Urban Computing using Call Detail Records: Mobility Pattern Mining, Next-location Prediction and Location Recommendation

by

Yan Leng

B.Eng., Beijing Jiaotong University (2013)

Submitted to the Department of Civil and Environmental Engineering
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science in Transportation
Master of Science in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2016

Author .....
May 19, 2016

Certified by .....
Jinhua Zhao
Edward H. and Joyce Linde Assistant Professor of City and Transportation Planning
Thesis Supervisor

Certified by .....
Larry Rudolph
Principal Research Scientist
Thesis Supervisor

Certified by .....
Haris N. Koutsopoulos
Professor of Civil and Environmental Engineering
Northeastern University
Thesis Supervisor

Accepted by .....
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Accepted by .....
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students.

Urban Computing using Call Detail Records: Mobility Pattern Mining, Next-location Prediction and Location Recommendation

by

Yan Leng

Submitted to the Department of Civil and Environmental Engineering
Department of Electrical Engineering and Computer Science
on May 19, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation
Master of Science in Electrical Engineering and Computer Science

Abstract

Urban computing fuses computer science with other fields, such as transportation, in the context of urban spaces by connecting ubiquitous sensing technologies, analytical models and visualizations to solve challenging problems in urban environment and operation systems. This paper focuses on Call Detail Records, one widely collected opportunistic sensing data source for billing purposes, to understand presence patterns, develop mobility prediction methods and reduce traffic congestions with location recommendations.

Understanding human mobility and presence patterns at locations are the building blocks for behavior prediction, service design and system improvements. In the first part, this thesis focuses on 1) understanding presence patterns at user locations with a proposed metric Normalized Hourly Presence, 2) extracting common presence patterns across the population with Principal Component Analysis; 3) and infer home and workplaces using K-means Clustering and Fuzzy C-means Clustering. The proposed method was implemented on MIT Reality Mining data, by which we demonstrate that with inference rates of 56% and 82%, the method can improve 79% and 34% in accuracy respectively in home and workplace inference comparing to the baseline model. In addition, it was implemented on the CDR data collected in a crowded city in China to prove its scalability and applicability in real-world applications. With Fuzzy C-means Clustering, we could flexibly trade-off between inference rate and accuracy to understand the interplay between the two and apply it for various purposes.

With an understanding of mobility patterns, the next crucial foundation in urban computing is mobility prediction, enabling transportation practitioners to take actions

beforehand and commercial organizations to send location-based advertisements, etc. Specifically, this paper focuses on next-location prediction from Call Detail Records. Mobility traces was analogized to language models, mapping cell towers to words and individual location traces to sentences. Recurrent Neural Network is a successful tool in natural language processing, which is applied in mobility prediction due to its acceptance of sequential input, variable input length and ability to learn the 'meaning' of cell towers. By implementing the method on Call Detail Records collected in Andorra, we show that the method improved more than 40% over the baseline model, with 67% and 78% accuracy in next location at cell tower and merged cell tower level respectively. The 'meanings' of the cell tower could also be inferred, the same as learning the meanings of words in sentences, from the embedding layer of Recurrent Neural Network.

The last project aims at tackling the challenge of severe traffic congestions with location recommendations. The availability of large-scale longitudinal geolocation data, such as Call Detail Records, offers planners and service providers an unprecedented opportunity to understand location preferences and alleviate traffic congestions. Location recommendation is a potential tool to achieve these two objectives. Previous research on location recommendations has focused on automatically and accurately inferring users' preferences, while little attention has been devoted to the constraints of service capacity. The ignorance may lead to congestion and long waiting time. We argue that Call Detail Records could help planners and authorities make interventions by providing personalized recommendations given the comprehensive urban-wide picture of historical behaviors and preferences. In this research, we propose a method to make location recommendations for system efficiency, defined as maximizing satisfactions toward recommendations subject to capacity constraints, exploiting travelers' choice flexibilities. We infer implicit location preferences based on sparse and passively-collected Call Detail Records. We then formulate an optimization model the defined system efficiency. As a proof-of-concept experiment, we implement the method in Andorra, a small European country heavily relying on tourism. By extensive simulations, we demonstrate that the method can reduce the travel time increased by congestion during peak hour from 11.73 minutes to 5.6 minutes with idealized trips under full compliance rates. We show that the average travel time increased by congestion is 6.17, 6.98, 8.37 and 10.98 minutes with 80%, 60%, 40% and 20% compliance rates. Overall, our results indicate that Call Detail Records can be used to make locations recommendation while reduce traffic congestion for system efficiency. The proposed method can be applied to other large-scale location traces and extended to other location or events recommendation applications.

Thesis Supervisor: Jinhua Zhao
Title: Edward H. and Joyce Linde Assistant Professor of City and Transportation Planning

Thesis Supervisor: Larry Rudolph
Title: Principal Research Scientist

Thesis Supervisor: Haris N. Koutsopoulos
Title: Professor of Civil and Environmental Engineering
Northeastern University

Acknowledgments

This thesis represents not only me at the keyboard, it is a milestone in three-year work at MIT transit lab and Human Dynamics Group at Media Lab pursuing two master degrees in Transportation and Computer Science. My experience at MIT has been nothing short of amazing. I start my journey at transit lab and I have been given unique opportunities by Professor Haris Koutsopolous, Professor Jinhua Zhao and have taken advantage of them. It is my fortunate to be advised by Professor Larry Rudolph and Sandy Pentland during my third year when pursuing dual degree in Computer Science. During this journey, I gradually realize and confirm my passion in big data and human behavior analytics. I appreciate the opportunity in spending three invaluable years at MIT, where I learned how to think, how to collaborate, how to interact with people, and how to view the world in a scientific way.

First, I wish to thank my advisors from Transportation group at CEE Department, Professor Haris Koutsopolous and Professor Jinhua Zhao. Haris's critical thinking, continual support and guidance has benfited me profoundly in shaping how to think and do research. Haris's questions towards the made assumptions, analysis and methodologies during meetings have great impacts in how I approach projects later on. I would like to thank Jinhua for his insightful comments and creative brainstorming sessions. Jinhua is suppportive and encourating for the various ideas. Jinhua is not only a great advisor but also a good friend with valuable suggestions in the choices I made.

Secondly, I wish to express my great gratitude to Larry Rudolph. Larry has helped me a lot in forming interesting research questions, such as my location recommendation idea. During the engaging discussions, Larry is both critical and supportive. Along with his disagreement at the arguments I brought up, Larry also help me find solutions. Larry taught me not only how to do research, but also how to think and

write critically and creatively.

I was fortunate to join human dynamics group at Media Lab directed by Professor Alex "Sandy" Pentland. Sandy is very encouraging and inspiring. He helped me in a deep way in telling compelling stories and mining shining gold nuggets in the data. Thanks to Sandy's trust, I've been given the opportunity to present *Andorra Living Lab Project* to the machine learning related groups in media lab. This was an invaluable experence that I will always remmeber.

Many people have helped and taught me immensely at transitlab, JTL and human dynamics group at media lab. I received insightful inputs and comments from the presentations during group meetings at these labs. It was a pleasure to discuss research ideas and receive feedbacks from my fellow students. Thanks to Zhan Zhao's input on both the technical and directions of my research, Saul Wilson's kind feedback on my master thesis, Xiaowen Dong's discussing on various research ideas, Luis Alonso's bridging me with Andorran experts, Alejandro Noriega Campero's helping on mining insights and visualizations for Andorra events analysis.

I would like to recognize all the friends I have made during my time at MIT. I want to express my deep appreciations to Zelin Li, Yiwen Zhu, Haizheng Zhang, Tianli Zhou, Yingzhen Shen, Zhuyun Gu, Shenhao Wang, Dan Calacci for the accompany during the down times and making my experience at MIT enjoyable.

Finally, I would like to express my profound gratitude to my parents, Song Leng and Yaqin Li for the constant support over the past 25 years, for always standing by me in all situations with their unending love and encouragement. I would also like to express my gratitude to Siyuan Liu, for the company and moral support during the past year. He has made life wonderful and look forward to many new and exciting experiences with him.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis focuses on three transportatoin-relaated research questions in the framework of urban computing: *understand* mobility patterns, *predict* mobility behaviors and *change* travel behaviors. The concept of ***urban computing*** was first formally introduced in 2007 [1, 2]. It is an interdisciplinary field fusing computing science and other traditional fields, such as transportation, in the context of urban spaces. It connects ubiquitous sensing technologies, analytical models, and visualizations to improve urban environments, city operation systems and qualities of human life. Urban computing can helps to understand urban phenomenon, predict the future of the cities, and make interventions in shaping the future in the correct direction [3]. The increasing storage, process, and model ability to handle massive amounts of data offers an opportunity to find better methods to interpret and transmit the data in a world and a smart city where people and machines are more inter-connected [1].

Motivated by the great opportunities of tackling big challenges in building smart cities[1] based on big data, this project empowers knowledge mined from Call Detail

[1] A *smart city* is an urban development vision to integrate multiple information and communication technology (ICT) solutions in a secure fashion to manage a city's assets. The city's assets include, but are not limited to, local departments information systems, schools, libraries, transportation systems, hospitals, power plants, water supply networks, waste management, law enforcement, and other community services. The goal of building a smart city is to improve quality of life by using technology to improve the efficiency of services and meet residents' needs [4].

Records collected in smart cities and comes up with a solution towards the major *transportation* and *mobility* issues our urban spaces face, as shown in Figure 1-1. There are multiple sources of big data, collected by sensing technologies and processed by large-scale computing infrastructures, which provide rich information about human mobility and urban spaces.

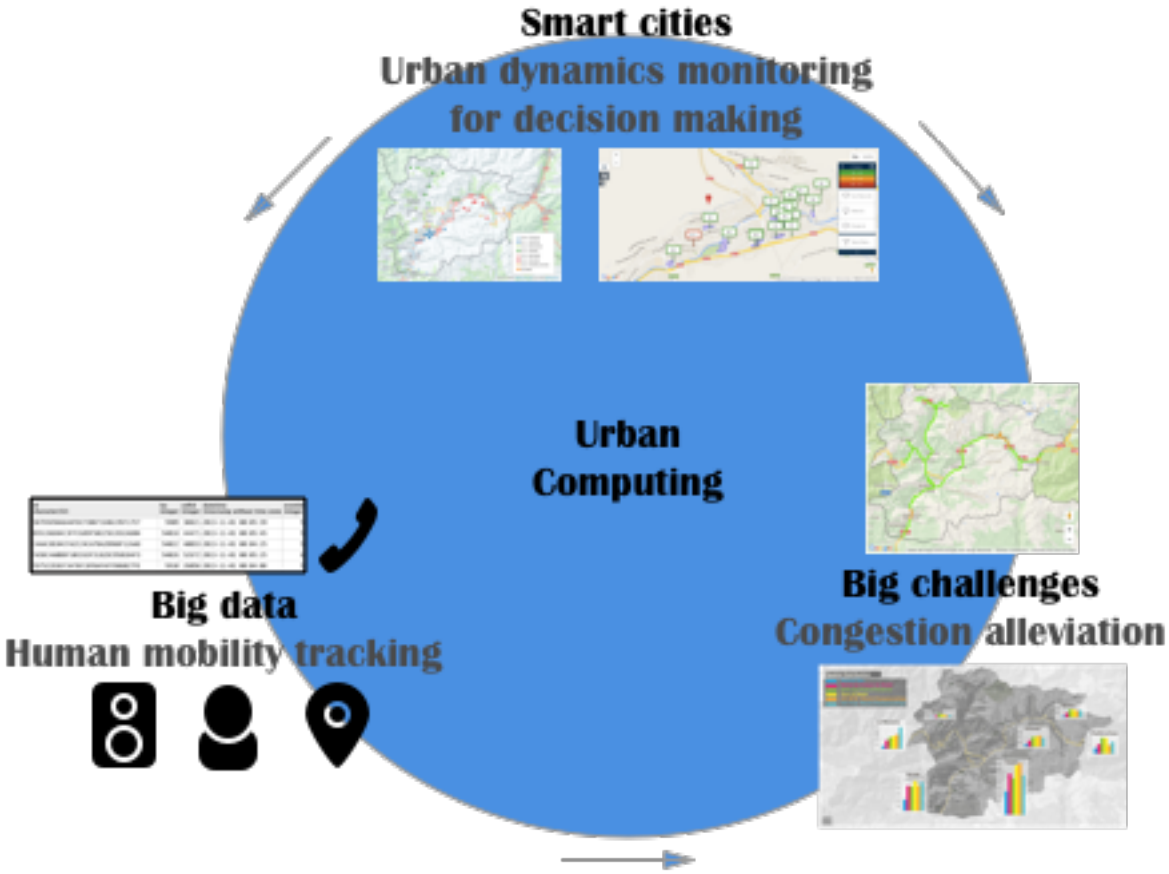

Figure 1-1: Big question: smart cities, big data and challenges

1.1 Background

Under the schema shown in Figure 1-1, Figure 1-2 depicts a general framework of this thesis, which focuses on a subset of urban computing and composed of urban sensing, urban data management, urban and mobility data analytics and service improvements [1]. This thesis aims to tackle big challenges in smart cities by using patterns inferred

and predictions made using big data.

Urban computing seeks to develop computational solutions that make cities more livable, more efficient, and better positioned for the coming decades [4]. There are three important aspects: understanding existing patterns, forecasting mobility behaviors and making interventions to shape a better future.

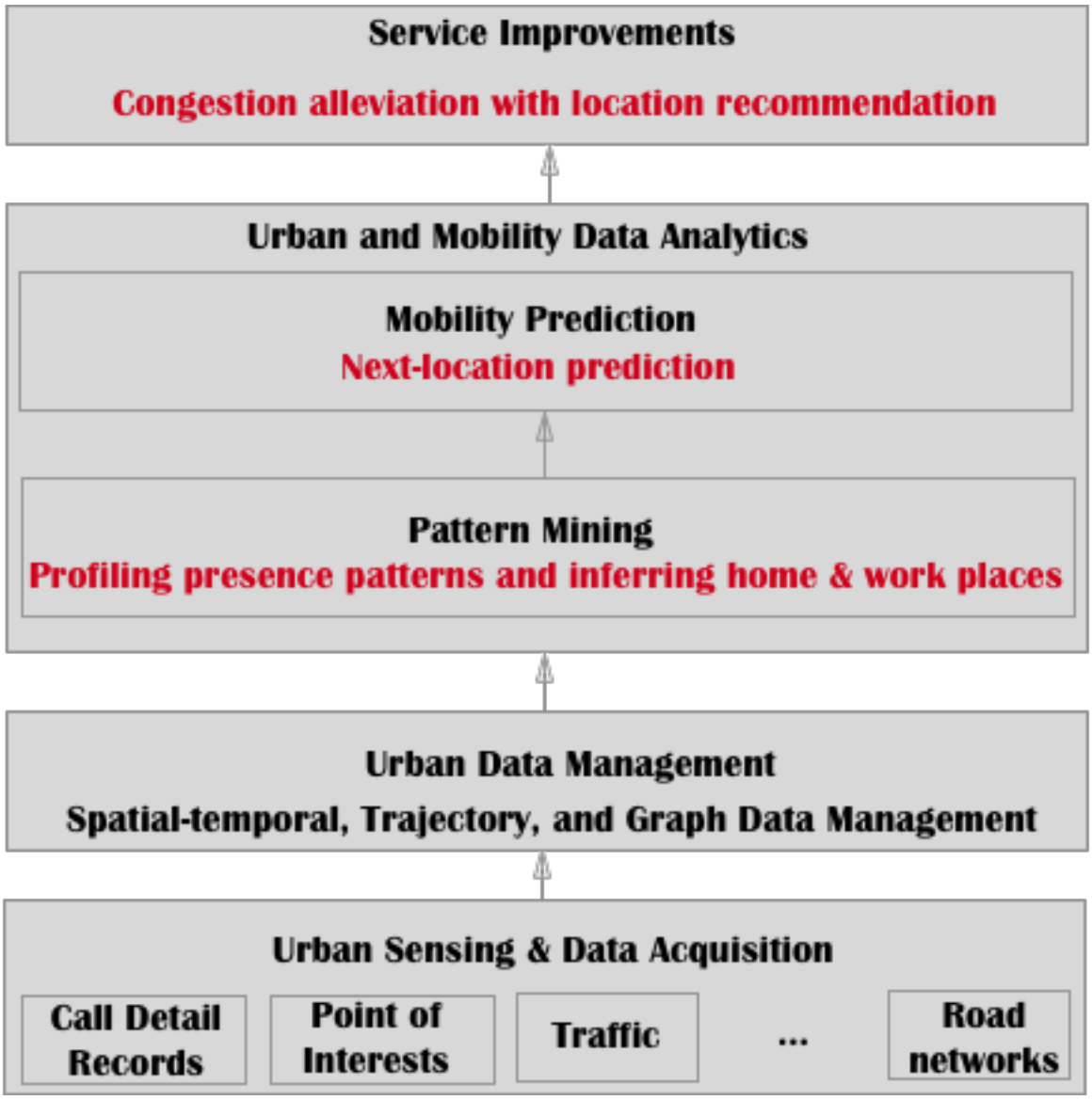

Figure 1-2: Framework of urban computing for transportation

Urban sensing data are collected through various methods. A small sample of frequently used data sources in transportation and mobility is listed. Mobile phone

records are passively collected by a mobile phone exchange for billing purposes. Points of Interest (POI) are collected by Yellow Page, , map data providers and manually logged POI information by social network users. Traffic data are collected by loop sensors, surveillance cameras, and floating cars. Map service providers and urban planning departments provide road network infrastructure information, containing length, speed constraints, road types and number of lanes. The dataset used in this thesis is Call Detail Records (CDR), which can reveal some salient aspects of human mobilities with some analysis.

In the *Data Management step*, the human mobility data are well organized by some indexing structure that simultaneously incorporates spatiotemporal information, the important characteristics of mobility data, for efficient data processing.

In the *Urban and Mobility Data Analytics* step, as shown in Figure 1-2, there are two components, *Mobility Pattern Mining* and *Mobility Prediction*. To make CDR useful for transportation planning and policy making, it is critical to understand human mobility and urban dynamics. Building upon this understanding, we could forecast population dynamics in the future.

The *Service Improvement* component of Figure 1-2 applies the reliable inference and prediction in the previous layer to solve congestion problems, which are big challenges in building smarter cities. One of the most appealing aspects of urban computing lies in its ability to reflect the congestion problems and provide urban planners and transportation practitioners with the opportunities to better build an efficient transportation system.

1.2 Motivations

In line with the backgrounds and framework described in section 1.1, this thesis has three motivations: to *understand* mobility patterns, to *predict* mobility behaviors and

to *change* travel behaviors. The detailed motivations in each category are decribed in this section.

1.2.1 Understand mobility patterns

Understanding human mobility is the basis that transportation operators and urban planners use to plan and design better location-based services. Large-scale behavioral datasets create new opportunities in delving into travel behaviors. It is critical to understand presence patterns at user location and labeling home and workplaces (the definitions are given in section 2.3.2).

Understanding the distributions of homes and workplaces helps service providers and planners to dynamically monitor the commuting trip pairs on a regular basis in the fast-developing regions with rapid emergence of new residential and commercial areas. This helps public transit operators better plan and optimize services in response to the changing demand. Due to the high stability and close connectivity with other locations, home and workplaces are the locations that should be inferred and profiled as the starting points and building blocks of understanding human mobility.

Home and workplaces are the two anchor points where individuals spend most of their time, representing the dynamic distribution of the population during different time period. They are also interconnected with other user locations and other activities by different trips. Moreover, the long dwelling time and high frequency also ensures the relatively higher accuracy in the inference process comparing to other random user locations. In addition, they are also the origins and destinations of commuting trips, main components of the traffic flow during peak hours.

An important application in inferring home location distributions lies in its linkage role in bridging CDR-based Origin-Destination (OD) matrix and actual OD matrix. Inferring multimodal OD matrix from CDR data provides a dynamic and low-cost method to understand the travel demand. However, the mobile phone penetration

and phone usage behavior create a gap between the actual travel demand and the observed CDR-based travel demand. To scale and expand the CDR-based demand to actual demand, comparing the population distributions is a way to calculate the expansion factors and scale up the OD matrix.

1.2.2 Mobility prediction in the future

With the explosion of location data from cell towers, Wi-Fi beacons and various mobile sensors, individual travel behaviors are more predictable and tractable than ever before. The availability of this data makes next-location prediction more accurate and predictable based previous footprints, which is the focus of the second research question. This is a significant building block benefiting many areas, including mobile advertising, public transit planning and urban infrastructure management [5, 6, 7]. For example, from the business side, predicting where people will go enables location-based service advertisers to distribute targeted coupons. Besides, from the systematic side, predicting travel demand enables public transit agencies and events organizers to take proactive actions responding to dynamic demand [8].

Predicting human mobility has been an increasingly popular topic in pervasive computing based on GPS, bluetooth, check-in histories, etc. Different data sources vary at spatial, temporal scales and the availability of contextual information [9]. Many researches build markov models and predicts the longitude and latitude as continuous variables based on previously travel trajectories [10, 11, 12]. Mathew (2012) [13] model location prediction as discrete varibles with contextual variables with activities and purposes with Hidden Markove Model. Domenico (2013) [9] and Alhasoun (2010) [5] uses mobility correlations, either measured by social interactions or mutual information, to improve forecasting accuracy. Though extensive researches have acceptable accuracy in predicting next locations, the performances of the current literature on CDR are poor [5]. Though CDR gives less accurate information, but

surprisingly are sufficiently accurate for Recurrent Neural Network.

1.2.3 Change travel behaviors: Interventions for system efficiency

The persuasive availability of large-scale geolocation data from mobile devices offers an unprecedented opportunity for location-based service providers, transportation agencies, tourism departments and governments to understand human mobility, provide personalized information and improve system operations. Location recommendation has been studied by researchers and applied in industry as a tool to recommend locations according to inferred preferences. However, unlike other recommendation problems, recommending locations simply according to travelers' preferences will lead to congestions and long waiting time due to the capacity constraints of road links or services.

First of all, large-scale location traces present a big picture of urban or country-wide behaviors for planners, transportation practitioners and service providers. The visibilities of population-wide behaviors,interests and the decision-making process across all the population enable authorities make interventions at a systematic level. Therefore, a recommendation system was built based upon not only satisfying personal preferences but also making the best use of the system capacity by balancing the demand.

Another critical component enabling location recommendations for system efficiency is the possibility in exploiting personal choice flexibilities at activity, location and temporal levels, specifically, what to do, where to go and when to go. Some travelers, especially for tourism purposes, care more about the activity instead of the locations for conducting the activities. This is more common for leiture trip purposes. For these groups of tourists, giving them recommendations on where and when to go would help them make informed decisions. Moreover, on the temporal side, when

travelers make decisions on when to carry out an activity, they will make decisions blindly within some constraint if no information is given [14]. Many travelers want to minimize the amount of congestion experienced on the routes and are willing to make destintion changes if heavy traffic is expected. The time for traveling are flexible within certain time constraint. Under these conditions, which is often the cases, the visibility of what other people's behaviors and congestions along road links will guide them make better decisions. This project defines travelers' freedom in deciding what to do, where to go and when to go as *choice flexibility*.

1.3 Research Objectives

In line with the motivations described above, this research identifies three main directions as shown in Table 1.1. The detailed research questions are depicted in the following session.

Table 1.1: Research directions

	Categories	**Research directions**
1	Mobility pattern mining	Segmenting and profiling user locations from cell phone data.
2	Mobility prediction	Next-location Prediction using Recurrent Neural Network.
3	Change behavior	Location recommendations exploiting personal choice flexibilities for system efficiency based on large-scale Call Detail Records.

Segmenting and profiling user locations from cell phone data The first research direction is to understand one dimension of mobility behaviors, the pres-

ence patterns [2] at user locations [3], which enable the identification of two important locations: home and workplaces, in daily routines more reliably and accurately.

Specifically, the main research questions this thesis aim to answer include:

1. Characterize user locations by hourly presence patterns based on the longitudinal observations from Call Detail Records.

2. Identify the underlying common behavioral structures at user locations across the urban population.

3. Develop a method to dynamically categorize user locations into home, workplaces and elsewhere based on the presences patterns.

4. Profile user location segmentations by the temporal presence patterns and spatial characteristics.

Next-location prediction Call Detail Records have been widely used in transportation research and industrial applications [3, 15]. Advances in location predictions is the foundation in pushing these researches in real applications. The second part of this thesis proposes a method to predict next location for each individual based only on historical location sequences from Call Detail Records. The accurate prediction method for CDR is also a critical foundation for other researches.

This project maps moblity model to language models based on the analogy between sentences and mobility traces, words and cell towers in movements. Recurrent Neural Network, a successful tool in natural lanauge processing, will be used to 1) learn the meaning of the cell towers, the basic unit of location traces, 2) make better next-location predictions. More concretely, the answer to the following questions are interested in:

[2]Presence is defined as the appearance of a user at a user location. Both terms are individual based.

[3]User location is defined as the weighted centroid of a cluster of cell towers that approximates the true locations of a user.

1. Whether mobility traces can be analogize to sentences.

2. Whether Recurrent Neural Network can be used in large-scale next-location predictions using Call Detail Records.

3. How does the performance of Recurrent Neural Network compare with widely-used markov models in next-location prediction on Call Detail Records.

Location recommendation exploiting choice flexibilities for system efficiency In line with the motivations mentioned in section 1.2.3, section 4 proves the usefulness of applying CDR data to mine implicit preference towards locations and make interventions for system efficiency with location recommendations. The detailed questions include the following:

1. How to proxy location preferences from mobile phone data and learn idealized preferences for new locations.

2. Whether Call Detail Records can be used in large-scale location recommendations with spatial and temporal information.

3. Whether and how location recommendations can be used to alleviate congestions and improve system efficiency.

4. To what extent can location recommendations help to reduce traffic congestions.

5. How to factor in the uncertainties in the performance the method.

1.4 Call Detail Record Data

The availability of spatial and temporal information from Call Detail Records lead to tremendous research and applications analyzing and modeling human mobility for

various purposes. Blondel [16] reviews the applications of CDR in many fields, including mobility, epidemics, spatial networks, dynamic networks, information diffusion, urban planning, event detections, etc. As stated in the previous sessions, this thesis mainly focuses on the potential of applying Call Detail Records in transportation and mobility analysis.

1.4.1 What is Call Detail Records

Call Detail Records (CDR) is a universal and opportunistic data source originally used for billing purposes by telephone service providers. Each record contains the transaction time and transaction location of the mobile phone user. The items in CDR data include encrypted user ID, Location Area Code (LAC), Cell ID, timestamp and event type, as shown in Figure 1-3. LAC and Cell ID jointly determine the geographical location (coordinates) of the cell tower and need to be linked to cell tower database to acquire the longitude and latitude of the cell tower. The event ID records the type of the transaction, usually consisting of call in, call out, messages and web-browsing.

	mcc integer	mnc integer	lac integer	cell integer	lng double precis	lat double precis	o_lng double precis	o_lat double precis	precision integer
1	460	1	0	33	117.053369	36.696597	117.059388	36.697001	3875
2	460	0	0	162	116.8876937	36.6253486	116.8876937	36.6253486	0
3	460	0	0	239	117.029929	36.654334	117.035824	36.654629	3875
4	460	0	0	447	117.050153	36.649177	117.056167	36.64955	3875
5	460	0	0	692	117.013695	36.67825	117.019553	36.678525	2440

Figure 1-3: Screenshot of CDR

In addition to the above mentioned spatial and temporal information, CDR stores other information that are useful to infer the socio-demographics of the individuals, as shown in Figure 1-4. For example, researchers can infer the phone type that the user is using, including the brand, the vendor, the model and the system. This can be used as the proxy of the disposible income of the user. In addition, the registration

country/city of the individual can be learned, enabling the inference the nationality or home city of the user. Inferring the characteristics of the user is useful in various applications.

