PhD Thesis

A Study on Human Mobility Using Cell Phone Traces

（携帯電話トレースを用いた人の移動に関する研究）

The University of Tokyo

Graduate School of Frontier Sciences

Department of Socio-Cultural Environmental Studies

47-137651 Dunstan Matekenya

ダンスタン　マテケニャ

Advisor Prof. Kaoru Sezaki

17th August 2016

Abstract

In today's urban spaces, people generate data everyday either consciously (e.g., by participating in participatory sensing projects such as OpenStreetMap) or inadvertently (e.g., by using metro systems or making phone calls). The result is that I have huge collections of digital traces telling *when* and *where* people go (this is what is considered as human mobility data). These huge collections of data often hide interesting information with high potential for decision making in many domains such as urban planning. However, there exists a gap between the raw data itself and the benefits which can be reaped from it. In other words, in order to leverage this kind of data in urban planning or disaster management or private businesses, I need to come up with robust tools to mine useful insights from the data. The work in this thesis contributes to the bridging of this gap by providing a set of techniques for mining mobility data. I develop these techniques by studying key aspects of human mobility using data generated from cellular transactions.

Although human mobility research has attracted huge attention, it is still open to further investigation. This is partly due to the fact that human behaviors are inherently fuzzy and dynamic. Also, most human mobility studies which leverage massive datasets make conclusions which are inevitably specific to characteristics of the datasets used in the study and therefore cannot be easily generalized. For instance, suppose a study uses GPS location data coupled with social network (SN) data to develop some algorithm for location prediction. This type of algorithm cannot be adopted wholesale in a scenario where the available location data is sparse (such as that from cellular networks) and also with no access to SN data. It is against this background that I tackle three problems related to human mobility as follows: visualization of mobility, discovering residence change and location prediction.

I first develop a web-based framework for interactive visualization of mobility patterns in order to allow easy and quick interpretation of trends. I then explore the use of Call Detailed Records (CDR) generated from mobile phones to infer change of place of residence (home). Finally, I undertake to improve performance of location prediction systems. In particular, our objective is to reduce training time of prediction models for individual users as well enhance prediction accuracy. In the visualization system I build, the objectives are two fold: first, I want to understand human mobility patterns based on peoples' calling habits; second, I want to identify city-wide events such as religious or sporting events based on call traffic of cellular towers.

Next, I investigate the potential to use CDR data as surrogate *residence history* with a purpose to discover residence change so that I can ultimately infer internal migration patterns. I first provide a rigorous definition of what I call *the residence change discovery problem.* I then propose a novel *sequential spatio-temporal clustering technique* which I call *MoveSense* to solve this problem. I carry out experiments to validate our technique. Results from the experiments show that our technique performed well with average *detection rate* of 71 percent, 68 percent and 72 percent across the three categories of datasets I tested it on.

In the location prediction task, the broad research question I address is this: can I leverage big data to enhance performance of location predictors without relying on external data sources? In order to answer this question, I first carry out spatio-temporal analysis of user call behavior and call activity and use the insights to propose an *enhanced bayes predictor* which leverages large scale data. Results from the experiments I conducted reveal that overall, the enhancements I propose improve the predictors' performance by 17 percentage points. Secondly, I investigate the potential to improve performance (accuracy and training time) of location prediction models by again leveraging large scale data. Given that users closer in space would exhibit similar mobility behaviors, our idea is to create what I are calling a *community model* for a group of users in a given geographic area and then use parameters from this model to enhance performance for individual users in the same community. I choose to experiment with logistic regression classifier. The results from our experiments show that our idea to use community-wide learned model parameters in individuals works very well and reduces training time for individual models by nearly 100 percent.

In summary, in this thesis I study human mobility using cellular phone data. The primary objective of our study is to develop techniques to mine insights from the data useful for urban planning and other application areas. To this end, I first proposed a simple but non trivial visualization system to ascertain mobility pattern of users. Second, I demonstrated a technique to automatically detect residence change which has useful applications in profiling internal migration. Finally, I conduct extensive study in enhancing perfomance (accuracy and training time) of location prediction models.

Acknowledgements

I could never have completed this work without the support and encouragement from many people. First and foremost, I would like to express my deepest gratitude to my advisor Kaoru Sezaki and co-advisor Masaki Ito. More than anyone else they have influenced my view of the academic research process, and instilled in me not only the discipline to adhere to ethical standards but also the inspiration to aim for high-quality research with the potential for impact. I truly value their honest opinions, their calmness and clear advice amidst challenging times, and their patience and understanding over the past three years. I am highly indebted to have had advisors that gave me all of the freedom, resources, guidance and support I could ever ask for during the period that lead up to this dissertation. In Sezaki Lab. they provide an environment where researchers can thrive and where the degree to which you succeed is ultimately in your own hands. Furthermore, my advisors also rendered to me all the support I needed in my personal life during the period of my PhD study.

As a member of Sezaki Lab. I feel fortunate to have been surrounded by a number of outstanding individuals who, at different stages of my PhD, were part of the Lab. Most of them were native Japanese students who were part of the Master program at the University of Tokyo. Although I cant mention them all, Japanese students were very kind and helped me with a lot of personal matters. Also, the secretaries in the Lab. Matsumoto-san and Naito-san generously helped me with personal issues. . I must make special mention of Guangwen Liu, a fellow international student in the Lab. He has been particularly helpful over the years in my academic work as well as personal life. We spent many hours in the Lab together and I learned a lot about academic research from him. I also would like to mention Yoshite Tobe from Aoyama Gakuin University who from time to time provided me with valuable advice and guidance on my research.

Finally, I cannot thank enough my wife, Mercy for being accepting, patient and loving irrespective of the many inconveniences caused by decision to undertake this doctoral course. I'm also grateful to my parents and siblings who have been morally supportive through out the course of my PhD reserach work.

June 2016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background: Human Mobility Studies

I can define human mobility as the trajectories which humans follow as they traverse space in their day to day activities. Characteristics of human mobility can be put into two broad categories: spatial-temporal and social. The spatial-temporal aspect refer to the trajectory patterns in physical space and time so that this feature can be thought to answer the question *where and when* about a person's mobility. On the other hand, the social aspect is broad and can encompass many attributes such as the activities associated with each trajectory as well as the contacts which this particular person makes. The goal of human mobility studies is not only to shed light on the aforementioned fundamental characteristics but also to understand how they relate to other aspects of human life. In the following, I present some of the specific questions which human mobility studies attempt to address: i) what physical laws govern human mobility [SKWB10]? ii) how far can I predict human mobility [SQBB10]? iii) what is the relationship between mobility and other aspects of human life (e.g.,diseases, economy)? iv) are there any statistical regularities in a person's mobility traces?

When I consider humans themselves as subjects, mobility can be studied at different scales where in this case scale refer to the number of persons. For example, one may focus on mobility of a group of people during special events such as religious events or emergency situations (e.g., earth quake disaster). This is a case of crowd modeling [JBH+12]. In other cases, I may be interested only in individual mobility patterns under normal circumstances (i.e., not in emergency situations or special circumstances such as celebrations) but just as a person go about their day to day life. In this thesis I focus on this kind *individual human mobility*. Further, when I consider spatial scales, I can classify mobility into two groups: that which occur within the boundaries of a country and that which involve crossing national borders as the case in intercontinental migration studies [WT08]. In our case, I limit our study to the former. Finally, I can also consider the temporal component of human mobility. In this regard, some studies focus only on long term mobility such as migration. For example, in the case of internal migration (which doesn't involve crossing national borders but rather some administrative boundaries within a country), the temporal scales (I can define this loosely as the amount of time required to observe a particular individual) in these kind of studies can range from months to years. On the other hand, in short term mobility, temporal scales range from seconds to hours, for example, in a traffic system in order to predict traffic levels, the objective may be to predict where some users will be in the next few minutes or in some systems predicting where a user will go next [MRJ12]. Its worth mentioning here that although I have separated spatial and temporal component for the sake of simplicity, in most cases these two are considered simultaneously. Majority of the research questions I tackle in this dissertation can be categorized under short term mobility, however, I also conducted an investigation which falls under medium-long term mobility.

The next question I address is what are the benefits of studying human mobility? Thorough understanding of laws that govern human movements has useful applications in public health [CBB$^+$07, BHG$^+$09, EGK$^+$04, KE05], urban planning [DLC11, MHS95, KCRB09], disaster planning and mitigation [LBH12], traffic engineering [WHB$^+$12] and marketing. For example, in the work in [EGK$^+$04] they demonstrate how population mobility and land use data can be used to simulate how a disease would progress within a single host and also how it would be transmitted across people in an urban area as they go to different locations in course of their daily activities. Lu and colleagues[LBH12] showed that understanding of people mobility during large scale disasters such as the the 2010 Haiti earthquake can enable prediction in advance of how people will move in future disasters which is useful for relief and recovery services. Regarding marketing and retail, location based systems such as Google maps, Yelp and Foursquare [Yel, Fou] as well as other contextual apps rely heavily on understanding user mobility in order to provide highly personalized services and also develop new products.

Before the advent of modern day sophisticated smartphones and wide spread usage of mobile phones, large scale human mobility studies relied on static, mostly government provided data such as that from census and surveys. Examples of such studies include those on trying to understanding relationship between diseases and mobility [SC93]. However, in smaller scale studies (e.g.,[Pro95]) where they were interested in mobility patterns during fire emergency, they could use cameras and other special equipment to study mobility. The obvious drawbacks were that the data was often static and therefore not up-to-date. In addition, it also means that mobility could only be studied up to some geographic level (e.g., census tracts level) and not at individual level. One consequence of this is that majority of the empirical studies in human mobility were actually limited to humanities and social sciences. For those few studies in other disciplines such as computer science, their investigations were mostly based on simulations [DS99] while other studies such as those in physics were theoretical but with useful applications in human mobility[LTH+92, FMG92]. All this changed, thanks to the advent of smartphones, they not only serve as the key computing and communication mobile device of choice but they also come with a rich set of embedded sensors, for example, an accelerometer, digital compass, gyroscope, GPS, microphone and camera. From GPS data I can extract human mobility data at very minute scale. Even for mobile devices without GPS it is still possible to get mobility data based on user cellular transactions though the location obtained from this kind of data is low resolution compared to that obtained from GPS based systems. One of the early mobility studies in computer science to make use of this kind of data is presented in [SKJH06]. Since then, there has been an explosion in human mobility studies using data generated from mobile devices. From here onwards, I will use the terms mobility and human mobility interchangeably. For instance, instead of human mobility data, I will often just say *mobility data*.

Because mobility data is central to the work in this thesis, I provide further details here. Due to its very nature most of this data is generated through mobile/portable devices such as cell-phones, smartphones, tablets, PDAs and laptop computers. In all mobility datasets, there must be an element of location. As mentioned earlier for smartphones, GPS provides the location information and it is perhaps the most accurate source of location information available at present for regular smartphones because it provides what can be considered as exact geographic coordinates of the device on the earth surface. I can also consider GPS location as high resolution because of this kind of accuracy. There are other techniques of localization, for example, cellular positioning, Wifi and bluetooth or a combination of any of these techniques. Datasets such as those based on cell-ID positioning has lower resolution as they only provide location of the device with respect to cells in the cellular network. Social networks can also provide some form of low resolution location information, for instance, geotagged tweets or home location of Facebook users. I need to point out two things about these kind of mobile datasets: first, due to sensitivity of the information, it is very hard to get access to these kind of data. Also, unlike in census or surveys where there is deliberate sampling and data is usually representative of the target population, the same cannot be said for this kind of data. Mobility datasets have biases from many sources such as ownership and researchers have already revealed such biases and in some cases suggest remedies [RZZB12, ZZZ^{+}13]. In my case, the mobility dataset I study comes from cellular phones. A mobile phone communicates wirelessly with a base station, usually with the one that is physically closest to it and using a technique known as *cell-id positioning*, we can assign to each cellular device in the system location of the nearest base station during a transaction such as call. Because its expensive to continuously track user location cellular companies only log details of user location when they make a transaction (e.g., phone call, data usage or text message). This kind of information can be considered as *cellular traces*, however the more technical term is *Call Detailed Records (CDRs)*.

Having provided an overview of human mobility, why I study it and how I study it (through mobility data), I now turn our attention to introducing the specifics of our study. In this thesis, I focus on three thematic areas of human mobility: data exploration and visualization; residence change discovery and location prediction. I first look at visualization and exploration of mobility data because it is a crucial preliminary step in understanding any mobility data. Data visualization is the process of encoding data as visual objects. Numerous studies (e.g., [TGM83]) have shown how thoughtful visualization of data can lead to easy discovery of insights from the data for both experts and non-experts. More recently, due to better graphics capability and bandwidth, interactive visualization has become popular. For interactive visualization, the user of data is able to interact with the data objects through, for example, web based interfaces. In our case, I tackle this challenge and come up with a simple but non-trivial web based interactive visualization which depicts user mobility. In addition, I also carry out exploratory analysis which feed into our more advanced studies.

Second, I study what can be considered as medium-long term human mobility. When I consider the temporal component of human mobility, long term mobility is the kind where I have to observe the persons of interest for periods longer than a month in order to determine whether what I are looking for is available or not. As mentioned earlier on, one example in this category is migration. More specifically, the task I tackle in this category is given CDRs for a single user automatically determine whether they changed place of residence or not. This is imperative in urban planning because given this kind of information, I can ultimately use it to determine whether a person migrated or not and internal migration information is crucial in city planning. In addition, even for movement which doesn't qualify as migration, this information could be informative for businesses considering I are in era when intelligence such as this is central to business operations.

Finally, most of our focus is on short term mobility studies. While there are many areas of interest in short-term mobility, in this work, I focus on location prediction. Our goal is to leverage large scale data to build heuristics which can enhance performance of location predictors (i.e., enhance accuracy) as well as improve scalability by reducing time required to train models. The key requirement in our work is that I want to achieve this without relying on external data sources. I have noted that in previous studies, most schemes which suggest performance enhancements usually do so by leveraging external data sources such as use social profiles from social networks. I argue that such data is not always available. In the developed countries this is the case due to privacy protection policies while in low income countries, the usage rates of social networks are still pretty low. Further in these low income regions, smart-phone adoption rates are also low, consequently, accessing any kind of data which would augment individual user location prediction would be hard. Therefore the methods that I propose and develop would be very useful in these kind of scenarios I just explained.

1.2 Problem Statement

In the previous section, I introduced the discipline of human mobility, and discussed the challenges and opportunities it presents. I now describe three specific challenges addressed in this dissertation.

First, in agreement with most scholars, I note that thorough exploration and visualization of mobility data is crucial in understanding the data as well as subsequent analysis. The challenge though with visualization is that off-the-shelf visualization software do not always work well. There are two main reasons for this: first, datasets are inherently different even though they could all be about human mobility. For instance, trends and patterns which could be visually appealing in dataset X would not necessarily be in dataset Y. Also, difference analysis objectives may call for significantly different visualization approaches. For our analysis objective and data, I didn't find off-the-shelf approaches which were useful and therefore I came up with our own. Overall, development of effective visualization systems is still an open problem in human mobility and for any particular dataset, its a must to develop such methods to aid in understanding the data. In response to this, I developed a web based interactive visualization of human mobility. Although I developed our solution in response to the current data set, I believe the sample principles can be applied in a different setting.

Second, I addressed the problem of predicting whether a person has changed place of residence or not from CDR data. Although this problem is old and has its origins from demography where it has been sufficiently adressed. The ability to develop methods to discover residence change from CDRs has many benefits. For example, this means that I can gain information needed by authorities such as internal migration, then reliance on large scale surveys or census could slowly be reduced because these new data sources offer several attractive advantages. For one thing, the effort and cost of data collection is nearly negligible when compared to that of a census. Most importantly, since this kind data is always available, it is possible to get on demand a picture of some population attribute of interest such as internal migration.

The final problem I address is about location prediction. The broad problem I tackle is how to leverage large scale data to improve accuracy of location predictors. Although this problem has been extensively studied most of the works are *context specific*. A good example of such contexts is the type of data used: *resolution of location data* and whether *additional data* (e.g., social relationships data) is used or not. For example, NextCell [ZXYG13] is a prediction system based on cell phone traces with the assumption that there is information about call patterns amongst users within the data. Other studies, e.g., Find me if you can [BSM10] use social relationships data to improve results of location prediction. Consequently, the conclusions drawn and the techniques developed cannot be transfered to new problems without substantial modifications. Furthermore, most of the previous works used very small datasets (in the range of thousands of users) to evaluate performance of their techniques and also study other issues related to performance. I argue that such small datasets limit the ability to study comprehensively behavioral factors which affect perfomance of location predictors. On the contrary, I use a dataset with millions of users.

1.3 Thesis Contribution and Outline

The contributions of the work in this thesis can be summarised as follows:

- *Visualization of mobility data.* In Chapter 2, I describe a web-based framework for interactive visualization of mobility data.

- *Residence change discovery.* In Chapter 3, I explore the use of Call Detailed Records (CDR) generated from mobile phones to infer change of place of residence (home). I first provide a rigorous definition of what I call *the residence change discovery problem* and then propose a novel *sequential spatio-temporal clustering technique* which I call *MoveSense* to solve this problem. Results from experiments show that our technique performed well with average *detection rate* of over 68 percent.

- *Enhancing Location Prediction with Big Data.* In Chapter 4, I tackle the problem of enhancing accuracy in location prediction systems. Specifically, the research question I ask here is can we leverage big data to enhance performance of location predictors? I choose to experiment with *Bayes based predictor.* First, I carry out spatio-temporal analysis of user call behavior and call activity and use the insights to propose an *enhanced bayes predictor* which leverages large scale data. Results from the experiments I conducted reveal that overall, the enhancements I propose improve the predictors' performance by 17 percentage points.

- *Using Community-wide models.* In Chapter 5 which is closely related to Chapter 4, I also study how to enhance location prediction systems. In this case, I propose the idea of using a community-wide model learned from data of multiple users to enhance models for individual users (reduce training time and improve accuracy). Results from the experiments I conducted show that using parameters from community model as lower bounds in the optimization process while training individual models drastically reduced the training time for individual models by almost 100 percent.

The rest of this thesis is organized as follows. In Chapter 2, I present a web based interface for visualizing mobility data. Chapter 3 tackles the problem of residence change discovery, the key research question I solve is how to automatically predict whether a person changed their place of residence using their call history. In Chapter 4, I introduce the problem of location prediction where our goal is to enhance location predictors using large scale data. I continue work in location prediction in Chapter 5. Finally, Chapter 6 gives concluding remarks.

Chapter 2

Visualization of Mobility Data

2.1 Introduction

Thanks to the pervasiveness of mobile phones, it is now possible to get unprecedented spatio-temporal digital footprints telling *when* and *where* people are at all times. If I can successfully mine from this data some of the information needed by city authorities such as internal migration, then I could end reliance on large scale surveys or census because these new data sources offer several attractive advantages. Numerous previous studies have demonstrated that its possible to establish the important places, in most cases *home* and *work* in peoples lives from Cellular Call Detailed Records(CDR). For instance, in [IBC+11] they used CDR data to identify home and work location while in [BCH+13] they were able to characterize human mobility at metropolitan scale by using a similar dataset. However, due to complexity and highly heterogeneous nature of human behavior this area is still open to further research. In this regard, I have implemented a *space-time visualisation system* which leverages ubiquity of mobile phones to understand not only individual cellular subscriber mobility behaviour but also traffic of cell towers.

CDRs are routinely collected by cellular network providers to help operate their networks. Each CDR contains information such as the time a voice call was placed or a text message was received, as well as the identity of the cell tower with which the phone was associated at that time. This information can serve as sparse and sporadic samples of the approximate locations of the phone's owner.

2.2 Related Work

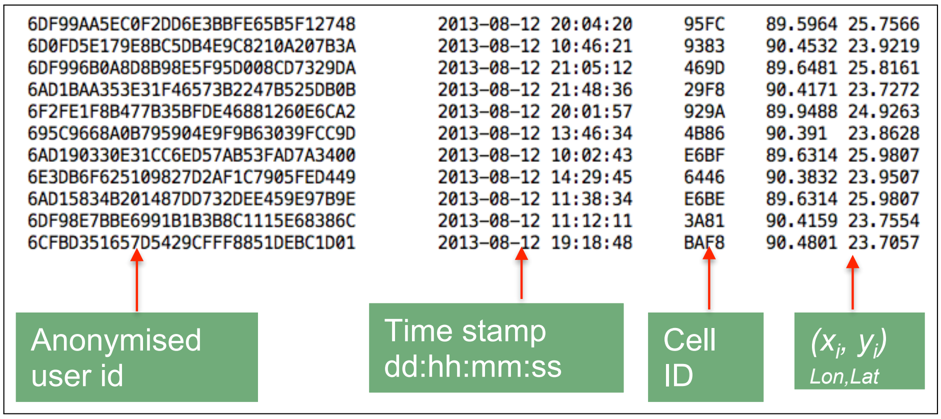
Recently, advances in statistical computation and graphic display have provided tools for visualization of data which was unthinkable a few decades ago. There has been corresponding advances in human-computer interaction which have also led to creation of completely new paradigms for exploring graphical information in a dynamic way, with flexible user control. However, the history of visual display of information dates back to as early as the 18th century (see the work in [FD01] for a detailed description of the history of visualization). While it is possible to visualize data from a wide range of topics, in this thesis my interest is in visualizing human mobility data. Consequently, I focus my review of previous research works in this area of visualizing movement.

