aims to present the findings of the experimental validation, and to address some limitations and future improvement of this method.

Figure 6.20 show a quick demonstration of regression graph visually depicting the relationship between the estimated population from the proposed method and real population numbers in 250 meters grid of 23 wards area. The results correlation coefficient showed that the estimated population had highly significant and positive correlation with the real population numbers.

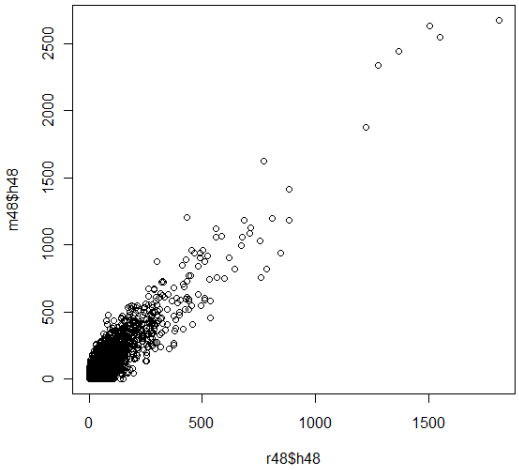


Figure 6.20 Plot of estimated population(y) and real population numbers (x).

In order to examine the accuracy of the model, the Root Mean Square Error (RMSE) had been used. Table 6.7 shows the R^2 and RMSE values at two representatives, one is from the estimated population using PTS magnification factor and the second is the estimated population using time-dependent O-D modification methods.

The estimated results of the O-D modification methods may be compared to those obtained using a previous PTS magnification factor. The correlation coefficient and root mean square error (RMSE) are computed. Obviously enough, the proposed method show significant improve of prediction for all area in Tokyo 23 wards. The correlation coefficient returned slight increases, while decreasing the RMSE mean of all zone.

As O-D matrices explain the overall mobility trends, our assumption is that more we can defined the O-D link between zones more the accuracy will be increased. We further investigate the results by increasing the population size in simulated space to 10^5 samples or about 12% of the total PTS population.

Again the results show a clear trend emerges in this test sample size, it was giving a better estimated of R^2 and RMSE. Figure 6.21, 6.22 illustrate graph in 15 minutes time slot from 5.00 to 24:00.

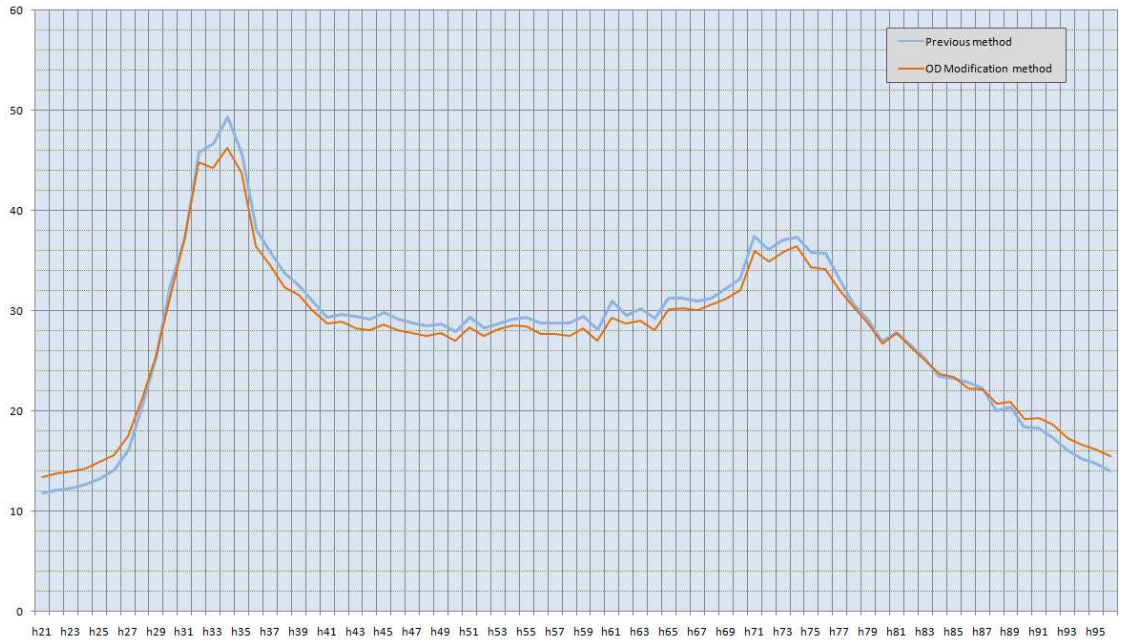


Figure 6.21 Plot of RMSE of the O-D modification methods(orange) compare with previous Person Trip Survey method (blue)