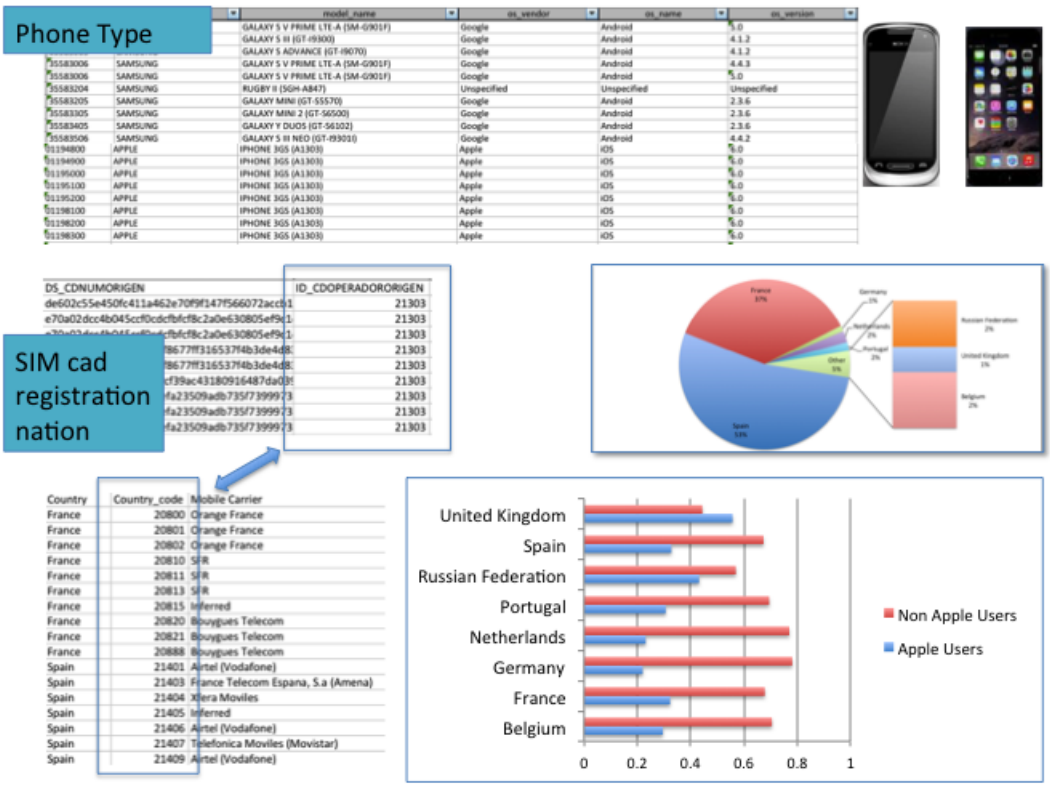

Figure 1-4: User information from CDR

Call Detail Records generates many new possibilities in travel behaviors at both individual and aggregate level. It is the most widely-penetrated data source to be used as a proxy for travel behaviors. There has been large amounts of efforts in using CDR in transportation to solve critical problems. Alexander (2015) [15] estimate Origin-Destination matrix and segment trips based on the inferred trips purposes, including home-based work, home-based other and non-home-based. Kung (2014) [17] compare commuting patterns, including travel time and distances, across different parts of the world at both country and city level. Phithakkitnukoon (2010) [18] analyze the corre-

lations in activity patterns for people who work in the same land use categories, which can further be used to estimate the most probable activities associated with certain regions of the city. Diao (2015) [19] and Calabrese (2014) [20] extract the embedded travel activity information and infer activity patterns by merging mobile phone data and travel diary surveys. Identifying home and workplace locations are the building blocks for not only the listed researces but also other real-world applications. From an application point of view, understanding the spread of individual daily activities, detecting emerging residential and commercial regions, help us better monitor urban dynamics and spatial-temporal distribution of the population. Knowing home and workplaces enable us to better deliver location-based service, such as targeting ads, localized services and etc. It is also important to combine CDR with census by scaling up mobile phone users to the whole population [21].

1.4.2 Challenges of Call Detail Records

Just as every coin has two sides, CDR has some drawbacks, making it challenging to be used in tranposrtation research and real-world applications.

First, CDR has *low spatial resolution*. The coordinates of the cell towers approximate the geographic location of the user, covering a 2000-meter radius area. The low spatial resolution generates great challenges in positioning the locations of individuals at compact and dense areas. The true locations will be obfuscated, which makes it hard for location-based service providers to pinpoint the spatial location of their customers. One way to deal with this is to process triangulation on raw CDR using three or more cell towers for each phone record. AirSage provides triangulated CDR data with a spatial resolution of 200 300 meters. However, this is not available for most mobile carriers.

Second, the call record is *event-driven*. The record is stored in the database only when a connection to the cell tower, such as a phone call, text message or web

browsing, is made. This sparse sampling characteristics results in a gap between travel behavior and presence pattern. One widely used and possibly problematic assumption made in researches and applications of CDR data is the highly correlation between mobile usage and presence pattern at different locations. It is known that different segmentations of mobile phone users have different usage patterns. Therefore, some sparse mobile usage groups may be under-sampled and frequent mobile users may be oversampled. Specifically for call behaviors are home and workplaces, people are likely to use landlines at these regular and fixed locations due to the low costs. One way to narrow the gap between mobile usage and actual travel behavior is to use frequent mobile users. However, this may generate bias due to the difference between the behavioral patterns of frequent and infrequent users. This project focuses on large-scale home and work detection and the method can be applied on frequent users, which will give a more accurate inference results within this specific group of people.

Third, it needs to be acknowledged that the mobile device ownership is another concern influencing the inference results of CDR data. Two concerns are related to mobile device ownership. For one thing, it is difficult to detect individuals who own more than one mobile device. For another, some tablets that are placed at a fixed location with no movements also confuse the presence patterns at home and workplaces.

1.5 Thesis Organizations

The thesis is organized into five chapters as follows. The first chapter introduces the background, motivations, research objectives and the specific data and Call Detail Records. In line with the motivations and research objectives, the successive chapters include three researches, include: *understand* presence patterns at user locations and

label home and workplaces, *predict* next location and *recommend* locations towards more systematic efficient level.

Chapter 2 is organized in the following way. Related works on inferring home and work locations based on CDR are introduced first. The next section describes the conceptual framework, the definition and calculation of Normalized Hourly Presence and eigenlocation, and the clustering techniques. The proposed method is tested on MIT reality mining data and compared with the state-of-art method in the literature. In addition, the method was implemented in real-world CDR, collected in a populated city in China, to validate the feasibility, practicability and scalability. The last section summarizes the methods and concludes with future works.

Chapter 3 is organized as follows. Recurrent Neural Network is introduced first, starting from problem formulation, followed by mapping to language models and descriptions on RNN. Next, RNN is implemented as a case study, tuning parameters and analyze the prediction results. Conclusions and future works are introduced in the end.

In Chapter 4, the framework of the methodology and detailed descriptions of each step are demonstrated, including preference inference, collective satisfaction maximization and traffic flow balancing and traffic flow inference. Next, a case study in Andorra is described. The required datasets is introduced. After that, the impact of the method on travel time decreasing compare to no intervensions and preference-based recommendation is analyzed. The next part evaluates the impacts and robustness of the method by varying allowed throughput and compliance rates. The last section summarizes results, points out some future works and potential applications.

Chapter 5 summarizes the methods and contributions of the thesis. The main limitations and future research directions are discussed in the end.

Chapter 2

Segmenting and Profiling User Locations from Cell Phone Data

Understanding human mobility and urban dynamics is critical for urban planners, transportation operators and location-based service providers. Call detail records (CDR) from cell phone data is an opportunistic, longitudinal and large-scale data source, creating new possibilities to track individual movements and monitor aggregate mobility patterns at flexible temporal and spatial scales. Home and workplaces are two most significant locations in individuals' daily lives and therefore they are the building blocks of various models in the transportation field, such as inferring commuting mode, trip purpose, etc. However, traditional ways of inferring these two user locations with household travel surveys and censuses, are labor-intensive, limited in scale, and infrequent. As a low-cost and regular complement to the surveys, we develop a method to infer home and workplace locations based on presence patterns extracted from CDR data. Current literatures, making simple and biased assumptions, are problematic in inferring home and workplace locations due to the large discrepancy between observed presence pattern from cell phone data and actual presence pattern of individuals. There are three questions we explore: (1) how to extract

individuals' presence patterns at user locations from the noisy CDR data, (2) whether there are common presence structures across the population, and (3) how to identify home and workplace locations more accurately. We implement the proposed method on MIT Reality Mining data and the CDR data collected in a fast-developing city in China. With the inference rates of 56% and 82% for home and workplace locations, we improve the accuracy by 79% and 34% respectively over other methods proposed in the literature. The application on real-world settings proves its feasibility in extracting common behavioral patterns and scalability in real-world implementations.

2.1 Introduction

Understanding human mobility and urban dynamics is the basis for transit operators, urban planners and location-based service providers in better understanding transportation demand, planing services and implementing urban policies. Traditionally, the mainstream datasource in understanding travel demand is household travel survey providing abundant transportation records including socio-demographic information, travel time, trip purposes, travel mode, etc. While it provides detailed travel logs and personal information, it is expensive to administer and participate in many aspects [22]. For example, the time between surveys conducted in developed countries are around 5 to 10 years, making it impossisble to keep pace with the rapid urban development. Furthermore, it is labor-intensive and costly to conduct the large-scale surveys.

The rise of ubiquitous digital data collection infrastructures, embedded in urban areas, leads to a dramatic increase in big data resources in monitoring urban dynamics and human mobility in an unprecedented large scale and finer granuality [16, 23]. Various urban sensors, such as cell towers, WiFi hotspots and bluetooth beacons, are exploding in the building of "smart cities" and making pervasive computing possible

[19]. Digital devices and sensors have "intruded" into our environments and monitored various aspects of our lives, including traveling, healthiness, transactions, and etc. With the explosion of information, machine learning and statistical tools have been applied by academia and industry to informational retrival, pattern recognition, insights generation and decision support systems. Overloaded with the amounts and variaties of data sources, we need statistical and machine learning techniques reveal salient behavioral features, inferring and generating hypothesis about behavior patterns, in order to optimize transportion systems, support planning decisions, dynamically monitor regional delineations or land use classifications [24].

Along with the exciting opportunities and wide applications, challenges are unavoidable. There are uncertainties, complexities and biases in the data collection as well as human behavior itself. Despite the importance of accurately inferring home and workplaces at individual and aggregate level, existing methods are not accurate and flexible enough in inferring home and workplaces, most of which are based on simple assumptions. This paper proposes a methodology to extract behavioral patterns at locations and infer home/workplaces in urban spaces based on Call Detail Records by mapping physical CDR coordinates to meaningful user locations with enriched interpretations. This is achieved by answering three questions: 1) what are the behavioral patterns at the user locations based on the longitudinal observations, 2) are there any common behavioral structures across the population, 3) add contextual information to user locations by labeling home and workplace locations.

The contributions of our work are threefolds.

1. We propose a universal method to infer home and workplaces on CDR with proved better accuracy.

2. We propose a feature, Normalized Hourly Presence, to extract behavioral characteristics from CDR-based user locations and extract shared behavioral patterns across the population using Principle Component Analysis.

3. The method is applied on the CDR data in a populated city in China, which proved its feasibility and scalability in revealing the behavioral patterns and labeling home/workplaces in real-world settings.

The remainder of the paper is organized as follows. Section 2.2 introduces related works on inferring home and work locations based on CDR. In Section 2.3, we describe the conceptual framework, the definition and calculation of Normalized Hourly Presence and eigenlocation, and the clustering techniques. In Section 2.4, we test the proposed method on MIT reality mining data and compare it with state-of-art method in the literature. We also implement the method in real-world CDR, collected in a populated city in China, to validate the feasibility, practicability and scalability.

.

2.2 Related works

With the increasing penetration and rising popularity of mobile phone and mobile communication, passive mobile phone location data has become a possible source of geographical data source to locate individuals and detect significant locations [25]. There are many researches into the method for inferring home and workplaces using CDR. The most widely-used method for inferring home and workplaces assumes that home and workplace locations are the two locations people visit the most frequently, measured by aggregating preferences by user locations. [26], [18] and [3] label user locations with the most presences during home and work hours as home and workplaces. The simple assumption that the amount of connections is proportional to stay length is problematic. People may use landlines at home or workplaces; in reality, user may not have home or workplaces that are detectable using CDR data. Also, some users may have multiple home and workplaces from CDR, which is not possible under such a simplistic algorithm.

Moreover, some methods use dwell time as the way to measure stay length. [17] assume that home and workplaces are those with longest dwell time during home and work hours. Similarly, [27] set a dwelling time threshold for user locations to be labeled as home and workplace locations. It should be noted that dwelling time at a specific user location is calculated as the time difference between the first appearance at the user location and first appearance at another user location. However, this is unreliable due to the event-driven characteristic of CDR data. User locations with no connections are ignored due to lack of observations. Moreover, the boundary thresholds to separate home and work hours are arbitrary.

Distance is used to capture the characteristics of workplace. According to Alexander (2015) [15], home location is assumed to be the most frequent location during home hours (before 8 am and after 7 pm) as in most other research. Meanwhile, work place is assumed to be the user location with the maximum distance from home. This is based on the assumption that for a given frequency of visits, longer trips are more likely to be commuting trips than shorter trips. This assumption, basing on rationale and historical empirical evidence, needs further validation on recent commuting patterns. Besides, the travel characteristics from travel surveys, which have high spatial granularity, are different from the untriangulated CDR observations with a spatial resolution of 2 kilometers, making the empirical distance learned from survey not appliable in CDR.

In addition, there are methods that use small-scale experiments to calibrate parameters and calculate thresholds. Ahas (2010) [25] proposed that the mean and the standard deviation of the earliest call to be used to differentiate home and workplace, which should be among the most frequent two user locations. The rationale is that people either call early in the morning or late at night at home. If the mean of the earliest call is later than 17:00 or the standard deviation of the earliest call is larger than 0.175, the user location is labeled as home. The thresholds were calcu-

lated from 2-month CDR data for 14 individuals. However, small-scale experiments unrepresentative, infrequent and unavailable in many cities.

From the above discussions, most of the methods proposed in the literature are problematic in ignoring the large gap between actual presence patterns and the locations from which people make phone calls and leave CDR traces. Further, location-independent time boundaries and thresholds are used to differentiate home and workplaces from other user locations, but they are arbitrary. When these values have been set by small-scale experiments, they are inherently location-dependent due to the cultural differences, and they are uneconomical to use for CDR analysis.

2.3 Methodology

In this section, we describe our method to map CDR-based user locations to home, workplace and other places based on longitudinal numeric coordinates. Two basic terms used throughout this paper are *user location* and *presence*. Both terms are individually based.

Definition 1. *User location.* We define user location as the weighted centroid of a cluster of cell towers that approximates the true locations of a user.

Definition 2. *Presence.* Presence is defined as the appearance of a user at a user location.

2.3.1 Data description

CDR records individual traces with timestamps and approximate locations of cell phone users whenever they initiate phone calls, send/receive SMS or browse the web. It is event-driven and does not cover the full picture of places that people have traveled to. The raw CDR data include encrypted user IDs, timestamps, Location Area Codes (LAC), cell tower IDs and event types. A seperate cell tower database furnishes the

location of the cell towers, approximating the locations of the individuals. This is why the spatial resolution of CDR is low.

2.3.2 Methodological ramework

The conceptual framework of the problem is shown in Figure 2-1. For each user, we observe a sequence of coordinates with timestamps, representing the digital footprints of the user across the observational period. Each record, which we refer to as "presence" in this paper, can be associated with an activity type, which is the trip purpose. However, there is no one-to-one correspondence between activities and user locations, meaning that a specific activity can be performed at different user locations and several activities can be conducted at the same user location. Therefore, the activity layer is unobservable from the passive-positioning and semantic-poor CDR data. Activities are conducted at a limited number of user locations, which can be roughly segmented into home/workplace, as the anchor points, and elsewhere. Presence patterns at user locations are one of the most basic aspects of human mobility. The proposed method skip the activity layer and identify home/workplace from user locations based on the observed longitudinal presences. CDR-based home and workplace are different from the traditional concept of home and workplace. *CDR-based home* is the user location with home-like normalized presences. Similarly, *CDR-based workplace* is the user location with weekday and weekend work-like normalized presences.

The method is shown in Figure 2-2. The detailed steps are introduced and explained in the following sessions.

2.3.3 Normalized hourly presence

We propose to use a feature called Normalized Hourly Presence (NHP) to capture when, how often and how long each user appears at each user location. This feature enables the extraction of not only the location-based presence frequencies, but also

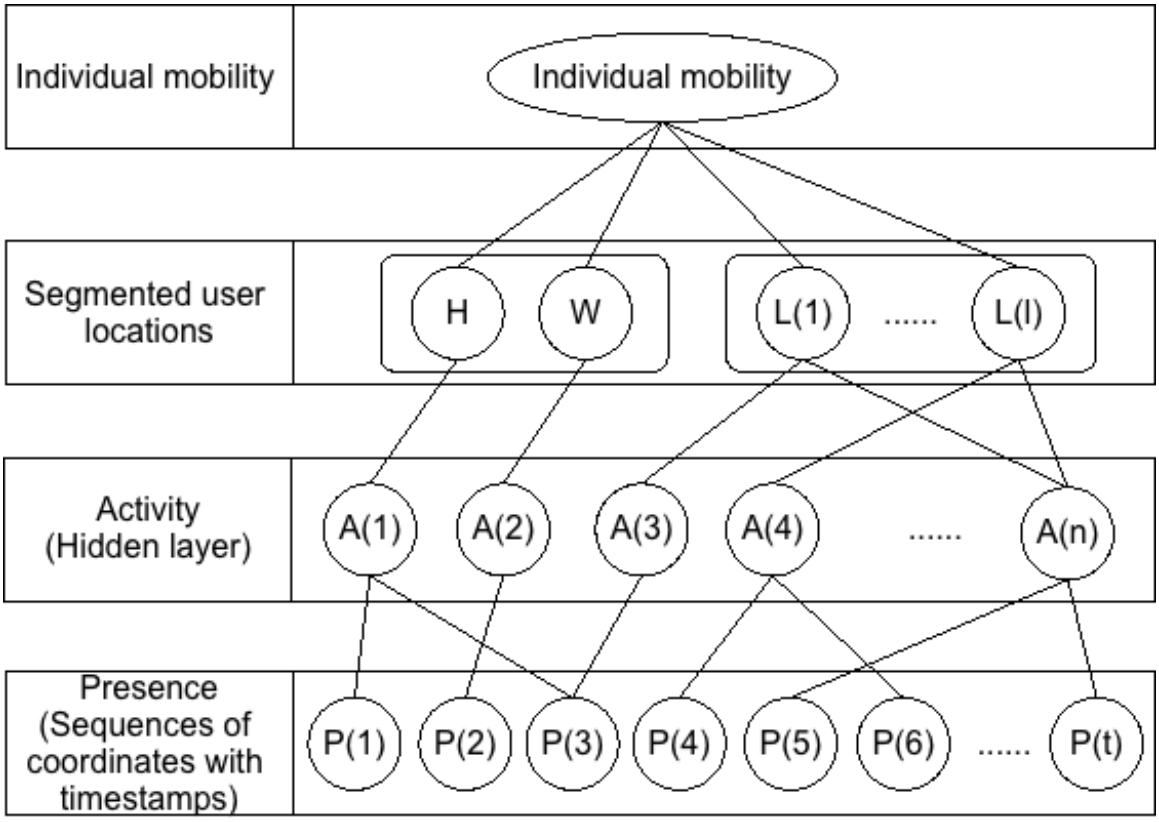

Figure 2-1: Conceptual Framework for Home (H), Workplace (W), and Other location (L)

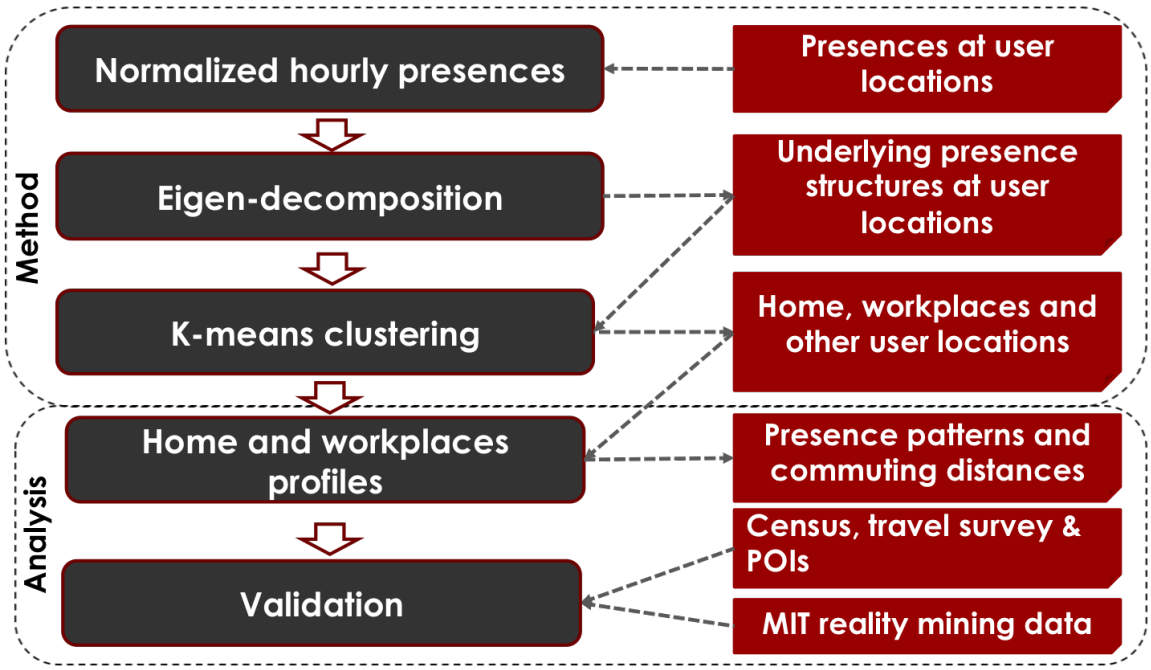

Figure 2-2: Framework of the method

the temporal variations in presence frequencies at user locations on weekdays and weekends.

We sum the number of presences in each hour of a weekday and a weekend across the observational period due to the sparsity of the data. Weekdays and weekends are aggregated seperately for the different presence patterns. The hourly presences are scaled to the percentage of presences with respect to the total number of phone connections of the individuals. This captures not only the frequencies for visit but also the heterogeneous call rates. Equation (2.1) shows the calculation of NHP, which is processed in PostgreSQL. In essence, each user location is characterized by a vector with forty-eight NHPs.

$$\widetilde{P}^{l}_{i,h} = \frac{P^{l}_{i,h}}{\sum_{j=1}^{L_i} P^{j}_{i,h}} \tag{2.1}$$

where $\widetilde{P}^{l}_{i,h}$ and $P^{l}_{i,h}$ are the NHPs and absolute hourly presence for individual i during hour h at user location l. $h \in [1, 48]$, representing 24 hours on weekdays and

Hour	4:00 - 5:00	5:00 - 6:00	6:00 - 7:00	7:00 - 8:00	8:00 - 9:00
User location 1	0	0	1	1	4
User location 2	0	0	0	0	1
User location 3	1	1	5	3	1
User location 4	0	0	1	0	0
User location 5	0	0	0	1	1

Figure 2-3: Variances explained by each principal components

weekends respectively. $L_{i,h}$ represents the unique set of user locations for individual i during hour h.

A made example is shown in Figure 2-3. User A presented at 5 user locations for 1, 0, 5, 1 and 0 times respectively during 6:00 – 7:00 on the weekday across the observational period ($L_A^6 = 5$). His normalized presence at user location 3 during this time period can be calculated as in Equation (2.2).

$$\widetilde{P}_{A,6}^{3} = \frac{P_{A,6}^{3}}{\sum_{l=1}^{5} P_{A,6}^{l}} = \frac{5}{1+0+5+1+0} = \frac{5}{7} \tag{2.2}$$

2.3.4 Eigenlocations

It is useful to reveal and extract the common behavioral patterns and daily routines at user locations across from the large-scale noisy user location dataset. The hypothesis is that there exists common behaviors.

Principal Component Analysis (PCA) has been proved to be useful to extract underlying structures from large-scale behavioral datasets by [28, 29, 30]. Therefore, we apply PCA to test our hypothesis. Based on the assumption that user's presence patterns at user locations with similar functions are similar across the population; that is, if the coefficients on the eigenlocations for two locations for two users are alike, these locations are likely to have similar function for the users.

Each eigenvector, named as eigenlocation, represents a typical presence pattern

by explaining a portion of the presences variances. It describes the common presence patterns over weekday and weekend across the urban-wide population. The eigenlocations are ranked by the explained variances, which is the associated eigenvalues.

The set of user locations for all users can be represented by $\Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_U$ where U represents the number of users. Γ_U is a 48 dimenstion vector characterized by the 48 NHPs. The average presence pattern of the user location is $\Psi = \frac{1}{U}\sum_{u=1}^{U}\phi_u$. And $\phi_i = \Gamma_i - \Psi$ is the deviation of a user location from the mean presence patterns. The below matrix $A = [\phi_1, \phi_2, \phi_3, ..., \phi_M]$. The calculations are shown in equation (2.3) and equation (2.4).

$$C = \frac{1}{48}\sum_{h=1}^{48}\phi_h\phi_h^T = AA^T \tag{2.3}$$

$$V'CV = \Lambda \tag{2.4}$$

where Λ =diag$\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_{48}\}$ and $V = [v_1, v_2, v_3, ..., v_{48}]$ is an orthogonal matrix where the j_{th} column v_j is the eigenvector correspondence to λ_j. For detailed proofs and derivations, refer to [31, 32].

2.3.5 Clustering

Clustering is a widely-used tool to discover underlying structures and group similar objects. It helps to find patterns in a collection of unlabeled samples by organizing items that are similar in some way. Because we hypothesize that home and workplaces should have similar NHP patterns, we use clustering techniques to identify home and workplaces.

Clustering requires a similarity measure. The similarity in our case should measure the closeness between presence patterns, more specifically, weekly and weekend normalized presences. The closeness is calculated by the Euclidean distance, the root

of the sum of square differences, between the coefficients of the eigenvectors from normalized hourly presences, with each hour equally weighted. The observed presence patterns are assumed to be an indicator of the types of the location to the individuals. In other words, the user locations within the same cluster cluster are the locations that have similar meanings or functions to users, such as home and workplaces.

There are two types of clustering: hard clustering (K-means clustering) and soft clustering (Fuzzy C-means Clustering (FCM)). Both of them are implemented in this project. K-means clustering assigns each user location to each cluster. FCM, on the other hand, assigns each user location a probability in each cluster, e.g., home, workplace and elsewhere, and each user location is assigned to the cluster where the probability is the highest or above some thresholds that we set. Because FCM permits flexibility in analyzing uncertain cluster membership, it accomadates the trade-offs between confidence in one cluster and the percentage of user locations assigned to a cluster.

As a conclusion, the alogirthm works in the following way. The inputs are the presences and the analysis unit is user location, which is characterized by NHPs on 24 hours of weekdays and weekends. Eigen-decomposition (Principal Component Analysis) is performed on the user locations to extract the underlying presence structures at the user locations. The output from this step is the eigenlocations, each representing a common presence structure. The coefficients of the eigenlocations, calculated from the projections onto the eigenlocations, can reconstruct the presence patterns and rule out redundant and noisy patterns at user locations. K-means clustering and Fuzzy C-means clustering are used on the coefficients of eigenlocations to cluster and segment user locations into home, work and other user locations.

2.4 Evaluation and applications

In this section, we apply our method on two datasets to: 1) evaluate its performance, inference rate and flexibility; 2) reveal the feasibility of using the proposed method in extracting common behavioral structures and revealing interesting patterns; 3) prove its practicality and scalability in real-world settings in inferring home and workplaces. We first test and evaluate our method on MIT Reality Mining data, an experiment conducted on about 100 individuals for more than one year in 2004. We compare our method with one baseline model, as described later in this section, and show that our method has a higher accuracy. We then implement the method on the real-world data collected in a crowded city in China to prove its feasibility and scalability on a large-scale dataset.

2.4.1 Data

Small-scale experiment

To prove the accuracy of our method, we use small-scale experimental data with labeled ground truth, the MIT Reality Mining data [33]. The Reality Mining project was conducted from 2004 2005 at the MIT Media Laboratory. The Reality Mining study followed about 100 subjects (including students and faculty), 73 of which can be used. The subjects were tracked by mobile phones pre-installed with software that recorded data about call logs, cell tower IDs, and phone status (idling or charging). Individuals reported their home, workplace, elsewhere seperately.

Real-world data

The large-scale CDR data we used were collected for two months in a populated and fast-developing city in China. The data is provided by one of the three mobile carriers

in China. We use a sample of 100,000 mobile phone users and 217,753 user locations as a case study to test the method.

Pre-processing is needed in real-world data to deal with the oscillation phenomenon. It happens when a user is within the coverage of two or more cell towers caused by the fluctuation of signal strengths from these cells or load balancing purpose [34]. In this step, user locations are clustered using the algorithm developed by [35]. The cell towers are first ranked according to the total number of days that they are connected to. The cell towers are then clustered according to Hartigan' leader algorithm with a spatial threshold of 1 km. This algorithm starts by setting the first cell tower in the sorted list as the center of a cluster. The subsequent cell tower is checked to see if they fall in the radius of 1 km. If it does, the cell tower is grouped into the existing cluster. Otherwise, it becomes a new cluster centroid. The algorithm completes when every cell tower is assigned to a cluster.

2.4.2 Results analysis and comparisons

In this section, we use the MIT Reality Mining data to test the proposed method. We compare it with the most widely-used method, named as the "Most Frequent" method, on MIT Reality Mining data. As stated in the literature review, the "Most Frequent" method makes the simple assumption that the user location with the most presences during the home hours (00:00 - 08:00 and 19:00 - 24:00) and work hours (09:00 - 18:00) are homes and workplaces respectively [3, 18, 26].

We compare the methods on two metrics: *accuracy* and *inference rate*. Accuracy is calculated as the percentage of inferences that are correct. Inference rate is calculated as the percentage of users for whom we identify home and workplace. The comparison results are shown in Table 2.1. The "Most Frequent" method identifies one home and one workplace for every user. The accuracies are 53% and 62% respectively. K-means clustering infer 56% home and 82% workplaces with 90% and 75% accuracies. FCM

infer 58% home and 84% workplaces with 88% and 74% accuracies respectively.

There exists a trade-off between the accuracies and inference rate, The larger number of home/workplaces inferred, the more mistakes made. As far as we are concerned, CDR data is at such a large scale that accuracy should be prioritized more than inference rates. Though the inference rates are relatively low compare to the "Most Frequent" method, the accuracy improved considerably with our method.

FCM has the flexibility to compromise accuracies and inference rate for different application purposes. If we increase the inference rate by relabeling 'elsewhere' whose membership[1] is less than median memership of the cluster as either home or workplace, the method can identify 78% and 100% home and workplaces respectively and the accuracies are 91% and 74%. The accuracy of workplaces stays the same comparing to simple FCM while the home accuracy improves 81%.