Mobility data has recently received heightened attention in the visualization community. Andrienko et al. [AA12] conducted a survey of what he called *Visual Analytics of Movement* and discussed various characteristics of movement data, and summarized three visualization categories: direction depiction, summarization and pattern extraction. Much of the works has focused on GPS based trajectory data, where the complete trace of the moving entities is recorded. For example, Ferreira et al.[FPV+13a] proposed a novel visual query model which allows user to interactively explore and compare results obtained from millions of taxi trips. In the work in [ZFAQ13], they proposed a novel visualization technique which can reveal interchange patterns in massive public transport trips. In addition, there are many other off-the-shelf softwares for data visualization such as MobiMap([mob16]) or Tableau ([tab16]).

While most of the previous works mostly used GPS trajectory data, in contrast, the data I consider in this thesis is sporadic with only data points when a person makes a phone call (see Section 2.3 for details about the data I use). Therefore, these previous methods cannot work out of the box without significant modification. Regarding the available off-the-shelf software, the key limitation is that they are built for generic purposes such as visualizing short term movement. However, in my case, I wanted to also have statistics for long term mobility. In summary, although the spirit of my work here is similar to all the previous studies (i.e., visualize human mobility data), in this thesis, I focus on generating mobility statistics which can allow better longer term comparison of human mobility.

Table 2.1: Background statistics of the dataset

Category	Description
Starting date	August 1, 2013
Ending date	December 31, 2013
Data gaps	October 1-October 30 2013
Number of cell towers	2101
Number of users	16,000,000
Number call events	3,5,000,000



```
6DF99AA5EC0F2DD6E3BBFE65B5F12748        2013-08-12 20:04:20    95FC    89.5964 25.7566
6D0FD5E179E8BC5DB4E9C8210A207B3A        2013-08-12 10:46:21    9383    90.4532 23.9219
6DF996B0A8D8B98E5F95D008CD7329DA        2013-08-12 21:05:12    469D    89.6481 25.8161
6AD1BAA353E31F46573B2247B525DB0B        2013-08-12 21:48:36    29F8    90.4171 23.7272
6F2FE1F8B477B35BFDE46881260E6CA2        2013-08-12 20:01:57    929A    89.9488 24.9263
695C9668A0B795904E9F9B63039FCC9D        2013-08-12 13:46:34    4B86    90.391  23.8628
6AD190330E31CC6ED57AB53FAD7A3400        2013-08-12 10:02:43    E6BF    89.6314 25.9807
6E3DB6F625109827D2AF1C7905FED449        2013-08-12 14:29:45    6446    90.3832 23.9507
6AD15834B201487DD732DEE459E97B9E        2013-08-12 11:38:34    E6BE    89.6314 25.9807
6DF98E7BBE6991B1B3B8C1115E68386C        2013-08-12 11:12:11    3A81    90.4159 23.7554
6CFBD351657D5429CFFF8851DEBC1D01        2013-08-12 19:18:48    BAF8    90.4801 23.7057
```

Figure 2.1: Screenshot of Data

2.3 The Dataset

In this work, I use a CDR dataset from a leading cellular phone operator in Bangladesh. It was collected in 2013 and covers the months of *August,September, November* and *December*. The data mainly include details of a call event: time of call, cell tower id and latitude and longitude of the cell tower involved in the call. The raw data came in CSV files with a total size of approximately 400 GB and contained over 3 billion call events. In Table 2.1, I present background characteristics about the dataset while in Fig. 2.1

2.4 Design Requirements and System Work-flow

The objectives of our system are two fold: first, I want to understand human mobility patterns based on peoples' calling habits; second, I want to identify city-wide events such as religious or sporting events based on call traffic of cellular towers. The workflow of our visualization platform is illustrated in Figure 2.2. The CDR data for all users is stored in a central database. When a

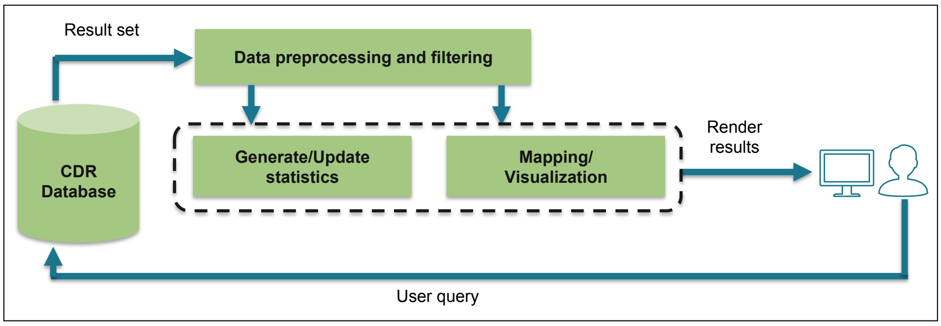


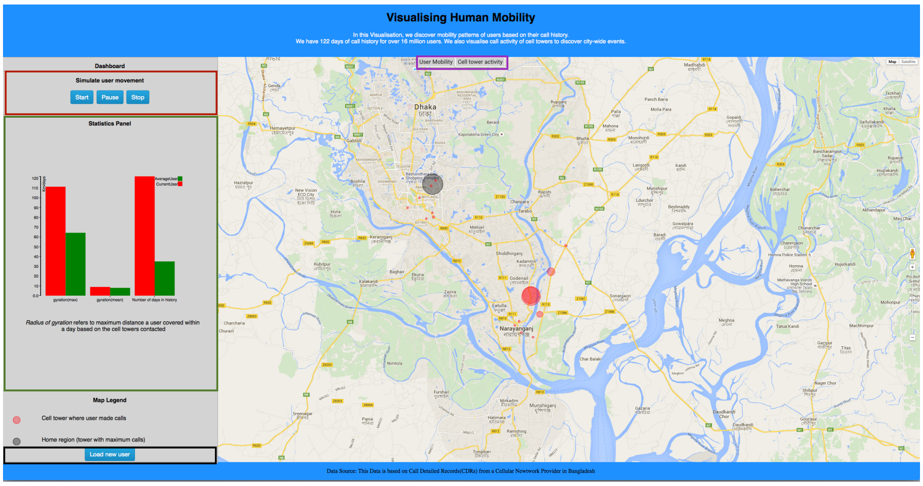
Figure 2.2: Workflow of the visualization platform

user sends a query, the system does a filter operation to get the required data. In addition, the system also pre-process the data in order to retrieve only relevant data. The results of this data is then used in computing mobility statistics and also updating the visualization to reflect the query. Finally, the results are rendered to the user.

2.5 The System-User Interface

The user interface of our system is presented in Fig. 2.3. The user interface consists of two main components: the left part is the dashboard while the right part is the map area. The dashboard is primary for the user to interact with the system (e.g., select to see data for a new user). The dashboard has four sub-panels: the top most part (shown in red in the screenshot) is the *timer*. The timer gets activated during animation of user movement and is used to indicate the time a user visited some place. The *statistics panel*, is where I display statistics for current user. In what I call the *navigation panel*, the user of the system can choose to *load a new user* and also whether they want to see *animation* of current user movement.

The map area is built on top of Google maps. In this area, I show the location of the cell towers which this user visited. The map area also contains buttons for switching between *individual mobility* and *cell tower activity*. In the map area, the system represents the places based on how often they were visited. The map legend is shown in the dashboard at the bottom.



Visualising Human Mobility

In this Visualisation, we discover mobility patterns of users based on their call history.
We have 122 days of call history for over 16 million users. We also visualise call activity of cell towers to discover city-wide events.

Radius of gyration refers to maximum distance a user covered within a day based on the cell towers contacted

Data Source: This Data is based on Call Detailed Records(CDRs) from a Cellular Newtwork Provider in Bangladesh

Figure 2.3: Screenshot of the visualisation system

2.6 Chapter Summary

In this chapter, I propose a web based interface that uses CDR data to visualise human mobility and cell tower activity. The objective of our system is to understand human mobility patterns based on peoples' calling habits and also to detect city-wide events based on cell tower call traffic.

Chapter 3

Residence Change Discovery-MoveSense

3.1 Introduction

In most countries, authorities are often interested to know when a person changes residence because such information is crucial for urban planning as it enables them to continuously update how many persons are staying within a particular geographic unit. Furthermore, it provides valuable insight on mobility patterns of people across administrative units within a country or city which could be useful in business intelligence. The change of residence within a country is referred to as *internal migration* if it is permanent and involves crossing some designated administrative boundaries. This kind of information is routinely collected by governments and city authorities through structures such as vital registration systems. However, periodically countries still need to conduct a *population census* or other large scale surveys to collect internal migration related data to complement the routine information. In fact, in most developing countries vital registration systems and other routine administrative data collection structures are almost nonexistent, consequently they rely entirely on population census for information on internal migration. The collection of migration data using these traditional methods have three important drawbacks:

- These traditional approaches can be time consuming as they often involve sending questionnaires to each and every household in a city or country. Even worse, in most developing countries this usually involves face-to face interviews.

- As expected from above, the temporal resolution of data collected in this way is very course. For example, in most countries censuses are conducted once in 5 years or 10 years.
- Subject to economy of a country the exercise can be very expensive.

Thanks to the pervasiveness of mobile phones and other mobile devices, it is now possible to get unprecedented spatio-temporal digital footprints telling *when* and *where* people are at all times. If I can successfully mine from this data the information needed by authorities such as internal migration, then reliance on large scale surveys or census could slowly be reduced because these new data sources offer several attractive advantages. For one thing, the effort and cost of data collection is nearly negligible when compared to that of a census. Most importantly, since this kind data is always available, it is possible to get on demand a picture of some population attribute of interest such as internal migration. It therefore comes as no surprise that recently there has been an influx of research work exploiting mobility datasets with intentions to generate various demographic characteristics of the population pertinent to urban planning. This far, most studies have demonstrated that its possible to establish the important places, in most cases *home* and *work* in peoples lives from Cellular Call Detailed Records(CDR). For instance, in [IBC+11] they used CDR data to identify home and work location while in [BCH+13] they were able to characterize human mobility at metropolitan scale by also using a similar dataset. I defer detailed discussion of previous work to the related work section, but for now it suffices to say that quite a number of researchers have achieved promising results regarding mining human behavior traits from mobile device generated data. However, due to complexity and highly heterogeneous nature of human behavior this area is still open to further research. Moreover, to the best of our knowledge none of the previous works has particularly contributed on discovering *residence change* from CDR data. This is the concern of this work.

In this work, I investigate the potential to use CDR as surrogate *residence history* of a person to discover residence change so that I can ultimately infer internal migration patterns. CDRs are routinely collected by cellular network providers to help operate their networks, for example, they can use them to identify congested cells in need of additional bandwidth. Each CDR contains information such as the time a voice call was placed or a text message was received, as well as the identity of the cell tower with which the phone was associated at that time. This information can serve as sparse and sporadic samples of the approximate locations of the phone's owner.

Although many algorithms have been developed to mine semantic places, including automatically discovering home and work location from trajectory data, the problem of discovering residence change despite being closely related is different from that of just identifying home location. Even demographers[LD] have reckoned that the task of discovering residence change is not trivial, particularly considering that the residence history to work from is usually incomplete. In this paper, the overall goal is to explore the use of CDRs generated by mobile phones to infer change of place of residence (home). Specifically, I investigate two research questions:

- Whether its possible to discover relatively permanent residence change from CDRs?
- If it is feasible to develop an algorithm to automatically discover such information?

In order to answer these questions, I first provide a rigorous definition of what I call *the residence change discovery problem.* Next, I propose a novel *sequential spatio-temporal clustering technique* which I call *MoveSense* based on *Hartigan Leader Clustering* to solve this problem. To validate the proposed technique, I carry out experiments using a massive spatio-temporal dataset covering four months of call activity for 16 million users from a leading Cellular provider in Dhaka, Bangladesh. I conducted the experiment in context of *unsupervised anomaly detection* after noting a close resemblance between the residence change discovery problem and that of unsupervised anomaly detection. Thus, users who changed residence are analogous to anomalous elements in a dataset while those who did not change residence are analogous to normal elements in the data. Based on this reasoning, I applied this technique on the data to classify users in the dataset into these two groups. Results from the experiments show that our technique performed well with average *detection rate* of 71 percent, 68 percent and 72 percent across the three categories of datasets I tested it on. The key contributions of this work can be summarized as follows:

- I adapt the problem of residence change from a traditional population census perspective and formally define it in the context of sparse trajectory data.

- I develop a novel spatio-temporal clustering technique based on Hartigan leader clustering requiring two parameters to automatically discover residence change.

- I carry out experiments on a real big spatio-temporal dataset to validate the proposed technique.

The rest of this chapter is organized as follows: in the next section I provide some background and formal definition of the residence change discovery problem. In section 3.3, I present details of the proposed technique. In section 3.4, I present details about the dataset as well as experimental set up and results. In section 3.5, I review previous research work. Finally, in section 3.6 I give concluding remarks and discuss future work.

3.2 Problem Formulation

Aside from CDRs which is the focus in this study there are other large scale spatio-temporal datasets depicting traces of human mobility. For example, GPS enabled devices generate what can be considered as high resolution trajectories because the sampling rate is usually high, for instance GPS on a mobile phone can be set to record location information every second. In fact, most of the mining of human mobility patterns has been applied on this kind of data, for example in [ZZXM09] [PBKA08] [CCJ10]. On the other hand, CDRs though equally large-scale can be considered as sparse, sporadic and low resolution human mobility traces because data is only recorded when a person makes a call or sends a text message. CDRs are collected by telecommunication providers when cell phone users use one of their services, most commonly when they initiate or receive a voice call or text message. Each time a user participates in a telecommunication interaction, his or her position can be approximately inferred by knowing the geographic coordinates of the nearby Base Transceiver Station(BTS) tower that has processed the call.

Clearly, from the preceding explanation CDRs can be considered as surrogate *residence history* for residents in a city. Thus, given this type of data the main intention is to automatically discover residence changes. On the surface, this is a seemingly trivial problem. However, even demographers and census practitioners [Man70] have long before acknowledged that the task of accurately capturing migration events can be notoriously hard. Some of the factors which can make this problem complicated are outlined below:

- Residence history is often incomplete. This is even more true in this case when I are considering spatially sparse CDRs. Loosely speaking, longer residence history would result in better estimations. For example, in a census questionnaire, the line of questions starts from a persons' place of birth inorder to determine if they changed residence at some point.

- Change of residence can be repeatable. This is a known phenomenon in demographic context. In other words, a person can change residence multiple number of times within a given period of time depending on motivation. As such there is a potential of missing some episodes of those changes.

- Though rare, there are some scenarios where a person has multiple residences. For example, students who stay at a boarding school, a commuter who stays in the city within the week and returns home over weekend.

- Behavior related to residence change can be highly heterogeneous with heavy variations across countries, regions and cultures. As such it may be difficult to devise an all-round approach to solve the problem.

3.2.1 Preliminary definitions

Before formally stating the residence change discovery problem, I first provide preliminary definitions and notations. I let H represent a persons' *residence history* and define it as a sequence of spatial-temporal points so that $H = \{(s_0, t_0), (s_1, t_1), \ldots, (s_{n-1}, t_{n-1}) \,|\, s_i = (x_i, y_i) \in \mathbb{R}^2\}$ where s_i represents location in Euclidean space at time t_i and $t_i \leq t_{i+1}$. Because I are considering CDR data, this definition assumes irregular sampling rate in both the temporal and spatial domain. For brevity, I will use h_i to represent the tuple (s_i, t_i) whenever need arises. I define $d(s_i, s_q)$ as a distance function between any two locations s_i and s_q in H. I define length of residence history as the number of elements in the residence history and denote it by H_d. It is dependent on the units of time used. This measure is mainly useful in the current case when data can be sparse in the temporal domain and also because I are interested in long term mobility. Given H, I can further define a time invariant set of *unique locations* $H_L = \{(s_1, f_1), (s_2, f_2), \ldots, (s_n, f_n) \,|\, f, i, q \in \mathbb{Z}, s_i \neq s_q\}$ where f_i represents frequency of occurrence of s_i in H. Considering that our focus is CDR data which is extremely sparse in the spatial domain this notation is justified and practical unlike in the case of GPS based trajectory.

I now define the notion of **Place of Residence**. Loosely speaking this represents a region where the person is usually found. Ideally, this is supposed to be a physical street address but in our case I represent this by a collection of points(cell towers) which is a subset of H. I use a collection of cell towers rather than one tower because I consider place of residence as a region rather than a single place. Thus, this place of residence includes all the regular places (including work) where the person visits. This is reasonable considering that in the dataset I consider (see Section 3.4) majority of people usually travel short distances. Moreover, for the purpose of detecting migration, use of a region rather than an exact point does not degrade accuracy.

Definition 3.2.1 (Place of Residence) *A place of residence $R = \{h_i^1, \ldots, h_j^m\} \subseteq H$ of length m is a set of temporally ordered points of H with index $1, \ldots, m$ defined based on two parameters; δ and ϵ as explained below:*

- ***temporal threshold*** $\delta \in (0, 1]$ *so that if I let $I = [a, b]$ represent the time interval over which R is defined then $I_d = \delta.H_d$ where I_d is length of the interval I over which R is defined and H_d is length of whole residence history.*

- ***distance/ neighborhood threshold*** $\epsilon \in \mathbb{R}$ *so that for some $c \in \mathbb{R}^2$ selected as centre $\{h_i \in R, d(c, h_i) \leq \epsilon\}$*

Example 3.2.1 (Illustration of temporal parameter δ and H_d) *Lets say I have CDRs covering a 30 day period in the month of August. I decide to use days as units of time. I have in the dataset a user X with residence history without gaps from $august - 1$ through to $august - 30$ as follows: $H^X = \{(s_0, aug - 1), \ldots, (s_{29}, aug - 30)\}$ so that $H_d^X = 30$ where s_i represents location. There is another user Y with disjointed history given by $H^Y = \{(s_0, aug - 1), (s_1, aug - 2), (s_3, aug - 8), (s_4, aug - 30)\}$ so that $H_d^Y = 4$. If I pick $\delta = 0.5$, then the choice of a legitimate interval user X would be $I = \{aug - 1, \ldots, aug - 15\}$ so that $H_d^X.\delta = 15 days$ while for user Y a legitimate interval would be $I = \{aug - 1, \ldots, aug - 2\}$ so that $H_d^Y.\delta = 2$. See Fig. 3.1 for illustration of both users.*

3.2.2 Problem formulation

Following the definitions above, I are now ready to provide a formal definition of the problem. The problem at hand is to determine if there has been residence

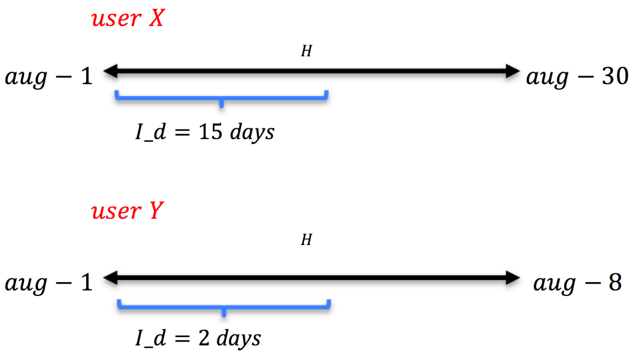


Figure 3.1: Illustration of setting of temporal parameter δ

change over a period defined by residence history based on parameters δ and ϵ.

Definition 3.2.2 (Residence Change Discovery) *Given residence history $H = \{h_0, h_1, \ldots, h_n\}$, lets suppose that I can find time intervals $I_1, I_2, \ldots, I_n$ such that $R_{I_1}, R_{I_2}, \ldots, R_{I_n}$ are the place of residences that correspond to these intervals, then I shall say there has been residence change if I can find at least one pair of residences that are spatially disjoint. More formally, $\{(R_{I_i}, R_{I_j}), i \neq j | R_{I_i} \cap R_{I_j} = \emptyset\}$*

3.3 Method

In this section, before I delve into the details of the approach I first provide a preamble on how a population census is conducted to facilitate understanding of some key terms and concepts which I borrow from the census approach.

3.3.1 Background: motivation from census

One recommended topic in most census questionnaires is *geographic characteristics of the population* [Div08]. The objective of this category is not only to determine number of people living in a given geographic unit but also to investigate migration, both internal (within country) and international. In order to achieve this, they ask questions about place of residence(usual residence, previous residence, duration of residence) and place of birth. Based on collected data and a further set of rules related to duration of residence and administrative boundaries the census Statisticians can establish if a particular person *changed place of residence* over the reference period and whether that change constitutes migration or not . For instance, they have to decide whether change of residence within a *county* should be considered migration or not. Clearly, this is the *spatial* component of migration. The *temporal* component is also important, for instance if authorities use a 6-month window, it means if a person for some reason changed addresses twice within a 6-month period it will not be considered as migration. In general, migration can be defined as a relatively permanent change of residence which involves crossing of some designated administrative boundaries.

The first important difference between this work and that of traditional census case is that in this work, I focus on detecting residence change and not necessarily migration although I are aware that residence change is the basis of migration. Secondly, I also recognize that it is relatively easier in the census case to solve the problem of residence change because they deal with a respondent either face to face through a personal interview or the respondent answers the census questionnaire by themselves, as such they have access to a more detailed residence history. On the other hand, in our case the user is anonymous, all I have is a dataset whose intention of collection was not even to capture residence change. In the next two sections, I use this understanding to formulate a solution for the residence change problem in the context of CDR data.