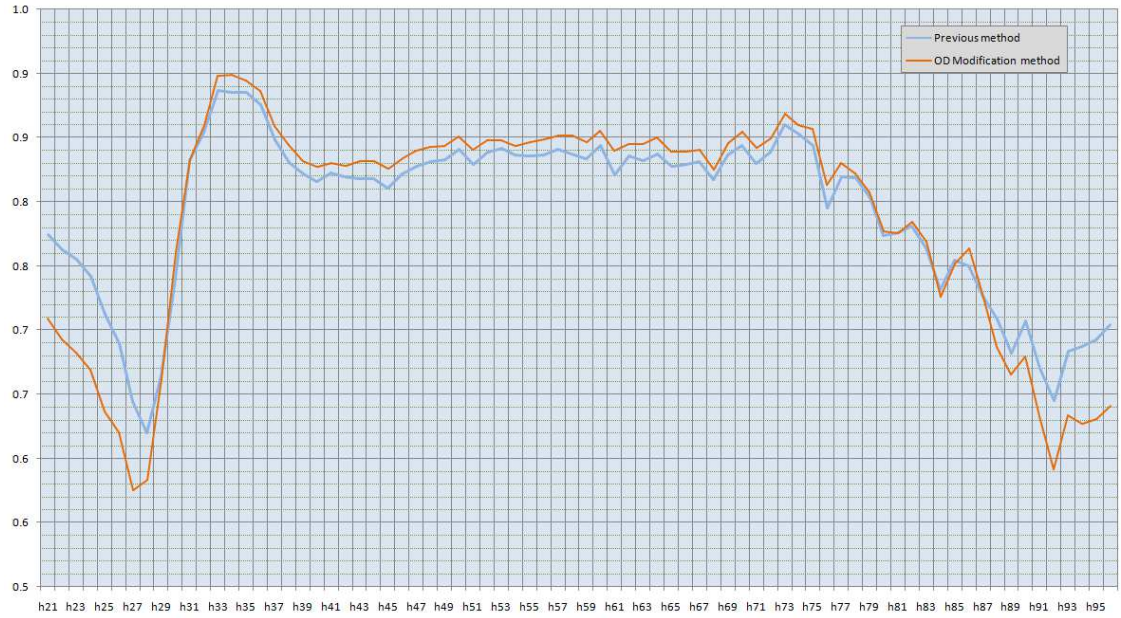


Figure 6.22 Plot of R^2 of the O-D modification methods(orange) compare with previous Person Trip Survey method (blue)

PREF	WARD	J CODE	CITYNAME	EST R2	EST STD ERR	EST R2(*)	EST STD ERR(*)	AREA
TOKYO	CHIYODA-KU	13101	千代田区	0.8884	24.5029	0.8949	23.75904	136
TOKYO	CHUO-KU	13102	中央区	0.9096	26.1743	0.9158	25.24061	113
TOKYO	MINATO-KU	13103	港区	0.8823	12.92391	0.8866	12.70453	242
TOKYO	SINJUKU-KU	13104	新宿区	0.9677	10.644817	0.972	9.902468	213
TOKYO	BUNKYO-KU	13105	文京区	0.7994	13.44848	0.8223	12.6389	131
TOKYO	TAITO-KU	13106	台東区	0.7145	16.74609	0.7229	16.43207	115
TOKYO	SUMIDA-KU	13107	墨田区	0.6352	9.62627	0.6462	9.50941	160
TOKYO	KOTO-KU	13108	江東区	0.6148	4.72715	0.6362	4.60034	426
TOKYO	SHINAGAWA-KU	13109	品川区	0.7589	9.62091	0.7606	9.59281	221
TOKYO	MEGURO-KU	13110	目黒区	0.632	9.12461	0.6722	8.81624	158
TOKYO	OTA-KU	13111	大田区	0.7285	2.86724	0.7362	2.82886	739
TOKYO	SETAGAYA-KU	13112	世田谷区	0.6439	3.02817	0.658	2.97619	781
TOKYO	SHIBUYA-KU	13113	渋谷区	0.8616	14.63037	0.868	14.2843	183
TOKYO	NAKANO-KU	13114	中野区	0.7176	7.13603	0.7265	7.01375	176
TOKYO	SUGINAMI-KU	13115	杉並区	0.7464	3.34208	0.7623	3.23188	448
TOKYO	TOSHIMA-KU	13116	豊島区	0.9253	14.5191	0.9283	14.20364	147
TOKYO	KITA-KU	13117	北区	0.7477	5.9944	0.7685	5.73527	248
TOKYO	ARAKAWA-KU	13118	荒川区	0.5276	13.78381	0.5357	13.55524	113
TOKYO	ITABASHI-KU	13119	板橋区	0.7492	3.79783	0.7485	3.80382	418
TOKYO	NERIMA-KU	13120	練馬区	0.5695	2.82645	0.5821	2.7933	652
TOKYO	ADACHI-KU	13121	足立区	0.6071	2.35355	0.6204	2.32141	706

TOKYO	KATSUSHIKA-KU	13122	葛飾区	0.6041	3.02952	0.611	3.00239	444
TOKYO	EDOGAWA-KU	13123	江戸川区	0.6318	2.39153	0.6341	2.38388	648

Table 6.7 Average Call activity of each occupation type

The results based on the correlation coefficients, Root Mean Square Error, and scatter plots tell us that the proposed model has a potential for further development and using the mobile phone log data for estimating dynamic population in the dense city area is a promising application. However, this model is sensitive to the given total number of estimated O-D matrix which refers to the overall mobile phone activities in the area thus needs to be identified as a requirement. Some limitations are explained as follow;

Limitations

① The simulation (experiment) has several limitations such as the user behavior models were created from a survey of 10^3 samples and it lacks of frequent generated activity users. The previous study suggested average traffic of ½ hour as an appropriated rate to reconstruct the mobile phone trajectory.