To lower the inference rates but improve accuracy, we adjust the threshold for cluster membership in FCM, the accuracy improvement are 79% and 55% for home and workplaces respectively by tagging user locations with membership that are less than the first quartile of the cluster membership as elsewhere. As shown in Table 2.1, the improvement in workplace inference is larger than that of home. Base on the results, we see that it is not appropriate to relabel the "home" to elsewhere whose membership is less than the first quartile of the cluster membership. Along with the trivial improvement in accuracy rate, inference rate drops to a large degree.

To draw a conclusion, the better way to infer home is to re-label home if the second-largest membership of the ones labeled as elsewhere is "home". For workplace detection, people who use the method can trade off between higher inference rate or higher accuracy.

[1]Membership in certain cluster is calculated by the probability of user location belonging to the user location cluster.

Table 2.1: Methods comparison

		Home	**Workplace**	**Improvement**
The "Most Frequent" Method	Accuracy	0.53	0.62	NA
	Inference rate	100%	100%	
K-means Clustering	Accuracy	0.90	0.75	79% & 34%
	Inference rate	56%	82%	NA
FCM (Balanced)	Accuracy	0.88	0.74	74% & 32%
	Inference rate	58%	84%	NA
FCM (Prioritizing Inference Rate)	Accuracy	0.91	0.74	81% & 32%
	Inference rate	78%	100%	NA
FCM (Prioritizing Accuracy)	Accuracy	0.90	0.83	79% & 55%
	Inference rate	42%	63%	NA

2.4.3 Real-world application

In this part, we prove the applicability and scalability of the method in real-world CDR data collected a populous city in China. We first analyze the pattern of eigen-locations, proving that it is possible to use PCA to reveal the common presence patterns at user locations. Then, we use Davies-Bouldin Index to determine the optimal number of user location clusters. After that, we analyze and interpret the presence patterns of the mean of each cluster. In addition, uncertainties of each user location cluster from FCM are analyzed. Finally, we comment on the scalibility and efficiency of the method base on computation time.

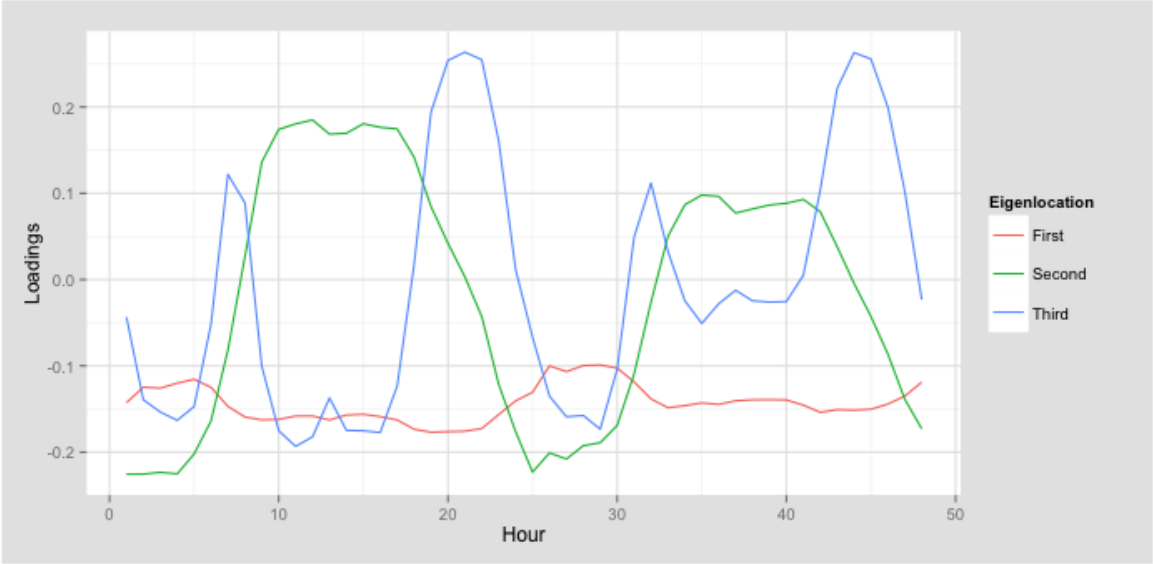

Figure 2-4: Top three eigenlocations

Eigenlocations

Eigenlocations, each represents a common pattern, characterize the pattern of user locations as shown in Figure 2-4. This confirms our hypothesis that there exist common presence patterns at user locations. The x-axis shows the 24 hours on weekday and 24 hours on weekend. The y-axis shows the loadings, which is the coefficients, of each hour on the eigenlocations. Large positive loadings indicate high presence frequency at user locations and small negative loadings (large-maganitude) indicates low frequency. Small magnitude, irrespective of the sign, indicates no clue for prediction. The red line, representing the first eigenlocation explaining 30.0% of variance, displays a pattern of infrequent-visiting location. The second eigenlocation (explaining 10.3% variance), as shown in the blue line, corresponds to a daytime-active location. Large positive cofficients on this eigenlocation are an indicator of workplace. The third eigenlocation, as illustrated in green line, represents home-like user location where individuals mostly stay during the evenings and early mornings.

We find out the first eight eigenlocations are intuitively interpretable. We then

choose eight to be the optimal number of eigenlocations since non-intuitive patterns are more likely to be noises. The first eight eigenlocations can explain 56% of the behavioral variances of user locations in total. A linear combination of the eigenlocations reconstructs the presence patterns at user locations. Therefore, original presence patterns at user locations are projected onto the first eight eigenlocations, each is reconstructed by the corresponding coefficients.

Optimal number of cluster

In the inference of home and workplace, our hypothesis is that four clusters should be used, including one workplace, one elsewhere and two home locations, one for normal-schedule workers and one for non-workers or short-distance commuters. To statistically test the validity and stability of the optimal number of cluster, we bootstrap Davies-Bouldin (DB) index for different number of clusters. DB index measures the scatter within the cluster and separation between clusters by the distances between each observation and its most similar ones [36, 37, 38, 39]. Therefore, the lower the DB index, the better the cluster configuration. The results of bootstrapping DB indexes are shown in Figure 2-5. Y-axis and x-axis show the DB index and the cluster size respectively. Each boxplot corresponds to the distribution of DB index for one cluster size. From the figure, we can see that there is an increase in DB index when the cluster number increases from 4 to 5. The decrease in DB indexes is trivial when cluster sizes are larger than 5. These observations indicate that the optimum number of cluster is four, which coincides with our hypothesis.

Clustering results

Figure 2-6 shows the clustering results from K-means clustering and FCM. X-axis represents 24 weekday hours and 24 weekend hours. Y-axis represents median NHP. Each line corresponds to the 48 median NHPs for one of the four clusters, includ-

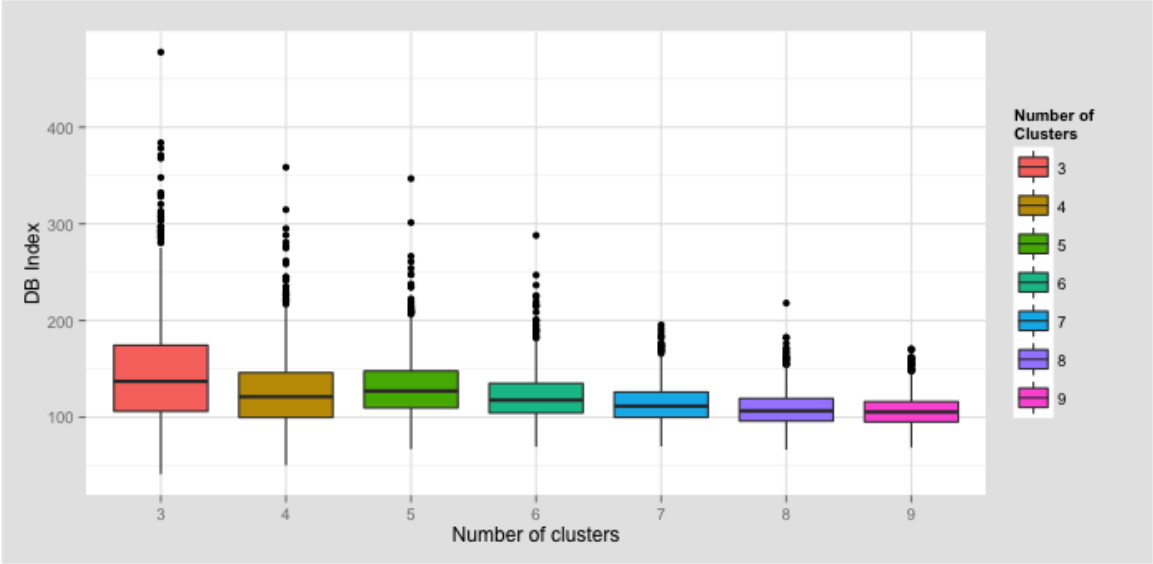

Figure 2-5: Bootstrapping DB index per number of clusters

ing two types of home locations (red and blue), one workplace cluster (green) and elsewhere cluster (purple).

Overall we can see that the results from the two clustering techniques are similar. The results are in line with the actual home and workplace patterns in our daily lives. The red line represents the home cluster for non-commuters or commuters who work near home. The percentages of presences at these locations are quite high throughout the week from 7:00 to 24:00. Note that for those who work near home, it is difficult to differentiate home and workplaces due to the low spatial resolution of CDR. The blue line represents the home for normal workers, who present at these user locations early in the morning and late at night. The presence frequencies are high before 7:00 possibly due to the automatic data fetching by some applications. The green line represents the work cluster. Travelers present at these locations more frequently during 9:00 - 20:00 on weekdays. The purple line represents an infrequent user location cluster, where individuals present at these locations with low frequency.

We also show the home and workplace distributions in the urban area in Figure 2-7. On the maps we show that the population and workplace density distributions of

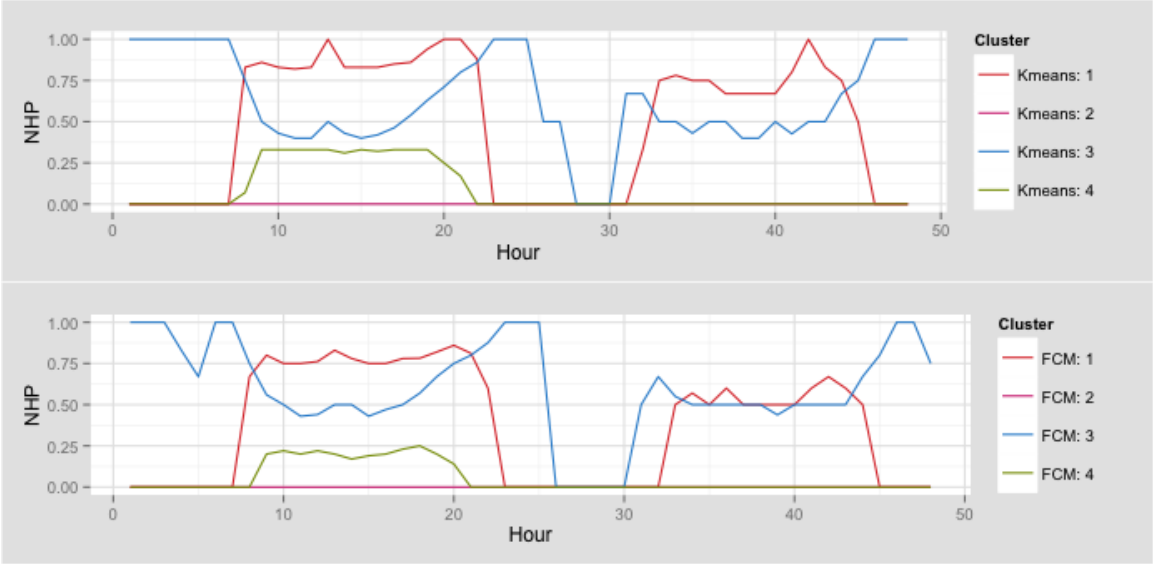

Figure 2-6: Clustering results from K-means clustering and Fuzzy C-means clustering

the Traffic Analysis Zones (TAZs). We scale up the 100,000 random sample to whole population and the values are calculated by taking the logarithmic transformation for better visualization. The color scales are positioned underneath the figures. From the map we can see that home and workplace locations are the densest in the center of the city, which is in line with the reality. The density of workplaces in eastern urban area is higher than that of the home density where there is a high-tech district with many new employment opportunities created.

We acknowledge that there are some limitations caused by this assumption. The group of mobile phone users with flexible work locations, such as shippers and drivers, or workers with abnormal work schedules, such as night shifters, will not be detected or even misidentified. Another limitation is that the inference result will be affected by phone usage patterns. For example, if individuals use landlines instead of mobile phones at home, it is hard to idenfity a home or workplace for these individuals due to the unobservable presence patterns solely based on CDR data.

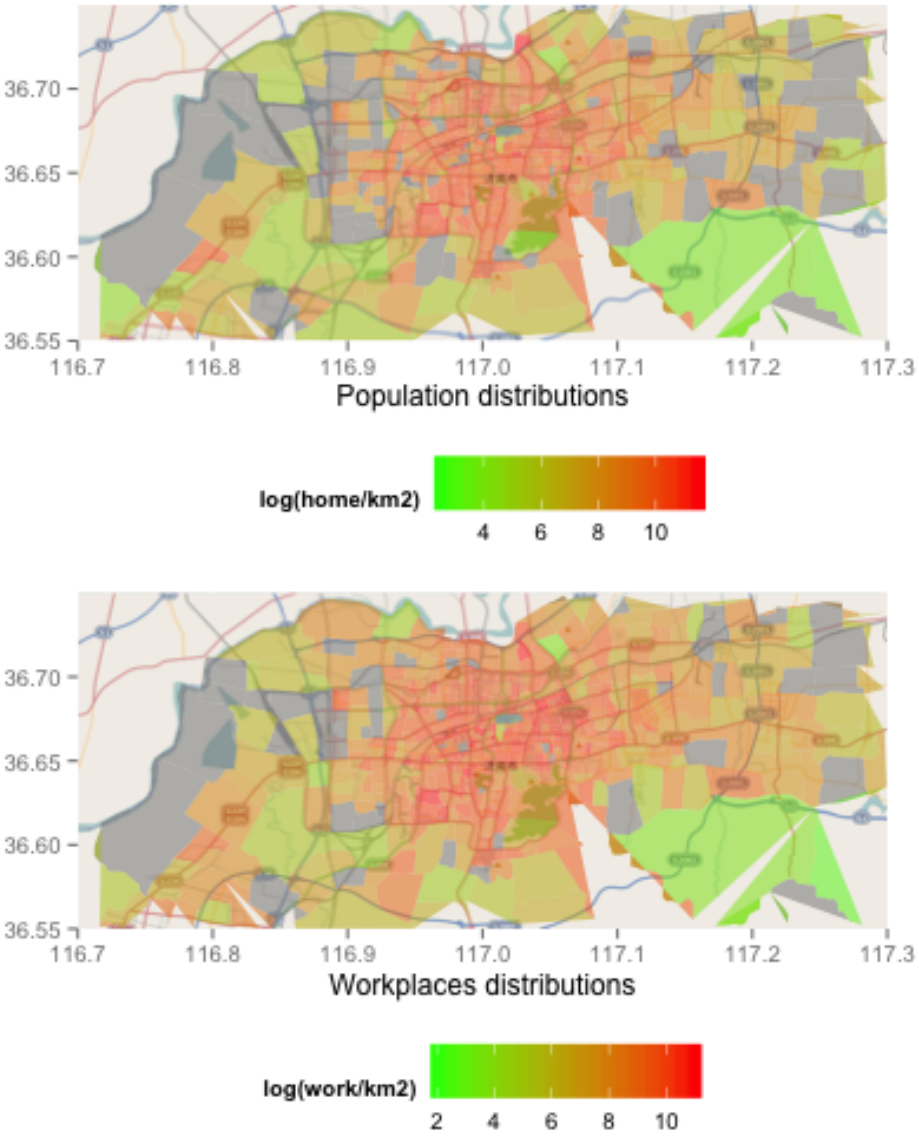

Figure 2-7: Home and workplace distributions

Uncertainty in behavioral inference

Confidence of inference results is represented by membership from FCM. The larger the membership, the more confident the results are. The membership distributions for each cluster are shown in Figure 2-8. The x axis are the membership and y axis are count of user locations in each membership range. Ligher color indicates more user locations. The median memberships are 0.56, 0.50 and 0.94 respectively for home, workplace and elsewhere. We can see that elsewhere has the highest confidence due to the small number of observable presences at these locations. The confidence for labeling workplaces is the lowest since people are more active during the daytime, making the inference more difficult.

The largest advantage of applying FCM in this research is the flexibility in trading-off between accuracies and inference rate using the membership. If we increase the accepted confidence level, less home/workplaces will be inferred. This can be done by setting an accepted membership threshold. Only home and workplaces with higher-than-threshold memberships are labeled as home/workplace. On the other hand, if we want to improve inference rate, we can lower the accepted threshold, which will help us to infer home and workplaces for more individuals.

Computationl time

Computational complexity is an important consideration in practical application. The computational complexity of K-means clustering is $O(ncdi)$ and that of FCM is $O(ncdi^2)$. n is the number of observations, which is the total number of user locations for all sampled mobile phone users. d is the number of features, which is 48. c is the prespecified number of clusters. i is the number of iterations until convergence. We test the method on 1,000,000 mobile phone users with 2,177,530 user locations, each has two-month presences. The running time for PCA is 15 seconds. The running time for K-means clustering is approximately 6.2 seconds and that for FCM clustering is

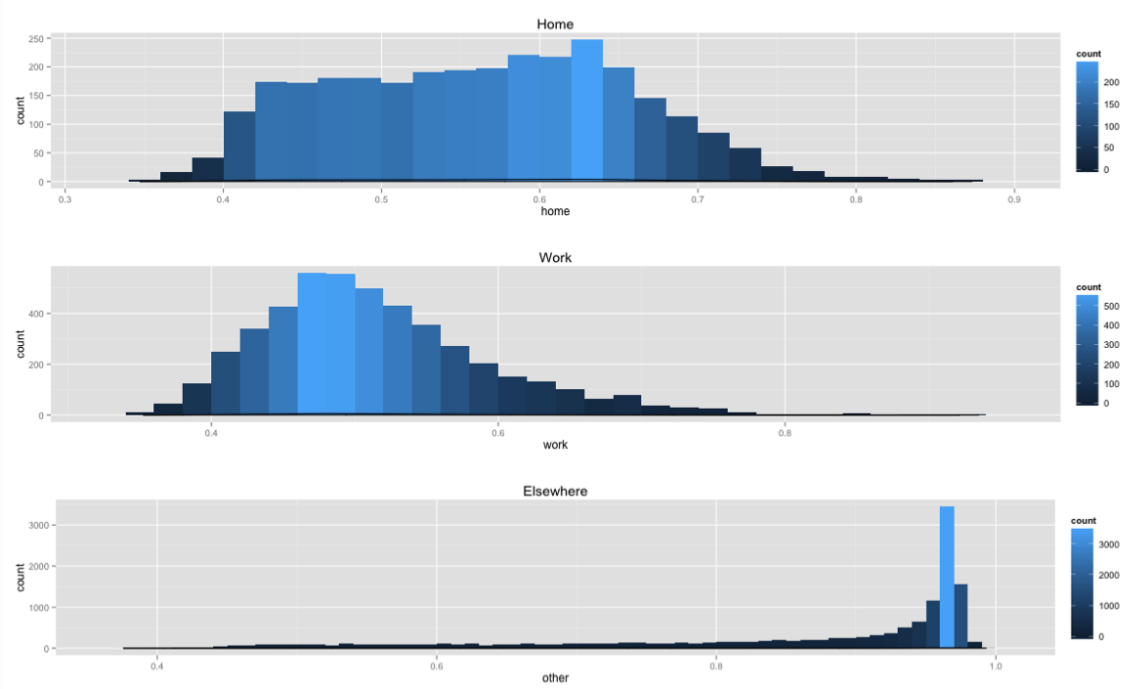

Figure 2-8: Confidence of the inference results

120.2 seconds. K-means clustering has shorter computational time than FCM, both of which can are endurable practical use.

2.5 Conclusions

The wide-penetrated mobile phone data provides longitudinal records for tracking human mobility and urban dynamics. Home and workplace, origins and destinations of commuting and other trips, are the most important user locations and are the foundations of many transportation researhes. However, the current literature, making simple and intuitive but biased assumptions are problematic in inferring home and workplace due to the large discrepancy between observed presence patterns from mobile phone data and actual presence patterns at user locations.

In this paper, we develop a method to extract behavioral patterns from the noisy CDR data and infer home/workplace with higher accuracy compare with the method

used in the literature. The paper answers the following questions: (1) how to characterize individuals' presence patterns at user locations, (2) whether there are common presence structures across the urban-wide population, (3) how to identify home and workplace from user locations. The method can be used for various applications. It can be used to monitor temporal and spatial distributions of population at both individual and aggregation level. It can also be applied in understanding some characteristics of commuting trips. Moreover, it can also be used to calculate the expansion factor to connect CDR-based OD matrices and actual OD matrices. Last but not the least, it can be applied to other large-scale behavioral datasets with location traces, such as GPS traces and WiFi beacons.

The proposed feature, Normalized Hourly Presences, is useful in characterizing individual presence patterns at user locations by revealing the frequencies and variations of presence patterns on weekdays and weekends from the noisy mpbile phone data. Principal Component Analysis can be used to extract the common behavioral structures at user locations across the population. Clustering techniques are useful in discovering the intrinsic groups of user locations and segmenting them into home, workplaces and elsewhere with higher accuracy. We test the method on the CDR data collected by MIT Reality Mining data and the real-world CDR data in a populous and fast-growing city in China. We show the feasibility, higher accuracy and scalability. We also demonstrate that with the inference rates of 78% and 100%, the method can improve home and workplace location inference accuracies by 81% and 32% respectively over other methods proposed in the literature. With FCM, we can flexibly trade off the inference rate and accuracies based on the confidence level of the inference results.

For future research, the proposed method can be extended to infer not only home and workplaces, but also other user locations, such as late night locations or weekend locations. The method is also useful to combine with other geographical data sources,

such as land use data, Point of Interest data, to infer other trip purposes and activities, or travel surveys with socio-demographic information. Additionally, it should be very useful for estimating commuting characteristics, such as commuting distances, departure and arrival times, etc. With more data available, we could also test the method on other big data sources in inferring home and workplaces, such as online-social networks check-ins (flickr, twitter), bank transactions, and etc.

Chapter 3

Recurrent Neural Network in Context-free Next-location Prediction

Location prediction is a critical building block in location-based services, transportation management and resource allocation. Various researches emerged in predicting next visited locations with acceptable accuracies on high spatial and temporal resolution data. However, the research on location prediction using sparse mobile phone data is limited with low accuracies. This project focuses on the issue of next-location prediction based on historical movements from sparsely and passively collected Call Detail Records (CDR). We model cell tower traces of each individual as a sentence and each cell tower as a word. Recurrent Neural Network has been extensively explored in statistical language models. Therefore, this research proposes to use RNN in next-location predictions, with which the real-valued vector representations are learned to characterize cell towers instead of the traditional numerical indices. With the implementation of RNN on the large-scale CDR collected in Andorra, this chapter proves the applicability of the method with accuracies of 67% and 78% at cell tower and merged cell tower levels, which has more than 30% improvements comparing with two baseline models.

3.1 Introduction

With the explosions of location data from cell towers, Wi-Fi beacons and various mobile sensors, individual travel behaviors are more predictable and tractable than ever before. Accurate next location predictions given previous footprints from these data sources is a significant building block benefiting many areas, including mobile advertising, public transit planning and urban infrastructure management [5, 6, 7]. For example, from the business side, predicting where people will go enables location-based service providers to distribute targeted coupons or advertisements. In addition, from the systematic perspective, predicting travel demand enables public transit agencies and events organizers to take proactive actions responding to the dynamic demand [8].

Predicting human mobility has been an increasingly popular topic in pervasive computing based on GPS, bluetooth, check-in histories, etc. Different data sources vary at spatial, temporal scales and the availability of contextual information [9]. Most research build markov models and predict the longitude and latitude as continuous variables based on previously travel trajectories [10, 11, 12]. Mathew [13] predicts next-location by Hidden Markov Model with contextual information, such as activities and purposes. Domenico [9] and Alhasoun [5] uses mobility correlations, either measured by social interactions or mutual information, to improve forecasting accuracy. Though extensive researches have acceptable accuracy in predicting next locations, the performances are poor with sparse Call Detail Records [5].

Call Detail Records have been widely used in transportation academia and industrial applications. Advances in location predictions is a foundation in pushing these researches. This research predicts next location for each individual based only on historical location sequences from Call Detail Records, a universal data source with largest coverage and least energy consumption. The accurate prediction method for

CDR is also a critical foundation for other research.

Deep learning has attracted extensive attentions with achieved excellent perfomances in speech and hand-writing recognition, image processing, time serious preiction etc. Recurrent Neural Network (RNN) is the state-of-art deep learning technique in temporal prediction [40]. In a nutshell, RNN is a network of neurons with feedback connections, manifested in learning vector representations of words that can handle arbitrarily long contexts [41]. This enables RNN to be used in temporal processing and sequences learning [42]. We hypothesize that RNN could be used in next-location prediction from mobile phone data due to its ability in inferring the meaning of the input units and handling sequential inputs and variable input. One characteristic of location prediction is the rich contextual information, such as the surrounding point of interests. It is difficult to incorporate this information with the added complexity even if these contextual information is available.

Due to the above reasons, this project models individual footprints as sentences, with each cell tower mapping to a word. By building a mobility model across all individuals with RNN, the "meanings" of cell towers can be learned, representing them by real-valued vector representations. The method is validated with the large-scale mobile phone data collected in Andorra. We test the mehtod in two experiment settings, cell tower and merged cell tower level. We compare RNN with two baseline model, including a naive and markov model.

To summarize, the contributions of this project are as follows:

- We propose a new perspective and prediction mechanism for location predictions by automatically learning the meaning of the cell towers for CDR. Recurrent neural network is explored in mobility prediction, mapping mobility models to language models.

- We implement the method in Andorra as a case study on sparse CDR at two spatial resolutions. The method clearly outperforms markov model and a base-

line model with an improvement of more than 30% accuracies. The accuracies are 67% and 78% at cell tower level and merged cell tower level respectively.

- The context of the cell towers can be inferred using the word-to-vector technique as used in language model. Instead of representing cell towers as indices, we use real-valued vector representations. The interpretation of cell towers is proved to be useful in location predictions.

The remainder of this chapter is organized as follows. In section 3.2, we introduce the RNN, starting from problem formulation, followed by mapping mobility models to language models and the architecture of our RNN model. In section 3.3, we implement RNN as a case study, tuning parameters and analyze the prediction results. We conclude this chapter with summarys and future work in Section 3.4.

3.2 Methodology

In this section, we describe the problem formulation and the structure of RNN. We first form the next-location classification problem as a language model. We claim that they have similar hierarchical structure. The prediction on next location is the same as next-word precition in language model speech recognition. The model utilizes the real-valued vector representation of cell towers to improve the prediction of human movements captured by CDR.

In Table 3.1, we show the mapping between mobility model and language model, as well as the applications. The cell towers of CDR correspond to words in language model, for which we try to learn the meaning of the cell tower in location traces. Short sequences of cell towers can be mapped to phrases, which may indicate the activities or trip purposes. The entire location trace is a sentence where the prediction of next location is the same as the prediction of next word. This mapping enables us to take advantage of RNN's excellent performance in statistical language model and to apply

Table 3.1: Mapping from language model to location traces

Language Model	Mobility	Applications
Word	Cell tower	Infer semantic meaning to cell towers
Phrase	Short sequences of cell towers	Infer activities
Sentence	Location traces	Location predictions

it in mobility predictions. Representations of the words, which is referred to as word embeddings, are basic units of language and also input of more complicated sentence semantic or syntactic understanding [43]. Each dimension of these representations captures latent information about a combination of syntactic and semantic word properties [44]. Learning the embedding of cell towers is the same as inferring the word semantics in language models. With this, we argue that the meaning of the unit of mobility (cell tower) in travel behaviors is critical, which is not taken into account in the models developed in the next location prediction literature.

3.2.1 Problem formulation

The problem of forecasting next locations can be formed as a classification problem of predicting the next location as the one that has the highest probability given the historical location traces, as shown in Figure 3-1. Given sequences of $(l_1, l_2, ..., l_{t^u})$ where t^u is the size of location sequences for user u, we predict l_{t^u+1} as the next location, which has the highest probability among all possible locations.

The input of the model is the location sequences, represented by cell tower index and the output of the model is the cell tower with the highest probability. One main advantage with modeling mobility model as language model using RNN is its ability in learning the "meaning" of the cell towers by representing them with embeddings in a dense vector space. This enables us to learn the similarities among cell towers and to enrich the context of locations. This helps us to deal with the sparseness of location

traces and make more accurate predictions even for travelers with few observations.

3.2.2 Recurrent Neural Network

Recurrent Neural Network is an adaptation of the traditional feed forward neural network, which can process variable-length sequences of inputs [45, 46]. It has been successfully used in language models, learning word embeddings, financial time series predictions, and etc.

RNN is advantageous in predicting next locations in the following senses:

- Location sequences.

  Travelers visit locations in a sequential way and RNN reads in data sequentially. In this way, the interdependencies and contexts among locations can be retained.

- Variable number of visited locations.

  Some individuals are more active than others and travelers' phone usage habits are different. The heterogeneity in the size of location traces and frequency of mobile phone usage makes traditional machine learning techniques inapplicable. Besides, traditional markov models can take only a fixed number of previous locations in predicting next locations. However, the ability to handle variable input lengths makes RNN appropriate in this situation ([42]).

- Interpretation of cell towers.