3.3.2 Preliminaries

Given the residence history, it is possible to generate as many intervals of this nature $I = \{[a, b] \ \mid \ t_0 \quad \leq a \leq b \leq t_n \subseteq [h_0, h_n]\}$ as our history can allow so that I could have $I_1, I_2, \ldots, I_n$. I now introduce **cluster** defined over any of the intervals I_q. A cluster $c_k = \{h_i^1, \ldots, h_j^m\} \subseteq H$ is a subset of the residence history whose members are systematically selected. I can also define a set of unique locations for cluster $L_k = \{(s_i, f_i), \ldots, (s_j, f_j)\}$. In addition, a cluster possesses three more important attributes : first, *Cluster centre* μ_k is defined based on equation 3.1 as the *weighted mean centre* weighted on f_i for each unique location s_i in the cluster while N_k is the count of members in the cluster c_k defined in equation 3.2. Finally, I also define a weight w_k based on total counts in the cluster as shown in equation 3.3.

$$\mu_k = \frac{\sum_{i=1}^{N_k} f_i s_i}{\sum_{i=1}^{n_k} f_i} \quad (3.1) \qquad N_k = \|\{s_i | d(\mu_k, s_i) \leq \epsilon\}\| \quad (3.2) \qquad w_k = \sum_{i=1}^{N_k} f_i \quad (3.3)$$

For brevity, I will denote a cluster $c_k = (\mu_k, N_k, w_k, L_k)$ to represent these key attributes of a cluster. A cluster is generated based on some time interval I_q as elaborated earlier and two parameters δ and ϵ which have been already introduced in section 3.2.1.

Next, I define *usual place of residence* and denote it with same notation as place of residence R. The addition of *usual* is to emphasize that this is a place where the person is usually found despite episodes of temporally absence. This can as well be considered as *home*. I now define R based on a cluster c_k. I further define the concept of *current place of residence* and *previous place of residence* which are simple extensions of usual place of residence based on time conditions.

Definition 3.3.1 (Usual place of residence) *For a nonempty set of clusters $C = \{c_1, c_2, \ldots, c_k\}$ defined on some temporal interval $I_q = [a, b]$, a usual place of residence R is a cluster c_k such that $w_k = max\{w_1, w_2, \ldots, w_k\}$.*

Definition 3.3.2 (Current place of residence) *This is a particular case of place of residence signifying where a person is currently staying. Given R_{I_q}, if the endpoint of interval I_q coincides with the endpoint of the whole residence history (h_n) i.e. $b = t_n$ then I shall call R the current place of residence and denote it by $R^c_{I_q}$.*

Definition 3.3.3 (Previous place of residence) *For a current place of residence $R^c_{I_q}$ defined on $I_q = [a, b]$. If I can define another interval $I_s = [u, v]$ such that $I_s \leq I_q$, further if I can find a usual place of residence R_{I_s} over this interval, then I shall call R_{I_s} a previous place of residence and denote it by $R^p_{I_q}$*

3.3.3 MoveSense: a sequential spatio-temporal clustering technique

I first reiterate that our primary goal is to probe a persons' residence history and determine if at some point over this period, they changed place of residence. Although its generally legitimate to define as many time intervals as possible over which to investigate residence, I are always limited by data availability. In fact, doing so would be advantageous because according to demographic literature e.g., see [WIL97], theoretically a person can change residence multiple times within a given period and in censuses they begin this kind of investigation by referring to place of birth. Ironically, the demographic community also recognizes that changing residence is a rare, so in some cases this phenomena modeled as a Poisson process to reflect this fact. Based on these two reasons, I design our solution assuming that our data is temporally sparse and therefore I consider the special case of splitting the history into two time intervals. I still want to stress that our method can be easily generalized to more than two intervals when data is available.

This approach is based on the *Hartigan Leader algorithm*[Har75]. This is a simple clustering approach which doesn't require predefining the number of clusters. Another advantage is that it only makes one pass through the data which is important considering scalability as our target is dealing with big data archives in the range of terabytes in size. Furthermore, the work in [IBC$^+$11] demonstrated that application of this algorithm worked well in identifying important locations in peoples lives from CDR data. In one of its original form, the Hartigan leader algorithm proceeds as below:

1. Choose a cluster threshold value
2. make the first item centroid of first cluster
3. For every new element:
4. Compute distance between the element and every cluster's centroid
5. If the distance between the closest centroid and the new element is smaller than the chosen threshold, then recompute the closest centroid with the new element
6. Otherwise, make a new cluster with the new element as its centroid

The idea behind this approach to pass through the data sequentially is what makes it suitable for our problem because our goal is to probe residence history, essentially I would want to start probing from the beginning of history and move sequentially to the end. Following this reasoning, I present our sequential spatio-temporal clustering technique which I call *MoveSense* as shown in algorithm 1. In the following, I present a description of the technique.

In most cases, the time element in trajectory data (including CDRs) is a time stamp. For example, the level of precision can go up to *seconds* or *milliseconds*. Considering that the nature of our investigation is to probe place of residence which is established over a relatively longer period, the first step is to pre-process the data to generate courser residence history. In this step, I aggregate location component (s_i) over the course time units. I use the function $aggregateBySecondaryTime(H, \hat{t})$ for this purpose. The inputs of this function are the raw residence history (H) and the preferred secondary unit of time $\hat{t}$. In this function, within a single secondary time unit, I pick location(s_i) with maximum frequency of appearance. Once I get secondary history $\hat{H}$, I use it as input in the main procedure $probeResHist(\hat{H}, \delta, \epsilon)$. This procedure emulates Hartigan algorithm as described earlier. I ensure that history $\hat{H}$ is sorted based on time. The initialization stage sets first cluster (c_1) to the first element in $\hat{H}$ and corresponding attributes (μ_1, N_k, w_k, L_k) of the cluster are also set accordingly. Next, I call the $expandCluster(C, \epsilon)$ procedure.

Algorithm 1 MoveSense-Residence Change Discovery

```
Require: $H, \delta, \epsilon, \hat{t}$                          ▷ $\hat{t}$ secondary unit of time
Ensure: $R$,current/current & previous place of residence
$N_i, j \leftarrow$ counter for cluster members
$\mu_k \leftarrow$ cluster centre
$w_i \leftarrow$ counter for incrementing cluster weight
$C \leftarrow$ set of clusters
procedure EXPANDCLUSTER($C, \epsilon, s_i$)
    $C = \{c_1, c_2, \dots, c_k\}$                               ▷ set of clusters
    $s_i$                                                        ▷ location under evaluation
    for all $c_i \in C$ do
        $dist \leftarrow \min d(s_i, c_j), j \in \{1, \dots, k\}$
        if $dist \leq \epsilon$ then
            $s_i \leftarrow c_j$                                 ▷ assign to closest cluster
            $N_j ++$                                             ▷ increment count
            $\mu_j \leftarrow updateWeightedCentre$
        else
            $j \leftarrow k + 1$                                 ▷ create new cluster
            $\mu_j \leftarrow s_i$
            $N_{k+1} = 1$                                        ▷ set count to 1
        end if
    end for
    update $C$
end procedure
function PROBERESHISTORY($H, \delta, \epsilon, \hat{t}$)
    $\hat{H} \leftarrow$ AGGREGATEBYSECONDARYTIME($H, \hat{t}$)
    $\hat{H} \leftarrow sortedByTime$
    $\hat{t}_{threshold} \leftarrow \delta.\hat{H}_d$            ▷ index of based on $\delta$
    $C \leftarrow \emptyset$
    $\mu_1 \leftarrow s_0$                                       ▷ Initialise cluster centre
    $w_1 \leftarrow 1$                                           ▷ Initialise cluster weight
    $C \leftarrow add(c_1)$
    $R \leftarrow \emptyset$                                     ▷ Residence(s)
    for $i \leftarrow \hat{t}_{start+1}, \hat{t}_{threshold}$ do
        $C \leftarrow$ EXPANDCLUSTER($C, \epsilon$)
        if $i = \hat{t}_{threshold}$ then
            $C \leftarrow$ FINDRESIDENCE($C, \epsilon$)
            $R \leftarrow add(c)$
        end if
    end for
    for $i \leftarrow \hat{t}_{threshold+1}, \hat{t}_{end}$ do
        $C \leftarrow$ EXPANDCLUSTER($C, \epsilon$)
        if $i = \hat{t}_{end}$ then
            $C \leftarrow$ FINDRESIDENCE($C, \epsilon$)
            $R \leftarrow add(c)$
        end if
    end for
    return $R$
end function
```

The objective of this procedure is to evaluate distance between the current point under examination and cluster centres of all existing clusters. If the minimum distance to centres of clusters is within ϵ then the current point is assigned to that cluster, otherwise a new cluster is created. This procedure is called repeatedly until I hit upper bound of the first interval; a point in time where $I_d = \delta.\hat{H}_d$. At this stage, I evaluate the available clusters to determine place of residence. I use the function $findResidence(C)$ for this. I are defining usual place of residence based on frequency of stay, therefore this function simply picks the cluster with maximum weight (w_k) and set it as *usual place of residence* .From here I drop the rest of the clusters and continue evaluating elements in the rest of the database(interval 2) in a similar fashion as before. When I hit upper bound of the second interval I again call $findResidence(C)$ to evaluate the clusters and determine place of residence in this interval. Finally, I evaluate these two based on definition 3.2.2 to determine if they are equal and consequently decide if there is change of place of residence.

3.4 Experiments

In this section I present details of the experiment I conducted to evaluate our technique. The experiment is based on a CDR dataset from a leading mobile cellular operator in Bangladesh. The primary task in the experiment was to classify users in the dataset as having changed place of residence or not. In the rest of the section, I present details of the dataset, experimental set up and results of the experiment.

3.4.1 Description of the dataset

The data is from a leading cellular phone operator in Bangladesh. It was collected in 2013 and covers the months of *August,September, November* and *December*. The data mainly include details of a call event: time of call, cell tower id and latitude and longitude of the cell tower involved in the call. The map in Fig. 3.2 shows location of cell towers.

The administrative structure of Bangladesh consists of seven(7) divisions at the top of the hierarchy. Below each division there are districts, currently there is a total of sixty four(64) districts which function as county. Then, there are there 488 Upazilas below the districts. There exists other structures below the Upazilas but I do not mention them here because they are outside the scope of our study. As shown in Fig. 3.2 the available data mainly covers two divisions; *Dhaka* and *Rangpur*.

The dataset has more than 3.5 billion call events. There are about 16 million unique users with varying levels of number of events over the four months period. I computed length of residence history for each user using the *aggregateBySecondaryTime* function with *days* as secondary units of time. I then plotted the empirical cumulative density of the number of days in history (see Fig. 3.3). I also characterized typical distances covered by a user in a single day. For this, I used a measure called *mean radius of gyration*. I computed it by getting maximum distance covered in a single day and then averaging over the whole history (see Fig. 3.4).

I present residence history of what I consider to be three representative groups of user categories based on geometric structure and complexity of their trajectories as shown in figures 3.5, 3.6 and 3.7. The first and second category consists of users whose trajectories show some kind of regular geometric pattern while the third category represents users whose trajectories has an irregular pattern. Figure 3.5 seem to suggest that the user may not have changed place of residence at all as most of the cell tower locations in their history are almost perfectly stacked around a small geographic region. Note that in scenario 1 (Fig. 3.5a) of this category, the user seem to have two concurrent locations with relatively same number of call events suggesting this user may be a commuter, however this does not suggest change of residence as the separate regions seem to be contacted concurrently. In the second category (see Fig. 3.6) both Fig. 3.6a and Fig. 3.6b clearly shows that there is a simultaneous temporal and spatial offset, which seem to indicate that the user may have moved permanently because there is almost no temporal overlap of the cell locations in history. The third group (see Fig. 3.7) on the other hand shows a very irregular pattern so that its visually impossible to discern whether this user may have changed place of residence or not. Incidentally, I use this understanding to generate a validation dataset for evaluation of our technique.

3.4.2 Unsupervised classification of users

In this experiment I apply our technique to the dataset described in the previous section. The main task is to classify all users into two classes; those who changed residence and those who did not. I carry out this experiment in the context of *anomaly detection* as I will show in the following section that the residence

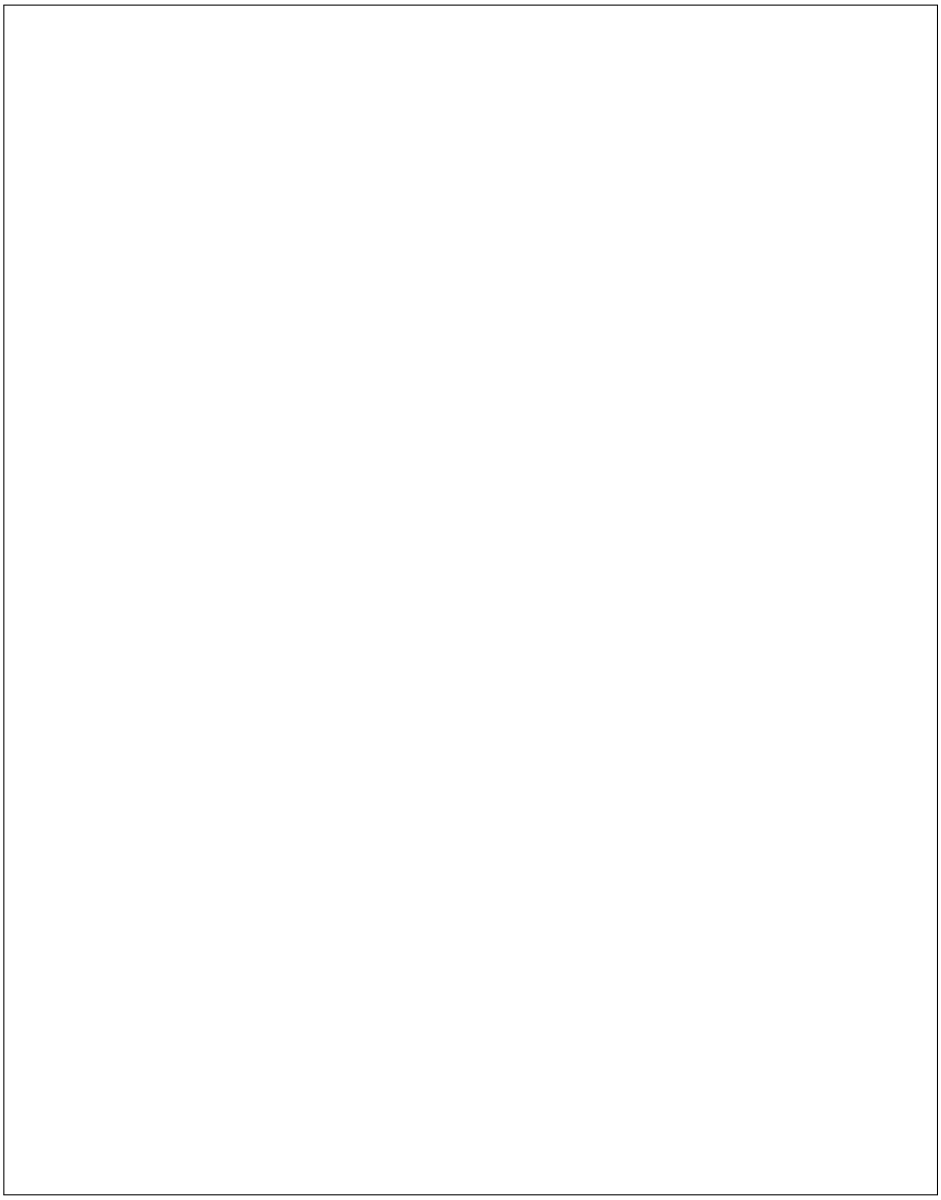


Figure 3.2: Location of Cell Towers in Bangladesh

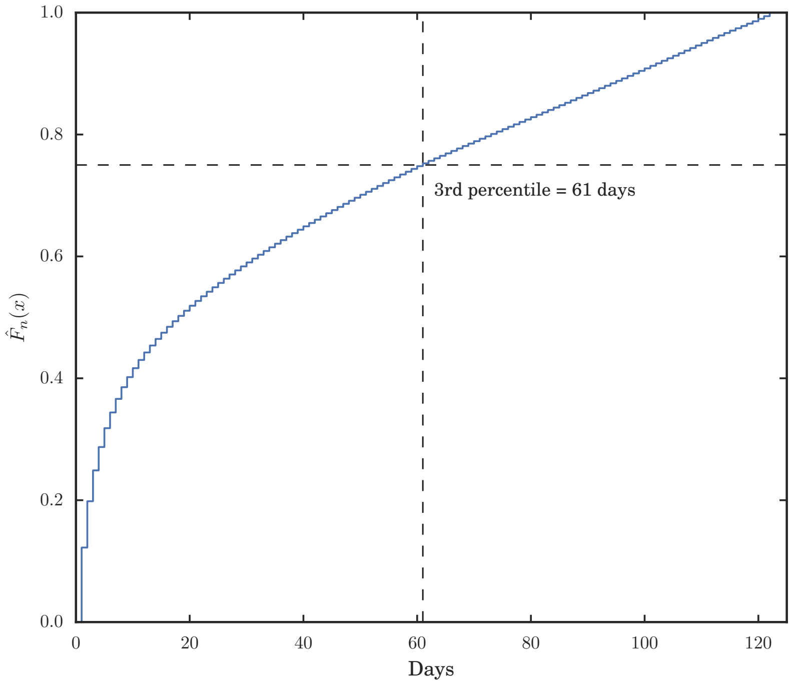


Figure 3.3: Cumulative frequency distribution of number of days

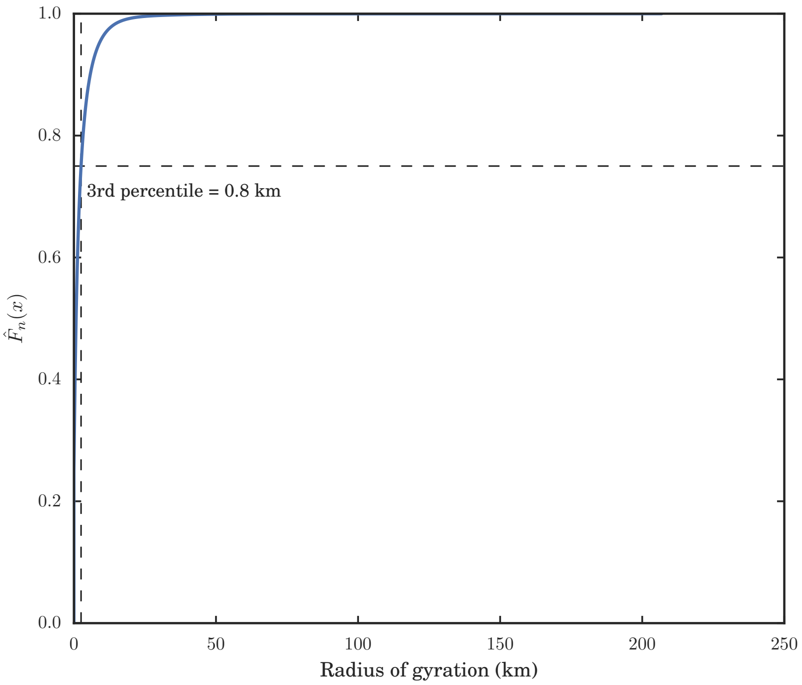
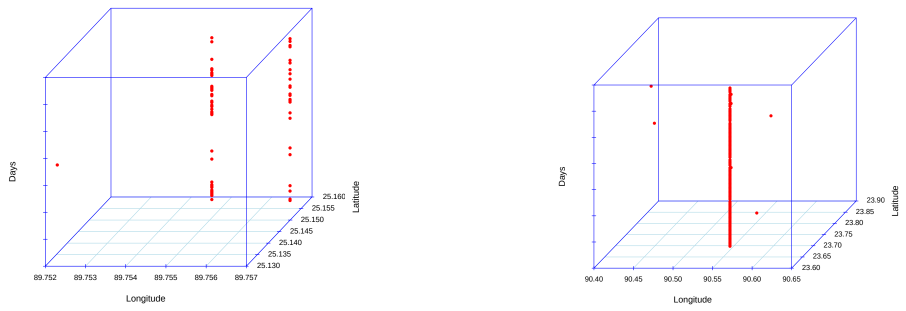


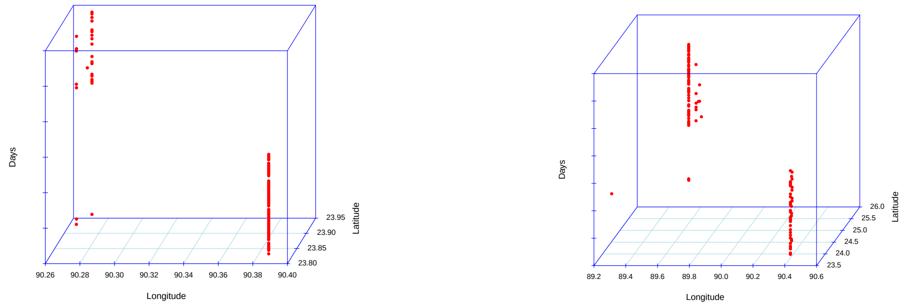
Figure 3.4: Cumulative frequency distribution of radius of gyration



(a) scenario 1

(b) scenario 2

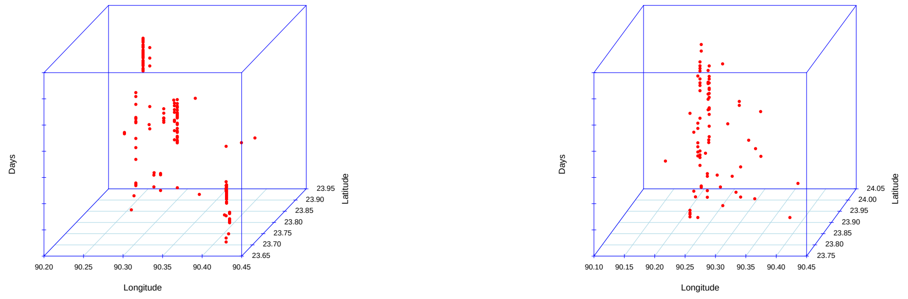
Figure 3.5: Regular mobility pattern: suggestive of no residence change



(a) scenario 1

(b) scenario 2

Figure 3.6: Regular mobility pattern: suggestive of residence change



(a) scenario 1

(b) scenario 2

Figure 3.7: Irregular mobility pattern: non-suggestive

change discovery problem has characteristics that resemble that of anomaly detection. In the following sections I present details about the approach, results and how I evaluate performance of our classification.