② From the experiment, we could not derived O-D matrix based on interval (t = 1 hour), A better results from this algorithm can be expected if revisit points could be captured in every time frame. This could be achievable by use either data mining or machine learning method with larger extent or longer period of mobile CDR data.

From the experiment, we give assumptions that the accuracy of estimated population will be increased if the specific conditions can be performed:

① The accuracy will be improved if the population is large (the best scenario is using the entire population)

② The accuracy will be improved if average area of calculated zone decrease, for instance, from district level to sub district level or the most micro scale at base station level but the complexity of calculation will be multiplied.

③ The accuracy will be improved if more specific O-D matrix can be determined. For instance, the time-dependent O-D matrix can be performed at suggested specific time frame:

06:00 - 09:00 as morning rush hours
09:00 - 12:00 as morning office hours

12:00 - 15:00 as afternoon (lunch, etc)
15:00 - 18:00 as late afternoon
18:00 - 21:00 as evening commuting time

From the initial experiment, we did test with limited numbers of population, and use ward area as an appropriated zone size of estimated O-D matrix. The results gave an estimated day time population of the area.

6.6 Data assimilation method

Data assimilation is a novel, versatile methodology for estimating variables. The estimation of a quantity of interest via data assimilation involves the combination of observational data with the underlying dynamical principles governing the system under observation. The melding of data and dynamics is a powerful methodology, which makes possible efficient, accurate and realistic estimations which might not otherwise be feasible. In case of dynamic population estimation, one important complexity is associated with the vast range of phenomena and the multitude of interactive scales of space and time.

6.6.1 Assimilation process

In this implementation, first, we give a state variable definition which holds a value that represents the real population in the area. We utilize magnification weight factor associated with PTS data as basic definitions in order to reconstruct the population. An approximate dynamics which govern the evolution of the scale of state variables are computed by applying the simulated mobile phone traffic.

In general, a data assimilation method consists of three components: a set of observations, a dynamical model, and a data assimilation scheme or melding scheme. A set of observations is basically referred to the calling probability of each base station grid-cell. The algorithm giving the dynamic model is detailed in the following steps. First we make an assumption with regard to the population parameter that each calling records can be associated with the user types, for instance occupation, sex, and age. By following this assumption, the PT data can be taken it into account. Next, we pick up all PT trajectories that share the same home place. By this we mean that the calling records can be compared within the

same user types which represented in the same base station grid-cell. The calling probability is computed by using a previously developed mobile phone simulation platform called "SIM Mobility". At each time step, we compare the computed calling records with the observed calling records. If the computed calling records are greater than the observed calling records, then we decrease the weights of all trajectories that passing through the subsequent grid-cell. In contrast, if the computed calling records less than the observed calling records, then we increase the weight of all trajectories. Figure 4 illustrate this approach in a simple scenario.

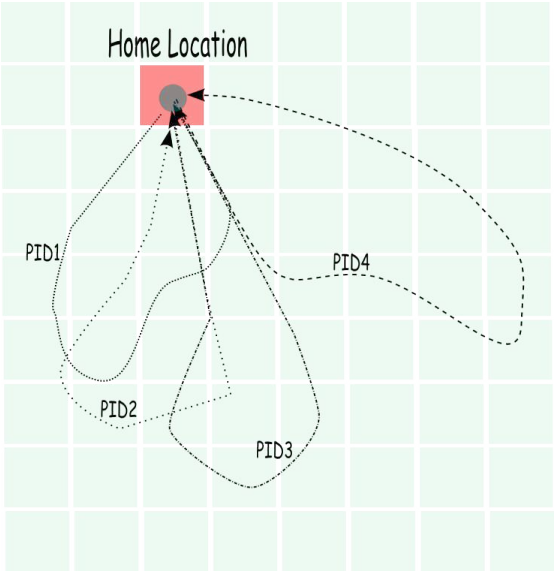


Figure 6.23 Configuration of the method, the selected samples share the same home location

The quantitative basis of melding is the relative uncertainties of the observations and the dynamics computed call probability. Thus, the aspect of assimilation schemes does not degrade the reliable information of the observational data but rather enhances that information content.