  CDR data is semantically poor, with each cell tower represented by a numerical index. However, RNN can learn the real-valued representations of cell towers based on location sequences of all individuals, relating "similar" cell towers close to each other in a vector space.

Our recurrent neural network model is built with Keras [47], which is a deep neural network python built on Theano [48]. Put it simply, the input of are sequences of cell towers, represented in indices. We embedded cell towers and passed them through

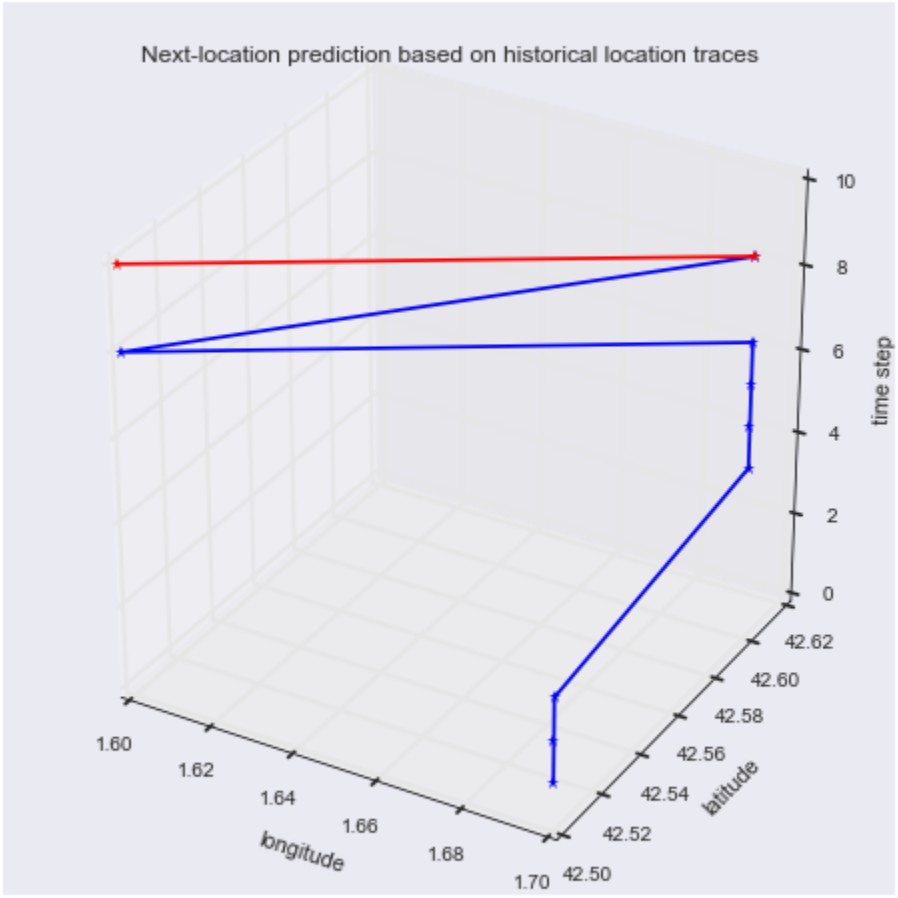

Figure 3-1: Next location prediction based on historical traces

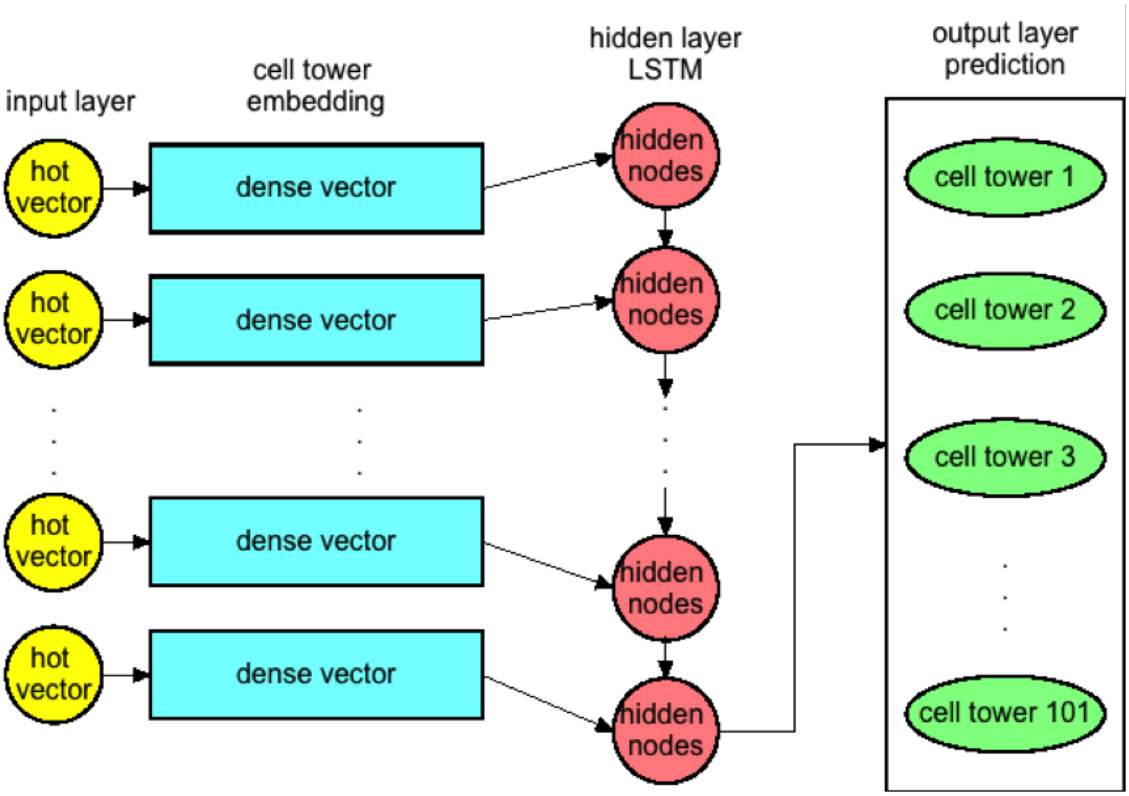

Figure 3-2: Architecture of RNN

LSTM layer. Siames neural network architecture is used to tie the embedding layer and LSTM layer weights together to ensure consistent embeddings [49]. After that, we concatenated the output of LSTM and fed them into a third LSTM before performing categorical classification in the output layer. The details of the layers are discussed as follows and are shown in Figure 3-2.

- The first layer is the input layer, using sequences of discrete index of each cell tower. As discussed in section (3.2.1), each training location sequence, represented as $L_u = (l_{u^1}, l_{u^2}, ..., l_{u^i}..., l_{u^t})$, is used as input. u is the index for user and u^t is the size of the historical locations for user u. l_{u^i} is the i^{th} location for user u, represented by the cell tower index. The sizes of location sequences vary across individuals. The location index space is the number of distinct cell towers.

- The second layer is the embedding layer. The embedding layer aims to transform discrete index of cell towers into real-valued dense vectors. "Similar" cell towers will be grouped close to each other in vector space.

- The third layer is the Long Short Term Memory (LSTM) layer. One appeal of applying RNN in next-location prediction is its ability to connect previous locations to the present locations, which is called long-term dependency. However, locations from long time ago may not be useful in inferring the next location.

  Conventional RNN without LSTM suffers from vanishing gradient, especially for long-sequence input. The parameters are estimated using gradient back propagation. Since the gradient ends up being multiplied for a large number of times by the weight matrix which is associated with the connections between neurons of the recurrent hidden layer. The magnitude of weights in the transition matrix has a strong impact on learning process. If the weights matrix is very small, it can lead to vanishing gradients where the gradient gets so small that the learning become small of stops. The learning of long-term dependences are therefore more difficult. On the other hand, if the weight matrix is larger, it will lead to a situation where the gradient is so large that the learning becomes diverge, which is called exploding gradients [50, 51, 52]. The calculations are shown in Equation (3.1) to Equation (3.6). The value i_j for the input gate is calculated with Equation .

$$i_t = \sigma(W_i l_j + U_i h_{j-1} + b_i) \tag{3.1}$$

  The candidate value for the states of the memory cells at j is calculated then.

$$\widetilde{C_j} = tanh(W_c l_j + U_c h_{j-1} + b_c) \tag{3.2}$$

We then compute the value for f_j, the activation of the memory cells' forget gates at j.

$$f_j = \sigma(W_f l_j + U_f h_{j-1} + b_f) \tag{3.3}$$

The memory cell's new state, C_j, at j can be caculated given the input gate activation, the forget gate activation the candidate state value $\widetilde{C_j}$

$$C_j = i_j \times \widetilde{C_j} + f_j \times C_{j-1} \tag{3.4}$$

We then compute the value of output gates and their output with the previous results.

$$o_h = \sigma(W_o l_j + U_o h_{j-1} + V_o C_j + b_o) \tag{3.5}$$

$$h_j = o_j \times tanh(C_j) \tag{3.6}$$

l_j is the j^{th} location to the memory cell layer. where W_i, W_f, W_c, U_i, U_f, U_c are the weight matrices. b_i, b_f, b_c are the bias vectors.

- The output layer is a categorical classification layer with the probability of all the locations. The output vector is L_{t^u+1}, the probability of next-location following the training location sequence with a dimension of m. m is the set of all possible locations. We use cross-extropy as the loss function since there are multiple potential next locations. L_{t^u+1} is a predictive distribution $P(L_{t^u+1}|L_u) = softmax(L_{i,j_{i+1}})$. More concretely, this can be modeled in Equation (3.7).

$$L_{j+1} = \sigma(W_o l_j + U_o h_{j-1} + b_o) \tag{3.7}$$

where W_o and U_o are the weight matrix of the output layer. b_0 are the bias vector of the output layer. L_{j+1} is the prediction of the $(j+1)^{th}$ location given j previous

locations.

3.3 Experiments

In this section, we conduct empirical experiments to demonstrate the effectiveness and applicability of RNN in large-scale next-location prediction. We first introduce the datasets and two baseline methods: the "Most Frequent" method and markov model. After showing the calculation of our evaluation metric, we compare RNN with the two baseline methods. We then show parameter tuning and empirical analysis of the results.

3.3.1 Data preparation

The CDR data was collected from Andorra for more around two years. As a case study, we evaluate different methods using CDR data collected over three weeks during May, 2015 in Andorra, a small country with 80% of GDP from tourism. We apply the method specifically on tourists, which can be filtered on the country code from CDR. We do not exclude travelers with too few observations as long as they have travel to more than one cell towers, which proves the generalarity of the method. There are 101 distinct cell towers in Andorra, as shown in Figure 4-3. We have 143420 travelers which are randomly split into 114736 training and 28684 test samples.

We use two settings with different spatial resolutions, at cell towers level and merged cell tower level. The second setting aims to deal with the oscillation phenomenon of CDR a connection is likely to happen within the coverage of two or more cell towers [34]. We use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to cluster nearby cell towers [53]. The maximum distance between two cell towers in the same cluster is set to be 1500 meters and the minimum number of cell towers in one cluster is 1. With this algorithm, we merged cell towers into

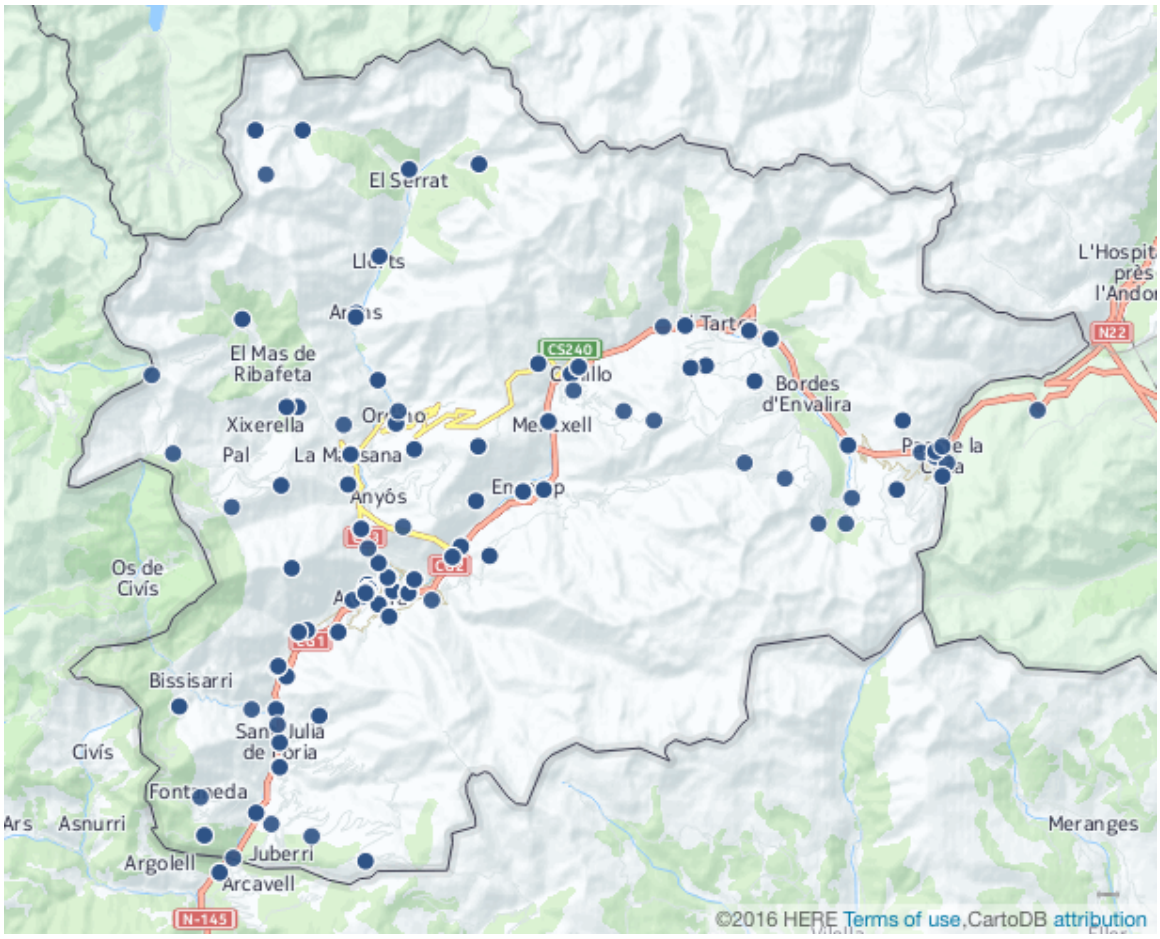

Figure 3-3: Tower distributions in Andorra

26 spatial clusters. Cell towers within the same cluster are represented in the same color, as shown in Figure (3-4).

Each Call Detail Record contains the encrypted user ID, starting and end time of the phone call, connected cell tower ID, nationality and the mode of the phone, as shown in Figure 3-5. The field "ID_CDOPERADDRORIGEN" records the country code, enabling us the recognize the registration nation of the user and filter out natives in this case study.

The sparseness of CDR data, an innate characteristic of it, makes the preference inference challenging. From Figure 3-6, we can see that 55% of tourists have one record and 22% of users have two records. Moreover, 62% of travlers have visited only one cities and 28% of travelers traveled to two cities.

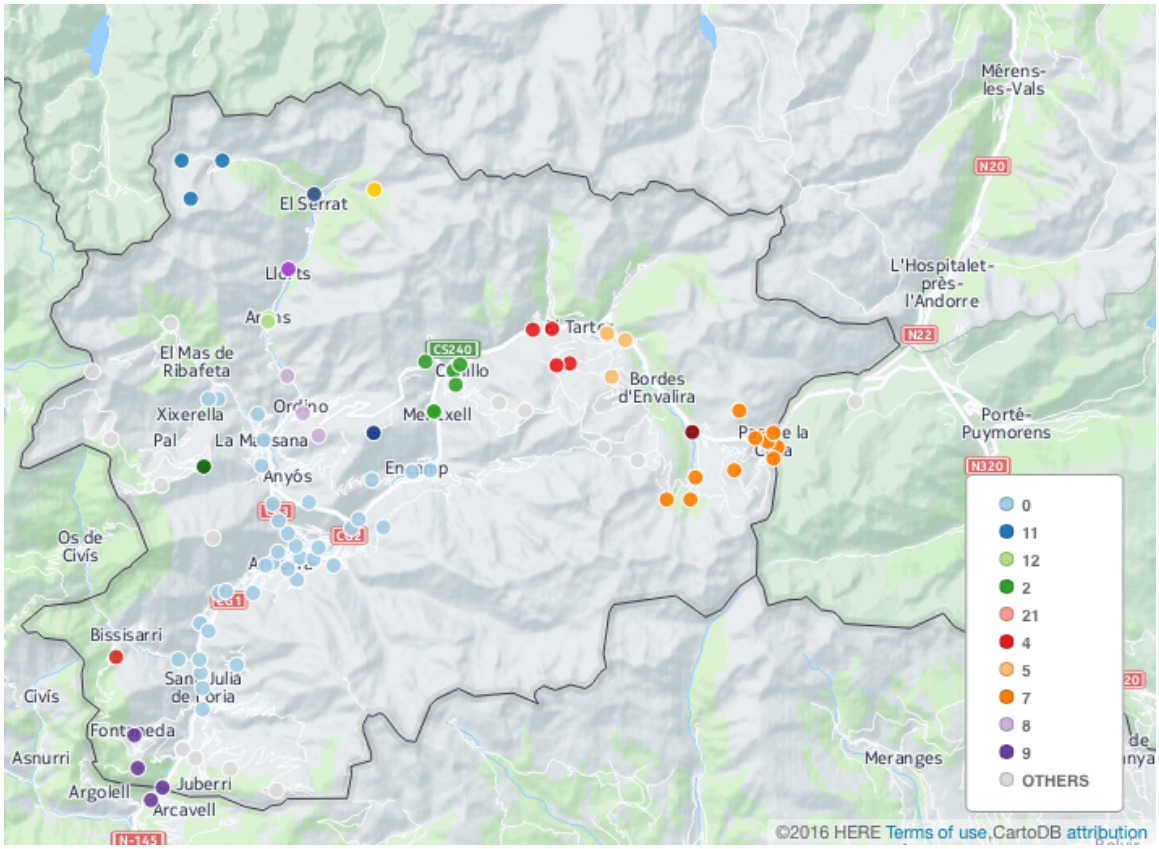

Figure 3-4: Merged cell towers
Each color represents each cluster. The cluster with only one cell towers are represented by in grey dots. Cluster '0', represented in light blue, is a large cluster. Most of the cell towers in the the capital area of Andorra are clustered together. Many cell towers in this cluster are very close to each other.

	id character(32)	lac integer	cellid integer	attribution integer	location integer	datetime timestamp without time zone	eventid integer
1	546E2D8D9D8693FDDA0D4A8CEC57A53E	5913	25185	531	531	2013-10-13 10:41:39	6
2	B398929835 71EF5CE723C93087BF09C6	5912	28334	531	531	2013-10-13 11:03:52	6
3	ABCE39A8BF1400AE4FDBF5589ACA3C9D	5915	49806	531	531	2013-10-13 11:04:51	6
4	574B269ABB965629E535E97960FAC0BE	54016	57728	531	531	2013-10-13 11:04:49	6
5	6CC7F4FFDA98F5FE177E6E3E373933EC	5913	25376	531	531	2013-10-13 11:04:31	6

Figure 3-5: Snapshot of Call Detail Records

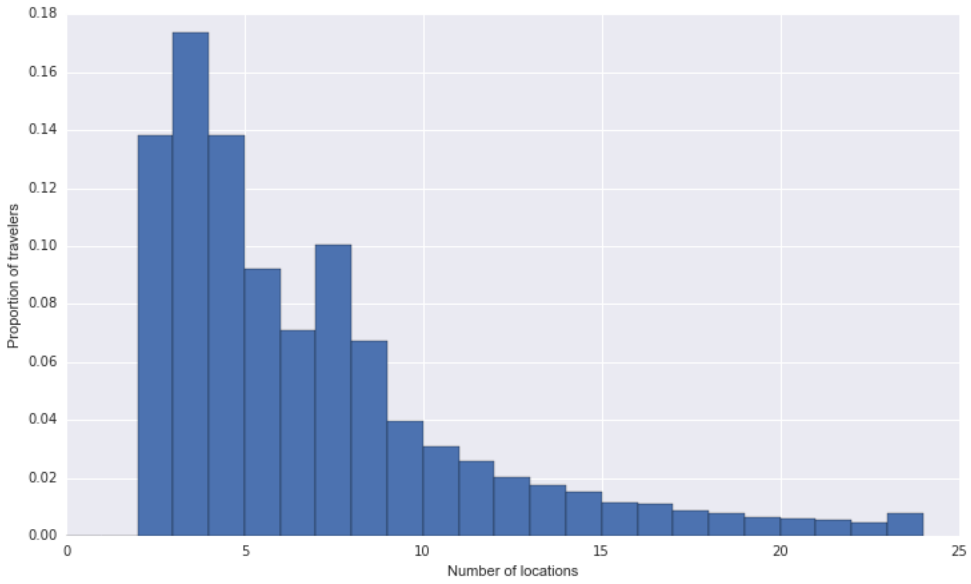

Figure 3-6: Histogram of number of cell towers users traveled to

Entropy is a powerful metric in quantifying the degree of predictability for a time series. In Figure 3-7, we show the histogram of entropy of our datasets, where entropy is calculated as in Equation (3.8) [34, 54].

$$S = -\sum_{j=1}^{N^u} j^u \log_2 p_j^u \tag{3.8}$$

where p_j^u is the probability of user u visiting location j.

3.3.2 Baseline models

We introduce two baseline models. The first one is "Most frequent" and the second on is the Markov Chain Model. "Most frequent" model is referred as a naive model, which predicts the next location using the most freuqenct location. The underlying assumption is the regularity of mobility behavior [55]. Markov model is built based on the contextual co-occurences between sequences of locations [7, 12, 56].

As briefly stated above, the "Most frequent" method predicts the next location as

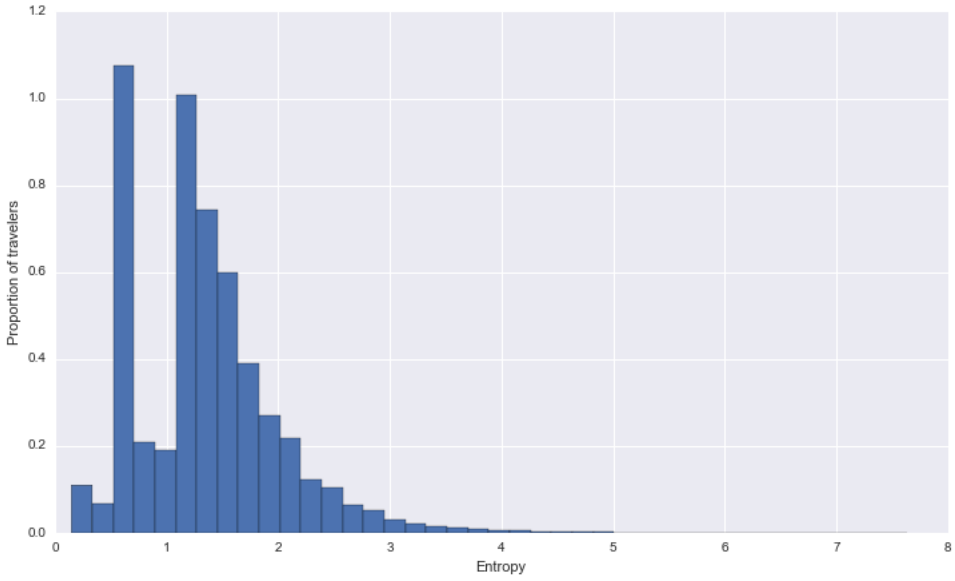

Figure 3-7: Entropy distribution

the location that is visited most frequently by the indiviudal, as shown in Equation (3.9).

$$P(l_{u^t+1}) = max(p_k|L_u) \tag{3.9}$$

where p_k is defined as the frequency at which the person visit the m^{th} cell tower among all cell towers.

The markov model couple location traces of the population and learn the transition probabilities to cope with the data sparsity issue. According to [5, 7], we can model the markov chain of order n using Equation (3.10).

$$p(L) = \prod_{u=1}^{U} p(l_u) = \prod_{u=1}^{U} \prod_{j=2}^{t^u} p(l_j^u | l_{j-1}^u, ..., l_{j-n+1}^u) \tag{3.10}$$

The transition probability matrix Tr can be calculated as in Equation (3.11)

$$Tr = p(l_t | l_{t-1}, ..., l_{t-n+1}) \tag{3.11}$$

Given the previous locations, the prediction is then determined by the transition matrix, Tr, choosing the destination l_{u^t+1} that maximizes the joint probability, as shown in Equation (3.12).

$$l_{u^t+1} = \arg\max_{l_{u^t+1}} p(l_{u^t+1}) = \arg\max_{l_{u^t+1}} \prod_{i=2}^{u^t} p(l_{u^t+1}|l_{u^t}, ..., l_{u^t-n+1}) \tag{3.12}$$

3.3.3 Results

In order to evaluate RNN, we compare the accuracies of RNN with the existing two baseline models as discussed in Section 3.3.2 on the same input under the two settings. The results are shown in Table 3.2.

The performance of each model was evaluated by prediction accuracy γ, which is the proportion of accurate predictions from all predictions [7] as shown in Equation (3.13).

$$\gamma = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \tag{3.13}$$

The improvements are calculated by the absolute improve normalized by the room for improvement as shown in Equation (3.14).

$$\text{improvement} = \frac{\Delta\gamma}{1 - \gamma_{base}} \tag{3.14}$$

The naive model has 50% and 63% accuracies in next location prediciton at cell tower and cell tower cluster level. Markov model across the population has 54% accuracy at the cell tower level, with 8% improvement. However, at the merged cell tower level, the population-wide markov model has lower accuracy than the naive model. Comparatively, RNN improves the accuracy of location prediction significantly, with an accuracy of 67% and 78% on the two settings. It improves 34% and 41% in accuracies compare with the two settings, indicating that RNN significantly improves the accuracy of the prediction of next location.

Table 3.2: Results analysis

	Accuracy	Computation time	Improvement
Cell tower level prediction			
Naive model (individual)	50%	1s	NA
Markov model (population-wde)	54%	3s	8%
RNN	67%	1287s	34%
Merged cell tower level prediction			
Naive model (individual)	63%	1s	NA
Markov model (population-wide)	57%	3s	-16%
RNN	78%	1244s	41%

The computational times are all calculated by a single-core Central Processing Unit (CPU) machine. We expect a great improvement in efficiency if processed on Graphics Processing Unit (GPU).

To understand the performance of RNN in each epoch (full pass through the training set [57]), we show the accuracy on validation set in Figure 3-8 using the second setting as an example. We can see that the improvement in accuracy is slight after the first epoch and reduce sharply after the eighth epoch. This shows the number of epochs required for training the model.

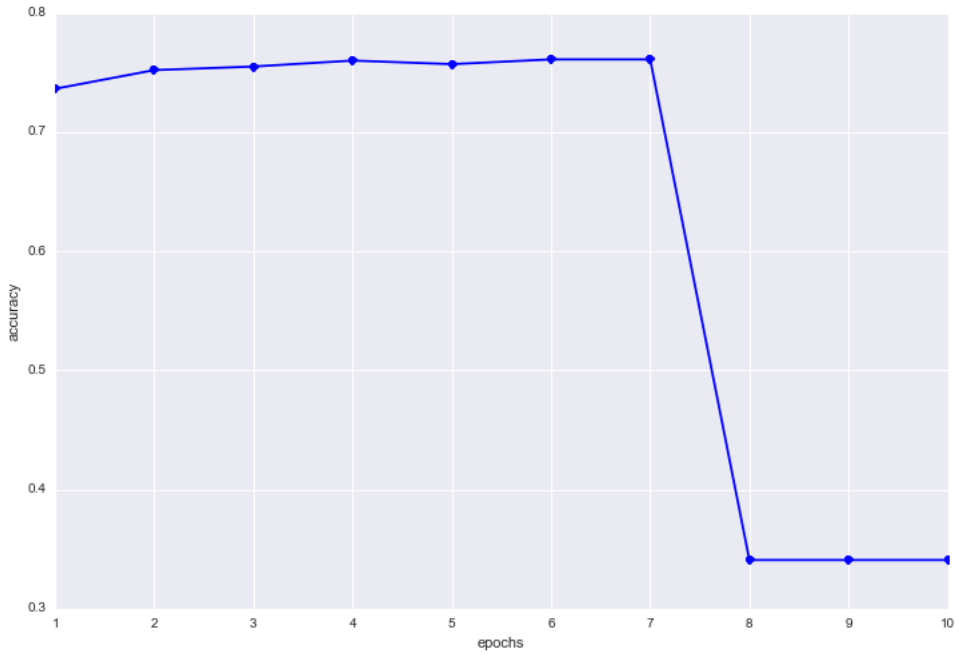

Figure 3-8: Loss per epoch - Merged cell tower level

3.3.4 Parameter tuning

In this section, we show the relationship between accuracies and various parameters critical for RNN. The parameters include: initialization of embedding layer, dimension of embedding layer, drop out rates and batch size. Besides, we also analyze the relationship between training sample size, computational time and accuracy. We use two activation functions, tanh and sigmoid, to capture the non-linear relationships in the model.

Dimension of embedding layer In this project, we claim that the cell towers correspond to words in sentences. We can therefore take advantage of RNN's ability in learning the "meanings" of the cell tower by representing each numeric index with a real-valued vector. In this chapter, we refer this real-valued vector as *cell tower embedding*. The dimension of the embedding layer is the size of the real-valued

vectors.