Approach and rationale

I noted that the residence discovery problem closely resembles that of *unsupervised anomaly detection*. In the unsupervised anomaly detection problem, the input is usually a large unlabeled dataset where most of the elements are normal and there are some anomalous elements buried within the dataset [Por00] and the task is to detect these anomalous instances. This is equally true in the present problem: while the probability to migrate or change place of residence varies greatly among individuals and societies, it is fair to say that this is a rare event. In this regard, it is plausible to invoke the same assumptions which are made in anomaly detection algorithms. Most unsupervised anomaly detection algorithms make two assumptions about the available data. The first assumption is that normal elements in the data hugely outnumber the anomalous ones. Secondly, they also assume that the anomalous elements are inherently(based on some qualitative characteristics) different from the normal ones. As a result of these two assumptions the idea is that since anomalous elements are rare and different they will appear as outliers of some form in the dataset, so that they can be flagged. This brief description of unsupervised anomaly detection is adapted from [EAP$^+$02].

In regard of the above, in this proposed technique, users who changed place of residence are analogous to anomalous elements in a dataset while those who did not change residence are analogous to normal instances of a dataset. In a two-class classification, the anomalous instances are usually labeled as positive ($y = 1$) while the normal ones are labeled negative($y = 0$).

Experimental set up

Our experimental set up was driven by the need to understand the effect of the spatial parameter (ϵ) and length of residence history on the results. Due to the fact that the available data covers only four months, which is 122 days in total, I did not experiment with parameter δ, I just fixed it at 0.5. Since I have a large dataset of users I repeatedly sampled subsets of the data as training set to understand setting of parameters.

First, to understand the effect of length of residence history I had to decide on the minimum length of residence history to include. Based on Fig. 3.3 and experimentation I decided to include users with at least 30 days of residence history. This represents length of the average user in the dataset but also its about one third of the total history. I then added two more categories so that I had three in total which I call *Case 1* representing users with residence history between 30-59 days, *Case 2* with history of 60-119 days and *Case 3* with complete history (120 days +).

Second, I set the spatial parameter ϵ as follows. First, this parameter is essentially a spatial a threshold, it is crucial in our technique because how large or small it is set has a direct effect on how many data instances will be classified as positive. The first task was how to fix the minimum value of ϵ. I decided to set the minimum value based on geographic extents of the lowest administrative region in Bangladesh. I computed spatial bounds of the *upazila level* administrative region and found the minimum to be 14 km. Therefore I set minimum at $\epsilon = 14$ and then incremented arbitrarily to get a list of $\epsilon = 14, 20, 25, 30$ in order to understand the effect of increasing the threshold.

Validation and discussion of results

As mentioned earlier on the dataset originally does not have any labels related to users home location, let alone information indicating whether they changed home. Due to this plus other logistical challenges, I could not get authentic ground truth data which ideally would be home location of some user over the whole four month period.

However, as illustrated in Fig. 3.5, 3.6 and 3.7 it is possible to make a judgment based on geometry of the trajectory as to whether a person changed residence or not though with less certainty. I used exploratory techniques to visualize the data and I were able to generate what I call a *quasi-ground truth data* to test performance of our data. I created three test datasets corresponding to three categories I introduced in the preceding paragraph. For each category, there were 80 users in the test dataset.

To evaluate our technique, I used two primary indicators of performance: the *detection rate* and the *false positive rate*. The detection rate is defined as the number of positive instances detected by the system divided by the total number of positive instances present in the test dataset. The false positive rate is defined as the total number of normal elements that were incorrectly classified as anomalous divided by the total number of normal elements. Overall, these two indicators are reasonable because they measure the ability of our technique to detect percentage of people who moved and at the same time determine how many incorrect classifications are made in the process.

I calculate these values over the labeled test dataset and the results are presented in Fig. 3.8, Fig. 3.9 and Table 3.1. Overall, the results indicate that our technique performed reasonably well over the three categories of the test dataset. As I expected, across the three categories there are strong variations dependent on the spatial threshold(ϵ). The spatial threshold has the natural effect of reducing the number of people detected as having changed residence which is reflected in the results. Fig. 3.8 reveals that in all categories the highest detection rate is seen when the spatial threshold is at its minimum ($\epsilon = 14$) this in turn corresponds to higher false positive rate. This trend though contradictory seems to be in line with experimental results from other unsupervised anomaly detection algorithms based on clustering (e.g see [EAP$^+$02]) for example.

For each category, I computed average of each of the indicator(see Table 3.1) and somehow surprisingly when I consider average, detection rate seems to be high for $Case3$ which has largest number of residence history. However, overall these averaged results seem to suggest that $Case3$ with large history has best results since it has highest detection rate and the false positive rate is also low compared to $Case1$. This is consistent with our expectations that the more history is available the more reliable the results would be.

Next, I discuss peoples calling patterns as it pertains to whether they call from home, work or somewhere else as this would potentially have a bearing on the results. This relates to the typical distances covered by individuals during their daily mobility pattern as presented in Fig. 3.4. In [GHB08] they indicated that radius of gyration follows a fat-tailed distribution which seems to agree with results in the aforementioned figure. In our dataset, 75 percent of the users have a mean gyration distance of around 3 km, which indicate that most people's daily movement is limited to a neighbourhood of less than 5km. Consequently, in this current work I safely assumed that whether a person calls from home or work would not have significant effect on results of our technique.

Nevertheless, I wish to quickly mention that the results should be interpreted with caution as I only used quasi-ground truth data rather than actual to compute the performance indicators, as such there is a possibility that some of the errors may be propagated from our visual interpretation of the trajectories when creating the test data.

Table 3.1: Performance indicators averaged over ϵ values

Category	Detection rate(%)	False positive rate(%)
Case 1	70.8	19
Case 2	67.9	14.2
Case 3	71.7	14.6

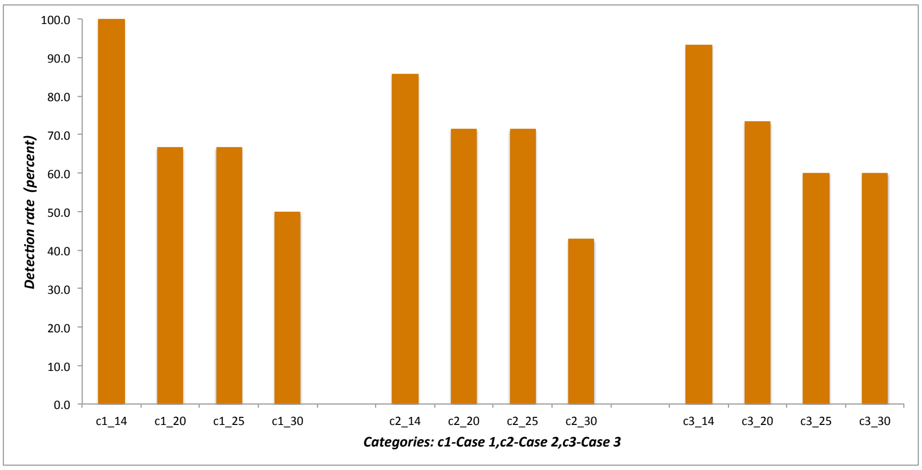


Figure 3.8: Performance Measures: Detection rates

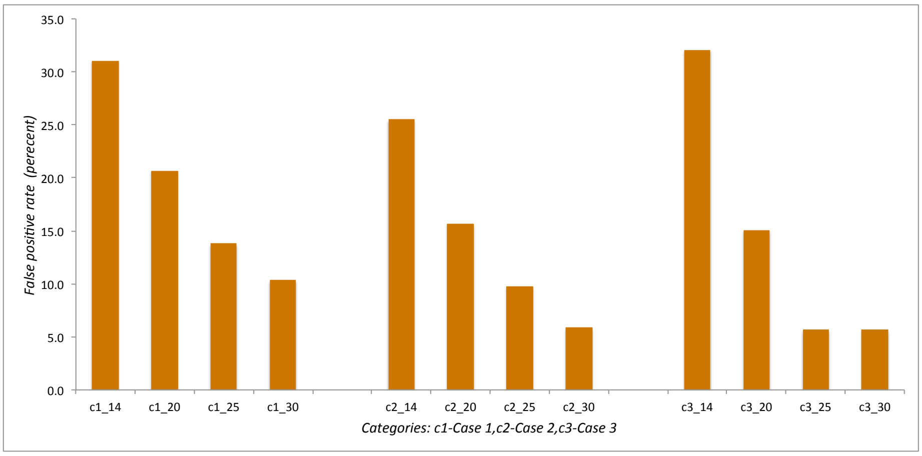


Figure 3.9: Performance Measures: False positive rates

3.5 Related Work

The approach I have presented in this work draws heavily from the conduct of *Population and Housing Census* particularly in developing countries as stipulated in [Div08]. I have taken this traditional approach and redefined the problem in the context of CDR data. As mentioned previously there are many differences between the traditional census approach and our method which I have proposed here. One important difference is that census uses questionnaire as such it is impossible to directly apply such methods on mobile phone dataset.

In the context of trajectory data, this work is closely related to [DIC14]. In their work, they used uncertain GPS trajectories to extract stay regions of animals. Their method is based on *dbscan algorithm* which is a widely used density clustering approach. They demonstrated the capability of their method to successfully discover stay regions. In fact, based on my preliminary publication [MMY+14] on this work, it seem to indicate that I started working on this similar problem around the same time. However, despite this similarity there are also key methodological differences. First, their work targets animal migratory behavior which although similar to human migration is inherently different. The most important difference being that animals migration depict more clear and obvious patterns than humans as such its relatively easy to deduce when animals migrate. Moreover, the problem formulation and the technique I propose in this work is based on *Hartigan leader clustering algorithm* and population census approach.

There are other works with weak connection to this current work. For instance, in [CDLLR11] [CDLR10] they consider origin destination patterns. In the latter work, they estimate origin-destination flows for weekday and weekend travel. This is related to this work but the difference is that they consider short-term trips in which the person returns to their home while in this case I'am interested in relatively permanent residence change. In demographic community, the work in [DLM+14] studied the use of CDR data to build dynamic population maps which is inherently a different task from what were are tackling. On the other hand, the work in [Blu12] is closely related to ours. In this work, they estimated internal migration from CDR data. However, although I have motivated this work by internal migration, in this current paper, I mainly focus on residence change which as previously mentioned does not always encompass internal migration. Moreover, in this work, I attempted to detect residence change from mid term data while in [Blu12] they had two years data.

3.6 Chapter Summary

In this work I first introduced the problem of residence change discovery in the context of CDR data from mobile phones. I argued that though this problem is loosely connected to that of *mining significant places* which has been heavily addressed in previous research works, it is inherently different and challenging when one consider spatially sparse CDR data which is case in the present research. As regards significance, the ability to automatically and reliably discover residence change from CDRs and mobile phone data in general would enable quick calculation of internal migration rates especially in low income regions where routine administrative data is nonexistent. Such information is highly valuable for urban planning and businesses. Moreover, even in developed regions residence change information could be useful in various other applications which deals with location based services.

I then formulated the residence change discovery problem by adapting from the census approach. I heavily borrowed from the terminology in censuses and adapted the problem in the context of CDR data. For instance, I noted that CDR data can be sparse both temporally and spatially. Also, in case of census they use administrative regions to make a decision on whether a person moved or not because their focus is strictly migration while in the current work I use a spatial threshold to enable us detect any change in residence. Based on this formulation I presented a sequential spatio-temporal clustering approach based on Hartiginan Leader clustering algorithm to solve the problem.

I conduct an experiment to evaluate the proposed technique. In the experiment, I carry out unsupervised classification of users to categorise them into two groups: those who changed residence and those who did not. I implemented this classification based on unsupervised anomaly detection approach. I used a large unlabeled spatio-temporal dataset from a leading mobile phone operator in Bangladesh. I experimented with a range of settings for the spatial parameter ϵ over three user groups categorized based on length of their residence history. I then validated the proposed technique on a quasi-ground truth dataset which I manually generated taking advantage of geometric characteristics of trajectories for representative users.

I used detection rate and false positive rate to measure perfomance of the proposed technique. The results from the experiments show that this proposed technique performed well with detection rates of 71 percent, 68 percent and 72 percent for the first,second and third group respectively. However, I are mindful of the fact that the labeled dataset I used to calculate these rates is not entirely ground truth and therefore the results should be interpreted with caution.

I set out to establish the potential to discover residence change from CDRs. I have demonstrated based on the experimental results it is possible to do so. I also demonstrated that the technique I developed can automatically discover users who changed residence under anomaly detection context with reasonale detection rates. Although my focus here is CDRs, I think the technique I have proposed is applicable with few modifications to other type of mobile phone datasets such as those from GPS because the algorithm only requires two parameters.

There are several directions for future work. First, in the introduction I mentioned that one important application of determining residence change from CDR data is the ability to obtain estimates of internal migration in a region, therefore as a case study I plan to use this approach to estimate internal migration rates from the CDR data from Dhaka, Bangladesh and compare results with official estimates from the Statistics Bureau. Secondly, I only measured performance of the proposed technique based on a quasi-ground truth data which I hand-generated. The drawback with this evaluation is that I did not test this algorithm on the more complicated cases, as such I would like to apply this algorithm on other mobility datasets such as that from Foursquare or other GPS based trajectories where it is relatively easy to obtain authentic ground truth data.

Chapter 4

Location Prediction

4.1 Introduction

During the past decade, *Location-based services*(LBS) have matured and become mainstay in most applications. Consequently, information about location of users has become a necessity. Thanks to the advent of cheap sensors, many mobile devices nowadays come with the capability to provide fine grained location trace of a user. However, beyond current location some LBS services can benefit from anticipating the location a user will visit in the near future. This is where *location prediction* comes in. For example, predicting a user's next location would be useful for urban planning in anticipating traffic levels. In addition and perhaps more importanly, location prediction can help to provide *approximate current user location* in cases where user device cannot provide exact current location because their device is incapable (device does not have GPS/Wifi capabilites) or due to GPS signal problems.

The problem of location prediction has been extensively studied[ZXYG13, PWJX14, SKJH06, DGP12, BSM10, CML11, KH06, NSLM12, KWSB04, SMM+11, WP12, EKK+13, Bur11, DDMGP14]. However, most of the works are *context specific*. A good example of such contexts is the type of data used: *resolution of location data* and whether *additional data* (e.g., social relationships data is used or not). For example, NextCell [ZXYG13] is a prediction system based on cell phone traces with the assumption that there is information about call patterns amongst users within the data. Other studies, e.g., Find me if you can [BSM10] use social relationships data to improve results of location prediction. Consequently, the conclusions drawn and the techniques developed cannot be transfered to new problems without substantial modifications. Furthermore, most of the previous works used very small datasets (in the range of thousands of users) to evaluate performance of their techniques and also study other issues related to performance. I argue that such small datasets limit the ability to study comprehensively behavioral factors which affect perfomance of location predictors. On the contrary, I use a dataset with millions of users.

In this study, I examine a *unique CDR dataset.* I believe the dataset is unique due to the following reasons: first, unlike majority of previous datasets studied in location prediction which come from developed countries, our data comes from Bangladesh which is a developing country. Secondly, it is very sparse in feature space (essentially, the data contains only three details about a call-event: caller/user identification, time of call and geographic location of base station station routing the call). Finally, it is a very *large data* with more than 16 million unique users. Given that smartphone penetration rate in Bangladesh is 6 percent [Luc14], clearly LBS services cannot thrive as most mobile devices do not have GPS. I envisage that location prediction techniques could be very useful in providing approximate user locations to various LBS applications thereby improving quality of life. However, the challenge with the present scenario is that in low income regions getting additional data (such as that from social networks) to improve location prediction is notoriously hard. It is against this background that I decided to explore ways to enhance performance of location predictors without requiring additional data.

The research question I tackle is how I can leverage big data to enhance performance of location predictors? I choose to experiment with *Bayes based predictor* because of ease of implementation and flexibility to add many beliefs without incurring huge computational expenses. In order to address this research question, I first carry out spatio-temporal analysis of user call behavior and call activity. I use insights from this analysis to propose an *enhanced bayes predictor* which leverages the large scale data. In summary, our work makes the following contributions.

- *Spatio-temporal analysis.* I conduct preliminary analysis to show lack of strong spatial autocorrelation of call activity. This result forms an important input into our approach.

- *Enhanced Bayes predictor.* I propose an enhancement to the naive bayes predictor under the context of *visit-revisit* modeling using a distance threshold and user *regular location* by levereging big data. This modelling approach builds on the *open-world modellling* introduced in [KH06] but in our case I do not require any external data and I also introduce other parameters. Results from the experiments I conducted reveal that although the results vary greatly across users, overall, the enhancements I propose improve the predictors' performance by 14 percentage points.

- *Large-scale evaluation of prediction perfomance.* This is the first time that such a big dataset has been studied in location prediction. I conduct extensive experiments to investigate performance of the predictor over a massive dataset with more than 3.5 billion calls and document how performance varies with various user characteristics.

The rest of this chapter is organized as follows. Section 4.2 provides a review of relevant literature. Section 4.3 reports on the results of spatio-temporal analysis and also characteristics of the dataset. Section 4.4 provides formal formulation of the prediction task. Section 4.5 presents details of the enhancement I propose. Section 4.6 reports on the experiments conducted and discussion of results . Section 4.7 concludes the work.

4.2 Related work

The increasing availability of human mobility datasets has led to an explosive growth of research work in location prediction. I have identified three core aspects of the single user location prediction research: *data, prediction task*, and *prediction approaches.* In this section, I review previous works on location prediction along these categories.

There are three important aspects which I consider as regards data-*resolution of location data*, *additional data* and *dataset diversity.* I can infer resolution of location data based on how the data was generated. Majority of the previous works[WP12, EKK$^+$13, KH06, DGP12, SMM$^+$11, MPTG09] conducted their experiments with location data generated from GPS/Wi-fi which I consider to be high resolution with only two works [KWSB04, ZXYG13] where they used CDRs which can be considered as low resolution. Regarding additional data, the inclusion of social relationships data like that from social networks has spurred a lot of interest with research works [PWJX14, BSM10, NSLM12] demonstrating that inclusion of such information enhances prediction results. I use $log_2 N$ to measure dataset diversity based on the number of unique users (N) in the dataset. In Fig. 4.1, I show that most of the previously studied datasets have far lower diversity compared to the data I study in this work.

The generic task is often to predict location of a user some time in the future. However, the subtle details such as what information I are starting with (current context), lead time or whether I are interested in temporal aspects (e.g., arrival time, stay duration) do matter. Most of the earlier works focused on the task of predicting the next place/destination a single user will visit. For example, in [NSLM12] they predict the next foursquare venue a user will visit while in [KH06] [MPTG09] they tackle the same task but the target user is a vehicle driver. Some of the studies e.g.,[DDMGP14] are specific on the lead time of prediction. In their study, they experiment with different lead times.

Several methods have been proposed for location prediction in different contexts (location data and prediction task). In an early analysis with GPS traces [AS03], Ashbrook et al. used Markov models to predict the next place that a user will visit. Song et al. [SKJH06] took a similar approach and evaluated domain independent predictors (including order-k Markov predictors) on symbolic location traces collected with Wifi. They concluded that order-2 Markov predictors gave the best results. Since then, Markov models and its variants (Mixed markov chain model, hidden Markov Model) have emerged as a popular choice for many research works[GKdPC12] [TCMA12] due to ease of implementation. Other probabilistic based techniques have also been used. In [KH06], they proposed a Bayesian based approach for predicting next destination based on predestinations. Similarly, in [DDMGP14] they use *naive bayes* and *kernel* based approach to predict user location on a smart-phone. The techniques in [KH06][DDMGP14] are similar to this work in that they are all grounded on Bayesian inference. However, the heuristics used to estimate prior probabilities do differ.

Clearly, the techniques developed and conclusions drawn from most of the works I have reviewed rely heavily on the nature of data studied (e.g., resolution of location data) and the prediction task. With respect to data and prediction task, our work is loosely connected to that in [ZXYG13]. However, there is a important difference regarding the data in that the CDR data they use contains *call pattern between users* which they use to generate social relationships between users which forms the basis of their method. As previously mentioned, I do not have such attributes in the dataset I study.