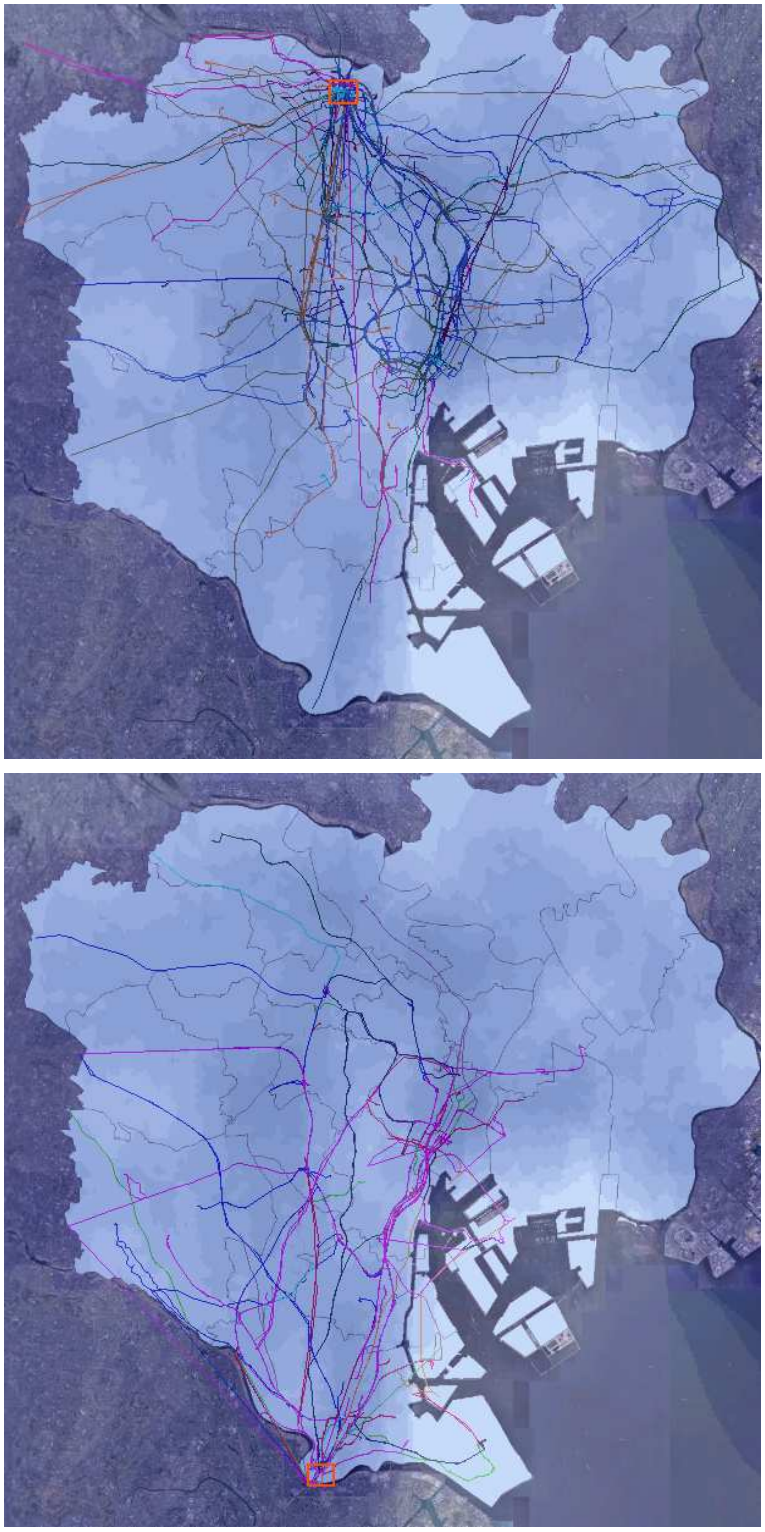


Figure 6.24 Giving sample traces of each occupation types in difference colors where all daily traces share the same home location (red square).

The initial implementation was conducted by observing from 75000 samples that live within Tokyo's 23 wards area which user types covered the person who work in sale, service industrial, transportation, service workers, business, technical, manager, student, housewife and unemployed. A dynamic model used to improve the state parameter was computed as:

$$W_k = f[c_k, o_k]$$

Where

W = weight magnification improvement
c = Computed call probability at the associated base station
o = Observed call probability at the associated base station
k = time from 5 to 24 in an hour interval
f(.) denotes the linear model

The updates can be performed several times, either as several iterations at a single time in order to enhance the smoothness of corrections, or as several corrections distributed in time.