In Figure 3-9, we show that the accuracies vs. embedding layer size. In general, *tanh* activation function has higher accuracies than *sigmoid* activation function. Within the tested dimension of embedding layer, we show that, for *tanh*, the accuracies dramatically decreases from four to eight and it gradually increases afterwards.

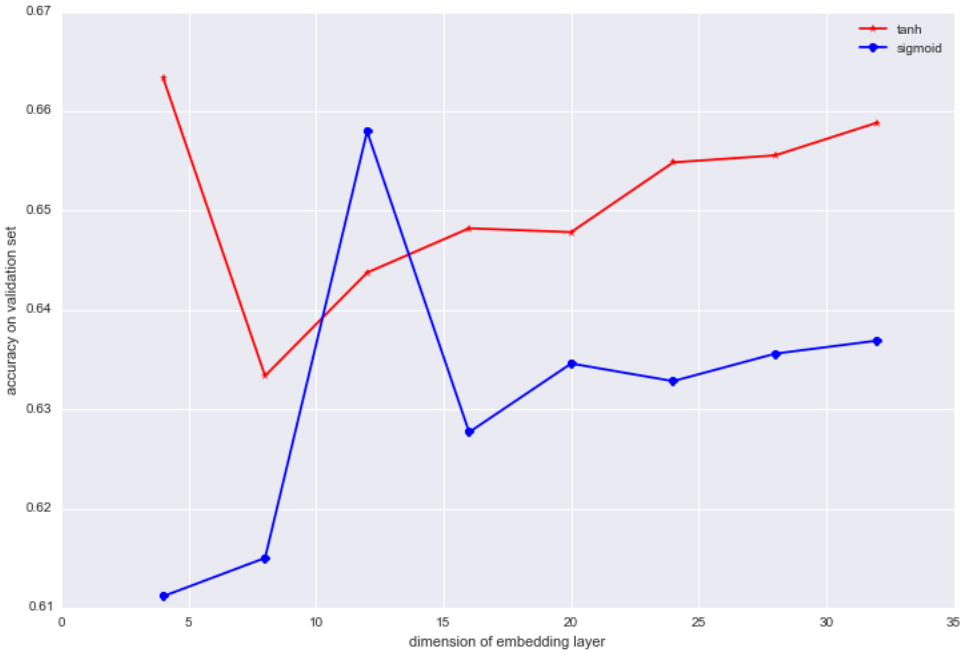

Figure 3-9: Parameter tuning: dimension of the cell tower

Drop out rates When training a network with a large number of parameters, an effective regularization mechanism is essential to combat overfitting. Dropout rate serves to regularize parameters and effectively reduce overfitting [58]. While training, dropout is implemented by only keeping a neuron active with some probability p [59]. As shown in Figure 3-10, the accuracy deceases when we increase dropout rates from 0.05 to 0.95,

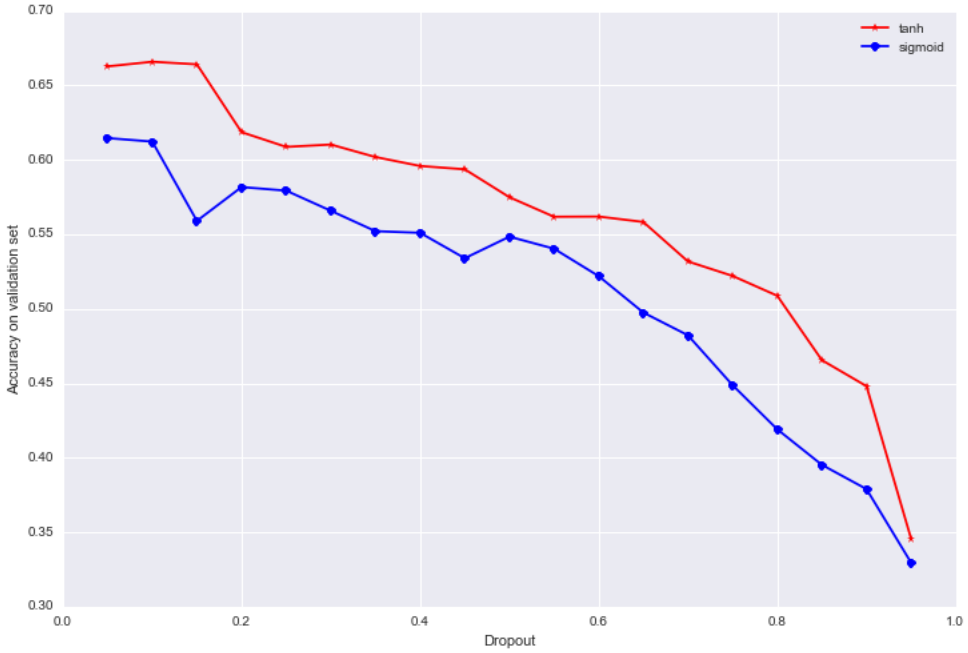

Figure 3-10: Parameter tuning: dropout rates

Mini-batch The size of mini-batch is important for decreasing the time to converge. To reach the desired level of accuracy, we need to balance the efficiency increase due to larger batch size and iterations increase to reach a desired level of accuracy [60]. As shown in Figure 3-11, with *tanh* as the activation function, the accuracy reduce gradually from 0.78 to 0.64 with the increase in batch size from 16 to 64. By further increasing batch size to 80, the accuracy reduces sharply to 0.35. By using sigmoid to activate, the slope for accuracy reduces slowly from 0.74 to 0.51. With Figure 3-12 we can see that the computational time decreases sharply from nearly 2500s to less than 1000s when the batch size increase from 16 to 32. However, with batch size higher than 32, the computational time does not vary much.

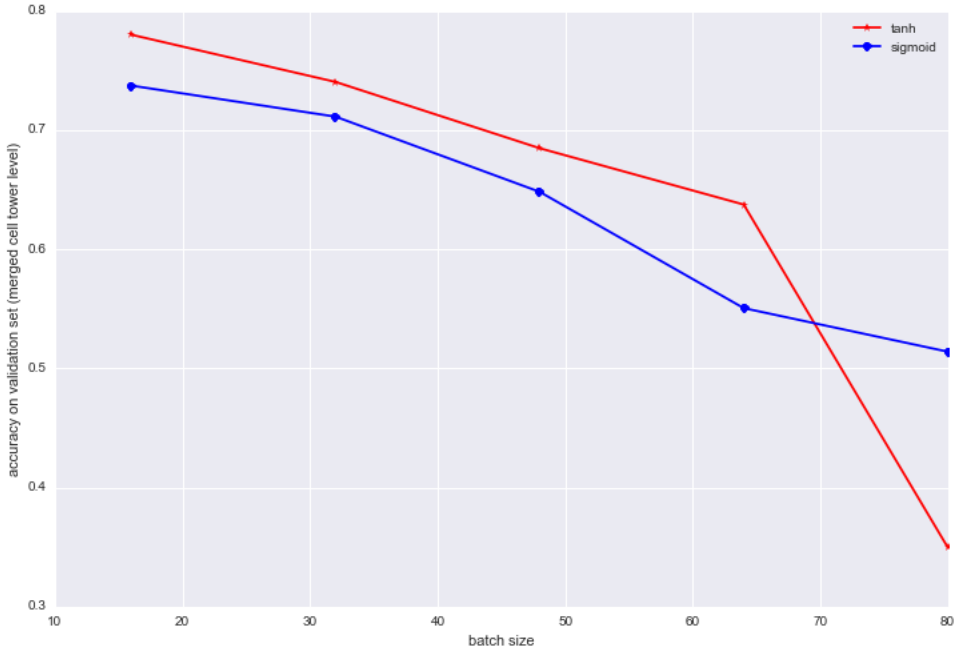

Figure 3-11: Patameter tuning: batch size vs. accuracy

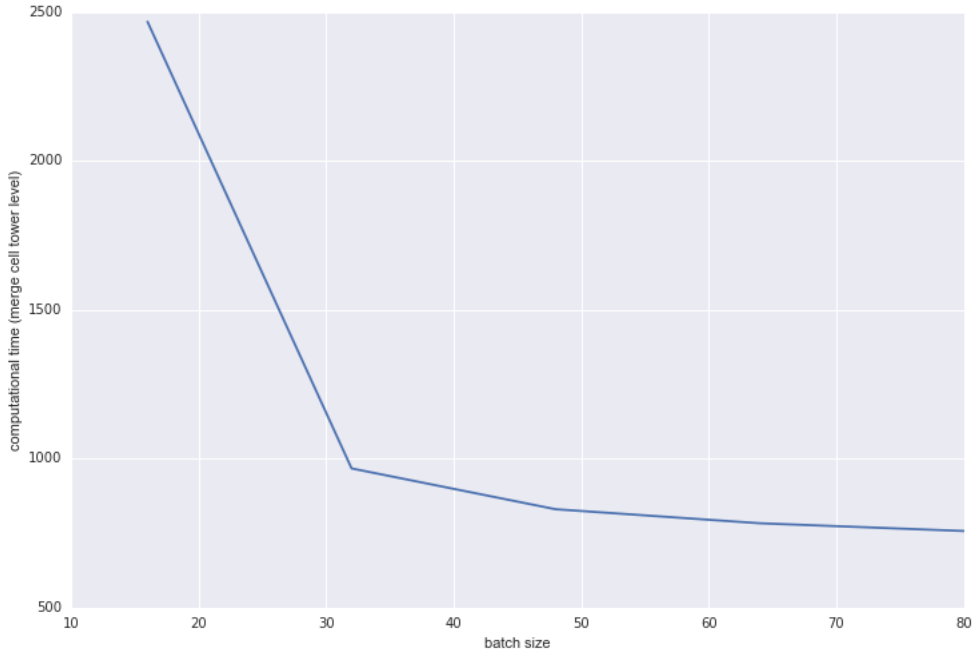

Figure 3-12: Patameter tuning: batch size vs. computational time

Training sample size In this part, we analyze the trade-offs between accuracy of validation set and computation time, mainly determined by training sample size. In Figure 3-13, we show the relationship between training sample size, computation time and accuracies. The y-axis is the accuracies of the validation set and x-axis is the number of the epoch. In the legend, we show the training sample size in the left column and computation time in the right column. In general, the larger the training sample size, the higher the accuracies. Large training sample size indicates long training time, however this is not always true. Larger observations per epoch may lead to few epochs before reaching a desired accuracy level.

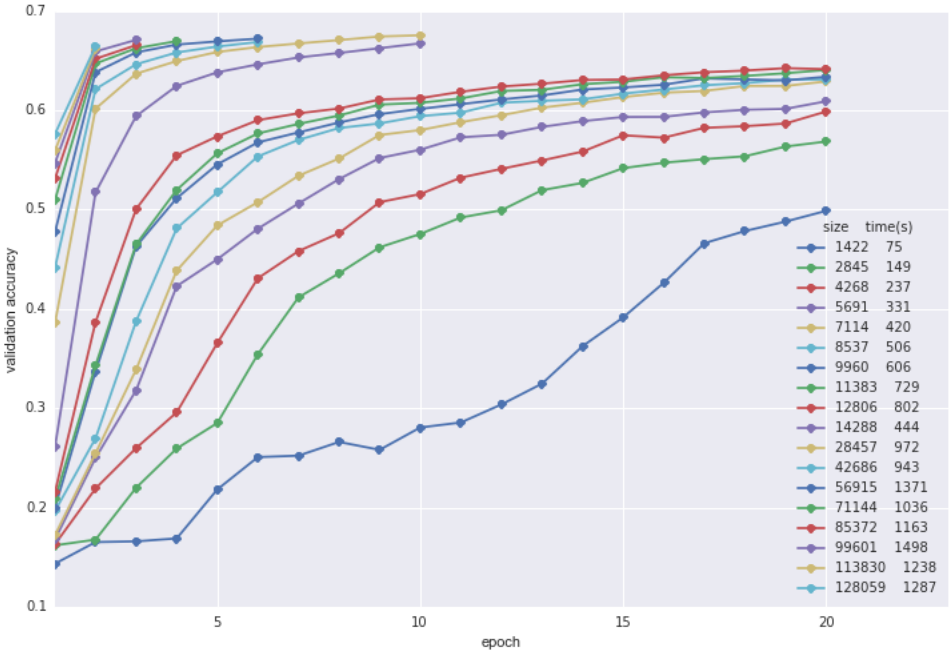

Figure 3-13: Accuracy for each epoch for different sample size and computational time

As a conclusion, the final parameters are determined as follows:

1. *tanh* is chosen as the activation function. It has higher accuracy than *sigmoid* in most cases.

2. The dimension of the cell tower is chosen to be 4. The accuracy decreases sharply when the dimension increases to 8.

3. The dropout rate is determined to be 0.05, since the larger the dropout rate, the lower the accuracy.

4. Batch size is determined based on the interplay between accuracy and computational time. The computation time decreases quickly when batch size increase from 16 to 32. And the accuracy does not vary much. Howevever, if the computational time is not an issue, smaller batch size is recommended for higher accuracy.

5. In terms of the training sample size, we recommend to use at least 57000 individuals, which reaches a comparable accuracy level with larger sample size. When the training sample is 70000, the computational time is the smallest for various sample sizes with similar accuracy.

3.3.5 Results analysis

Accuracy vs. Nationality

The average next-location prediction accuracy varies across different nations as shown in Figure 3-14 and Figure 3-15. The average accuracy of all the nations is shown in the red reference line. At the cell tower level, Belgium, French, Germany and Russian has higher than or approaching average accuracies while the performances on other nations are relatively lower. At the merged call tower level, average accuracies do not vary much.

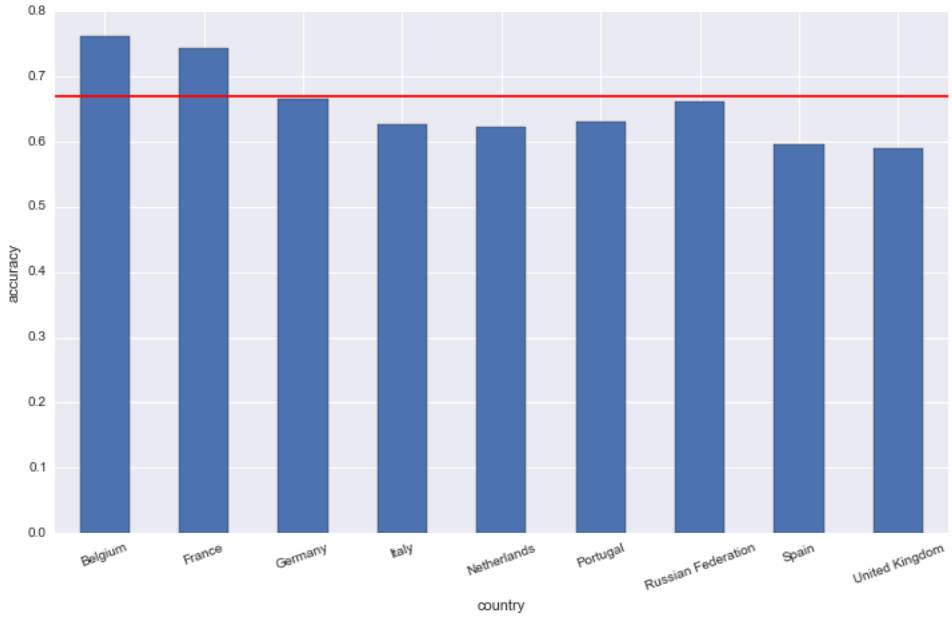

Figure 3-14: Cell tower level: country vs. accuracy

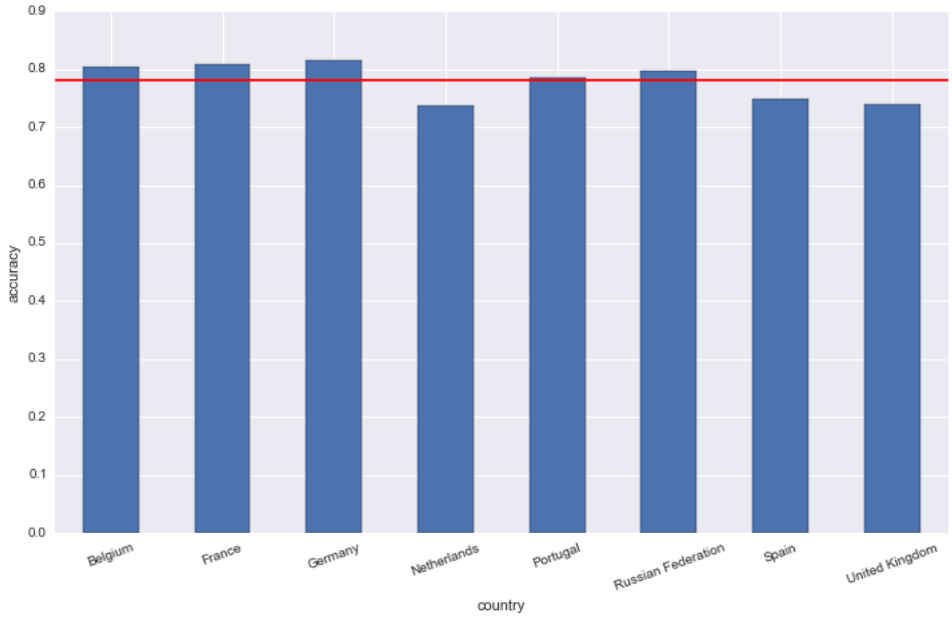

Figure 3-15: Merged cell tower level: country vs. accuracy

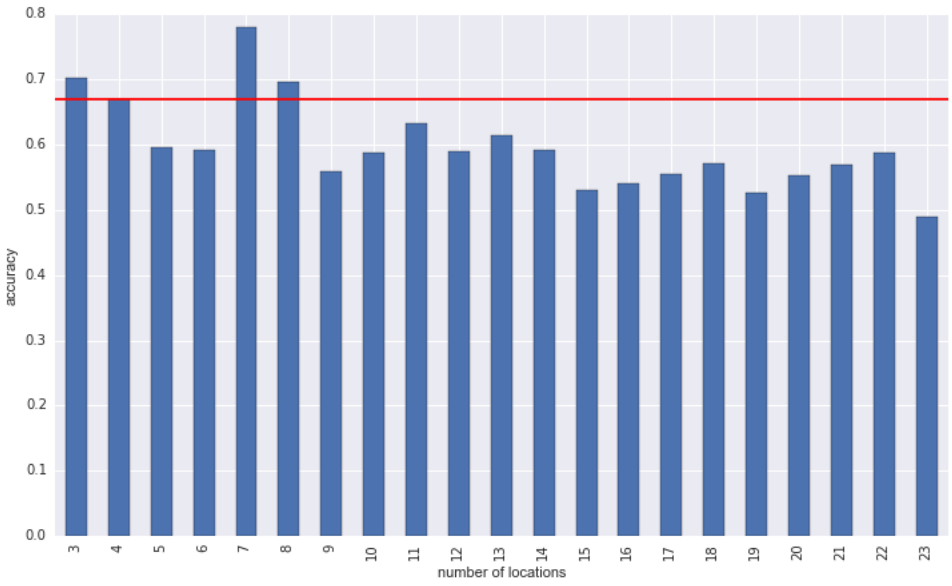

Figure 3-16: Cell tower level: Number of locations vs. Accuracy

Accuracy vs. Frequency

The average accuracy of prediction the next-location varies according to the number of historical locations as shown in Figure 3-16 and Figure 3-17. The average accuracy for all the predictions is shown in the red line for reference.

At cell tower level, when the number of historical location is three, four, seven and eight, the average accuracies are higher than the overall average accuracy. The prediction on users with larger number of historical locations are less accurate. Comparatively, the average accuracies do not fluctuate much at the merged cell tower level. This indicates the prediction accuracy with different number of historical locations varies according to the spatial resolution.

Location embedding and clustering

The embeddings of the cell towers are learned as the output from the embedding layer. Each cell tower is represented by a real-valued vector with a dimension of four.

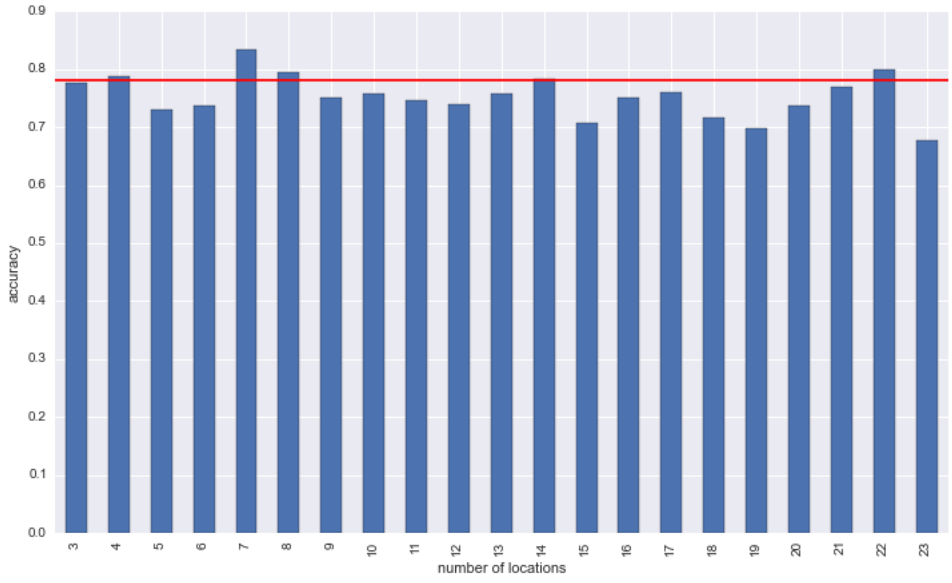

Figure 3-17: Merged cell tower level: Number of locations vs. Accuracy

With the cell tower embeddings, we cluster them into eight categories based on the cell tower embeddings, as represented in different colors in Figure 3-18.

3.4 Conclusion

This project discusses the applicability and performance of Recurrent Neural Network built upon the analogy between mobility behaviors and language models, which appears to be well-posed as a natural language processsing problem. The observations from the case study suggest that RNN can be applied to predict next location that manifest the characteristics of location prediction - specifically sequential, variable number of locations and cell tower interpretations. This chapter provides three key contributions in next-location prediction using CDR. First, we propose to model location traces using RNN and show that location traces greatly fit the settings and architecture of RNN. Second, we enrich the context of cell towers by representing each with location embeddings, dense vector with real-value representations. Thirdly, we

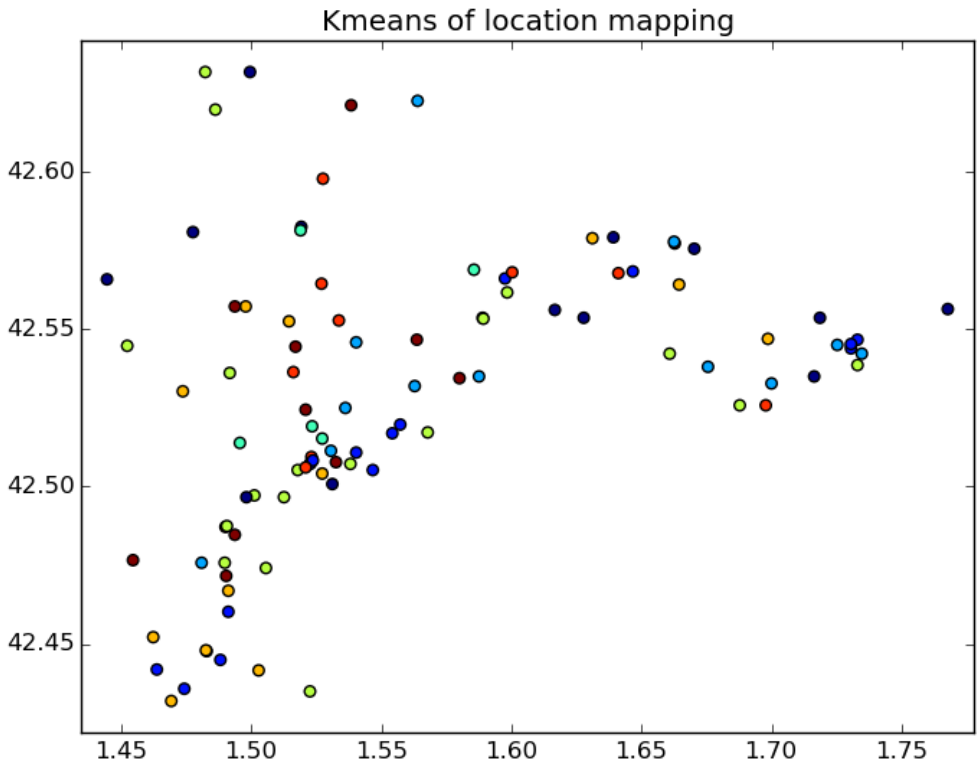

Figure 3-18: Cell tower clustering based on learned embeddings

implement the method in a large-scale case study of Andorra, showing that it has more than 30% accuracy improvement in two settings with different spatial resolutions. Specifically, it has 67% and 78% accuracies in next-location predictions at cell tower and merged cell tower levels respectively. The high accuracies indicate that the method is ready to be applied in many real applications.

Future directions to enhance and enrich mobility predictions using RNN include: 1) forecasting sequences of locations; 2) predicting next location with time stamps. One step further building upon next-location is to predict several locations in the future. We can make location predictions several steps further based on the previous prediction. This fits into the framework of language model well by predicting not only the next word, but the following sequences [61]. At the temporal level, we can use another continuous string describing the timestamps as input to predict the arrival time at the next location. This is helpful in real applications to understand the dynamics of the demand.

Chapter 4

Location recommendations exploiting personal choice flexibilities for system efficiency based on large-scale Call Detail Records

Traveling is an increasingly significant part in our lives. The availability of large-scale longitudinal geolocation data, such as Call Detail Records, offers planners and service providers an unprecedented opportunity to understand location preferences and alleviate traffic congestions. Location recommendation is a potential tool to achieve these two objectives. Previous research on location recommendations has focused on automatically and accurately inferring users' preferences, while little attention has been devoted to the constraints of service capacity. The ignorance may lead to congestion and long waiting time. Call Detail Records could help planners and authorities make interventions by providing personalized recommendations given the comprehensive urban-wide picture of historical behaviors and preferences. This research proposes a method to make location recommendations for system efficiency,

defined as maximizing satisfactions regarding recommendations subject to capacity constraints, exploiting travelers' choice flexibilities. We infer implicit location preferences based on sparse and passively-collected Call Detail Records. An optimization model is formulated with the defined system efficiency. As a proof-of-concept experiment, the method is implemented on the CDR data in Andorra, a small European country heavily relying on tourism. By extensive simulations, we demonstrate that the method can reduce the travel time increased by congestion from 11.73 minutes to 5.6 minutes with idealized trips under full compliance rates. We show that the average travel time increased by congestion is 6.17, 6.98, 8.37 and 10.98 minutes with 80%, 60%, 40% and 20% compliance rates. Overall, the results indicate that Call Detail Records can be used to make location recommendations while reduce traffic congestion for system efficiency. The proposed method can be applied to other large-scale location traces and extended to other location or events recommendation applications.

4.1 Introduction

The persuasive availability of large-scale geo-location data from mobile devices offers an unprecedented opportunity for location-based service providers, transportation agencies, tourism departments and governments to understand human mobility, provide personalized information and improve system operations. Location recommendation has been studied by researchers and applied in industry as a tool to recommend locations according to inferred preferences. However, unlike other recommendation problems, recommending locations simply according to travelers' preferences will lead to congestions and long waiting time due to the capacity constraints of road links or services. We argue that location recommendations should be made for *system efficiency*, which is defined as the maximization of satisfactions regarding location rec-

ommendations given road capacity constraints. Large-scale behavioral data sources, such as Call Detail Records, generate great opportunities to achieve this objective.

First of all, large-scale location traces present a big picture of urban or country-wide behaviors for planners, transportation practitioners and service providers. The visibilities of population-wide behaviors, interests and the decision-making process across all the population enable authorities make interventions at a systematic level. Therefore, we can build a recommendation system based upon not only satisfying personal preferences but also making the best use of the system capacity by balancing the demand.

Another critical component enabling location recommendations for system efficiency is the possibility in exploiting personal choice flexibilities at activity, location and temporal dimensions, specifically, what to do, where to go and when to go. Some travelers, especially for tourism purposes, care more about the activity than the locations for conducting the activities. This is more common for leisure purposes. For these groups of tourists, giving them recommendations on where and when to go would help them make informed decisions. Moreover, on the temporal side, when travelers make decisions on when to carry out an activity, they will make decisions blindly within some constraint if no information is given [14]. Many travelers want to minimize the amount of congestion experienced on the routes and are willing to change destinations if heavy traffic is expected. The time for traveling is flexible under certain time constraint. Under these conditions, which is often the cases, the visibility of what other people's behaviors and congestions along road links will guide them make better decisions. In this paper, we define travelers' freedom in deciding what to do, where to go and when to go as *choice flexibility*.

With the increasing pervasion of mobile phones, Call Detail Record (CDR) has the largest penetration rate compare to all other location traces data. However, it is rarely explored in location recommendation yet due to its sparse, poor-semantic and

coarse spatial resolution characteristics [62]. In this paper, we explore and prove the usefulness of using CDR data to mine implicit preference towards locations. With matrix factorization, we infer the preferences regarding locations where the individuals have not traveled to based on what similar travelers have visited and similar locations. The method we propose to infer location preferences from location traces with no explicit review is not restricted to CDR data. It can also be applied to other location traces data, such as WiFi, check-ins, bluetooth, etc.