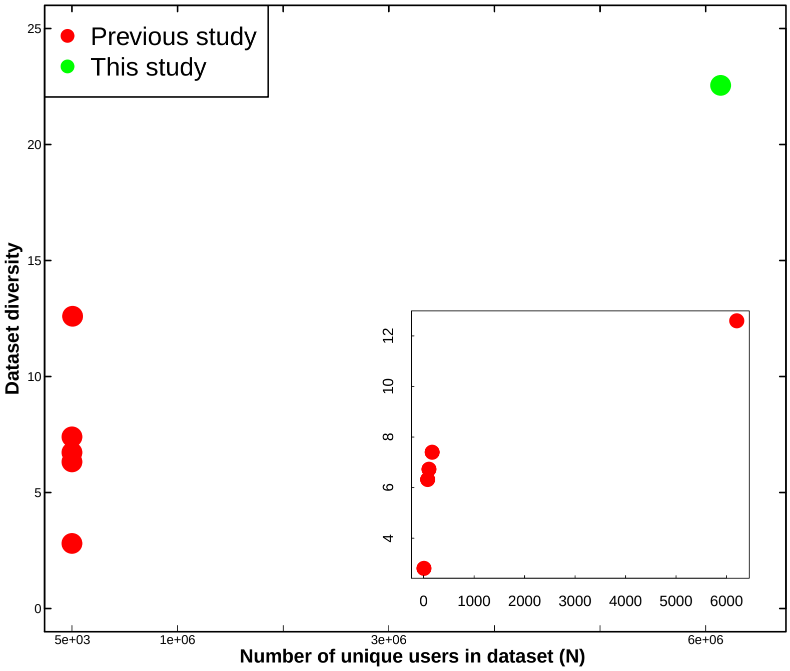


Figure 4.1: Comparison of dataset diversity in location prediction studies

4.3 The dataset

4.3.1 A primer on cell-ID positioning and Call Detailed Records(CDR)

In a cellular network, a *cell* is a geographical area covered by a *base station* -a piece of equipment that facilitates wireless communication between user device (UD) and a network. The cell covered by a base station can be from one mile to twenty miles in diameter, depending on terrain, population density and transmission power. A UD is always receiving message broadcasts by these base stations; thus, I can approximate its actual location using the geographical coordinates of the corresponding base station. Hence, the UD is assumed to be located at the base station coordinates independently of its actual position within the cell [TV04]. In this work, cell location refer to base station coordinates. CDRs constitute a sequence of places, related to a single user, where a call (or text message) was made thus describing the user's hop movements as they make calls. Due to personal nature of this data, the richness of available

Table 4.1: Background statistics of the dataset

Category	Description
Starting date	August 1, 2013
Ending date	December 31, 2013
Data gaps	October 1-October 30 2013
Number of cell towers	2101
Number of users	16,000,000
Number call events	3,5,000,000

attributes in a CDR dataset vary. For instance, the MIT Reality Mining dataset [ZXYG13] is a publicly available CDR dataset with rich information about call events (including caller and recipient of calls). On the other hand, in this work, I have access to what I consider to be a very sparse CDR dataset. In our case I define *sparsity in feature space* in terms of features (not considering those which can be derived) in a dataset available for prediction. Each call event has the following attributes: anonymised user ID, time stamp and BTS tower coordinates.

4.3.2 Background characteristics

The data is from a leading cellular phone operator in Bangladesh. It was collected in 2013 and covering the months of *August, September, November* and *December*. For some logistical reasons, I do not have data for October. As pointed out earlier on, I consider this data to be a case of sparse CDRs. For example, for a call event, I do not have attributes indicating where the call is directed to. In Table 4.1, I present background characteristics of the data. In Fig. 4.2 and Fig. 4.3, I present calling pattern of two random users in order to emphasize diversity of users in the dataset (user shown in Fig. 4.2 seems to be a more frequent user while the one depicted in Fig. 4.3 makes very few calls). Although the total number of unique users is very large, usage patterns and length of history varies greatly with the mean history being *35 days*. For this study, I select *6,141,508* users whose history length is greater than 35 days to ensure that I have consistent users.

4.3.3 Spatio-temporal analysis of call activity

I carried out exploratory analysis to understand how call activity (number of calls made at a base station per unit of time) varies in both space and time. Strictly speaking, in this current work, I study the spatial and temporal compo-

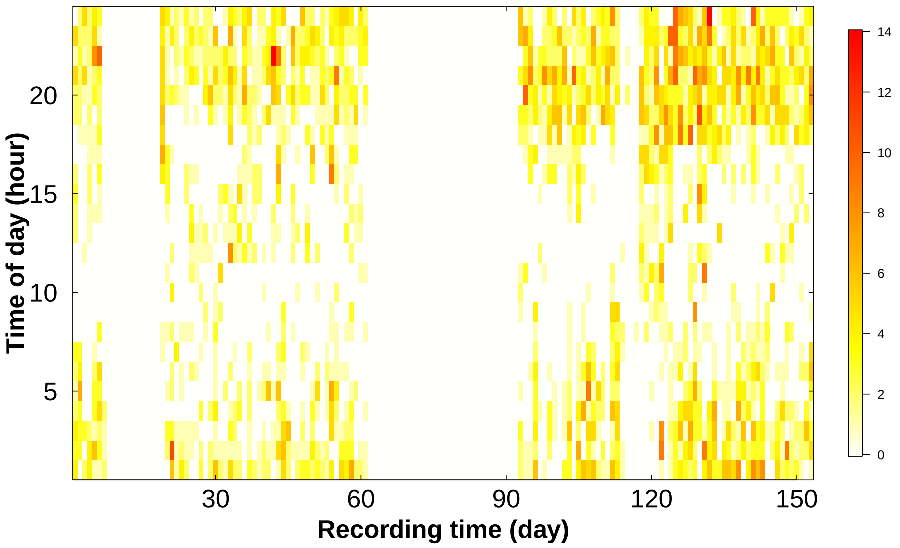


Figure 4.2: Call pattern of sample user: Frequent user (The intensity of the color ramp represents count of calls)

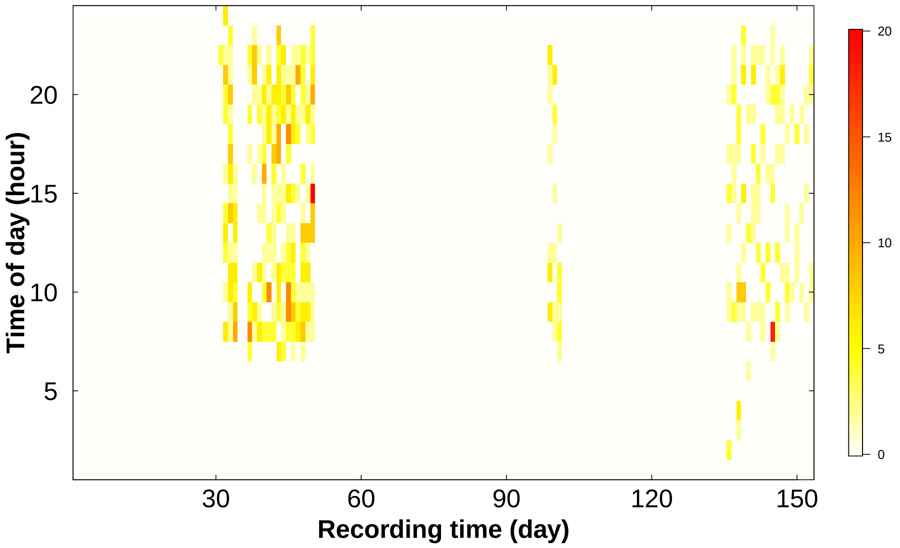


Figure 4.3: Call pattern of sample user: sparse use (The intensity of the color ramp represents count of calls)

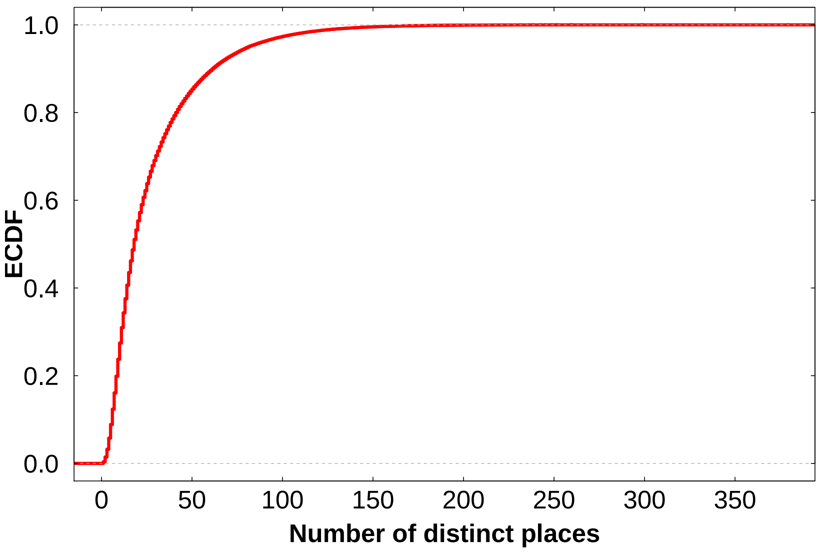


Figure 4.4: ECDF of number of distinct cells visited by a single user

nents separately although the spatial analysis do have a temporal component. In order to understand variations in space, I measure spatial autocorrelation of call activity. Spatial autocorrelation measures how a variable is correlated to itself in space. I choose to use *Moran's I* because it is one of the most commonly used indices for this purpose based on our literature search. I consider call activity at two levels of temporal resolution: *hour* and *day*. For this work I did not do an exhaustive analysis but rather as a proof of concept, I picked arbitrarily a single day *August 1,2013*. I computed total calls made on this day disaggregatd by base station. For hour, I picked *12 noon* and computed call counts in a similar fashion. The choice of 12 noon is because it is one of the peak hours. I then just computed Moran's I on these two datasets and present the results in Fig Fig. 4.5 and Fig. 4.6. Due to the clear spatial separation between the greater Dhaka region (shown in green) and rural northern regions (shown in red) which would influence the results of Moran's I, I computed this measure separately for each of them.

For the northern region, for both *hourly* (see Fig. 4.5) and *daily* (see Fig. 4.6) results, the p-value is very large (when considered at 95 percent significance level) which which suggests lack of spatial autocorrelation. On the other hand, for Dhaka region, in both cases the p-value is very low and therefore suggests positive (due to positive value of Moran's I) spatial autocorrelation. However,

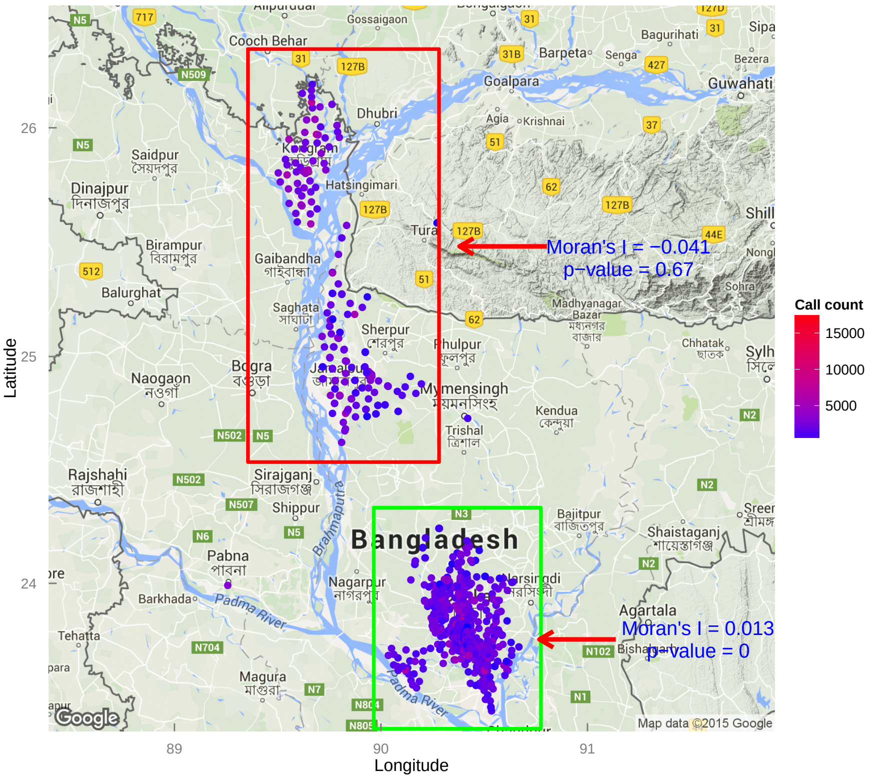


Figure 4.5: Total calls made at 12 noon of August 1 2013

I note that the values 0.013 and 0.074 for hourly and daily respectively indicate that the strength of spatial autocorrelation is very weak. Due to this lack of strong spatial autocorrelation, I think it is reasonable to think of base station as points of interest (POI) which attract people differently at different times. This is the insight I incorporate into our scheme for enhancing Bayes based predictor.

In the temporal dimension, I use hour as the unit of time. The question I ask is how much does call activity vary across the hours of the day. I computed for the whole dataset, total count of calls made at each hour disaggregated by day of the week. The choice to disaggregate by day of the week was simply to understand variations across days as well. I utilise an image plot shown in Fig. 4.7 to understand this question. As expected, peak call times seem to coincide with regular working hours. I observe that calling peaks from late morning around 11 to 17 hours. Across the days, Monday and Tuesday seem to have more calls than the rest of the days.

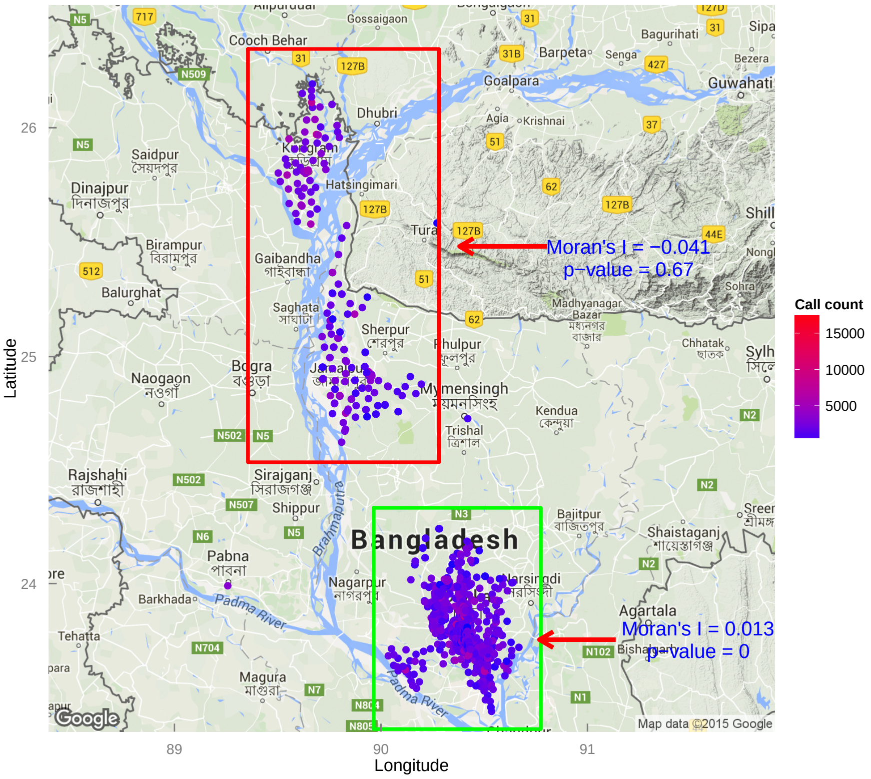


Figure 4.6: Total calls made on August 1

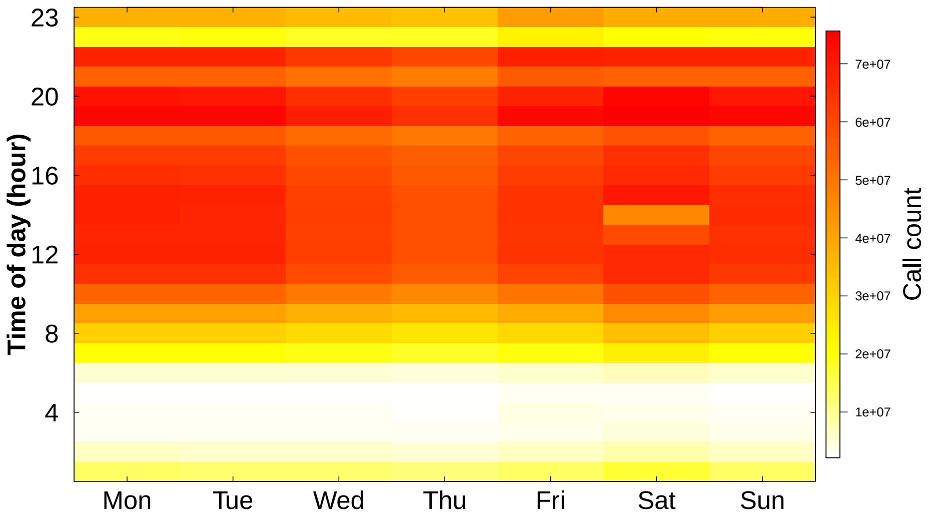


Figure 4.7: Hourly variation of calls across the week

4.4 Prediction task formulation

I assume there is a cellular network with many subscribers and I have access to their call history in the form of CDRs. Our interest is to know which *cell* a single user will visit next given their history of visits. Therefore the spatial resolution of user location is cell level. In this current work, I do not concern ourselves with the *lead time of prediction*, rather I just want to predict the next location regardless of when they may appear there. This is because I are dealing with CDRs which are sparse and sporadic.

4.4.1 Preliminary definitions

A cell c represents a discrete *geographic region* where a user can visit. I denote a set of all cells in the cellular network as $\mathcal{C}$. For convenience, I label them as $c_1, c_2, \ldots, c_K$. Strictly speaking, it is difficult to establish the exact geographic extents of a cell but it is helpful to think of cells as *voronoi cells* generated from all base station coordinates. Thus, I can associate each cell with geographic coordinates of a base station, I denote the geographic coordinates of cell c_k as $c_k^s \in \mathbb{R}^2$ because in our case I have only horizontal coordinates (latitude and longitude). Secondly, I characterize each cell by call count (number of calls made per unit of time). For example, if I set unit of time for call count as *hour* then the notation d_{c_k} means call count per hour for cell c_k. I denote the set of users in the network as $\mathcal{U}$ so that $\{u_i \in \mathcal{U} | i = 1, 2, \ldots, N\}$. A CDR represent location history for a given user. I denote it as $H_{u_i} = \{(t_j, c_k) | c_k \in \mathcal{C}, t_j \in \mathbb{R}, j = 1, 2, \ldots, n\}$ where t_j is a time stamp so that this is user history up to time t_n. I characterize each user by a set of features (e.g how often they make calls, the distances they travel every day) in order to learn about their mobility patterns. I represent these attributes as an $N \times p$ dimension matrix of features X_{u_i} where p is the number of features as shown in (4.1).

4.4.2 Problem definition

I now formally define the *prediction task.* Given a single user location history up to time t_n as follows: $(t_1, c_1), (t_2, c_2), \ldots, (t_n, c_k)$ I would like to make a prediction of the *next cell* the user will visit at t_{n+1}. In this work, I do not specify the look ahead time ($\Delta t = t_{n+1} - t_n$) of prediction. However, I assume that I are dealing with a regular user (who on average makes a phone call at least once in 2 days). I cast this problem in a probabilistic fashion as follows: $P(\mathcal{C} = c_k | X)$-estimate probability of user visiting cell c_k given their observed location history. For brevity, I drop the user subscript, so that when have this kind of construction I are always referring to a single user. I then invoke *Bayes theorem*, and decompose probability of a location c_k conditional on the observed history which can be represented by the feature vector X as shown in (4.1).

$$X = \begin{bmatrix} x_{11} & x_{12} & , \ldots, & x_{1p} \\ x_{21} & x_{22} & , \ldots, & x_{2p} \\ .. & .. & .. & .. \\ x_{N1} & x_{N2} & , \ldots, & x_{Np} \end{bmatrix} \tag{4.1}$$

$$P(\mathcal{C} = c_k | X) = \frac{P(X | \mathcal{C} = c_k) P(\mathcal{C} = c_k)}{P(X)} \tag{4.2}$$

In (4.2) above, the term to be estimated $P(\mathcal{C} = c_k | X)$ represents posterior probability distribution over the set of cells $\mathcal{C}$ while $P(C = c_k)$ represents the prior probability that cell c_k will be visited, finally $P(X | C = c_k)$ is the likelihood of cell c_k given that I observed features vector X. The likelihood can be estimated by using any of the available methods such as maximum likelihood. The challenge though is to put together a good scheme for estimating prior probabilities which translates into reasonable and accurate posterior probability distribution. In this formulation, the output of the prediction task is a set of cells with corresponding probabilities. In the next section, I provide details of our proposed scheme for computing reasonable prior probabilities and eventually inferring the posterior distribution $P(\mathcal{C} = c_k|X)$.

4.5 Enhanced bayes predictor

In the previous section, I presented a probabilistic formulation of the prediction task at hand. Here, I first present a set of further variables and corresponding notation I use in our approach, I then provide conceptual background and motivation for our technique. Also, I describe the set up of a prediction system based on our proposed technique.

4.5.1 Preliminaries

From each event in the user location history, I can generate features which are informative of his/her mobility patterns. The quantity and richness of features which can extracted obviously depend on the nature of the available data. In our approach, I assume that the only data I have is the location history itself. In Table 4.2, I present a full list of these features and other variables I use in our technique. I believe most of them are self explanatory, so, I elaborate on those I deem require further clarification: a) **Time of the day**-*td*: I stratify the day into three blocks based on time as follows: day(0800hr-1700hr) *dy*, evening(1800hr-1900hr) *ev* and night (2000hr-0700hr) *ng*. b) **Gyration** g: In this work, I actually use *mean gyration* which is the average of the maximum daily distance covered by a user. I include this variable because previous research [SQBB10] revealed significance of this variable to predictability of human mobility. I model most of the features as discrete random variables. However, geographic coordinates of a cell c_k^s, cell call count d_{c_k}, gyration g and regular location r^s are modeled as continuous random variables.