Regarding post validation, we use root mean square error (RMSE) to test the accuracy of this approach by evaluating two different scenarios which is supposed to take part and offer the key to distinguish the effect parameters. The two configurations were configured like so;

Scenario 1: Number of iterations

1. 20 days observation period
2. 40 days observation period
3. 60 days observation period

Scenario 2: Number of population

1. Full population (100%)
2. 50% of population
3. 25% of population

6.6.2 Results and discussions

The RMSE of the estimates decreases significantly with this method at all time. Assimilation process reduces the error more at high mobility stage during 7-8 am. in the morning and around 6 pm. in the evening. As for the first scenario, the accuracy improvement after increasing the simulated iteration can be seen. However, there was no major effect to the improvement of this estimation by increasing the simulation date. (Figure 6.25)

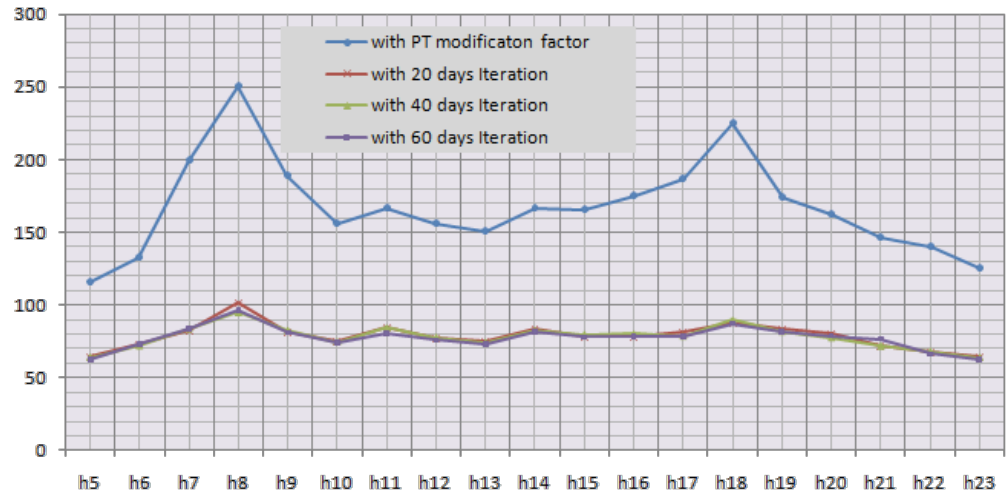


Figure 6.25 Compare the RMSE results of the previous PT method and the assimilation method on different iterated scenario

pid	code	work	magfac	weight
533976	53393692	15	50	0.971
533996	53393692	8	69	1.204
534085	53393692	4	82	0.691
534118	53393692	9	83	0.547
534144	53393691	5	74	0.568
534223	53393681	13	52	0.958
534289	53393692	14	46	1.069
534290	53393690	8	52	0.638
534352	53394508	15	47	0.632

534397	53394508	9	41	0.505
534612	53394508	14	47	0.763
534644	53393596	15	88	0.828

Table 6.8 The results of the individual weight modification through the data assimilation process

code	True_value	PT_magfac	Data_assim
53393523	280	325	288
53393524	229	298	247
53393525	221	338	300
53393526	296	233	209
53393527	314	528	479
53393528	414	622	463
53393529	185	322	234
53393530	238	308	249
53393531	162	91	102
53393532	164	353	241
53393533	334	585	472
53393534	154	77	77

Table 6.9 Compare the results of estimated population

We performed the additional test by changing the configuration of the number of population. The simulation results in the 2nd scenario described a greater degree of improvement when the observed population increase and using the entire observation yield the best estimation at all time. (Figure 6.26)

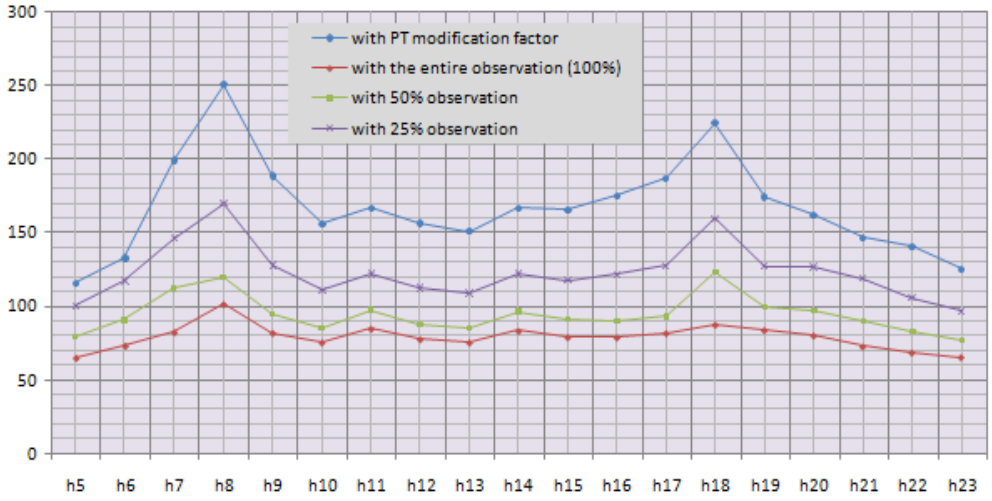


Figure 6.26 Compare the RMSE results of the previous PT method and the assimilation method on different observed population scenario

In this chapter, we have developed new methodologies for population estimation by integrating mobile phone calling records and Person Trip data. The results suggest a considerable improvement of estimated accuracy over the prior method giving by Person Trip Survey. Our future direction is to provide broader scenarios to cover a range of issues that currently limit the feasibility of using this valuable mobile phone data for this kind of research.