With the inferred preferences, we formulate an optimization problem to make best use of system capacity at a pre-determined congestion level with the objective of maximizing the satisfaction towards the recommendations. The method is implemented in Andorra, a European country heavily relying on tourism. The performance of our method is evaluated by the travel time saving comparing with no interventions and preference-based-only recommendation. We analyze the impact of the method at different allowed throughput to understand the trade-off between idealized trips, which we will formally define in section 4.3.1, and increase in travel time. Furthermore, we analyze the efficiency of the method at various compliance rates of the recommendations.

To summarize, the contributions of the paper are as follows:

- We propose a spatial-temporal location recommendation method base on longitudinal and comprehensive Call Detail Records, demonstrating the potential of using CDR in large-scale location recommendations.

- We propose a new perspective in making location recommendations for system efficiency by exploiting users' choice flexibilities.

- We demonstrate the applicability and effectiveness of the method by implementing it on CDR data collected in Andorra. With simulations, we show that we could reduce average travel time increased by congestion during peak hour

from the current 18 minute to 5.6 minute with the same level of travel demand.

- We conduct extensive simulations to evaluate the impact of the proposed methods under various situations. With these experiments, We show the interplay of satisfactions regarding recommendations and incurred congestions. We also simulate various situations with varying compliance rates towards the recommendation.

This paper is organized as follows. Section 4.3 demonstrates the framework of the methodology and detailed descriptions of each step, including preference inference, collective satisfaction maximization and recommendations for system efficiency and traffic flow inference. Section 4.4 describes a case study in Andorra. We first introduce the required datasets. After that, we analyze the impact of the method on traffic congestion alleviation compare to the status quo with no interventions and preference-based recommendation. We also evaluate the impacts and robustness of the method by varying allowed throughput and compliance rates. The last section summarizes results, points out some future works and potential applications.

4.2 Related works

Back to mid-1990s, recommender systems have attracted much attentions to computer scientists [63]. Traditional recommender system deals with user-item rating matrix, where each user has an explicit rating some of the items, with the goal of predicting 'missing' ratings. Collaborative filterings (CF) are very popular in practice. Two main types of CF strategy include user-based CF and item-based CF. The first method assumes that users prefer items similar in the past will like the similar items in the future. The second method assumes that a user will prefer items that are similar to other item he has liked in the past [64, 65, 66]. However, user-based and item-based CF have high time complexities and poor scalability. They also perform poorly when

data is highly sparse. To deal with these problems, matrix factorization has been proposed and applied in recommendation systems recently, which approximates the rating matrix with user-profile and item-profile matrix [67, 68].

Similar to the recommendation in other fields, location recommendations, especially using CDR, have the characteristics of no available explicit ratings, highly sparse and no contextual information. In additional, location recommendation has the salient feature of strict service capacity constraints.

- No explicit ratings available. Unlike the recommendations for books and movies where user rate for past items, we only have frequency data for location recommendations from GPS trajectories, check-ins, and etc [69].

- Sparsity. Data sparsity is a critical problem where we may have multiple records for one location while no records for others [70, 71].

- Context-free. Unlike movies or books recommendations where we may have some contextual descriptions or categories that we could group items in, recommended locations solely based on CDR are only characterized in longitudes and latitudes with no textual descriptions available [72, 73].

- Limited service capacity. Physical presences at the locations are constrained by service capacities. It is no wonder that people may rush to the same location blindly if we only recommend popular locations. This is not economically efficient with large amounts of congestions created, deteriorating travelers' experiences and system performance.

In early-2010s, several studies introduced traditional recommender engines to personalized location recommendation. Ye (2011) [74] introduces user-based and item-based CF to location recommendations using user check-in data based on the assumption that similar users have similar tastes and users are interested in similar

POIs. They found out that user-based CF has higher performance than item-based CF. Berjani (2011) [69] employs the more effective and efficient matrix factorization in POI recommendations on check-in history. To deal with the lack of explicit rating issue, the paper binarizes or bins frequencies to obtain pseudo ratings. Regularized matrix factorization is used to learn the latent feature matrix. They show that matrix factorization has better performance than item-based CF.

Currently, many research focuses on utilizing additional information for more accurate location recommendations. Geospatial factors, social network, and temporal influences are three main types of used used. Some researchers argue that users prefer nearby locations rather than distant ones, which is defined as geographical clustering phenomenon. They assume the distance of visited locations follow certain distribution [75]. Yuan (2013) [76] assumes that the willingness to move from a POI to another POI is a function of distance. Cheng (2012) [77] assumes that users tend to check in around several centers. In social-influenced recommendation, researchers make the assumption that friends shared more common interests than non-friends [78, 79]. To make use of temporal influence, researchers make better location recommendations for different temporal states, which may carry the information about the activities [73, 76, 80]. However, research shows that user preference itself, other than additional information, plays a more important role than spatial and social influences [70, 74]. Even if additional information are useful in making better recommendations, many of them are not always available.

Quercia (2010) [81] is the only work that make recommendations using mobile phone data. However, this paper makes recommendations only base on item-based CF and the spatial closeness to the residential areas. This method is computational inefficient and hard to scale.

Regardless of how sophisticated these methods are in making accurate recommendations, they ignore the constraints of service capacity. As stated above, recom-

mending locations without accounting for the status of demand and supply of the transportation system often leads to congestion, large energy consumption and quality deteoriation of the environment. Therefore, we aim to propose a method to utilize system efficiency while making accurate location recommendations base on historical location traces due to large-scale mobility data.

4.3 Methodology

In this section, we describe the proposed method in details. We start with some definitions used specifically in context of location recommendation. We layout the framework of the method. After that, we introduce how to infer location preferences, followed by traffic flow balancing. These two sections are the core component of the proposed method. In the end, we describe how to infer traffic flows from Call Detail Records.

4.3.1 Definitions

In this section, we define six terms used in the context of CDR-based location recommendations, including user profile, realized trips, location preference, realized trips matrix, allowed throughput and idealized trips.

Definition 1. *User profile.* Each CDR contains the longitude, latitude, timestamp and nationality of the user. User profile $l_{u,g}$ is generated for each user based on individual mobility traces ($l_{u,t_1,g}$, $l_{u,t_2,g}$, ...). g is the user group of user u based on his characteristics directly obtained or inferred from CDR. t_i is the number of presences at the location i. The more characteristics we infer from the data, the more targetted the market is.

Definition 2. *Realized trips.* Realized trips measure the nubmer of times individuals have traveled to each location. $v_{u,i}$ is realized trips of user u at location

i.

Definition 3. *Location preference.* Unlikely traditional recommendation systems where users express their preferences with explicit ratings, scores or reviews, CDR is passively collected with only timestamps at the locations. Given no explicit rating available, we assume that **realized trips**, $v_{u,i}$, measures the implicit preferences of user u towards location i.

Definition 4. *Realized trips matrix.* The realized trips matrix are calculated by the visiting frequency matrix, which is denoted in $V \in N^{m \times n}$ and where there are m users and n locations. Each entry $v_{u,i}$ in the matrix V records the realized trips of user u to location i. As stated above, CDR has the sparsity issue resulting in realized trips matrix as a sparse matrix.

Definition 5. *Allowed throughput.* To improve the performance and efficiency of traffic flow from travelers perspective, it is important for transportation agencies and planners to address the congestion challenges. This can be controlled and managed base on the tolerated congestion, captured by throughput/traffic volumes of each road link.

Definition 6. *Idealized trips.* Idealized trips infer the number of times travelers may visit a location (though actually not) based on the realized trips of similar travelers and similar locations. Idealized trips measures the extent of individual satisfaction regarding the recommended locations in a quantitative way. The idealized trips for new locations are inferred based on realized trips in the historical traces from similar tourists and similar locations with Matrix Factorization, as illustrated in section 4.3.3.

4.3.2 Framework

The proposed location recommendation system aims to recommend locations that satisfy individuals' preferences while making efficient use of service capacity. Figure

4-1 shows the architecture of the proposed location recommmendation system. The inputs to the framework are represented in yellow ellipticals, including CDR, traffic counts, road network and allowed throughputs (determined by the authorities and planners). The output is the personalized recommendations with spatial and temporal information, as shown in the green elliptical.

As shown at the bottom of Figure 4-1, The final recommendation requires two main components to build a optimization model, including the objective function and the capacity constraint. The objective function is captured by location preferences and the road throughput constraints are determined by the road charateristics, fixed demand and variable allowed throughput.

From path ①1, we obtain the preferences matrix of users receiving recommendations, such as tourists from CDR, for the objective function. We first use realized trips, calculated by visit frequencies, as a proxy for location preferences. We infer hidden factors that characterize both locations and travelers inferred from visiting frequencies. High correspondence between locations and travelers leads to a recommendation. This is matrix factorization, one type of latent factor model. This enables us to formulate the objective function.

Following path ②, we infer fixed traffic demand, scaling factor that maps CDR-based OD matrix to actual vehicle flows, and share of peak hour traffic flows. The fixed demand is the base traffic flow that can not be changed, such as that of locals. In path ③, we identify the designed road capacity based on the attributes of the road network. The combination of path ②, path ③ and allowed throughputs determine the throughput constraints, which enable planners and transportation practitioners to trade-off between congestion and satisfaction regarding recommendations.

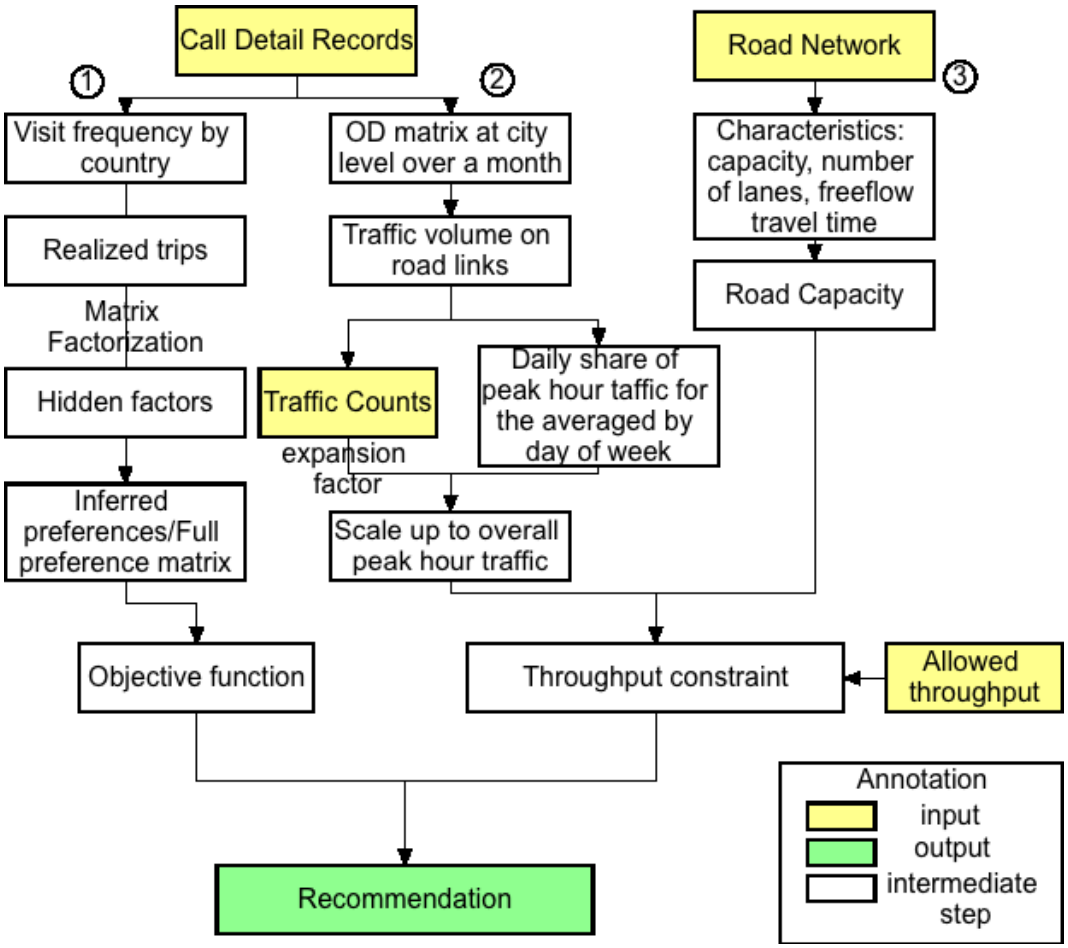

Figure 4-1: Methodology framework

Table 4.1: Data comparisons in location recommendation

	CDR	GPS [82]	Social network check-ins (Foursquare, Gowalla , twitter, etc.)[80]
Sampling bias	Relatively low	High	High
Spatial resolution	Low	High	High
Explicit review	No	No	Yes
User information	Nationality, type of phone	Some has socio-demo	Some has socio-demo
Passively collected	Yes	Yes	No
Energy cost	Low	High	High

4.3.3 Preference inference

In this section, we describe how to infer individual location preferences regarding all the locations based on CDR. We then show how to use matrix factorization to infer location preferences on sparse CDR data.

In this thesis, we developed methods to learn the preferences from CDR. The data has seldomly been used in location recommendations. In the literatures, most large-scale location recommendations are implemented on social network check-in data and a small part is on GPS histories. CDR has the largest spatial coverage and saves more energy than other location traces. Comparing to check-in data, it is passive collected with no active communication needed. Furthermore, CDR data has the largest coverage scale with no requirements of application installed, active logging and etc. This is critical for planning and management at a systematic level. We compare the characteristics of CDR and other data sources in the location recommendation research as shown in Table 4.1.

Matrix factorization

There are two main strategies for understanding individual preferences, content filtering and collaborative filtering. Content filtering requires a profile for each user or

product to characterize its features, such as demographic information and questions about the user and the characteristics about the product [67]. Collaborative filtering make recommendations based past user behaviors by analyzing the relationships between users and the items. User or item-oriented filtering usually requires dense observations. We use low-rank matrix factorization, one type of latent factor model, to infer travelersâĂŹ preference regarding new locations. This model characterizes both the locations and users by vector of factors inferred from location frequency patterns, mapping both travelers and locations to a joint latent factor space of dimensionality k. The latent factor space determines why/how traverlers like each location. In our problem, for example, we attempt to interpret hidden factors as hidden personal interests towards activities, such as outdoor activitis lovers, shopping lovers, and etc [67]. By characterizing both the location and the travelers automatically inferred from the frequency matrix, we could predict the preferences regarding new locations with respect to a certain traveler and a certain location. High correspondence between location and user factors leads to a recommendation.

In this problem, each user u is characterized by a vector p_u measuring the extent to which the tourist is interested the locations. The dimension of p_u is the number of hidden factors, the size is represented by k. The resulting dot product $l_i^T p_u$ captures the interaction between tourist u and location i, as shown in Equation (4.1). This predicts tourists' interests regarding the location with no visits.

$$x_{ui} = l_i^T p_u = \sum_{k=1}^{K} l_{ik}^T p_{ku} \tag{4.1}$$

To represent Equation (4.1) in a matrix operation, we have Equation (4.2).

$$X \approx L \times U^T = \hat{X} \tag{4.2}$$

where V is realized preference matrix and $\hat{V}$ is idealized preference matrix.

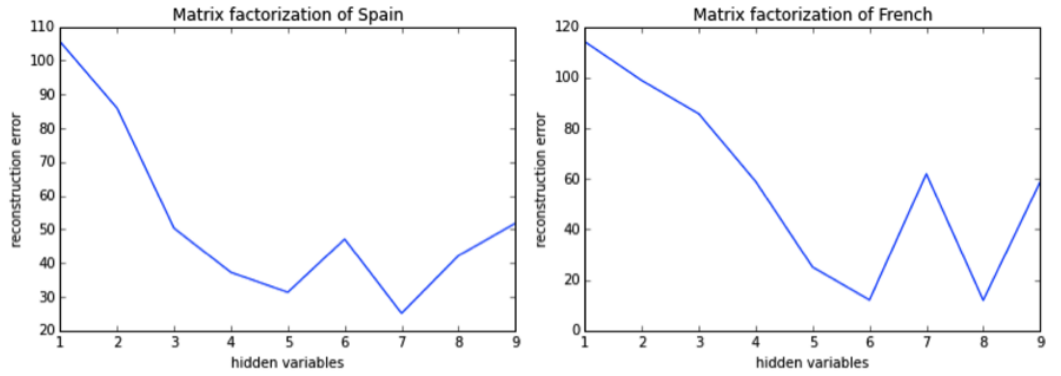

Figure 4-2: Illustration of determining the number of hidden variables in matrix factorization

Gradient descent is used to iteratively minimize the difference between the realized preference matrix and the idealized preference matrix [83]. Assume O contains all the traveler-location pairs together with the realized preferences. The objective is to minimize the total sum of errors between the two matrices. The idealized preferences regarding locations with no realized preferences can be determined based on the associations between traveler and location preferences.

$$Error = \sum_{(p_u, l_i, v_{ui}) \epsilon O} (v_{ui} - \sum_{k=1}^{K} l_{ik}^T p_{ku})^2 \tag{4.3}$$

To determine the number of hidden factors, we minimize the root square mean error (RMSE) between V, realize preference matrix, and $\hat{V}$, idealized preference matrix, as calculated in Equation (4.3). As an illustration, Figure 4-2 shows the determination of hidden factors for two tourist group, French and Spanish, by plotting the number of hidden variables on the x-axis and RMSE on the y-axis.

4.3.4 Recommendations for system efficiency

To recommend locations for system efficiency, an optimization model is built to maximize the preferences regarding location recommendations subject to road capacity constraints. This model can be easily extended to the case where capacity constraints

Table 4.2: Key notations in this chapter

Notation	Meaning
x_{ijt}^{g}	Decision variable representing whether recommend tourist i in group g to travel to city j on day t
x_{i}^{g}	Decision vector of tourist i in group g
p_{ij}^{g}	Inferred preferences of tourist i in group g regarding location j
$R(x_{i}^{g})$	Map tourist i in group g to road links
C_{road}	Vector of designed road capacity determined by the characteristics of the road links, where $C_{road} \in N$
D_{fixed}	Vector of fixed demand that is changeable, where $D_{fixed} \in N$
$AETH$	Vector of allowed excess throughput, where $C_{remain} \in N$
I^{g}	Set of indices for tourists in group g, where $I \in N$
G	Set of indices for tourist groups, where $G \in N$
J	Set of indices of locations, where $J \in N$
T	Set of indices for days to be recommended, where T $\in N$

exist by modifying the constraint accordingly.

To illustrate the idea and simplify the model, the location optimization model is built based on the following assumptions:

- To simplify the road capacity constraints, we assume that travelers will travel to no more than k location in a day.

- We assume certain compliance rate for all the tourists. The definitions and simulations base on the compliance rates are described in section 4.4.2.

- The customers could be segmented into several groups for more targetted marketing. From CDR, they can be segmented based on nationality, phone type, frequency, temporal characteristics, and etc.

Notations

In order to formulate the optimization model, we define the following notations in Table 4.2.

Model formulation

With the notations above, the mathematical formulation of the optimization model is as follows. The objective function is shown in Equation 4.4 and the constraints are show in Equation 4.5 to Equation 4.7.

The objective function of this optimization model (4.4) represents the maximization of overall satisfaction regarding the recommendations, the total idealized preferences.

Constraints (4.5) and (4.6) are personalized constraints of the number of locations individuals are recommended, according to the first assumption. We recommend only one location across the time period. However, this can be easily modified based on the applications.

Constraints (4.7) is the crucial constraint, the collective constraint, based on allowed throughput, road characteristics and the unchanged traffic demand. This is a flexible constraint that can be modified by the agencies, government and planning department to control the congestion level. It reflects how much delay in travel time the authourities can tolerate or how much road performance they are willing sacrifice.

$$\text{maximize} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{g=1}^{G} \sum_{i=1}^{I^g} p_{ij}^g \times x_{ijt}^g \tag{4.4}$$

s.t.

$$x_{ijt}^g \in \{0,1\}, \quad \text{for} \quad t \in T, j \in J, g \in G, i \in I^g \tag{4.5}$$

$$\sum_{t=1}^{T} sum_{j=1}^{J} x_{ijt}^g \in \{0,1\} \quad \text{for} \quad g \in G, i \in I^g. \tag{4.6}$$

$$\sum_{t=1}^{T} (\sum_{j=1}^{J} \sum_{g=1}^{G} (\sum_{i=1}^{I^g} R(x_{ijt}^g))) - D_{fixed} \leq AETH + C_{road} \tag{4.7}$$

4.3.5 Traffic flow inference

The following session describes how to infer traffic flows along road links during peak hours from Call Detail Records. In a nutshell, we first infer tower-to-tower Origin-Destination matrix. We then scale it up to capture the population-wide movements. Finally, we distribute traffic flows to road links and calculate the peak period traffic flow.

OD matrix inference has been widely studied in the literatures [84, 85, 86]. However, it needs careful scaling up to the actual population using whatever data is available in real applications to map the CDR-based OD Matrix to the actual ones. In this paper, traffic count collected by cameras at key locations is used. Other useful datasets with the same functions include household travel surveys, census, etc.

- Tower-to-tower OD matrix

  Each call detail record contains unique anonymized caller id, initiating and ending time of the call, cell tower id, registration nation of the phone and phone mode. It could be linked with another cell tower database for the approximated location of the caller at the timestamp the call is initiated. Tower-to-tower OD matrix captures the aggregated movements from one cell tower to another cell tower.

- Map OD pairs to road links

  For further calculation of travel time and congestion analysis, the traffic flow needs to be assigned to road links. The tower-to-tower OD matrix can be mapped to road links based on the road networks using Equation (4.8).

$$road_i = \sum_{jk} O_j D_k \tag{4.8}$$

  where $road_i$ is the sum of the traffic flows of all tower-to-tower OD pairs passing

through road link i.

- Scaling factor

  CDR could only track individuals when they are connecting to cell towers, calling, messaging or using data services. Users with no cell phones, trips with no phone usage and users not traveling with vehicles all lead to the mismatching between CDR-based taffic flow and actual traffic flows. Therefore, expansion factors are needed to scale the CDR-inferred traffic flow to the actual traffic flows on road links.

$$TC_i = R_i \times \beta_i \tag{4.9}$$

  The subscript i is the index of the road link. TC_i is the actual traffic counts and R_i is the sum of all the traffic flows on the road link at the same temporal scale. Here, β_i captures both the penetration rates of mobile phone (not every individual have mobile phone and some may have multiple phones), the non-usage of mobile phone (travelers may not call before they depart and after they arrive) and vehicle usage issue (several travelers may share the same vehicle and travelers may not use automobile when traveling).

- Peak hour traffic flow

  Traffic volumes vary a lot throughout the day. An hour is chosen to be the time window for the calculation of traffic flow, the commonly used unit in traffic flow analysis. The hourly traffic flow varies to a large degree and on different road links. There is a large variation in traffic flow for different hours and they vary differently across road links. To capture the most severe congestion period, we use the peakest hour of each each road link.

To summarize, we decribe the proposed method in this section. We start with the definitions developed for the proposed methodology and the methodological frame-

work. After that, we introduce the application of Matrix Factorization in location preference inference. Next, we develop an optimization model maximizing preference regarding location recommendations subject to capacity constraints. In the end, we describe how to infer traffic flows from Call Detail Records.

4.4 Case study and experiments

To demonstrate the applicability and evaluate the method, we make personalized location recommendationss for a weekend in August, 2015 collected in Andorra. With an estimated amount of 10.2 million tourists annually, tourism accounts for roughtly 80% of its GDP [87]. Therefore, Andorra developed active summer and winter resorts and holds various events to attract more foreign tourists to come, stay and spend.

In the following parts of this section, we first introduce the data used in the experiments, including CDR, road network and traffic counts. By comparing the proposed method with the status quo with no intervention and the preference-based only recommendations, we show the effectiveness in reducing traffic congestions. We then do extensive simulations to test the effective and robustness of the method at different scenarios. We first test the method with different allowed throughput, which enables agencies and planners to trade-off between tolerating more congestions and accomodating more trips. We then test the method on various compliance rates. This shows the impact and effectiveness of the method under different accepatance rate scenarios.

4.4.1 Data

For this case study, we rely upon three data sources, including CDR, traffic counts and road network as summarized below.

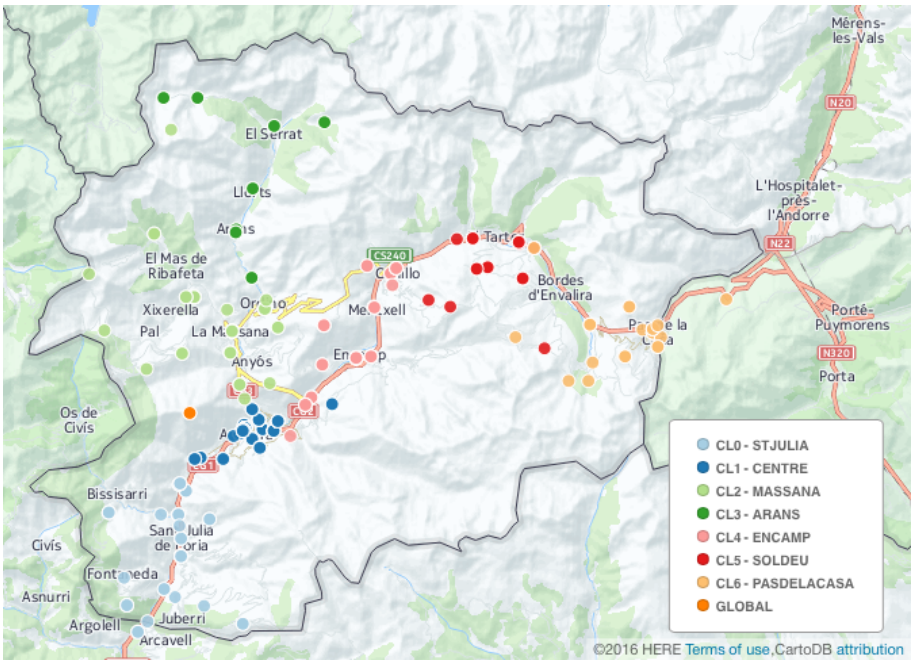

Figure 4-3: Tower distribution in Andorra
Each color represents one cluster. The legend consists of cluster index and the shortage of the city name.

Call Detail Records

Andorra is situated between France and Spain. These two nations, our target markets, take up 90% of travelers visiting Andorra. Spanish enter Andorra from Saint Julia, south of Andorra and French tourists enter from El Pas dela Casa, north-eastern of Andorra. The road links connecting these two boundaries to other parts of the countries are different. The locations are recommended at the city level, including St Julia, Massana, Arans, Encamp, Centre, Soldeu, El Pas dela Casa, as shown in Figure 4-3. The cell towers of different cities are represented in different colors as shown in the legend. We aim to provide recommendations to 47743 tourists, including 20311 French tourists and 27432 Spanish tourists. We make the assumptions that these tourists are the targeted market of tourists who decide to travel to Andorra. Note that identifying this group of tourists is not a trival issue in real application.

Each Call Detail Record contains the encrypted user ID, starting and end time of

DS_CDNUMORIGEN	DT_CDDATAINICI	DT_CDDATAFI	NUM_DURAI	ID_CELLA_IN	ID_CDOPERADORORIGEN	TAC_IMEI
de602c55e450fc411a462e70f9f147f566072accb190e9686ad83e356927be65	2015.08.28 23:04:00	2015.08.28 23:04:09	9	9061	21303	35829106
e70a02dcc4b045ccf0cdcfbfcf8c2a0e630805ef9c1e9487e4c7ca06a39429d9	2015.08.29 00:52:55	2015.08.29 01:33:28	2433	1582	21303	35858805
e70a02dcc4b045ccf0cdcfbfcf8c2a0e630805ef9c1e9487e4c7ca06a39429d9	2015.08.28 21:20:35	2015.08.28 21:22:00	85	9061	21303	35858805
c2fc0dc7cb2c622a149bf8677ff316537f4b3de4d831ce7b7b1cc9e4c169da8d	2015.08.29 04:06:14	2015.08.29 04:06:19	5	9112	21303	35206706
c2fc0dc7cb2c622a149bf8677ff316537f4b3de4d831ce7b7b1cc9e4c169da8d	2015.08.29 04:28:00	2015.08.29 04:28:05	5	9112	21303	35206706
26245a191d00d617449cf39ac43180916487da03977ef374ff1fc623d3ebb5df	2015.08.28 21:40:53	2015.08.28 21:44:11	198	9010	21303	35982605
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 22:08:18	2015.08.28 22:08:57	39	1032	21303	35584706
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 23:00:41	2015.08.28 23:00:42	1	1071	21303	35584706
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 22:56:05	2015.08.28 22:57:08	63	9502	21303	35584706

Figure 4-4: Snapshot of Call Detail Records

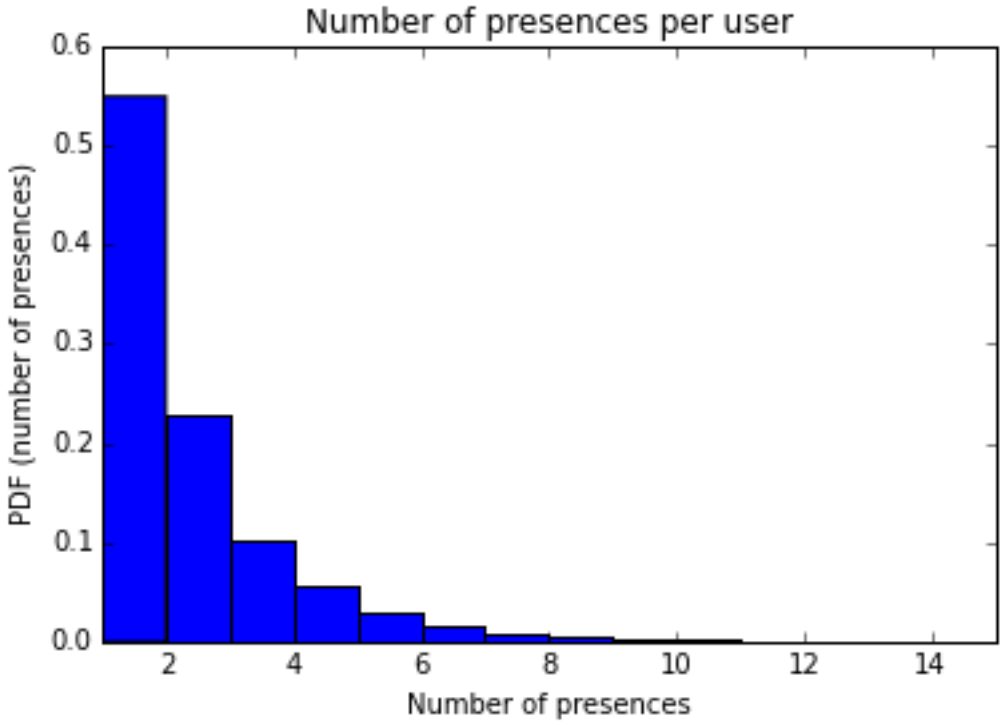

Figure 4-5: Number of presences per user

the phone call, connected cell tower ID, nationality and the mode of the phone, as shown in Figure 4-4. The field "ID_CDOPERADDRORIGEN" records the country code, enabling us the recognize the registration nation of the user. The spatial distribution of the cell towers are shown in Figure 4-3, with different color representing different cities.