I cast our approach in the context of *supervised classification*, so that the observed location history up to time t_n is for training the model. For convenience, I maintain an n dimension column vector $H_L = \Big[l_1, l_2, \ldots, l_n\Big]^T$ where $l_i \in \mathcal{C}$ which stores cell visited in observation i corresponding to feature matrix X so that (x_{ij}, l_i) represents value of jth feature in observation i. I also maintain a set of distinct cells visited by the user $\mathcal{L}$ generated from H_L. I also maintain a matrix of cell traffic data D (I don't show the matrix here due to space limitations). For call count, I choose unit of time to be *hour* because in CDR

Name	Notation	Type	Domain	Description
cell	c_k	discrete	C	a cell in network
cell location	c_k^s	continuous	$\mathbb{R}^2$	geographic coordinates of cell
call count	d_{c_k}	continous	$\mathbb{N}$	cell call count per unit time
location	l_i	discrete	C	user location from history
unique cells	$\mathcal{L}$	discrete	$\mathcal{L} \subseteq \mathcal{C}$	unique locations visited by user
dw	dw	discrete	$\{sun, \dots, sat\}$	day of the week
dc	dc	discrete	$\{week - end, holiday, week - day\}$	day category
tod	tod	discrete	$\{day, evening, night\}$	time of the day
h	h	discrete	$\{0, 1, \dots, 23\}$	hour of call
gyration	g	continous	$\mathbb{R}$	average maximum daily distance
regular location	r^s	continuous	$\mathbb{R}^2$	most visited location

Table 4.2: Description of features

data user location history is not updated frequently so I believe this resolution is suitable. Therefore, the dimensions of D will be $K \times 24$.

4.5.2 Naive Bayes predictor

There two possibilities to consider when I think about which cell a user will visit next: the user can either *revisit* the cells they have been to before, or they can *visit* a cell they have never visited before. For now, I ignore the latter possibility and assume that users will only revisit places they have been to before. Although clearly unrealistic, this is a reasonable assumption considering that people usually visit a few set of places. Moreover, in some situations where data isn't available, this may be the only feasible model. I call this approach the *revisit model.*

With the above assumption and an additional assumption of independence of features in matrix X, I can easily estimate the posterior distribution over the set of cells in $\mathcal{L}$. I can estimate the prior probability for a cell to be revisited-$P_{revisit}(\mathcal{C} = c_k)$ based on frequency of visits in the location history. I write (4.2) to reflect this new terminology as shown in (4.3). I use maximum likelihood to estimate the class densities $P(X = x_j|C = c_k)$ can be estimated from frequencies. As expected, the output from (4.3) is a probability distribution over the visited cells $\mathcal{L}$. In order to get a single best estimate, I use *maximum a posteriori probability-MAP).*

$$P(\mathcal{C} = c_k|X = x_j) = \frac{P(X = x_j|\mathcal{C} = c_k)P_{revisit}(\mathcal{C} = c_k)}{\sum\limits_{q=1}^{||\mathcal{L}||} P(X = x_j|\mathcal{C} = c_q)P_{revisit}(\mathcal{C} = c_q)} \quad \text{where } c_k \in \mathcal{L} \tag{4.3}$$

4.5.3 Enhanced Bayes predictor

I arleady pointed out that the *revisit* model presented in the previous section is naive because contrary to its assumption, people always visit new places. I now take into account the possibility that a user can visit new places and call this the *visit-revisit model*. Clearly, the simple naive bayes predictor presented in (4.3) cannot work here because it will fail to return a prediction for those cells which are not in $\mathcal{L}$ because they will have zero prior probability. In the rest of this section, I describe a scheme which solves this problem by incorporating a set of unvisited cells which satify a predetermined distance threshold and thereby potentially allowing us to be able to predict a location which a user has never visited before.

First, I would like to have an estimate of prior probability for each cell regardless of whether a user has visited it or not. I denote this as $P_{visit}(\mathcal{C} = c_k)$ to indicate prior probability of a user visiting a cell. In this work, I assume I have access to data of all users in a cellular network, which in turn means I have access to comprehensive and exhaustive data on call activity for all cells in the network. Recall that I maintain this data in D. I can use this data to estimate prior probability for each cell based on call count. In Section **??** I demonstrated that cells can be thought of as P.O.I's so that the more calls a cell routes, the popular it is and the higher the prior probability it gets.

Does this mean for each user, when I want to make a prediction, I should consider all cells-$\mathcal{C}$? I can think of three options: a) **Consider all cells**: I can use this naive approach and compute prior probabilities and every time I want to make a prediction, I loop through all cells $\mathcal{C}$. The obvious disadvantage is unnecessary computational costs without actually enhancing results so much because some cells are far away and irrelevant to the user. b) **Consider all cells but use a weighting function**: This will have similar computational costs as in the previous scenario but it will take into some weighting. c) **Select unvisited cells within a distance δ of some designated location a**: In the current work, I adopt this approach because it reduces the computational costs. There are two components to the inclusion criteria: a distance threshold $\delta \in \mathbb{R}$ and most regular place r^s. This is not a cell, I rather compute this as a geographically weighted mean center based on visited cells (I weight by frequency of visit to each cell). I introduce V as a set which holds cells which have not been visited but are within δ km of r^s. More formally, I define $\{c_k \in V | c_k \notin \mathcal{L}, d(c_k^s, r^s) \leq \delta\}$ where $d(.)$ is a distance function.

The next problem is how to blend the prior probabilities. For instance, consider a cell $c_k \in \mathcal{L}$, I have to consider both the prior based on history of visits $P_{revist}(\mathcal{C} = c_k)$ as well as the *base* prior probability from call count which I denote by $P_{base}(\mathcal{C} = c_k)$. I use a simple convex combination of probabilities to combine these two to obtain an overall prior probability to visit cell c_k as shown in (4.4).

$$P_{visit}(\mathcal{C} = c_k) = \begin{cases} P_{base}(\mathcal{C} = c_k) & c_k \in V \\ \alpha_1 P_{revisit}(\mathcal{C} = c_k) + \alpha_2 P_{base}(\mathcal{C} = c_k) & c_k \in \mathcal{L}, \alpha_1 + \alpha_2 = 1 \end{cases} \tag{4.4}$$

Therefore, I can rewrite the posterior distribution from (4.3) to incorporate this blended prior probability as shown in (4.5). The rest of the terms can be estimated exactly in the same way as in the revisit model and I also employ MAP technique to obtain a single prediction. Next, I have to decide how to choose the values of the parameters α_1, α_2. In the current work, I do not use any optimisation procedure but rather learn these parameters from data.

$$P(\mathcal{C} = c_k | X = x_j) = \frac{P(X = x_j | \mathcal{C} = c_k) P_{visit}(\mathcal{C} = c_k)}{\sum_{q=1}^{||\mathcal{L} \cup V||} P(X = x_j | \mathcal{C} = c_q) P_{visit}(\mathcal{C} = c_q)} \tag{4.5}$$

4.6 Evaluation

In this section I present details of a series of experiments I conducted to evaluate the proposed technique. In particular, I investigate the following three key aspects;

- *Overall performance of the predictor.* Here, I ask how do accuracy of the enhanced bayes predictor compare with the naive bayes version?
- *Influence of parameters.* They are three important parameters in the proposed technique: $\delta, \alpha 1, \alpha 2$. In the current work, I only experiment with δ and fix the values for $\alpha 1, \alpha 2$ as explained in Section 4.6.1.
- *Performance across users.* Different users have varying location histories and mobility patterns in terms of length, number of distinct places visited. Therefore, I investigate how accuracy varies across these factors.

4.6.1 Experimental Design

In Fig. 4.8, I illustrate the two key components of the prediction system I set up: *the preprocessing and feature extraction unit* and *the location predictor*. The feature extraction unit receives input in the form of raw CDR data which comes on demand from cellular network. In this unit I transform the raw data into cell traffic data D and the feature vector X. This set up imitates *on-line processing* so that I use the available data to train the model and keep updating the matrices D and X and retraining the model as more data comes. I choose to update the matrix D every hour as opposed to updating it for every single call event because I noticed that the probabilities computed from the counts do not change significantly within one hour. Also, this updating this matrix for every call would result in heavy computation costs. As mentioned earlier, I do not optimize values of the parameters $\alpha 1, \alpha 2$ in (4.5) but rather provide fixed values based on the reasoning that people are more likely to *revisit* previously visited places rather than visit new places. Therefore, I tried a couple of combination of values and ended up setting $\alpha 1 = 0.7$ and $\alpha 1 = 0.3$.

For a quick evaluation, I extract a subset of *6,141,508* users with at least 35 days long location history. For each new location in the history, the predictor updates the matrices X and D and then makes a prediction about the next location. Our experimental framework tracks the *accuracy* of the predictor for each user. I define the accuracy of a predictor for a particular user to be the fraction of times the predictor correctly identified the next cell. This is the evaluation metric for our technique.

4.6.2 Results

Overall performance

I use *median* as the overall measure of prediction accuracy across all users. Overall, I found that the median accuracy for enhanced bayes is *54 percent* while that for naive bayes predictor is *37 percent* which is an improvement of 17 percentage points. In Fig. 4.9, I present empirical cumulative distribution of the experimental results comparing the accuracy of these two schemes.

In Fig. 4.10, I investigate how accuracy vary as recording time increase progresses. Naturally, I expect that perfomance should improve as I get more data about the user. I choose *three random users* out of the pool of users to enable easy visualization of the results and show how accuracy for these users change with increase in history length. I also show how *median* and

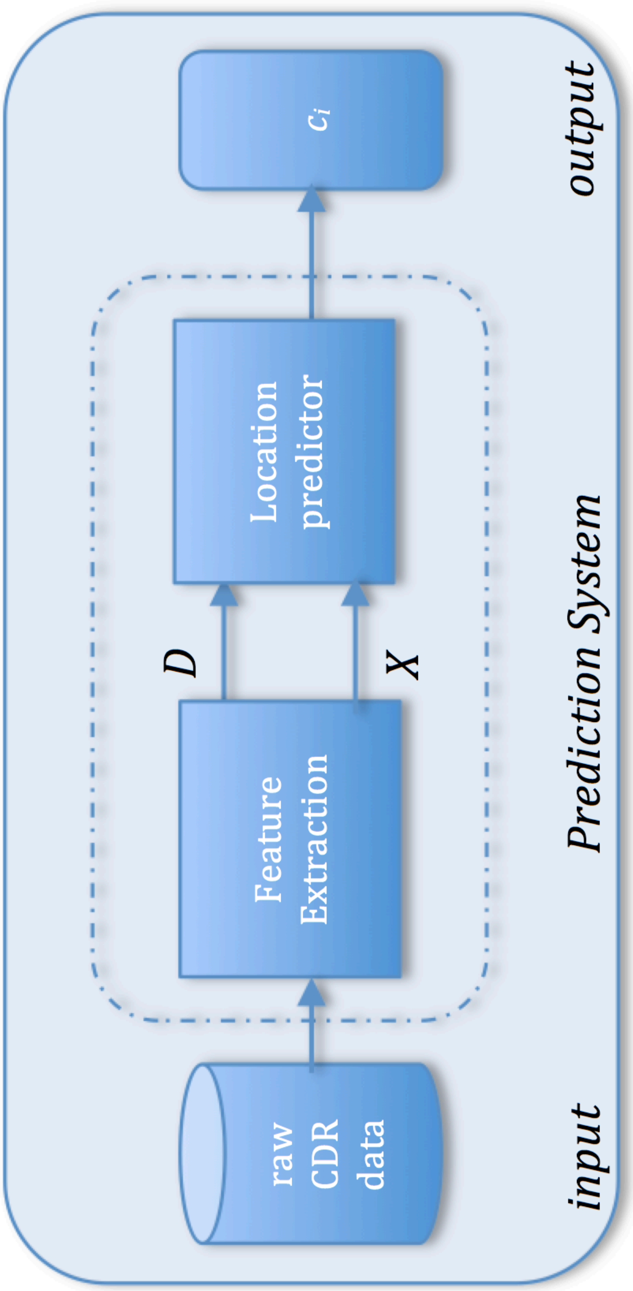


Figure 4.8: Illustration of set up of the enhanced Bayes prediction system

mean accuracy for all the users change over time. The results indicate that for individuals users, accuracy of the predictors clearly improve with time. However, mean and median accuracy do not show strong rise over the recording days.

Influence of parameters

The distance threshold δ is one of the most important parameters in this technique. I first analyzed the distribution of *gyration*. Recall that I defined this as maximum distance a user covers every day. I computed *mean gyration* and found that across all users, then mean gyration is *3 km*. Such a small value may look surprising but considering that the data is from Bangladesh which is a low income country with rudimentary transport infrastructure, as such people do not travel long distances on a daily basis. I therefore chose to experiment with four values: 2,4,6 and the user mean gyration g. I updated user gyration everyday when new data is added. In Fig. 4.11, the results clearly show that setting δ to user gyration provide best results.

Accuracy Vs. user attributes

I have shown in the previous sections how the performance of the predictor varies greatly across users. I therefore investigated this further. I generated the following user characteristics: length of history (total number of calls), average calls per day, average call interval, total number of distinct cells visited and entropy. I compute the entropy of a given user based on the distinct number of cells visited using the following formula: $H(\mathcal{L}) = -\sum_{i=1}^{||\mathcal{L}||} P(l_i)log_2P(l_i)$. In Table 4.3 I present *Pearson correlation coefficient* between these characteristics and accuracy based on the whole dataset. Although most of these variables representing user characteristics have very weak correlation with accuracy, I note that the direction of association agree with our intuitive reasoning. For example, it is reasonable to say that a person who makes more calls is also more likely to make calls in different places which loosely speaking would mean their location is more hard to predict. This is reflected in the negative correlations for length of history and average of calls. On the other hand, for average call interval, the story is different, I suppose that a user with wider call intervals makes few calls and therefore is more likely to stay at fewer places thereby making it easy to predict their location. This again is reflected in the positive correlation of call interval with accuracy. The largest correlation is observed for the entropy variable. Entropy calculated based on cells measures predictability

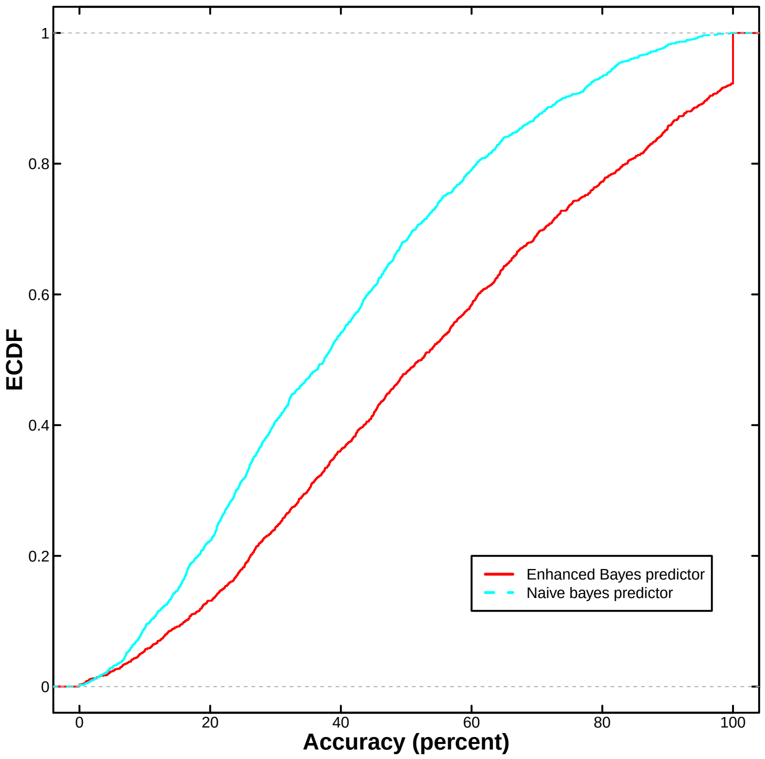


Figure 4.9: Accuracy-the enhanced predictor and naive bayes predictor

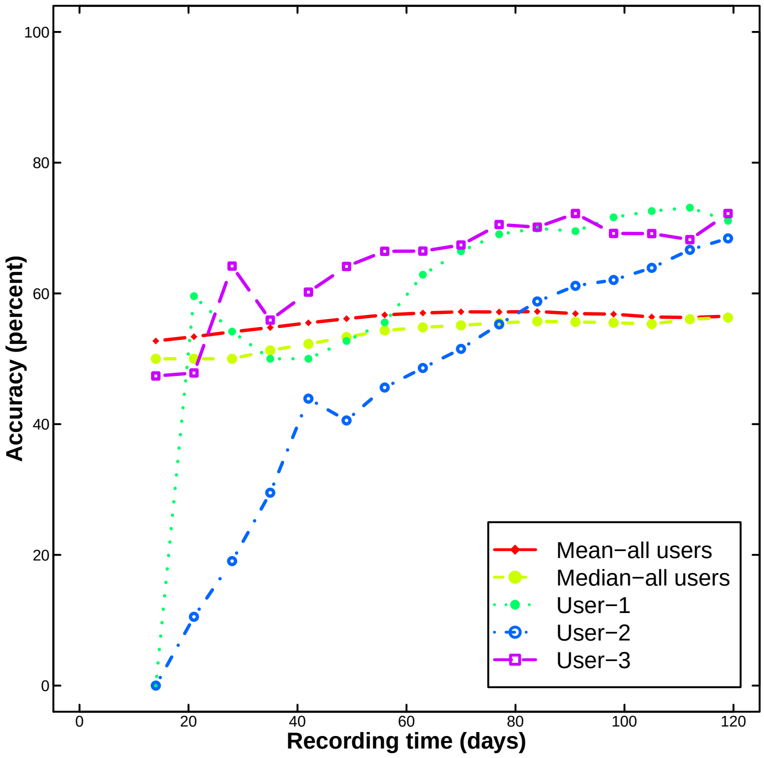


Figure 4.10: Accuracy vs. days in history

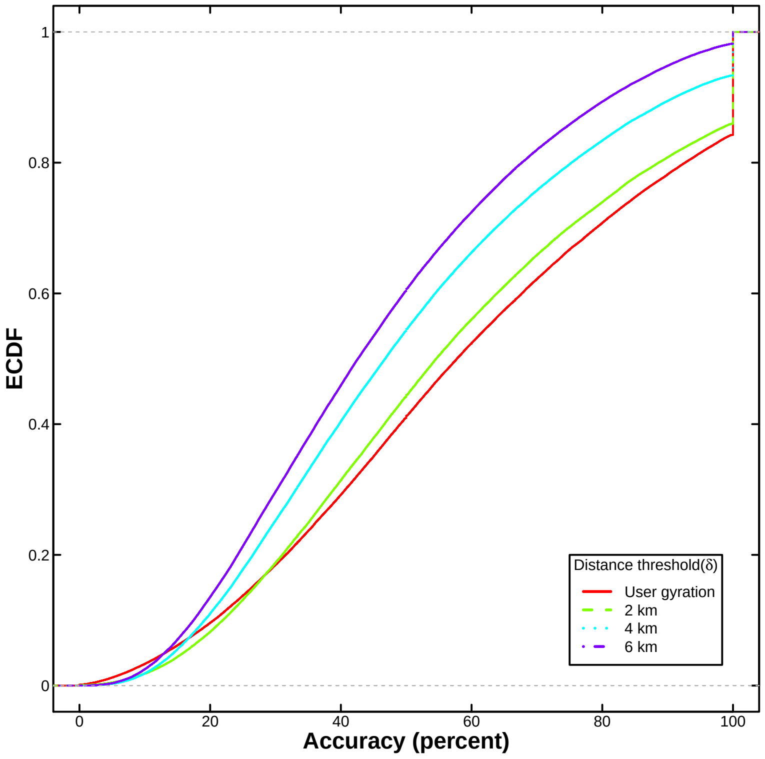


Figure 4.11: Accuracy vs. distance threshold

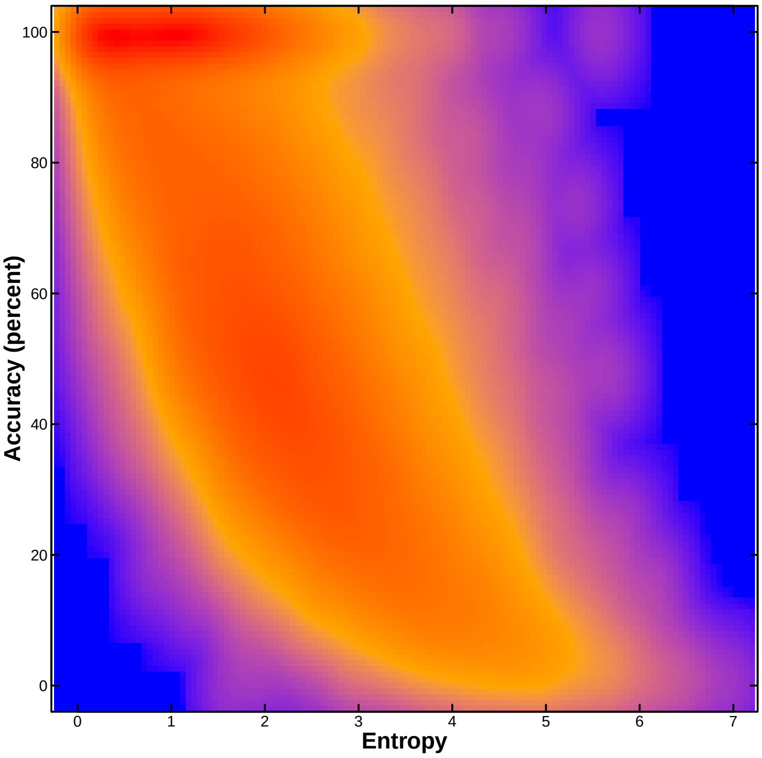


Figure 4.12: Correlation between accuracy and user entropy

of user. The interpretation of entropy of a variable (and in this case) is that higher values correspond to higher uncertainity which agrees with our results presented in Table 4.3 and also shown in Fig. 4.12.