Chapter 7

Conclusions and Future Prospect

7.1 Conclusions

This study explores the potential of using mobile phones in the new context and broader advances towards the understanding of today's excessive mobility. The research has contributed to the underlying cause initiative of evaluating mobile phone footprint as a new component of urban system. This dissertation presents a combination of system development, statistical analysis, design of algorithms, data mining, survey study, simulation framework and population estimation model. The rest of this chapter we are going to summarize the major contributions of the research, discuss the future prospect and collaboration.

7.1.1 Summary of contributions and main finding

This research designs to express a novel approach of using digital footprints or user-generated mobile phone data to draw the answers of the questions of this study.

To help answer the research question, Mobile Sensing system, a standard platform for visualizing city-wide human activities and human flow in the complex build environment has been suggested in chapter 3. Ultimately, the world has a common platform only if the key platform elements are open and standard. The mobile sensing system was implemented on web-based interface in order to maximize compatibility and interoperability. We developed standard webAPIs, Internet-based services; to presents the way to visualize time series of aggregated mobile phone sources in pseudo-color contour maps and grid density surfaces. The interface synchronized with interactive graph to give analyses and to explain the activities of mobile phone use in the specific area. The system dedicates a great

deal of effort to the initiated study of using mobile phone for exploring a wide range of urban dynamics.

The great concern on utilizing this mobile phone data lies on the first research question that is "How can actual population at a specific time-frame be estimated from calling records of mobile phones". We have investigated the real mobile phone Call Detail Records (CDRs) data of central Bangkok and Boston area (Chapter 3 and 5). The initial analyses were evaluated using interpolated time series of datasets through the visual analysis. The diagnostic task was done by observing particular land use at specific time frame. The results confirmed the correspondent between urban activity and mobile phone usage.

A more informative case was conducted by analyzing CDRs data from millions of mobile phone users in Massachusetts. Data mining and clustering techniques were applied to examine the calling behavior patterns. The call characteristic in Massachusetts can finally be classified into 5 clusters, which the results minimized the range of sum of squared errors and we discovered that 59% of the population in Massachusetts has very low average call in the weekdays. 15% of population only uses a mobile phone for receiving call and text. Apart from that, 25% of the population has as average call only once a day. The similar finding was found in the result of questionnaire survey in Tokyo. (Chapter 6) Furthermore, the number of mobile phone users has been confirmed with the Tract level planning of census 2000 database. The results showed a good correlation between number of mobile phone users and population in census data.

Additional finding of this remarkable dataset is to analyze human mobility. We find a strong correlation in daily activity patterns within the group of people who share a common work area's profile (Chapter 4). It was report that, within the group itself, the similarity in activity patterns decreases as the distance between them increases.

The finding in chapter 6 has been answered to the problem of how dynamic people mobility can be reconstructed from calling records of mobile phones. The design methodologies used statistical mobility collected from mobile phone Call Detail Records (CDRs) data, incorporated with Person Trip Survey (PTS) to calibrate the estimation of population density. We demonstrate how this approach can be

applied to real world application by developing a simulation system call "SIM Mobility" which generates virtual mobile phone activities. The first design methodology utilized a stochastic global mobility of the O-D trip as flow distribution trends and therefore modifies the magnification factor in the original PTS data. The outcome demonstrated its ability to provide results of satisfactory accuracy with fast implementing and, hence, potential to support the deployment of dynamic urban population estimation procedure. The second design methodology used assimilation process to fine-tune the weight factor where the benefits of assimilation are much longer lasting and essentially giving a better predictor of longer term estimation. These studies provide a preliminary prediction result and would certainly benefit for further analysis.

Among the new idea that emerged in this period, this paper describes an important scenario in which mobile devices with user-driven contents are fully changing the way of urban mobility research as well as the way to estimate the population dynamics. A city could be understood as a seamless system and fully integrated within the co-creation process of new services, products and societal-urban infrastructures.

7.2 Future prospect and collaboration

Building telecommunications networks was mostly a private sector activity from planning, construction and operation. This was in sharp contrast to urban planning, water, power and transportation infrastructure, which were exclusively a public sector manage by government or local government. More specifically, in terms of future policies, government and mobile phone operator need to begin develop an intellectual framework for looking at telecommunication system as real massive mobility and interactive city's sensor.

Generally, mobile phone providers tend not to retain mobile phone CDRs data, in part because there's no business reason to store the data, and in part because the storage costs would be prohibitive. However, a coming increasingly aware of this data, it will eventually encourage the mobile phone providers to open and collaborate with researcher since they could see the benefits from their early involvement in this new area.

However, one of the most discussed topics in doing this research is how to avoid privacy violations problems. The problems of privacy concerns are not new; people have been talking about privacy since the being use of surveillance cameras or GPS phones. Nevertheless, we need to consider the future use of this great and promising data sources from the fast developed mobile technology and this may require the government to obtain a warrant that mobile phone information will not be used unless it is trusted for privacy leak free.