From Figure 4-5, we can see that 55% of tourists have one record and 22% of users have two records. Besides, as shown in Figure 4-6, 62% of travelers have visited only one cities and 28% of travelers traveled to two cities. The sparsity problem, an ibnate characteristic of CDR data, makes the preference inference challenging.

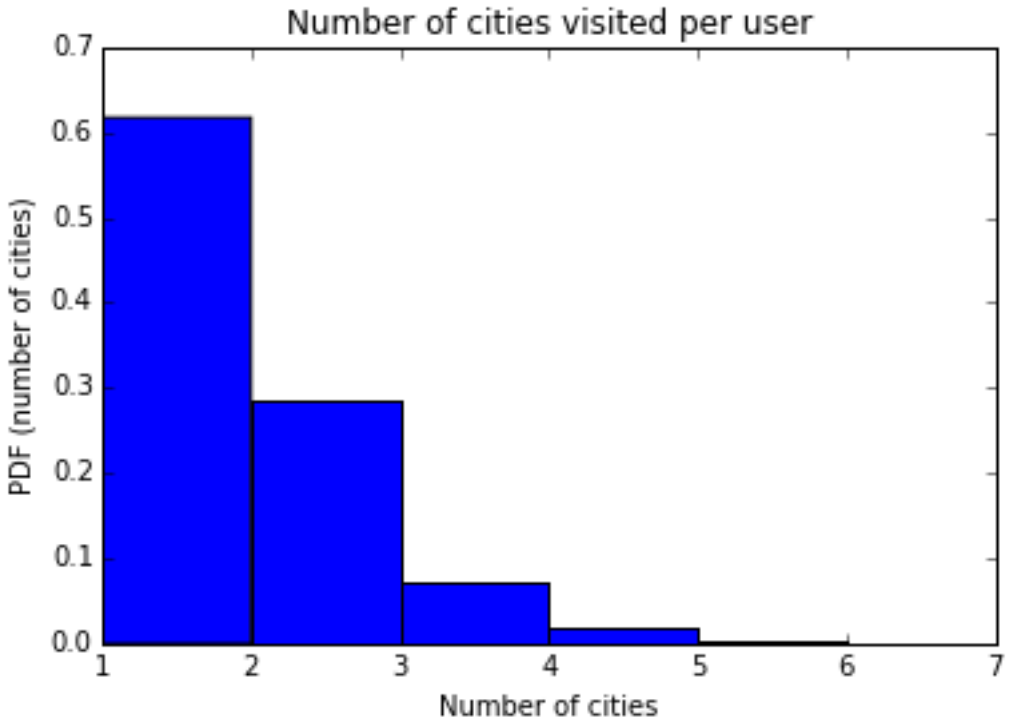

Figure 4-6: Number of cities per user

Traffic counts

Traffic counts are collected at key locations by Andorra government to monitor the internal mobility by cameras. The monthly data are publicly available. The monthly traffic counts of six key locations are selected as the ground truth to scale up CDR-based traffic flows to actual traffic flows.

Road network

To calculate the travel time, we need to map Origin-Destination matrix to road links, which is obtained from the road network scraped from Google Map API. Free-flow travel time and number of lanes are key inputs to model travel time as a function of traffic flow, as shown in section 4.4.2. The road network linking different cities in Andorra are shown in Figure 4-7. The attributes of each road link, including name, connecting cities, number of lanes, capacities and free flw travel time, are shown in

Table 4.3: Road Characteristics

Road Name	Connecting Cities	Number of Lanes	Capacity per Lane (vehicle per hour)	Free Flow Travel Time
CG1 (FEDA)	StJulia - Centre	2	846	9
CG3 (Serra de l'Honor)	Massana - Centre	1	1686	11
CG2 (Pleta Engolasters)	Encamp - Centre	1	1686	3
CG3 (Ordino North)	Massana - Arans	1	1742	9
CG2 (Meritxell)	Encamp - Soldeu	1	1742	19
CG2 (Border Pas de la Casa)	PasDeLaCasa - Soldeu	1	1742	18

Table 4.3.

4.4.2 Evaluation

In this section, we assess the effectiveness of the method using two metrics. **overall idealized trips** measure the satisfaction of all the individuals regarding the recommendtions. **Increased travel time caused by congestion** measures the externality of the recommendations based on congestion level on road links. Table (4.4.2) summarizes congestion level, average travel time increased by congestion and idealized trips of the status quo with no interventions, the preference only recommendation and the proposed method under various scenarios.

In transportation, practitioners and planners are interested in modeling the relationships between traffic flow and travel time based on the characteristics of the road infrastructure. One of the most simplistic and widely used one is Bureau of Public Road function, which models travel time as a function of the ratio between actual traffic volume and road maximum flow capacity, volume-over-capacity (VOC) [88], as

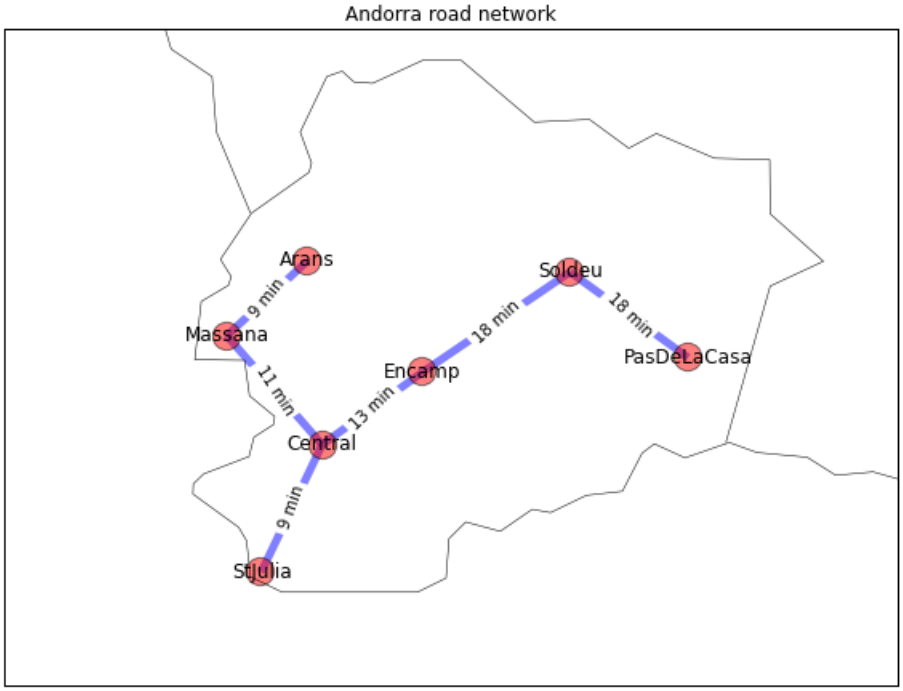

Figure 4-7: Andorra road networks

Table 4.4: Evaluation Results [a]

Scenario	Congestion (V/C)	Increased travel time during peak hour (min)	Total idealized trips
Status quo (No interventions)	1.67	18.58	NA
Preference only	1.43	11.73	64925
Proposed (100% compliance)	1.16	5.61	44930
Proposed (80% compliance): exp. 1	1.22	6.17	49997
Proposed (60% compliance): exp. 1	1.27	6.98	53680
Proposed (40% compliance): exp. 1	1.33	8.37	57442
Proposed (20% compliance): exp. 1	1.40	10.40	61219
Proposed (80% compliance): exp. 2	1.18	5.75	43836
Proposed (60% compliance): exp. 2	1.22	6.13	41042
Proposed (40% compliance): exp. 2	1.24	6.54	38186
Proposed (20% compliance): exp. 2	1.26	6.97	35326

[a] *exp.1* is based on the setting in Experiment 1 and *exp.2* is based on the setting in Experiment 2. as discussed in section 4.4.2

shown in Equation (4.10).

$$t_{current} = t_{free\,flow} \times (1 + \alpha(V/C)^{\beta}) \tag{4.10}$$

where $t_{free\,flow}$ is the free flow travel time on the road segment, α and β are parameters that are used to characterize the non-linear relationship between V/C and $t_{current}$. According to Bureau of Pubic Roads, the default value are $\alpha = 0.15$ and $\beta = 4$.

We compare our method with two baseline models. The first one is the current condition with no interventions. The second is the preferenece-based only recommendations where we recommend locations simply based on personal preferences taking

no system efficiency into account. According to Table 4.4, we observe a dramatic impact in reducing the increased travel time caused by congestion from the current 18 min to 5.6 min under full compliance rate, which we will define in section 4.4.2. By comparing with the preference-based only method, it reduce nearly half of the congestion time sacrificing one third of the preference. We will discuss the two experiment settings in section 4.4.2.

Interplay anlaysis among asllowed excess throughput, idealized trips and increase in travel time

In this session, we perform extensive simulations to gain insights on the interplay between congestion and satisfaction regarding recommended locations under various conditions. The allowed excess throughput measures the part of additional throughtput that is larger than designed capacity of the road link. It is up to the transportation agencies or tourism departments to determine the congestion level they can tolerate, giving them the freedom to intervene and manage traffic demand beforehand.

The relationship between allowed excess throughput and travel time increased by congestion is shown in Figure 4-8. The relationship between the two factors are roughly linear, as shown in Equation (4.11). With one additional traffic flow, the average increase in travel time is 0.00055 minute. In other words, 1818 traffic flows on the entire road network would increase the travel time by one minute.

$$\Delta t = 0.00055 \times TH + 4.80. \tag{4.11}$$

where Δt is average travel time increased by congestion (min). TH represents the allowed excess throughput (vehicle/lane).

We then analyze the relationship between allowed excess throughput and idealized trips. The larger the allowed throughput, the higher traffic volumes are allowed on

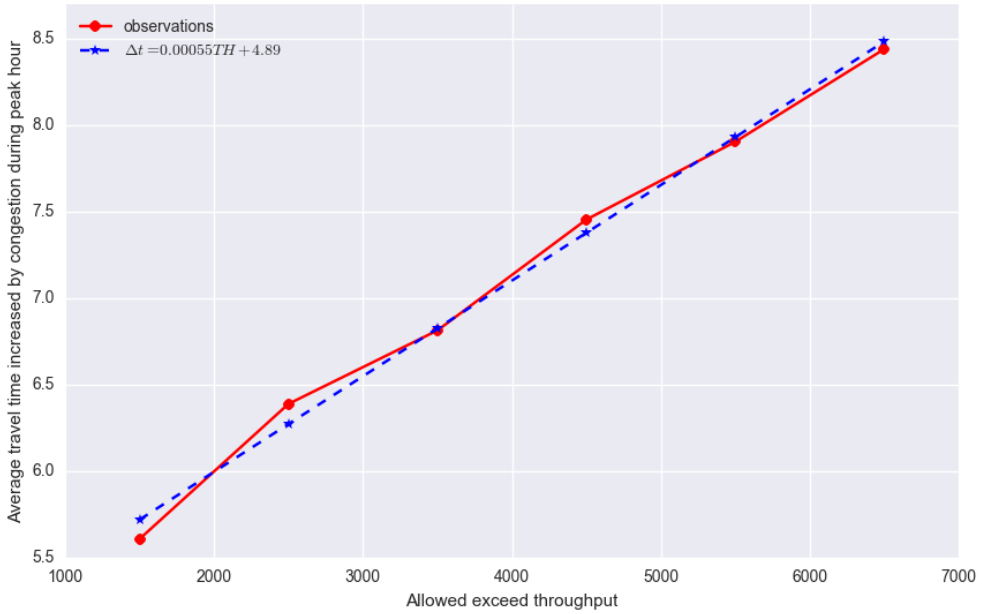

Figure 4-8: Allowed exceed throughput vs. average travel time increased by congestion

the road links, and the larger are the satisfactions regarding the recommendations. The increase in idealized trips is faster when allowed throughput is small. We use a logarithmic regression to model the relationship between the two, as shown in Equation (4.12) and Figure 4-9.

$$IT = 5741 \times ln(TH) + 2907 \tag{4.12}$$

where IT represents overall idealized trips based on the recommendations.

Idealized trips and congestion are the two factors we try to balance. It is useful to learn the interplay between the two and thereby make better informed decisions. We use a two-degree polynomial regression to model the interplay between idealized trips and increase in travel time cased by congestion, as shown in Equation (4.13) and Figure 4-10. If the average travel time increase by 1 minute, we can increase 12097 idealized trips. If the average travel time increase by 2 minutes, we can increase 21181

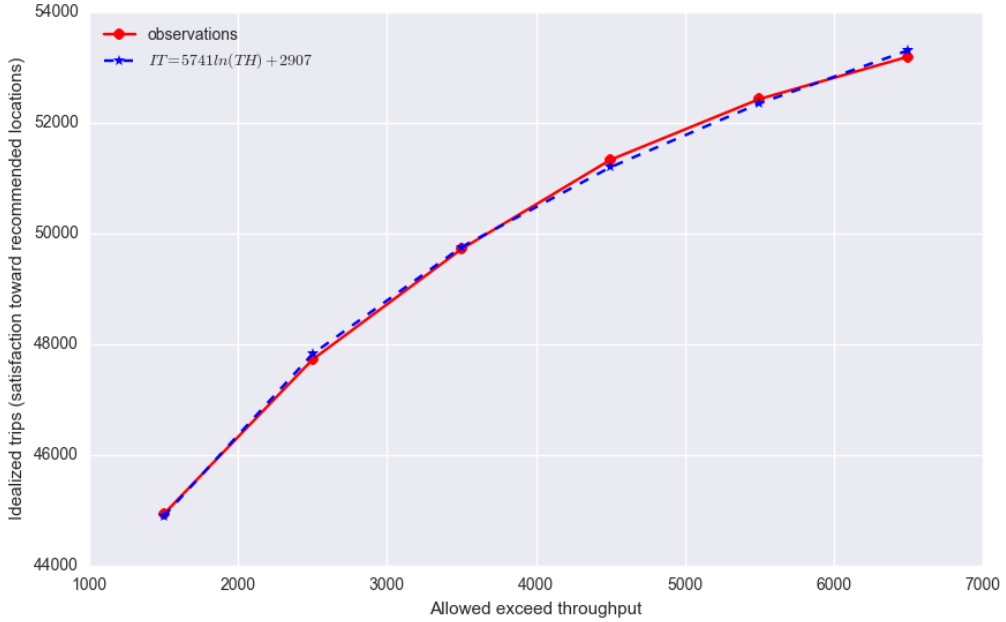

Figure 4-9: Allowed exceed throughput vs. idealized trips

idealized trips.

$$IT = -553 \times \Delta t^2 + 10743\Delta t + 1907 \tag{4.13}$$

Compliance rate analysis

The synergy among different individuals and the authorities play a fundamental role in intelligent tourism management, event planning and building smart cities. If every tourist can take part, the system will be closer to ideal mobility status. However, in reality, due to time availability, budget constraints and customers' trust in the recommendation, travelers could either follow, still stick to their own preferences or be deterred away by recommendation. We analyze the impact of our method under various compliance rate situations to understand the robustness of the method with two experimental settings.

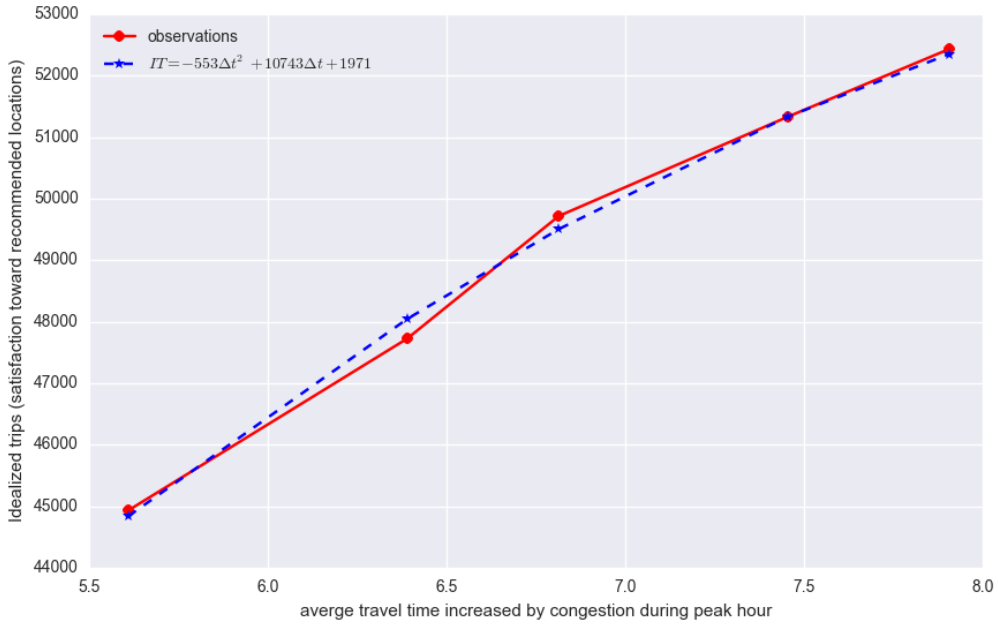

Figure 4-10: Increased travel time vs. idealized trips

Experiment 1 In this experiment, we assume certain compliance rate for all the individuals. For the individuals who do not comply, we assume that they will follow their own preference, meaning that they will travel to the location where they are most interested in. This group of toursits is those who have a hard preference regarding the locations and is reluctant to make behavioral change.

Experiment 2 As Experiment 1, we assume certain compliance rate for all the individuals. A slight difference lies in the individuals who refuse to adopt the recommendation. To understand the impact of poor recommendations that drive people away, we make the assumption that 50% will not change their behavior and the remaining 50% will not come.

Allowed throughput vs. increase in travel time We model the relationship between allowed exceed throughput and average travel time increased by congestion under the two experiment settings, as shown in Figure **??**. In general, we can see

that the higher the compliance rate, the less congestion. This implies the importance in successful marketing. Both experiments show that the higher the compliance rate, the steeper the slope. Since higher compliance rate indicates larger control power of the authorities. The relationship between these two factors are different in the two experiments. In experiment 1, we observe a linear relationship while experiment 2 shows a logarithmic relationship. From the slopes of both experiements, we show that the higher the compliance rate, the larger impact the model has. Besides, allowed throughput has larger control in experiment 1 than in experiment 2. Both observations indicating the importance of successful marketing strategies.

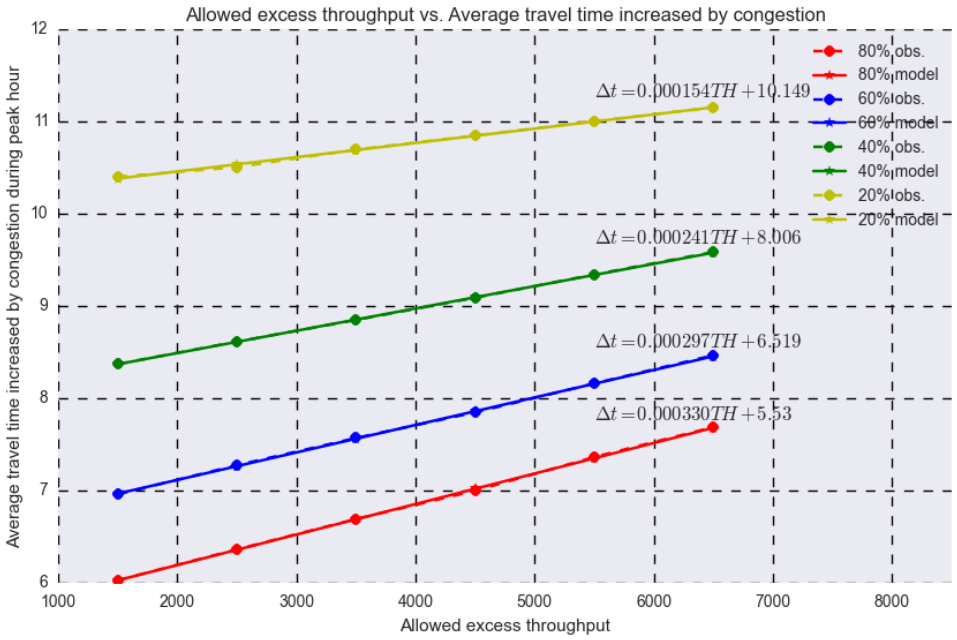

Figure 4-11: Setting 1

obs. is the simulation results. *model* is the best line of fits.

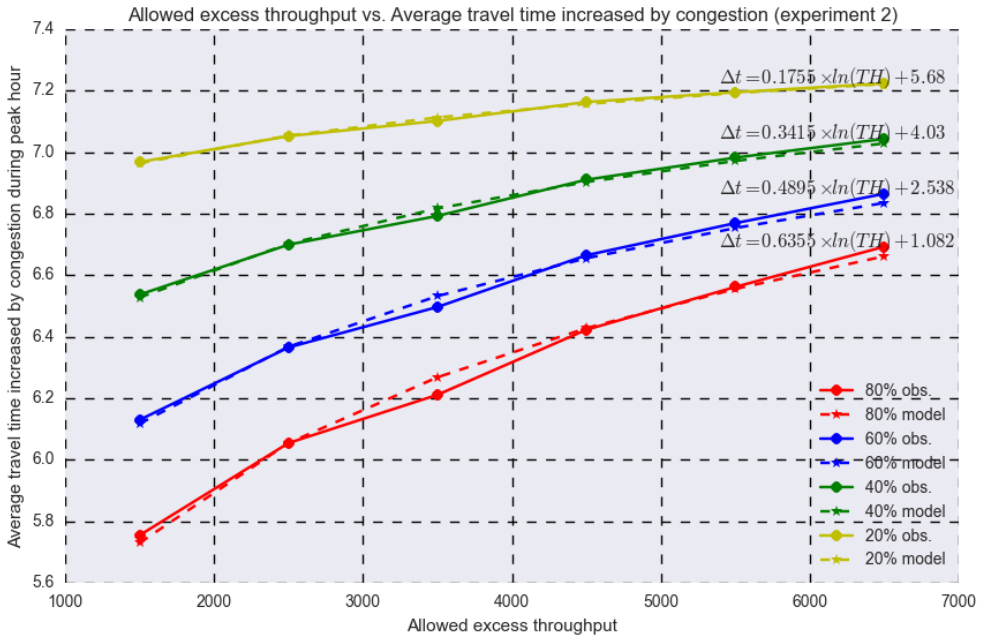

Figure 4-12: Setting 2

obs. is the simulation results. *model* is the best line of fits.

Allowed throughput vs. idealized trips There exists a positive relationship between idealized trips and allowed exceed throughput, as shown in Figure **??**. Both experiments are modeled with one-degree logarithmic model with steeper slope in the beginning and flatter slop at the end, which indicate that as the allowed throughput increase, the increase in idealized trips decreases with a larger portion of travelers satisfied. In experiment 1, the smaller the compliance rate, the higher the idealized trips since travelers will follow their personal preferences. On the contrary, in experiment 2, the smaller the compliance rate, the lower the idealized trips since travelers are turned away by the dissatisfied recommendations. In experiment 1, large allowed throughput decrease the differences in idealized trips of various compliance rates since people are more likely to follow personal preferences. However, in experiment 2, the differences in individualized trips among various compliance rates increase as less restrictions will be exerted on individual recommendations.

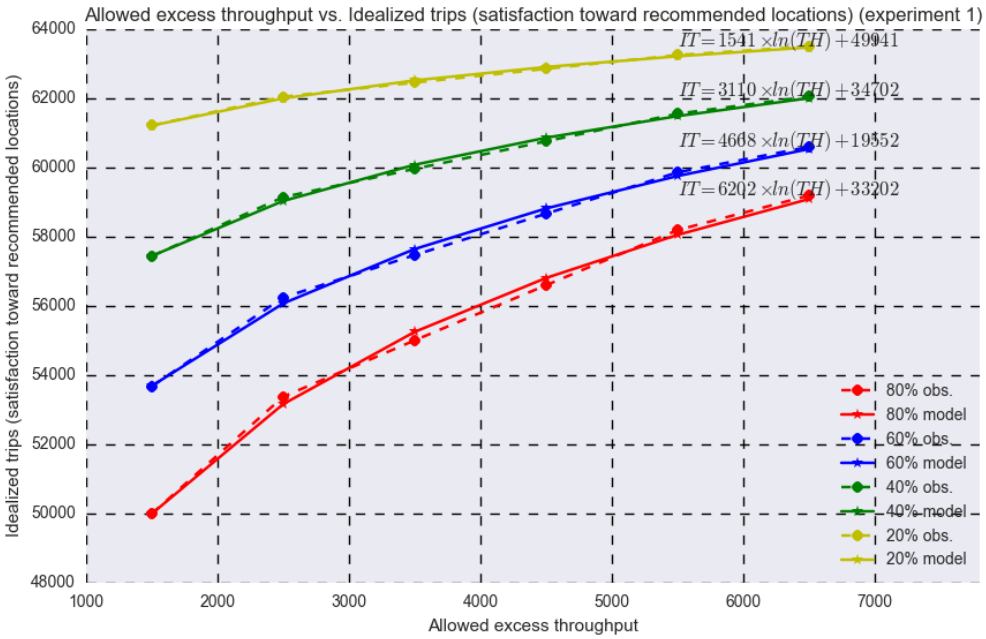

Figure 4-13: Setting 1

obs. is the simulation results. *model* is the best line of fits.

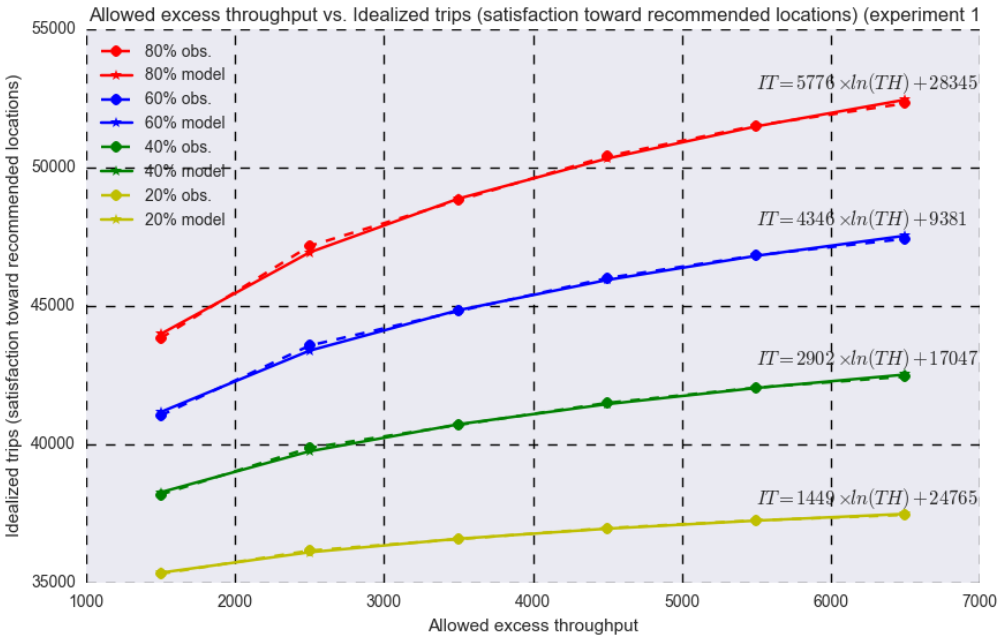

Figure 4-14: Setting 2

obs. is the simulation results. *model* is the best line of fits.

Idealized trips vs. increase in travel time The relationships between idealized trips and increase in travel time under different compliance rate are shown in Figure 4-16. Both of the experiments are modeled by two-degree polymoial regressions. In general, the more idealized trips satisfied, the larger increase in travel time. However, with balanced traffic flow, we could satisfy more idealized trips with smaller increase in travel time. For example, as shown in Figure 4-15, the fourth point (from the left) of 80% observation line and first point of the 60% observation line both generate an increase in travel time by around 7 minute. However, the difference in idealized trips is around 2500. Morever the right-most point of 40% compliance rate and second left-most point of 20% has similar idealized trips with 1 minute difference in average travel time caused by congestion.

As the compliance rate decreases, the slopes decreases in both experiments. The implication is: the higher the compliance rate, the more control we have for the idealized trips and average increased travel time. The impact also becomes more predictable. This indicates the importance in improving marketing strategy.