Table 4.3: Correlation between user attribute and accuracy

User Attribute	R^2
Length of history	-0.21
Average calls per day	-0.23
Average call interval	0.22
Number of distinct cells	-0.54
Entropy	-0.72

4.7 Chapter Summary

The ability to predict future location of a cellular network users is crucially important with numerous applications. For instance, given a low income region with extremely low smartphone penetration rate, I envisage (though clearly ironical) that location prediction techniques have a potential to provide approximate *current* user location which would be useful in LBS services in these regions. In this work, I experiment with a Bayes based predictor in order to find ways to enhance their performance by leveraging big data. To this end, I study a large CDR dataset. I first explore the dataset and find that I can use call activity to generate prior probabilities for use in Bayes predictor. With this reasoning, I develop an enhanced Bayes predictor which uses a distance threshold and the users' regular location to improve generation of prior probabilities. Experimental results show that the enhancements I propose increase accuracy of the Bayes based predictor by 17 percentage points. I also find a novel result that accuracy is highest when the distance threshold is set to the users gyration. In conclusion, I demonstrated that it is feasible to leverage big cellular data to enhance location predictors without relying on external data which is encouraging for low income regions. There are several directions for future work. First, I would like to carry out a more comprehensive and exhaustive analysis of spatio-temporal variation of call activity. Second, I intend to explore more sophisticated prediction models albeit with less demanding computation requirements and see if similar enhancement can be done. Further, I intend to optimize parameters $\alpha 1, \alpha 2$ and δ.

Chapter 5

Location Prediction-Communal Models

5.1 Introduction

Thanks to the advent of cheap sensors, many mobile devices nowadays come with the capability to provide fine grained location traces of a user. However, beyond current location some applications can benefit from anticipating the location a user will visit in the near future. This is where *location prediction* comes in. For example, predicting a user's next location would be useful for urban planning in anticipating traffic levels.

The problem of location prediction has been extensively studied[ZXYG13, SKJH06, DGP12, KH06, SMM+11, EKK+13]. However, most of the works are *context specific*. A good example of such contexts is the type of data used: *resolution of location data* and whether *additional data* (e.g., social relationships data is used or not). For example, in [BSM10] they use social relationships data to improve results of location prediction. Thus, the conclusions drawn and the techniques developed cannot be transfered to new situations (e.g, cases where additional data isn't available) without substantial modifications. The aforementioned observation plus the fact that human mobility is inherently dynamic means that the area of location prediction is still open to further explorations.

In this study, I explore the potential to improve prediction models for individuals (reducing model training time and increasing prediction accuracy) by leveraging community-wide data. To this end, I carry out a systematic experiment with real life mobility dataset. I choose to experiment with logistic regression classifier because although simple, it is robust. My broad experimental objectives are two fold: investigate whether use of community-wide learned model parameters in the training process of an individual model can reduce training time and whether systematic combination of community wide model parameters with individual model parameters can improve accuracy of predictions. I demonstrate that when I use parameters from community model as lower bounds in the optimization process while training individual models, the training time is drastically reduced by almost 100 percent. However, regarding accuracy, current results do not show considerable improvement of accuracy when I combine community wide model parameters and individual model parameters. The main contribution of this work can be summarized as follows:

- *Experiments and evaluation.* I conduct a systematic experiments with a large scale data to understand how well community level mobility patterns approximates individual users.

- *Simple heuristic for logistic regression.* I demonstrate that given a large scale data for many users, the idea of using community models can drastically speed up the process of training models and making predictions for individual users which is a desirable property for any system.

5.2 Related work

I have identified three core aspects of the single user location prediction research: *data, prediction task*, and *prediction approaches* and I review previous work along these categories. Majority of the previous works[EKK$^+$13, KH06, SMM$^+$11] conducted their experiments with location data generated from GPS/Wi-fi which I consider to be high resolution with only two works [KWSB04, ZXYG13] where they used CDRs which can be considered as low resolution. In [BSM10] they demonstrated that inclusion of social network data enhances prediction results. The generic task is often to predict location of a user some time in the future. However, the subtle details such as what information I are starting with (current context), lead time or whether I are interested in temporal aspects (e.g., arrival time, stay duration) vary across studies. Most of the earlier works focused on the task of predicting the next place/destination a single user will visit. Several methods have been proposed for location prediction in different contexts (location data and prediction task). One of the popular approaches is use of Markov models [SKJH06, AS03] which is natural owing to the sequential nature of location prediction tasks. However, other non sequential models have also proved to perform even better than Markov models as demonstrated in [EKK$^+$13, MRJ12].

As regards data, this study is different from most of the previous works because I use low resolution CDR data. However, my approach and philosophy is most similar with the work in [MRJ12] where they use collaborative filtering to improve prediction accuracy for new users. However, the important difference is that they try to match new users with similar old individual users while in my case the intention is to learn from a large community which also turns out be computationally cheaper as compared to computing similarities for individual users. This work is different in spirit from the rest of the works reviewed because I also interested in reducing training time while they were only focusing on accuracy. There are many situations where accuracy can be traded for fast results.

5.3 The dataset

In a cellular network, a *cell* is a geographical area covered by a *base station* -a piece of equipment that facilitates wireless communication between user device (UD) and a network. The cell covered by a base station can be from one mile to twenty miles in diameter, depending on terrain, population density and transmission power. A UD is always receiving message broadcasts by these base stations. Thus, I can approximate its actual location using the geographical coordinates of the corresponding base station. Hence, the UD is assumed to be located at the base station coordinates independently of its actual position within the cell [TV04]. In this work, cell location refer to base station coordinates. CDRs constitute a sequence of places, related to a single user, where a call (or text message) was made thus describing the user's hop movements as they make calls. Due to personal nature of this data, the richness of available attributes in a CDR dataset vary. For instance, the MIT Reality Mining dataset [ZXYG13] is a publicly available CDR dataset with rich information about call events (including caller and recipient of calls). On the other hand, in this work, I have access to what I consider to be a very sparse CDR dataset. In this work, I define *sparsity in feature space* in terms of features (not considering those which

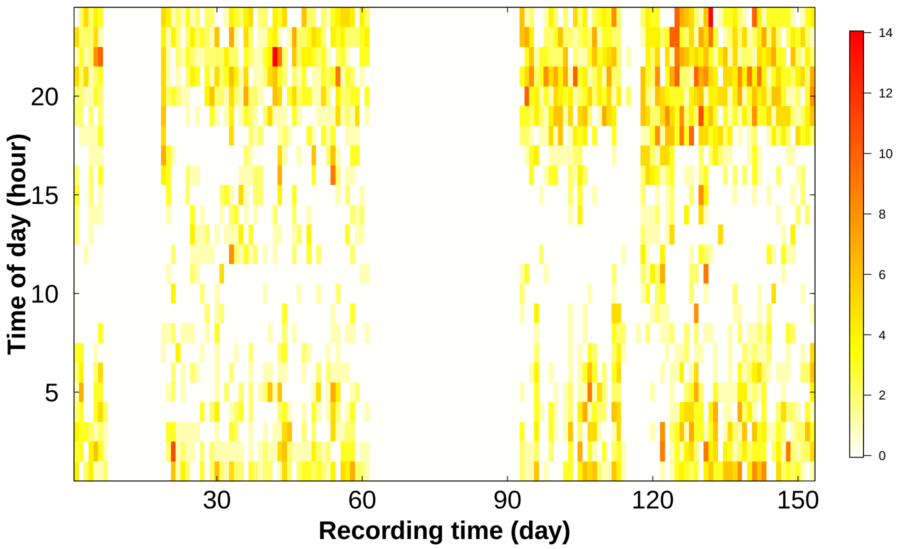


Figure 5.1: Call pattern of sample user: Frequent user

can be derived) in a dataset available for prediction. Each call event has the following attributes: anonymised user ID, time stamp of call event and base station coordinates.

The data is from a leading cellular phone operator in Bangladesh. It was collected in 2013 and covering the months of *August, September, November* and *December*. For some logistical reasons, I do not have data for October. The data was acquired through an agreement between the cellular company operator and The University of Tokyo's Shibasaki Lab. This agreement doesn't allow public access to the data. In Table 5.1, I present background characteristics of the data. There are a total of 16 million unique users in the dataset. However, majority of the users are sparse and non-consistent users as demonstrated in Fig. 5.2 and Fig. 5.1, in which I show the calling pattern of two random users in order to emphasize diversity of users in the dataset (user shown in Fig. 5.1 seems to be a more frequent user while the one depicted in Fig. 5.2 makes very few calls).

5.4 Location prediction

In this work, I assume there is a cellular network with many subscribers and I have access to their call history in the form of CDRs. My interest is to know

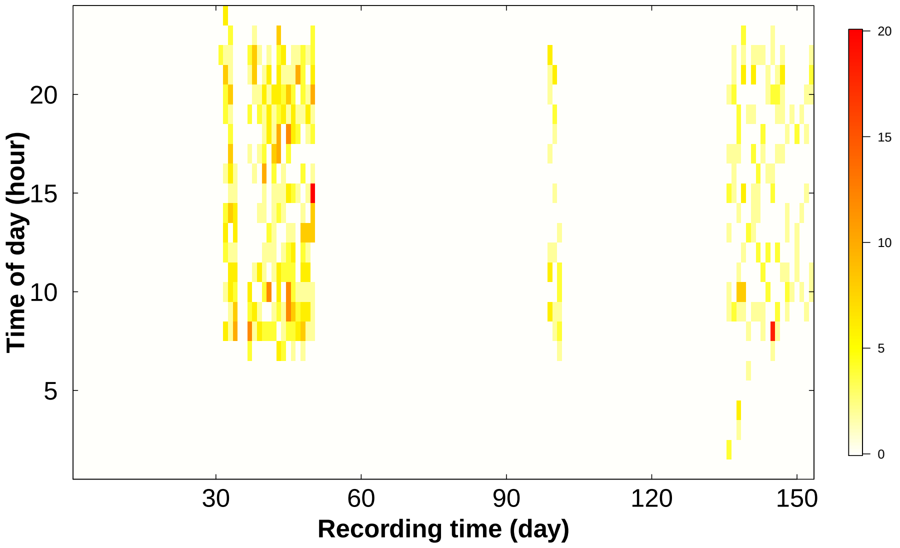


Figure 5.2: Call pattern of sample user: sparse user

Table 5.1: Background details of the dataset

Category	Description
Starting date	August 1, 2013
Ending date	December 31, 2013
Data gaps	October 1-October 30 2013
Number of cell towers	2101
Number of users	16,000,000
Number call events	3,5,000,000

Name	Notation	Type	Domain	Description
cell	c_k	discrete	C	a cell in network
cell location	c_k^s	continuous	$\mathbb{R}^2$	geographic coordinates of cell
call count	d_{c_k}	continous	$\mathbb{N}$	cell call count per unit time
location	l_i	discrete	C	user location from history
unique cells	$\mathcal{L}$	discrete	$\mathcal{L} \subseteq \mathcal{C}$	unique locations visited by user
dw	dw	discrete	$\{sun, \ldots, sat\}$	day of the week
dc	dc	discrete	$\{week-end, holiday, week-day\}$	day category
tod	tod	discrete	$\{day, evening, night\}$	time of the day
h	h	discrete	$\{0, 1, \ldots, 23\}$	hour of call
gyration	g	continous	$\mathbb{R}$	average maximum daily distance
regular location	r^s	continuous	$\mathbb{R}^2$	most visited location

Table 5.2: Description of features

which *cell* a single user will visit next given their history of visits so that the spatial resolution of user location is cell level. In this current study, I do not concern ourselves with the *lead time of prediction*, rather I just want to predict the next location regardless of when they may appear there. This is because I are dealing with CDRs which are sparse and sporadic.

5.4.1 Features for prediction

A cell c represents a discrete *geographic region* which a user can visit. I denote a set of all cells in the cellular network as $\mathcal{C}$. For convenience, I label them as $c_1, c_2, \ldots, c_K$. Strictly speaking, it is difficult to establish the exact geographic extents of a cell but it is helpful to think of cells as *voronoi cells* generated from all base station coordinates. Thus, I can associate each cell with geographic coordinates of a base station, I denote the geographic coordinates of cell c_k as $c_k^s \in \mathbb{R}^2$ because in this case I have only horizontal coordinates (latitude and longitude).

I denote the set of users in the network as $\mathcal{U}$ so that $\{u_i \in \mathcal{U} | i = 1, 2, \ldots, N\}$. A CDR represent location history for a target single user. I denote it as $H_{u_i} = \{(t_j, c_k) | c_k \in \mathcal{C}, t_j \in \mathbb{R}, j = 1, 2, \ldots, n\}$ where t_j is a time stamp so that this is user history up to time t_n. From each call event in the user location history, I can generate temporal features (I are limited to temporal features because of sparsity of data in feature space) which are informative of his/her mobility patterns (e.g., how often they make calls and when do they usually make calls). In Table 5.2, I present a full list of these features and other variables I use in this work. I represent these features as an $N \times p$ dimension matrix of features X_{u_i} where p is the number of features as shown in (5.1).

$$X = \begin{bmatrix} x_{11} & x_{12} & ,\dots, & x_{1p} \\ x_{21} & x_{22} & ,\dots, & x_{2p} \\ .. & .. & .. & .. \\ x_{N1} & x_{N2} & ,\dots, & x_{Np} \end{bmatrix} \tag{5.1}$$

I cast this approach in the context of *supervised classification*, so that the observed location history up to time t_n is for training a prediction model. For convenience, I maintain an n dimension column vector $H_L = \Big[l_1, l_2, \dots, l_n\Big]^T$ where $l_i \in \mathcal{C}$ which stores cell visited in observation i corresponding to feature matrix X so that (x_{ij}, l_i) represents value of jth feature in observation i. I also maintain a set of distinct cells $\mathcal{L}$ visited by the user generated from H_L.

5.4.2 Location prediction task

I now formally define the *prediction task.* Given a single user location history up to time t_n as follows:
$(t_1, c_1), (t_2, c_2), \dots, (t_n, c_k)$ I would like to make a prediction of the *next cell* the user will visit at t_{n+1}. I do not specify the lead time ($\Delta t = t_{n+1} - t_n$) of prediction. However, I assume that I are dealing with a regular user (who on average makes a phone call at least once in 2 days). Clearly, this task can be solved using most of the classification algorithms where the target output is cell location.

5.4.3 Prediction with logistic regression

Background

In logistic regression, the objective is to estimate posterior probabilities of K classes through use of linear functions in the features. Although it is probably most popular with binary classification, this model has an efficient generalization for the case of multi-class classification often referred to as *multinomial logistic regression.* For $k = 1, \dots, K-1$ and parameter vector θ_k^T it can be represented as in 5.3 [FHT01]. In this case, the interpretation is that it is the probability that a user visits cell c_k given their location history.

$$P(\mathcal{C} = c_k | X = x) = \frac{\mathrm{e}^{\theta_{k0} + \theta_k^T x}}{1 + \sum\limits_{j=1}^{K-1} \mathrm{e}^{\theta_{j0} + \theta_j^T x}} \tag{5.2}$$

Model training is the process of finding optimal values for the parameter vector θ_k^T. An error function $E(\theta)$ in the model parameters as shown in 5.2 is usually derived using maximum likelihood.

$$E(\theta_1^T, \dots, \theta_{K-1}^T) = -\sum_{i=1}^{N} \sum_{k=1}^{K-1} l_{ik} \log y_{ik} \tag{5.3}$$

where $y_{ik} = P(\mathcal{C} = c_k | X = x_i)$ and N is the total number of training examples. Next, in order to determine optimal values of θ_k^T the error function $E(\theta)$ is minimize with respect to θ. The general form of optimization problem is shown in 5.5.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & f_i(x) \leq b_i, \ i = 1, \dots, m. \end{aligned} \tag{5.4}$$

This error function can be minimized by a family of *Newton-Raphson* methods. In the quasi newton Raphson technique, the algorithm takes *initial values* (θ_0) and then makes updates to θ using first and second order derivatives of $E(\theta)$ for a specified number of iterations until convergence.

Community Vs. individual models

For any target individual user, I can use their location history to create an *individual model.* In addition, I also introduce the concept of community M which I define as a set of cells selected based on geographic proximity such that $M \subset C$. I can split a geographic region containing the cell towers into equal size grids and create communities based on grids so that all cells which fall into some grid j will belong to community M_j where $j = 1, \dots, N$. The goal is to create a *community model* which captures high level patterns as opposed to individual patterns. In order to achieve this, I first create what can be considered as *community data* by simply pooling together all data from each cell in the community and then train a model based on this data. Next, I discuss inclusion criteria for an individual user to a community. Since users make phone calls across different cells, it is almost impossible to find a user with only a single cell tower in their location history. I use a simple inclusion procedure based on proportions and parameter γ as follows: $\frac{|M \bigcap \mathcal{L}|}{|\mathcal{L}|} \geq \gamma \implies u_i \in M$ where $\gamma \in [0, 1]$. For example, if I set $\gamma = 0.9$, then a user u_i belongs to community M_j if 90 percent of the cells in his location history belong to M_j.

Proposed heuristics

The overall rationale of my proposals is that since a community model captures group level mobility patterns, I believe some of the randomness in individual level mobility can be reduced by taking into account group level patterns.

Custom initial values: Given a community model M_j^θ characterized by its parameters θ (I omit the subscripts and superscripts for simplicity), then for any individual user in community M_j, instead of using random guesses in the model training, I can use M_j^θ. Theoretically, since $E(\theta)$ is convex, different initial values still result into same solution. However, model training time could possibly change because rate of convergence during the minimization process is sensitive to initial values.

Constrained optimization: The minimization process as explained in Section 5.4.3 can be subject to some constraints. In this case, I would like the parameters in the individual models to be bounded by the parameter values in the community model so that the minimization can take the form shown in 5.5

$$\begin{aligned} &\underset{\theta}{\text{minimize}} && E(\theta) \\ &\text{subject to} && \theta \leq b,\ k = 1, \ldots, K-1. \end{aligned} \tag{5.5}$$

where b is coming from M_j^θ. Again, in normal cases, this would not change the optimal values since its a convex optimization problem but it would definately affect the rate of convergence.

Ensemble model: The approach of combining models is common in machine learning and the resultant model is often called an *ensemble*. I apply this technique here by combining parameters from community model with those from individual model through a simple weighted averaged. Given two constants $\alpha_1, \alpha_2, \alpha_1 + \alpha_2 = 1$, I derive an ensemble model H as follows: $H^\theta = \alpha_1 M_j^\theta + \alpha_2 I^\theta$ where I^θ represents parameters from individual model belonging to community M_j.

5.5 Experiments

5.5.1 Experimental set up

Experiment objectives

The goal of the experiment is twofold: to test whether parameters from community model can be used to reduce training time of individual models and also test if an ensemble model derived from community and individual model would give better prediction accuracy. Specifically the objectives are as follows:

1. Use community-wide model parameters in training of individual model (the target here is model training time).
    - Use community-wide model parameters as initial values in the minimization of $E(\theta)$ for individual model. I are calling this scenario *Custom initial values model.*
    - Use community-wide model parameters as lower bounds in the minimization of $E(\theta)$ for individual model. I are calling this scenario *Constrained optimization model.*
2. Ensemble model (the target here is accuracy). Create an ensemble model from community and individual models and use it in place of individual model. I are calling this scenario *Ensemble model.*

Evaluation measures

Although most of the prediction systems are online in nature, I do the experiments in batch mode. I have four months of location history (August, September, November, December) and I use 70 percent of this data for training and the rest for testing. I use two evaluation measures: *accuracy* and *training time.* I define accuracy simply as the proportion of correct predictions made out of the total number of predictions a model was tasked with. Training time of a model is the amount of time it takes for a model to go through the training data and find optimal model parameters. For each target individual in the experiment, I create four models: Standard model (normal logistic regression model), custom initial values model, constrained optimization model and ensemble model. I measure training time for all models except ensemble. I also record accuracy for all the models. I group users by community and compute summary evaluation measures : *median accuracy* and *median training time* within each community which I report on in the results section.