Presence awareness on human mobility also appears likely to be a very promising in adapting this research idea to the real society. Crowd management and public emergency spot forecasting could be one of good examples. A long term capture of mobile phone data would create a city-wide native signature where the anomaly events can be detected or forecasted in real time. In next few years, we believe that user-generated mobile phone sources will become one of the most valuable contents to improve our understanding of urban dynamic, human mobility, real time population density, large-scale social networks and social interactions.

Another aspect in utilizing this research idea is by using different available dynamic information such as internet traffic, email or social media data sources. The results would give extra values with global scale image since we can break the limitation of distance or country boundary that normally happens in case of mobile phone.

Finally, the importance of this research ultimately lies in how it can be practically applied and utilized massive amount of mobile phone CDRs data to understand the city in motion. This raises important questions about how to collect, store the data and aggregate them then analysis and transforming the findings into solutions. Methodologies such as data visualization, data mining and statistical analysis could be used, possibly applied within a specific thematic domain. This dissertation has answered some crucial start up questions of this area, namely, studying and uncovering the hidden aspect of human mobility and estimated dynamic population distribution from the user-generated mobile phone sources. I, the author, therefore encourage all potential readers to continue explore the specific contributions of this novel approach since using mobile phones as sensing devices and aggregating "crowdsourced" data for urban analysis is still in the early

phases of theoretical development. Mobile sensing can potentially paint a complex and dynamic portrait of the urban environment in which users are based.

REFERENCES

A. Akkerman, The Urban Household Pattern of Daytime Population Change, The Annals of Regional Science, Volume 29 No.1, pp1-16, 1995

A. Okabe, B. Boots, K. Sugihara, S.N. Chiu, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd Edition, John Wiley, 2000

A. Pashtan, 2005 Mobile Web Services (Aware Networks, Illinois)

A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson, "People-centric urban sensing," in ACM/IEEE WICON 2006. Boston, MA, USA.

A. Tatem, Y. Qiu, D. Smith, O. Sabot, A. Ali and B. MoonenThe, use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents, 8:287, Malaria Journal 2009.

B. Bhaduri, LandScan USA: A High Resolution Population Distribution Model , the Interface 2002 Conference, 2002.

B. Friedrich, Y. Wang, Improvement of OD Estimation Based on Disaggregated Flow Information, In Proceeding: the 16th Mini - EURO Conference and 10th Meeting of EWGT, 2005.

C. Kaiser and M. Kanevski, Population Distribution Modelling for Calibration of Multi-Agent Traffic Simulation, In proceeding 13 th AGILE International Conference on Geographic Information Science 2010, Portugal

C. Ratti, "Go with the flow" The Economist Technology Quarterly 10 March, page 12-13, 2007.

C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman, Mobile landscapes: Using location data from cell-phones for urban analysis, Environment & Planning, 33(5), pp. 727–748, 2006.

C. Song, Z. Qu, N. Blumm, A.L. Barabási, Limits of Predictability in Human Mobility, Science. 2010 Feb 19;327(5968):1018-21.

D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. Nature, 439:462–465, January 2006.

D. Brocklehurst, People Flow Modelling Benefits and Applications within Industry, Loughborough University, 2005

D. Mountain and J. Raper, Modelling human spatio-temporal behaviour: A challenge for location-based services, In Proceedings of the 6th International Conference on GeoComputation, Australia, 24-26 Sep, 2001

D. Pariente, "Geographic interpolation and extrapolation by means of neural networks" Proceedings of the Fifth European Conference and Exhibition on Geographic Information Systems, volume 1, 684 – 693, 1994

E. Chung, and M. Kuwahara, Mapping personal trip OD from probe data. International Journal of Intelligent Transportation Systems Research, 5(1). pp. 1-6, 2007.

F. Aurenhammer and R. Klein, " Voronoi diagrams" In: Handbook of computational geometry Eds J.-R. Sack and J. Urrutia (North-Holland, Amsterdam) pp 201 – 290, 2000.

I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy walk nature of human mobility. In Proceedings of INFOCOM 2008, Phoenix, AZ, April 2008.

I. Yasuyuki, K. Shin'ichi, T. Niwat, S. Ryohei, S. Kaoru and T. Yoshito., An Implicit and User-Modifiable Urban Sensing Environment, International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems November 4, 2008, Raleigh, NC, USA.