Data preparation and the experiment process

In order to work with more consistent users, I extracted a subset of data from *City of Dhaka* and then picked users with 500 or more call events. Next, I needed a reasonable basis for optimal grid size for community demarcation. I use *gyration* for this. Gyration is the maximum distance traveled by a user within

Table 5.3: Experimental data details

Community	Number of cells	Number of users
com254	54	150
com363	130	124
com387	46	176

a single day, I found the mean gyration to be less than 3km. The significance of this is that majority of the people usually don't go beyond a radius of 3km from their homes during their daily travels. Therefore, I set grid size to $3km \times 3km$ and found a total of 212 communities . However, due to processing requirements, I decided to select only those communities with more than 15 cells within them which reduced the target communities to only 23. Out of these, I report on results for three communities with varying number of users (see Table ?? for details about experiment data). The whole process flow of the experiment is as follows: I trained a community model for all the target communities and store their parameters. I ensured that I only trained these models with 70 percent of the data. Then for each individual user, trained four models in turn as explained in the previous sections and recorded training time and prediction accuracy. For the ensemble models I gave more weight to the individual model and therefore after some experimentation set $\alpha_1 = 0.3$ and $\alpha_2 = 0.7$. I carried out this processing using Weka machine learning library [HFH+09]. The number of iterations was set at 200 for all the models.

5.5.2 Results and discussion

In the first investigation, I investigated the effect of using custom initial values (as opposed to using default initial values-in most cases zeros) in the minimization of $E(\theta)$. I also investigated the effect of constraining the minimization of $E(\theta)$ (as opposed to the default unconstrained minimization) by providing custom lower bounds taken from community parameters. I present the results of this investigation in Fig. 5.3. One very interesting result is that when I use custom lower bounds in the minimization process, there is a nearly 100 percent reduction in training time without significantly affecting accuracy (see Fig. 5.4). Training of a logistic model has $O(n)$ time complexity, as such training time increases linearly with number of training examples and number of classes (in this case cells). Consequently, community $com363$ has much longer training time as compared to the rest of the communities because it has the largest num-

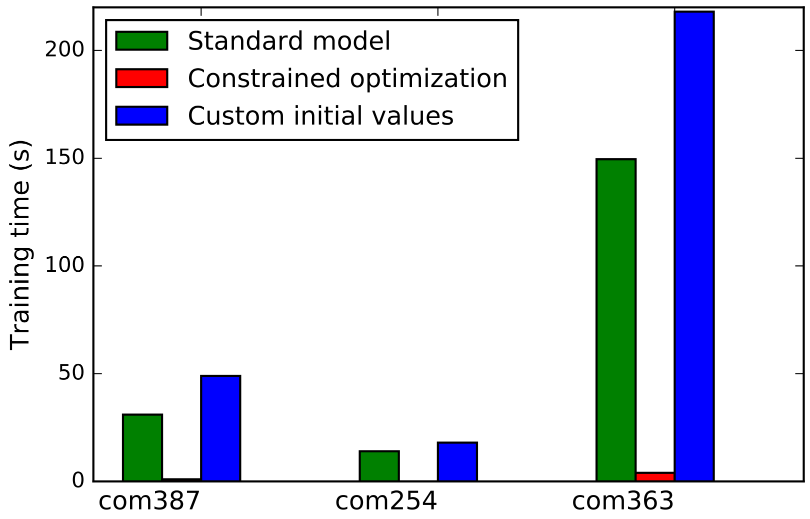


Figure 5.3: Comparison of model training time

ber of cells. For such kind of scenarios where I have many classes and training instances, this result shows that I could significantly reduce training time by using lower bounds from community model. On the other hand, use of custom initial values doesn't seem to work well as it is increasing the training time rather than reducing it. I don't consider training time in the case of ensemble model because training time is same as in the standard model (I have to train a standard model first and then use the parameters to derive ensemble model).

I also investigated accuracy of all the categories of the models I derived. In Fig. 5.4, I compare accuracy of different variations of the models across the target communities and also accuracy when I combine the results from all the communities. Although I didn't expect accuracy difference among the standard, constrained and custom initial values when considered within each community, the results shows slight accuracy differences. This is practically possible even when minimizing a convex function. However, I expected differences between standard and ensemble model because they are derived differently. The results show that the standard classifier has better performance than the ensemble model.

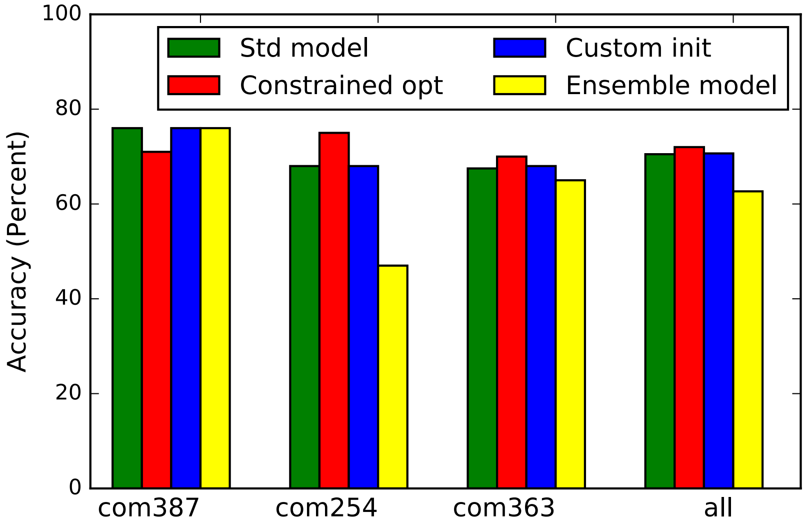


Figure 5.4: Comparison of model accuracy

5.6 Chapter Summary

I set out to investigate how I can leverage large scale data to come up with community-wide model parameters which I believe captures group level mobility patterns common to all individuals within this particular community. I argue that this kind of community model would be useful in location prediction for individuals in two scenarios: to reduce training time of models and perhaps even improve accuracy. I carried out experiments to test my hypothesis using a real life dataset of human mobility. The results of the experiments support this hypothesis that indeed if I take parameters learned from systematically selected community-wide data and use them as lower bounds in the minimization of the logistic regression error function for individual models, the training time is reduced by nearly 100 percent. In any computation system, the ability to reduce computational time without compromising on accuracy is desirable. On the contrary, when I use community wide learned parameters as initial values in training individual model, I find that training time actually increases. Also, the ensemble model I created from community and individual models doesn't improve prediction accuracy as I expected. I believe this is due to the way I created communities, I could have created more personalized communities by considering individual mean gyration rather than the overall mean gyration.

However, this kind of approach requires much more computations.

I still believe incorporation of community-wide parameters has potential to improve accuracy in individual models. For example in cases of bayesian logistic regression where I need to provide prior distribution of parameters, these community wide parameters could be useful. In future work I want to consider this approach. I also would like to try different approaches to demarcating communities and see what effect they have on accuracy.

Chapter 6

Conclusion and Future Work

In this thesis, I conducted a study of human mobility using data generated from call detailed records. I undertook this study because thorough understanding of laws that govern human movements has useful applications in public health, urban planning, disaster management, traffic engineering and marketing. Nowadays, a study like ours is relatively easy to conduct due to the high penetration rate of smartphones and cellular phones, which has led to generation of massive datasets in urban spaces depicting when and where people go every day. The primary research questions I addressed in this study can be summarized as follows:

1. How can we visualize spatio-temporal mobility data in order to easily discern daily as well as aggregated mobility patterns for individual users?
2. Given location history of an individual user, can we detect if a user has changed place of residence (home) or not? If so, can we develop an algorithm to accomplish this task in an automated fashion?
3. How to enhance location prediction models for individual users? In the current study, I define enhancement as reducing the training time of models as well as increasing prediction accuracy.

In response to the questions above, I first developed a simple but useful web-based framework for interactive visualization of daily mobility patterns in order to allow easy and quick interpretation of trends. In the system, I show summary statistics regarding user movement such as how far they travel everyday. I also visualize commonly visited places and identify home region of the user. I then explored the potential to use call detailed records generated from mobile phones as surrogate residence history to infer change of place of residence (home) for individual users. Finally, I undertook to improve performance of location prediction models. In particular, my objective was to reduce training time of prediction models for individual users as well enhance prediction accuracy.

The simple visualization framework I developed in Chapter 2 is very useful for discerning patterns in mobility and consequently in application of these insights in decision making. In Chapter 3, I proposed a clustering technique to automatically identify users who change place of residence given their residence history in the form of CDRs. Results from the experiments I conducted show that this proposed technique performed reasonably well. In Chapters 4 and 5, I tackled the problem of enhancing location prediction models. My guiding principle was to improve performance without relying on any form of external data. I first experimented with a Bayes based location predictor and managed to come up with an *enhanced Bayes predictor* which outperforms the regular naive Bayes predictor without use of external data. Next, I targeted to reduce model training time and suggested the use of parameters from a community wide model. Here again, results from the experiments show that my idea to use community-wide learned model parameters in individuals works very well and reduces training time for individual models by nearly one order of magnitude.

With regard to the work in Chapter 3, although I'am confident the results are promising and have potential to benefit urban planners in cities, this particular research was strongly limited by a number of factors which could hinder application of this technique in other settings. First and foremost, I didnot have access to ground truth data, in this case, the ideal ground truth data would have been a user's place of residence at different points in time. Consequently, I generated what I call a quasi-ground truth data to evaluate the proposed technique. Secondly, the data used in this study spans four months, a much longer history would have been desirable and would have definitely resulted in more reliable evaluation. The overall consequence of these two factors is that the results need to be interpreted with caution. There are several directions for future work. First, in the introduction I mentioned that one important application of determining residence change from CDR data is the ability to obtain estimates of internal migration in a region, therefore as a case study I plan to use the approach I proposed to estimate internal migration rates from the CDR data from Dhaka, Bangladesh and compare results with official estimates from the Statistics Bureau.

For the work on location prediction there are certainly many aspects that could be improved with further work. For example, I only experimented with what can considered as none-complex and linear algorithms. I chose to do so because algorithms such as logistic regression and Bayes based are easy to implement and are also computationally cheap. The latter is very important for quick evaluation and also considering I had 16 million users to evaluate. However, it is important to note that the problem I solve in Chapter 4, which essentially reduces to *how to handle new unseen classes in the case of multi-class classification problem* is inherent to all classifiers when faced with a multi-class problem regardless of whether they are linear or not. In some Machine Learning literature, this problem is referred to as *open set classification* [SJB14]. In this regard, the heuristics I proposed in Chapter 4 can equally work even on complex classifiers. Nevertheless, as future work, I still would like to apply these proposed enhancements to other family of classifiers such as neural networks or tree based methods.

The situation is slightly different for the success I saw in reducing training time of models in Chapter 5, in this case, the process is sensitive to whether the optimization problem being solved is convex or not. So in future work, I intend to apply the technique of using communal parameters in non-convex optimisation settings such as in neural networks and my belief is that this approach may work even better in this setting because due to the none-convexity of the optimization problem, choice of initial value for parameters is crucially important and usually results in different results.

Appendices

Appendix A

Publication List

A.1 Journals

1. Arai, A.; Fan, Z.; Matekenya, D.; Shibasaki, R. Comparative Perspective of Human Behavior Patterns to Uncover Ownership Bias among Mobile Phone Users. ISPRS Int. J. Geo-Inf. 2016, 5, 85.

A.2 International Conferences

1. Dunstan Matekenya, Masaki Ito, Yoshito Tobe, Ryosuke Shibasaki,and Kaoru Sezaki, "MoveSense: A spatio-temporal Clustering Technique for Discovering Residence Change in Mobile Phone Data ", ACM SIGSPATIAL International Workshop on GeoStreaming 2015 , Seattle, WA, USA, November 3 - 7, 2015
2. (Poster) Dunstan Matekenya, Masaki Ito, Yoshito Tobe, Ryosuke Shibasaki,and Kaoru Sezaki, "Communal Parameters: A Study into Using Community-wide Learned Prediction Models in Individual Users", 2nd EAI International Conference on IoT in Urban Space, Tokyo, May 2425, 2016
3. Guangwen Liu, Masayuki Iwai, Yoshito Tobe, Dunstan Matekenya, Muhammadi Asif Khan and Kaoru Sezaki, "Beyond horizontal location context: measuring elevation using smartphone's barometer", ACM Ubicomp 2014, AwareCast 2014: Third Workshop on Recent Advances in Behavior Prediction and Pro-active Pervasive Computing.

A.3 Domestic Conferences

1. Dunstan Matekenya, Guangwen Liu, Masaki Ito, Kaoru Sezaki, “ Energy Efficient Sensing with Spatio-temporal Predictioni ”, 2014 IEICE General Conference, Niigata, Japan, March, 2014.
2. Matekenya Dunstan, Masaki Ito, Yoshito Tobe, Teerayut Horanont, Ryosuke Shibasaki and Kaoru Sezaki, ”Event Detection in Individuals Using Mobile Phone Traces ”, 2014 IEICE Communications Society Conference, vol. BS-6, Tokushima, Japan, 19-26 September, 2014.

Bibliography

[AA12] Natalia Andrienko and Gennady Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, page 1473871612457601, 2012.

[AS03] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[BCG$^+$09] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[BCH$^+$11] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):0018–26, 2011.

[BCH$^+$13] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[BHG$^+$09] Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Vittoria Colizza, et al. Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1):1, 2009.

[Bis07] C Bishop. *Pattern Recognition and Machine Learning*, volume 1. Springer, New York, 2007.

[Blu12] Joshua E. Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125, 2012.

[BM08] Ingrid Burbey and Thomas L Martin. Predicting future locations using prediction-by-partial-match. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, pages 1–6. ACM, 2008.

[BSM10] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[Bur11] Ingrid Burbey. Predicting future locations and arrival times of individuals. 2011.

[CBB$^+$07] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Med*, 4(1):e13, 2007.

[CCJ10] Xin Cao, Gao Cong, and Christian S Jensen. Mining significant semantic locations from gps data. *Proceedings of the VLDB Endowment*, 3(1-2):1009–1020, 2010.

[CDLLR11] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, pages 36–44, 2011.

[CDLR10] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Human mobility prediction based on individual and collective geographical preferences. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pages 312–317, 2010.

[CML11] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference* *on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[Cre15] Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 2015.

[DDMGP14] Trinh Minh Tri Do, Olivier Dousse, Markus Miettinen, and Daniel Gatica-Perez. A probabilistic kernel method for human mobility prediction with smartphones. *Pervasive and Mobile Computing*, 2014.

[DGP12] Trinh Minh Tri Do and Daniel Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 163–172. ACM, 2012.

[DIC14] Maria Luisa Damiani, Hamza Issa, and Francesca Cagnacci. Extracting stay regions with uncertain boundaries from gps trajectories: a case study in animal ecology. In *22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014).* ACM, 2014.

[Div08] United Nations. Statistical Division. *Principles and recommendations for population and housing censuses.* United Nations Publications, 2008.

[DLC11] Giusy Di Lorenzo and Francesco Calabrese. Identifying human spatio-temporal activity patterns from mobile-phone traces. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pages 1069–1074. IEEE, 2011.

[DLM$^+$14] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[DS99] Sajal K Das and Sanjoy K Sen. Adaptive location prediction strategies based on a hierarchical network model in a cellular mobile environment. *The Computer Journal*, 42(6):473–486, 1999.

[EAP$^+$02] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[EGK$^+$04] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[EKK$^+$13] Vincent Etter, Mohamed Kafsi, Ehsan Kazemi, Matthias Grossglauser, and Patrick Thiran. Where to go from here? mobility prediction from instantaneous information. *Pervasive and Mobile Computing*, 9(6):784 – 797, 2013. Mobile Data Challenge.

[FD01] Michael Friendly and Daniel J Denis. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *U RL http://www. datavis. ca/milestones*, 2001.

[FGP11] Katayoun Farrahi and Daniel Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):3, 2011.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *Information Theory, IEEE Transactions on*, 38(4):1258–1270, 1992.

[Fou] FourSquare. https://foursquare.com. Accessed: 2016-05-16.

[FPV$^+$13a] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, Dec 2013.

[FPV$^+$13b] Nuno Ferreira, Jorge Poco, Huy T Vo, Juliana Freire, and Cláudio T Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2149–2158, 2013.

[GHB08] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[GKdPC12] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[Gos00] James Gosling. *The Java language specification*. Addison-Wesley Professional, 2000.

[Har75] John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

[HFH$^+$09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[HJL04] David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[HU14] Samiul Hasan and Satish V Ukkusuri. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, 44:363–381, 2014.

[IBC$^+$11] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, pages 133–151. Springer, 2011.

[ICWG14] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[JBH$^+$12] Anders Johansson, Michael Batty, Konrad Hayashi, Osama Al Bar, David Marcozzi, and Ziad A Memish. Crowd and environmental management during mass gatherings. *The Lancet Infectious Diseases*, 12(2):150 – 156, 2012.

[Jor02] A Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

[KCRB09] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[KE05] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[KH06] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *UbiComp 2006: Ubiquitous Computing*, pages 243–260. Springer, 2006.

[KWSB04] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM, 2004.

[LBH12] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[LD] WILLIAM LJXu-DOEVE. Measurement of internal and international migration.

[LHK$^+$09] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863. IEEE, 2009.

[LTH$^+$92] Hernan Larralde, Paul Trunfio, Shlomo Havlin, H Eugene Stanley, and George H Weiss. Number of distinct sites visited by n random walkers. *Physical Review A*, 45(10):7128, 1992.

[Luc14] Barbara Arese Lucin. Country overview: Bangladesh. Technical report, GSMA Intelligence, 2014.

[Man70] VI Manual. Methods of measuring internal migration. *Population Studies*, (47), 1970.

[MBY$^+$15] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *arXiv preprint arXiv:1505.06807*, 2015.

[MHS95] Hernán A Makse, Shlomo Havlin, and H Eugene Stanley. Modelling urban growth patterns. *Nature*, 377(6550):608–612, 1995.

[MMY$^+$14] Dunstan Matekenya, Ito Masaki, Tobe Yoshito, Horanont Teerayut, Shibasaki Ryosuke, and Sezaki Kaoru. Event detection in individuals using mobile phone traces. In *Proceedings of the 2014 IEICE Communications Society Conference,vol. BS-6*, 2014.

[mob16] August 2016.

[MPTG09] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.

[MRJ12] James McInerney, Alex Rogers, and Nicholas R Jennings. Improving location prediction services for new users with probabilistic latent semantic analysis. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 906–910. ACM, 2012.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[NSLM12] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1038–1043. IEEE, 2012.

[PBKA08] Andrey Tietbohl Palma, Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 863–868. ACM, 2008.

[PHDL$^+$10] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, pages 14–25. Springer, 2010.

[Por00] Leonid Portnoy. Intrusion detection with unlabeled data using clustering. 2000.

[Pro95] Guylene Proulx. Evacuation time and movement in apartment buildings. *Fire safety journal*, 24(3):229–246, 1995.

[PWJX14] Poria Pirozmand, Guowei Wu, Behrouz Jedari, and Feng Xia. Human mobility in opportunistic networks: Characteristics, models and prediction methods. *Journal of Network and Computer Applications*, 2014.

[RZZB12] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[SC93] Elías Sevilla-Casas. Human mobility and malaria risk in the naya river basin of colombia. *Social Science & Medicine*, 37(9):1155–1167, 1993.

[SCV$^+$05] Ian Smith, Mike Chen, Alex Varshavsky, Timothy Sohn, and Karen Tang. Algorithms for detecting motion of a gsm mobile phone. In *ECSCW'05 Workshop on Location Awareness and Community*, 2005.

[SJB14] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2317–2324, 2014.

[SKJH06] Libo Song, David Kotz, Ravi Jain, and Xiaoning He. Evaluating next-cell predictors with extensive wi-fi mobility data. *Mobile Computing, IEEE Transactions on*, 5(12):1633–1649, 2006.

[SKWB10] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[SMM+11] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[tab16] August 2016.

[TCMA12] Le Hung Tran, Michele Catasta, Lucas Kelsey McDowell, and Karl Aberer. Next place prediction using mobile data. In *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*, number EPFL-CONF-182131, 2012.

[TGM83] Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[Tod80] Michael Todaro. Internal migration in developing countries: a survey. In *Population and economic change in developing countries*, pages 361–402. University of Chicago Press, 1980.

[TV04] Emiliano Trevisani and Andrea Vitaletti. Cell-id location technique, limits and benefits: an experimental study. In *Mobile Computing Systems and Applications, 2004. WMCSA 2004. Sixth IEEE Workshop on*, pages 51–60. IEEE, 2004.

[WCDLR10] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE, 2010.

[WHB+12] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

[WIL97] FRANS J WILLEKENS. Probability models of migration: complete and incomplete data. *Southern African Journal of Demography*, pages 31–43, 1997.

[WP12] Jingjing Wang and Bhaskar Prabhala. Periodicity based next place prediction. In *Nokia Mobile Data Challenge 2012 Workshop. p. Dedicated task*, volume 2. Citeseer, 2012.

[WT08] Fleur Wouterse and J. Edward Taylor. Migration and income diversification:: Evidence from burkina faso. *World Development*, 36(4):625 – 640, 2008.

[Yel] Yelp. http://www.yelp.com/. Accessed: 2016-05-16.

[YZ14] Ouri Wolfson Hai Yang Yu Zheng, Licia Capra. Urban computing: concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology (ACM TIST)*, 2014.

[ZFAQ13] Wei Zeng, Chi-Wing Fu, Stefan Müller Arisona, and Huamin Qu. Visualizing interchange patterns in massive movement data. In *Computer Graphics Forum*, volume 32, pages 271–280. Wiley Online Library, 2013.

[ZN12] Jiangchuan Zheng and Lionel M Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 153–162. ACM, 2012.

[ZXYG13] Daqiang Zhang, H Xiong, L Yang, and V Gauither. Nextcell: predicting location using social interplay from cell phone traces. 2013.

[ZZXM09] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[ZZZ$^+$13] Zengbin Zhang, Lin Zhou, Xiaohan Zhao, Gang Wang, Yu Su, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. On the validity of geosocial mobility traces. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, page 11. ACM, 2013.