J. Conway, 2003, "PostgreSQL-embedded Statistical Analysis with PL/R" O'Reilly Open Source Convention, 7 – 11 July.

J. Hutchins, A. Ihler, and P. Smyth, 2008, Probabilistic Analysis of a Large-Scale Urban Traffic Sensor Data Set. Second International Workshop on Knowledge Discovery from Sensor Data (ACM SIGKDD Conference, KDD-08)

J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, 2007, "Cellular Census: Explorations in Urban Data Collection" IEEE Pervasive Computing 6(3) 30-38

J. Schlaich, T. Otterstätter , and M. Friedrich, Generating Trajectories from Mobile Phone Data, TRB 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies, 2010

J. Yoon, B. D. Noble, M. Liu M. Kim, Building Realistic Mobility Models from Coarse-Grained Traces, in proceedings of the 4th international conference on Mobile systems, applications and services, pp 177 - 190, 2006

K. Farrahi and D. Gatica-Perez, Discovering Human Routines from Cell Phone Data with Topic Models, IEEE International Symposium on Wearable Computers (ISWC), 2008

K. Lee ,S. Hong, S. Joon Kim, I. Rhee, S. Chong, SLAW: A Mobility Model for Human Walks, , In Proceeding, INFOCOM, Rio de Janeiro, Brazil, 2009,

K. Lee, S. Hong, S. Joon Kim, I. Rhee and S. Chong, Demystifying Levy Walk Patterns in Human Walks

K. Sakamura and N. Koshizuka, Ubiquitous Computing Technologies for Ubiquitous Learning, Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'04), IEEE 2004.

K. Sato, Y. Sekimoto, R. Shibasaki, Study for Reconstruction of People Flow in the City with Person Trip Data, In Proceedings of AsiaGIS, 2008

L. Bertolini, M. Dijst, 2003, "Mobility Environments and Network Cities" Journal of Urban Design 8(1) 27–43.

L. Mitas, H. Mitasova, 1999, "Spatial Interpolation", In: Geographical Information Systems: Principles, Techniques, Management and Applications Eds P.Longley, M.F. Goodchild, D.J. Maguire, D.W.Rhind, (Wiley, New York) pp 481 – 492.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. Nature, 453:779–782, June 2008.

M. DeMers, 2000 Fundamentals of Geographic information system (Wiley and sons, New York)

M. Erwig, R. H. Guting, M. Schneider and M. Vazirgiannis, Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases GeoInformatica, 3:269–296, 1999.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, L. H. Witten; The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol. 11, Issue 1, 2009.

M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In Proc. of IEEE INFOCOM 2006.

M. Kawabata, A GIS-Based Analysis of Jobs, Workers and Job Access in Tokyo, CSIS Discussion Paper No. 57, July, 2003

M. Piorkowski, Sampling Urban Mobility Through Online Repositories of GPS Tracks, ACM Hot Planet, June 2009

M. Piorkowski, N. SarafijanovocDjukic, M. Grossglauser A Parsimonious Model of Mobil e Partitioned Networks with Clustering, COMSNETS, January 2009.

M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. 2008 "Mobile call graphs: beyond power-law and lognormal distributions" Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 596 – 604, AMC

M. Srivastava et al., "Network system challenges in selective sharing and verification for personal, social, and urban-scale sensing applications," in HotNets '06, Irvine,

California, USA, November 29 2006, pp. 37–42.

N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, A. T. Campbell, 2008, "Urban Sensing Systems: Opportunistic or Participatory" Proceedings of the 9th workshop on Mobile computing systems and applications 11-16

N. Eagle, A. Pentland , D. Lazer, Inferring social network structure using mobile phone data. PNAS (2007)

N. J. Boucher, 2001 The cellular radio handbook : a reference for cellular system operation. 4th ed (Wiley and sons, New York)

N. Ohmori, Application of Information on Human Activity-Travel Behavior in Urban Space and Time in the Information Age, Spatial Data Infrastructure for Urban Regeneration, pp 127-145, 2008

P. Sistla, O. Wolfson, S. Chamberlain and S. Dao, Modeling and Querying Moving Objects, IEEE International Conference on Data Engineering, Apr 1997, Birmingham, U.K.

R. Neuwirth, Shadow cities: a billion squatters, a new urban world, New York: Routledge, 2006.

S. Ng, C. Cheung and S. Leung, Magnified Gradient Function With Deterministic Weight Modification in Adaptive Learning, IEEE Transaction on Neural Networks, Vol.15, No. 6, 2004.

T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich, "Mobiscopes for human spaces," IEEE Pervasive Computing - Mobile and Ubiquitous Systems, vol. 6, no. 2, April - June 2007.

T. Hengl, 2007, "A Practical Guide to Geostatistical Mapping of Environmental Variables" JRC Scientific and Technical Reports

T. Tsekeris, L.Dimitriou, and A.Stathopoulos, Simultaneous Origin-Destination Matrix Estimation in Dynamic Traffic Networks with Evolutionary Computing, Lecture Notes

in Computer Science, Volume 4448, pp 668-677, 2007.

Y. Asakura and E. Hato, Tracking survey for individual travel behaviour using mobile communication instruments, Transportation Research Part C 12 (2004) 273–291

W. Hsu, K. Merchant, H. Shu, C. Hsu and A. Helmy, Weighted waypoint mobility model and its impact on ad hoc networks, ACM SIGMOBILE MC2R 9, pp. 59-63, January 2005

Working Group, 2005, "Urban Mobility Initiative" European Conference of Transport Research Institutes, Report ECTRI number 2005-06.