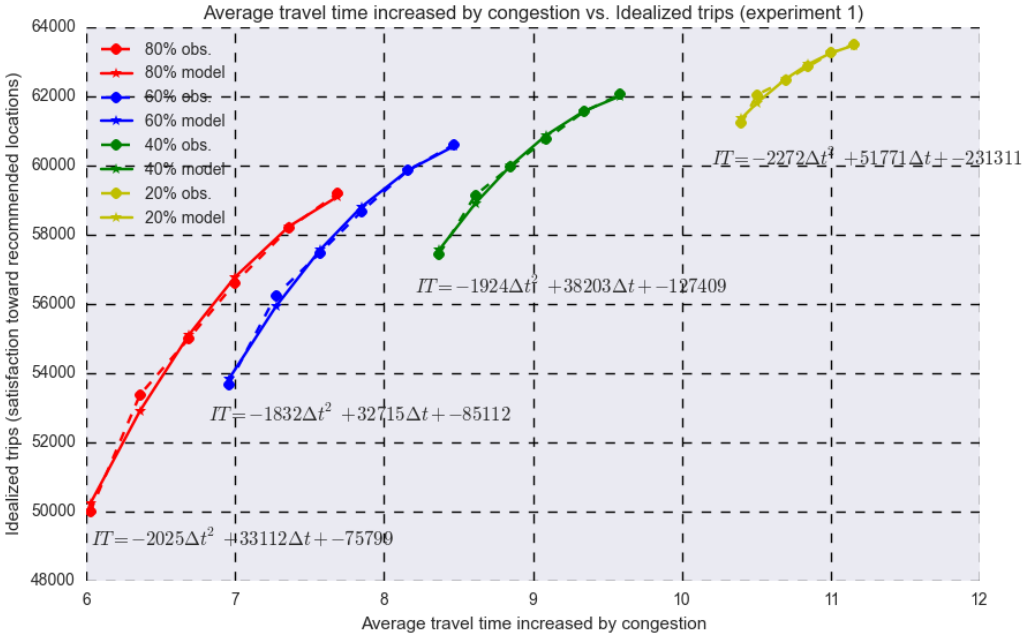

Figure 4-15: Increased travel time vs. increased idealized trips - Setting 1

obs. is the simulation results. *model* is the best line of fits.

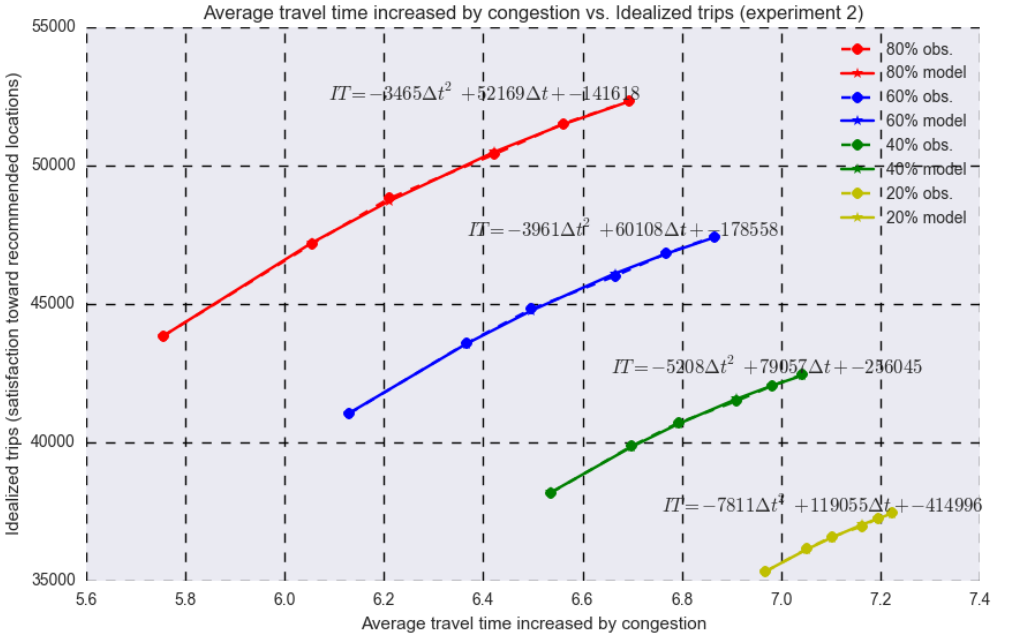

Figure 4-16: Increased travel time vs. increased idealized trips - Setting 2

obs. is the simulation results. *model* is the best line of fits.

In this section, we apply the proposed method in Andorra. The data required for the experiments is described first, including CDR, road netowrk and traffic counts. With the same number of tourists attracted to Andorra, we compare the increase in average travel time caused by congestion. We show that the travel time increased by congestion of status quo with no intervention and preference-based recommendation is 18 minutes and 11.73 minutes respectively. With the proposed method, we can reduce dramatically to 5.6 minutes. We also test the performance of the methods under various compliance rates, showing that the average travel time increased by congestion is 6.17, 6.98, 8.37 and 10.98 minutes with 80%, 60%, 40% and 20% compliance rates.

4.5 Conclusion

In this paper, we introduce a method for making location recommendations for system efficiency by exploring users' choice flexibilities. Though CDR has been widely used in human mobility and transportation research, no study has touched on its application in location recommendations. We demonstrate how to infer location preferences using semantically-poor location records. We then formulate an optimization problem with the objective of maximizing overall idealized trips to satisfy individual preferences. We model the allowed throughput as an important constraint for traffic demand management. To demonstrate the usefulness of the method, we apply it to CDR data from Andorra, a small country relying heavily on tourism, We evaluate and compare the method with the status quo with no intervene and the preference-only recommendation. We show that the method can reduce the average travel time increased by congestion from 18 minutes to 5.6 minutes with 13688 idealized trips and 8.4 minutes with 20807 idealized trips both under full compliance rate.

We perform multiple simulations to understand the interplay between idealized trips and increased travel time caused by congestion. Our method has the flexi-

bility in location recommendation and traffic demand management enabling service providers and transportation planners to adjust the allowed throughput according to their needs. There is a trade-off between the travel experiences on the road/road performances and location satisfactions. It is up to transportation agencies, tourism departments and the government to decide the trade-offs. In addition, we do experiments on various compliance rates to understand the robustness of the impact of the method.

The contributions of the paper are as follows:

- We propose a spatial-temporal location recommendation algorithm base on longitudinal and comprehensive Call Detail Records, proving the potential of using CDR in large-scale location recommendations.

- We integrate location recommendations and systematic service efficiency (traffic congestion reduction) by exploiting users' choice flexibilities.

- We prove the applicability and effectiveness of the method by implementing it on CDR data collected in Andorra, showing that we could reduce the average travel time caused by congestion during peak hour from 18 minutes to 5.6 minutes.

- We conduct extensive simulations to evaluate the impacts of two key factors, allowed throughput and compliance rates. Different levels of allowed throughput reveal the interplay between traffic congestion and satisfactions regarding the location recommendations. Compliance rates indicate the robustness of the method when different numbers of individuals adopt the recommendations.

4.5.1 Implementations

The proposed method can be well applied in tourism locations or events recommendations. There are some issues in real implementations that we need to solve as part

of the recommendation strategy. Tourism department and transportation authorities are the two agencies that will could use the method to coordinate travelers to achieve social optimum that will benefit the society as a whole. There are two channels that the recommendation can be sent, either via text message or mobile App notifications. With mobile App, the method can be applied after some pre-processing steps with the higher-spatiotemporal-resolution location traces. The recommendation message or notification include both the spatial location but also the time (at day level) to visit.

Take the case of Andorra as an example. There are two markets for the impelmentation of the method. The implementation on the two markets can also be viewed as two different type of applications.

1. Travelers that are interested in visiting Andorra but have not planned to visit Andorra. The main goal for this market is to make recommendations that will attract them to visit Andorra.

2. Travelers that are planned to visit Andorra but their next location (predicted) are not in line with social optimal learned from the proposed method. The main goal for this market is to intervene their traveling by recommending them to visit on a different day, or diverting them to a location that they will be similarly interested in, or both. With this intervention, we can achieve system efficiency.

For the first market, it is important to identify the relevent market. The relevancies can be learned based on different features, such as the type of event, temporal preference (weekend vs. weekday, seasons, and etc.). We first need to identify users that are in line with the relevancies. We then formulate a large-scale optimization problem to make spatial-temporal recommendation. There are two rates that are critical to be estimated in order to understand and simulate the impact of the rec-

ommendation. Whether they will visit Andorra and whether they will comply to the date and location of the recommendation.

For the second market, the next-location prediction method proposed in chapter 3 should be used to decide whether to send the recommendation. If the predicted next location is not in line with the recommended location, the recommendation will be sent. In this way, the individuals that perform ideally are not spammed. An experiment is simulated on the same data used in section 4.4. We found out that under full compliance rate, around 18% of individuals will travel to the same location to be recommended. This percentage is relatively consistent when we very the allowed exceed throughput. Note that for this group, it is important to predict whether the individual will visit on certain day. The prediction of the arrival date is out of the research scope of this thesis. It is identified as a future step.

There are some

4.5.2 Future works

We show one potential application of the method for location recommendations in tourism planning. However, there are other cases where the model is useful and applicable. For example, this method can be applied to the recommendations of social events to city dwellers under service capacity or road congestion constraints. It can also be applied to location-based coupon distributions in shoping malls with the store capacity as the constraint.

A number of limitations of the research presented in this paper provid a basis for further investigation.

1. Real-world applications and human behaviors are complex. Uncertainties arise due to the gap between the providing and the adoption of the recommendations. Though we use compliance rate to capture the adoption of the location recommendations, this does not model the real situations. The actual compli-

ance rate from surveys or real experiments within different market segments will help with more accurate modeling.

2. We focus mainly on pre-visit recommendation. However, secondary congestion may arise as a post-recommendation problem if some travelers follow suggestions while others do not, which is called Information Braess Paradox [89]. The use of real-time data would make it possible for real-time responses and better recommendations..

3. In the proposed method, only two measures are taken into account to measure system efficiency, satisfactions regarding the recommendations and congestion level. However, there are other metrics at the system level that are valuable to be considered, such as environmental impact, etc. To handle multiple objectives in the optimization model and analyze the interplay among them is a significant direction step into.

Chapter 5

Conclusions

This final chapter reviews and summarizes the work presented thus far. Motivated by the great opportunities of urban computing and availability of large-scale data sources, the general research question this thesis asks is: How can we use the power of knowledge mined from Call Detail Records (CDR) to understand mobility patterns and solve the major mobility issues our urban spaces face? This paper explores three areas of the application of Call Detail Records in Transportation and mobility in three directions.

The first project is to understand presence pattern from mobile phone data and infer home/workplaces from CDR. Second, based on the understanding of mobility patterns, this paper explores how to use Recurrent Neural Network in next-locations predictions. The last project aims at using mobile phone data to recommend locations with the goal of traffic congestion alleviation. In the following sessions, Section 5.1 describes the research summaries and contributions for the three sections regarding home/workplace inference, next-location prediction and location recommendation for congestion alleviation. Finally, section 5.2 describes several ways to continue or expand this work in future research.

5.1 Research summary

This paper focuses on three aspects in urban computing in transportation, including mobility pattern mining, mobility prediction and mobility interventions as shown in Figure 1-2. Specifically, the three research questions this thesis addresses include: segmenting and profiling user locations, next-location predictions and location recommendations for service efficiency.

5.1.1 Segmenting and profiling user locations

In this project, we developed a new method to extract behavioral patterns from the noisy CDR data and infer home/workplace with higher accuracy. This project answers the following questions:

1. how to characterize individuals' presence patterns at user locations.

2. whether there are common presence structures across the urban-wide population,

3. how to identify home and workplace from user locations.

We use Normalized Hourly Presences to characterize individual presence patterns at user locations, including the frequencies and variations of presence patterns on weekdays and weekends from the noisy mobile phone data. Principal Component Analysis is used to extract the common behavioral structures at user locations across the population. Clustering techniques is used in discovering the intrinsic groups of user locations and segmenting them into home, workplaces and elsewhere with higher accuracy than the method used in the literatures. We test the method on the CDR data collected by MIT Reality Mining Project and the real-world CDR data in a populous and fast-growing city in China. We show that the feasibility, high accuracy

and scalability. We also show that with the inference rates of 78% and 100%, the method can improve home and workplace location inference accuracies by 81% and 32% respectively over other methods proposed in the literature. With Fuzzy C-means Clustering, we can flexibly trade off the inference rate and accuracies and can also assign confidence levels to the results.

In short, the contributions of the work in segmenting and profiling user locations are threefolds:

1. We propose a feature, Normalized Hourly Presence, to extract presence patterns from CDR-based user locations and extract shared behavioral patterns across the population.

2. We propose a universal method to infer home and workplaces on CDR with demonstrated higher accuracy.

3. The method is applied on the CDR data in a populated city in China, which proved its feasibility and scalability in revealing the behavioral patterns and labeling home/workplaces in real-world applications.

5.1.2 Next-location predictions

In this project, this thesis discusses the application of Recurrent Neural Network in mobility prediction based on the analogy between mobility behaviors and language model. Mobility traces prediction appears to be well-posed as a natural language processing problem. The observations from the case study suggest that RNN can be applied to predict next location that manifest the characteristics of location prediction - specifically sequential, variable number of locations and cell tower interpretations. In short, the contributions of next-location prediction are three-folds:

- We propose a new perspective and prediction mechanism for location predictions by automatically learning the meaning of the cell towers for CDR. Recurrent

neural network is explored in mobility prediction based on the mapping between mobility models and language models.

- We implement the method in Andorra as a case study on sparse CDR at two spatial resolutions. The method clearly outperforms markov model and a baseline model with an improvement of more than 30% accuracies. The accuracies are 67% and 78% at cell tower level and merged cell tower level respectively.

- The context of the cell towers can be inferred using the word-to-vector technique as in language model. Instead of representing cell towers as indices, we use real-valued vector representations. The interpretations of cell towers are proved to be useful in location predictions.

5.1.3 Interventions with location recommendations

In this paper, we introduce a methodology for making location recommendations for system efficiency by exploring users' choice flexibilities. Although CDR has been widely used in human mobility and transportation research, no study has touched on its application in location recommendations. We demonstrate how to infer location preferences using semantically-poor location records. We then formulate an optimization problem with the objective of maximizing satisfactions regarding location recommendations subject to capacity constraints. We model the allowed throughput as an service constraint for traffic demand management. To demonstrate the usefulness of the method, we apply it to CDR data from Andorra, a small country relying heavily on tourism, We evaluate and compare the method with the status quo and the preference-only recommendation. We show that the method can reduce the average travel time increased by congestion from 18 minutes to 5.6 minutes with 13688 idealized trips and 8.4 minutes with 20807 idealized trips under full compliance rate.

We perform multiple simulations to understand the interplay between satisfac-

tions regarding recommendations and travel time caused by congestion. Our method has the flexibility for service providers and transportation planners to adjust the allowed throughput according to their needs for traffic or demand management. There is a trade-off between the travel experiences on the road/road performances and satisfactions regarding recommendations. It is up to transportation agencies, tourism departments and the government to decide the trade-offs. In addition, we do experiments on various compliance rates to understand the robustness and effectiveness of the method under various adoption rates.

The contributions of the paper are as follows:

1. We propose a spatial-temporal location recommendation algorithm base on longitudinal and comprehensive Call Detail Records, proving the potential of using CDR in large-scale location recommendations.

2. We bring up a new recommendation perspective by integrating location recommendations and system efficiency by exploiting users' choice flexibilities.

3. We prove the applicability and effectiveness of the method by implementing it on CDR data collected in Andorra, showing that we could reduce the average travel time caused by congestion during peak hour from 18 minutes to 5.6 minutes under full compliance rate.

4. We conduct many simulations to evaluate the impacts of two key factors, allowed throughput and compliance rates. Different levels of allowed throughput reveal the interplay between traffic congestion and satisfactions regarding the location recommendations. It gives authorities the flexibility in managing the congestion level and satisfaction regarding recommendations. The simulations on compliance rates models the systematic situations when different numbers of individuals adopt the recommendations. We show that the average travel time

increased by congestion is 6.17, 6.98, 8.37 and 10.98 minutes with 80%, 60%, 40% and 20% compliance rates.

5.2 Future work

There are a number of opportunities to continue or extend this research:

Segmenting and profiling user locations Understanding presence patterns at user locations and inferring home/workplaces is a starting point in understanding human mobility. There are further research directions that need to be explored.

1. For future research, the proposed method can be used to infer not only home and workplaces, but also other user locations, such as late night locations or weekend locations. This can be done by increasing the number of clusters to identify more types of user locations for different applications purposes. To combine with other geographical data sources, such as land use data or travel surveys, is also helpful in categorize user locations.

2. Additionally, estimating commuting characteristics is useful and important in real applications, such as commuting distances, departure and arrival times, etc.

3. The method could also be tested on other big data sources in inferring home and workplaces, such as online-social networks check-ins (Flickr, Twitter), bank transactions, etc.

Next-location predictions Future directions to enhance and enrich mobility predictions using RNN include: 1) forecasting sequences of locations; 2) predicting next location with time stamps. One step further building upon next-location is to predict several locations in the future. We can make location predictions several steps further

based on the previous prediction. This fits into the framework of language model well by predicting not only the next word, but the following sequences [61]. Predicting user locations several steps far into the future is helpful for planners and transportation authorities in taking better actions to respond to and manage the system.

At the temporal level, we can use another continuous string describing the timestamps as input to predict the arrival time at the next location. This is helpful in real applications to understand the dynamics of the demand.

Interventions with location recommendations In this project, we show one potential application of the method for location recommendations in tourism planning. However, there are other cases where the model is useful and applicable. For example, this method can be applied to the recommendations of social events to city dwellers under service capacity or road congestion constraints. It can also be applied to location-based coupon distributions in shopping malls with the store capacity as the constraint. A number of limitations of the research presented in this paper provide a basis for further investigation.

1. Real-world applications and human behaviors are complex. Uncertainties arise due to the gap between the providing and the adoption of the recommendations. Though we use compliance rate to capture the adoption of the location recommendations, this does not reflect the real situations. The actual compliance rate from surveys or real experiments within different market segments will help in better modeling.

2. We focus mainly on pre-visit recommendation. However, secondary congestion may arise as a post-recommendation problem if some travellers follow suggestions while others do not. The use of new and real-time data would make it possible for real-time responses and better recommendations.

3. In the proposed method, only two measures are taken into account, satisfactions regarding the recommendations and congestion level. However, there are other metrics at the system level that are valuable to be considered, such as environmental impact, etc. To handle multiple objectives in the optimization model and analyze the interplay among them is a significant direction step into.

4. We demonstrate the effectiveness of the proposed method in improving system efficiency based on simulations in real-world settings. One natural future work is to implement the method in actual situations. The first problem in real application is how and when to distribute the recommendation. Text messages and mobile apps are two potential ways for distribution. Still take the Andorra case as an example. The recommendations can be messaged either on their way to Andorra or several days beforehand. If travelers will be receiving the message on their way to Andorra, we only need to infer compliance rate, whether they will follow the recommendation or not. On the other hand, if we want to send the recommendation several days beforehand, it is critical to identify the targeting and relevant travelers that are interested in visiting during certain events or time periods. Their preferences, time and budget constraints can be inferred from historical visits. However, during actual modeling, in addition to compliance rate, another parameter is whether they will visit Andorra or not. The model becomes more complicated in this setting.

Appendix A

Call Detail Records in Tourism Analysis

A.1 Events analysis

A.1.1 Summer events

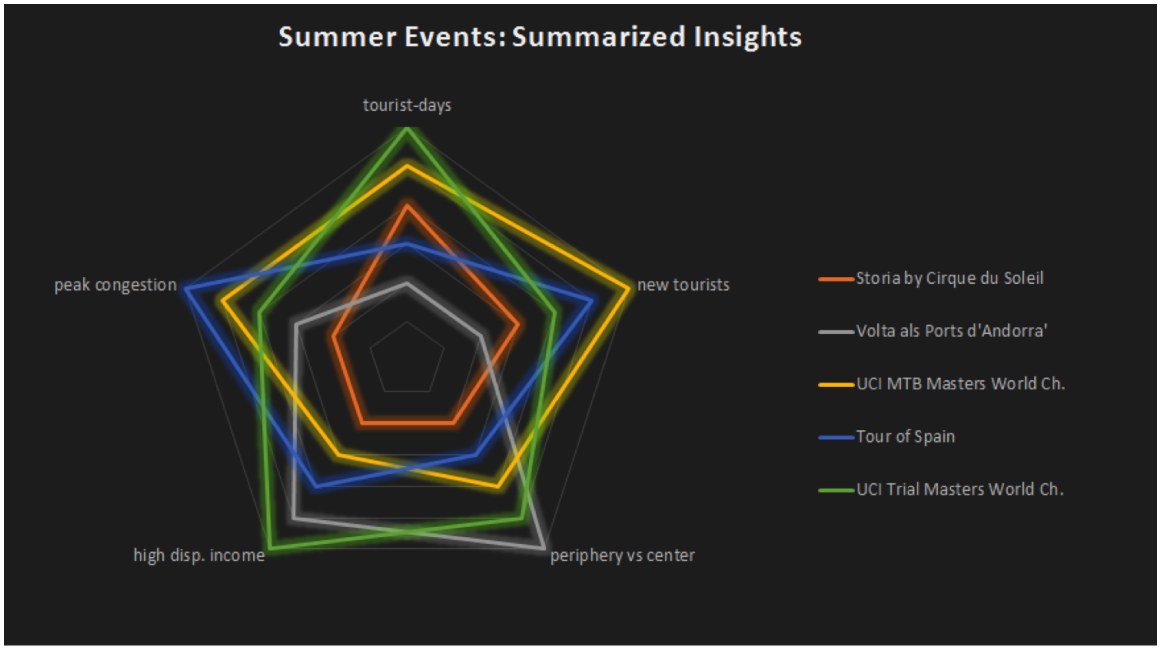

Figure A-1: Summer events analysis

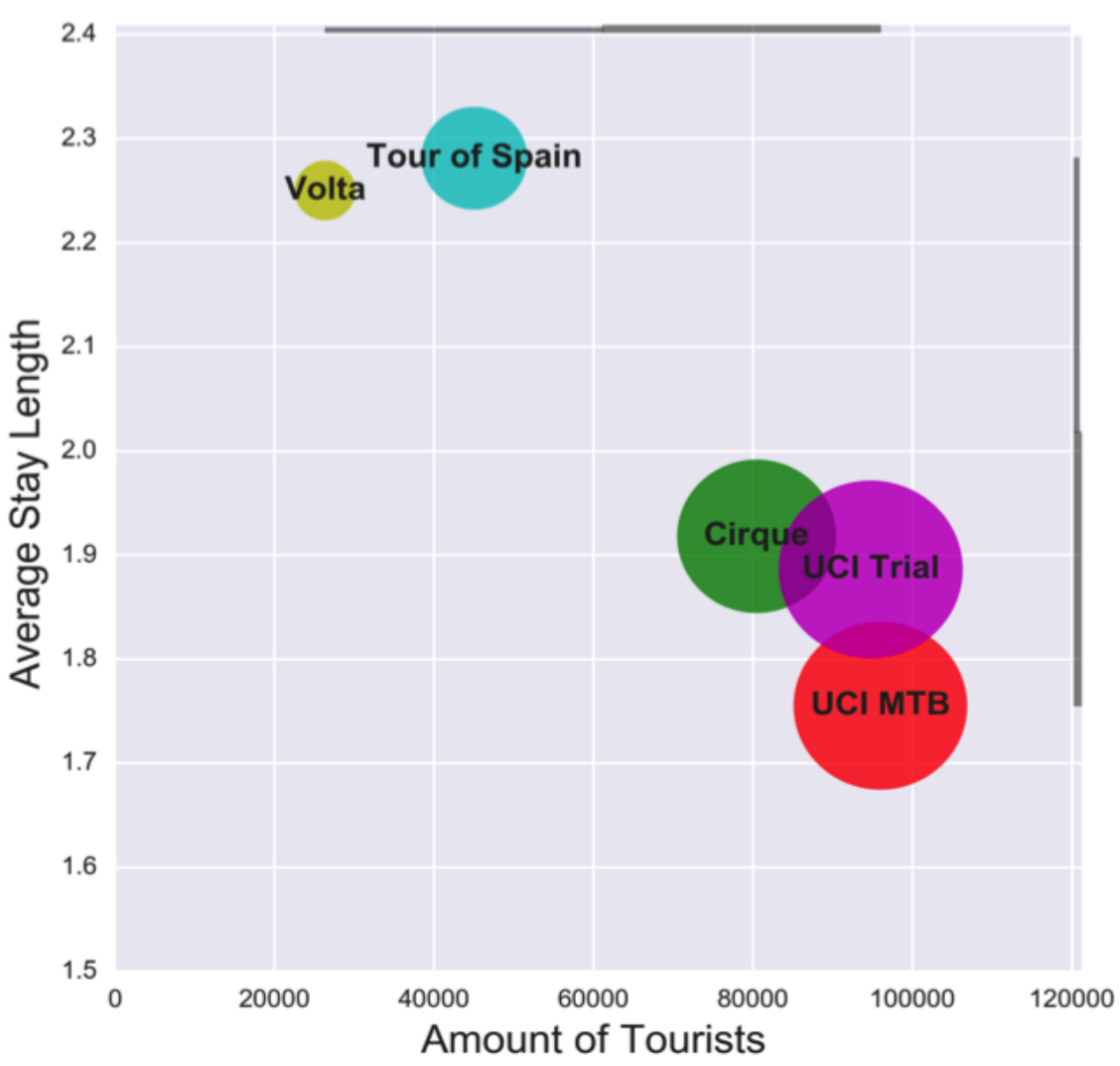

Figure A-2: Summer events analysis: tourists amounts vs. average stay length

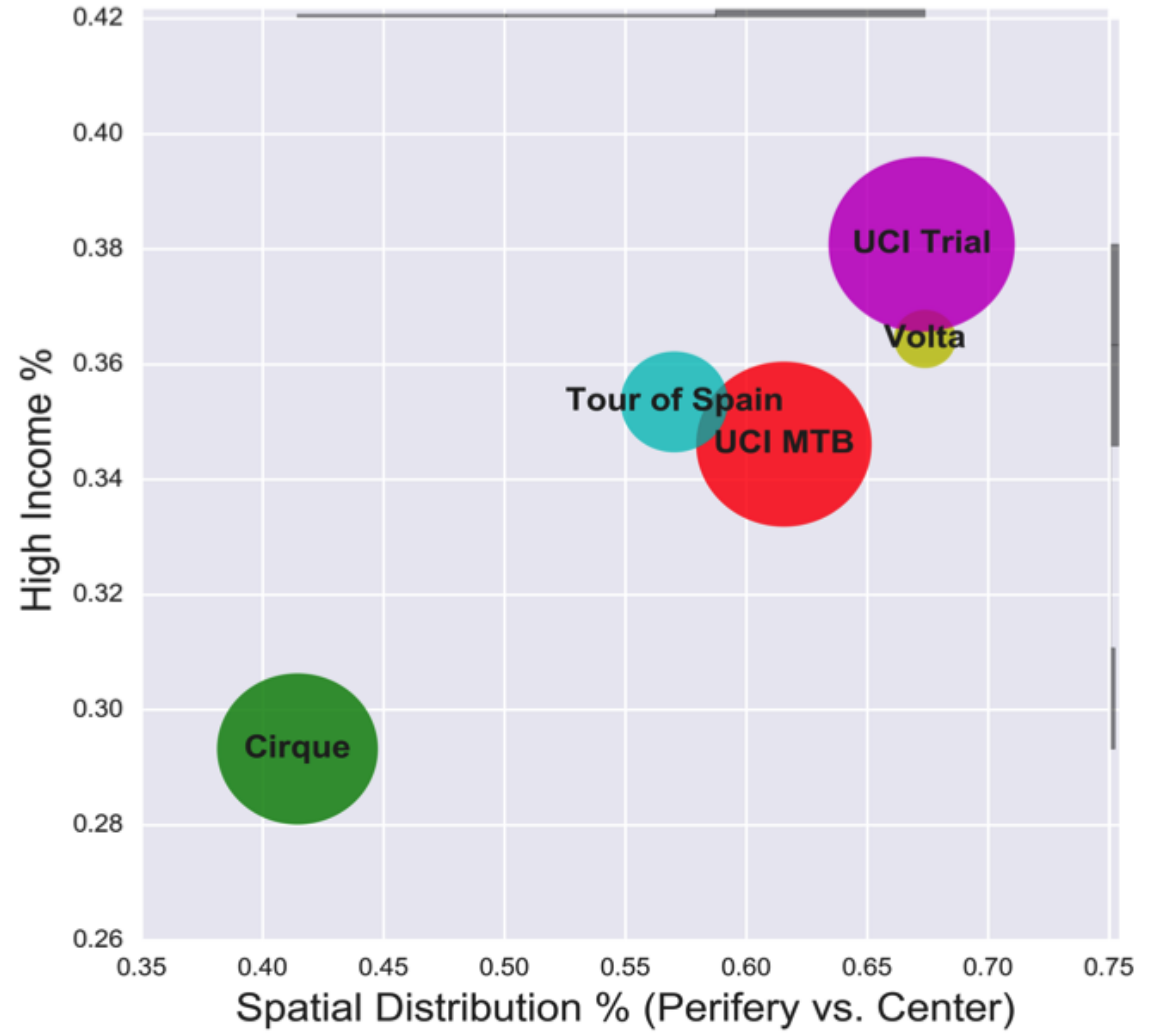

Figure A-3: Summer events analysis: spatial distribution vs. income level

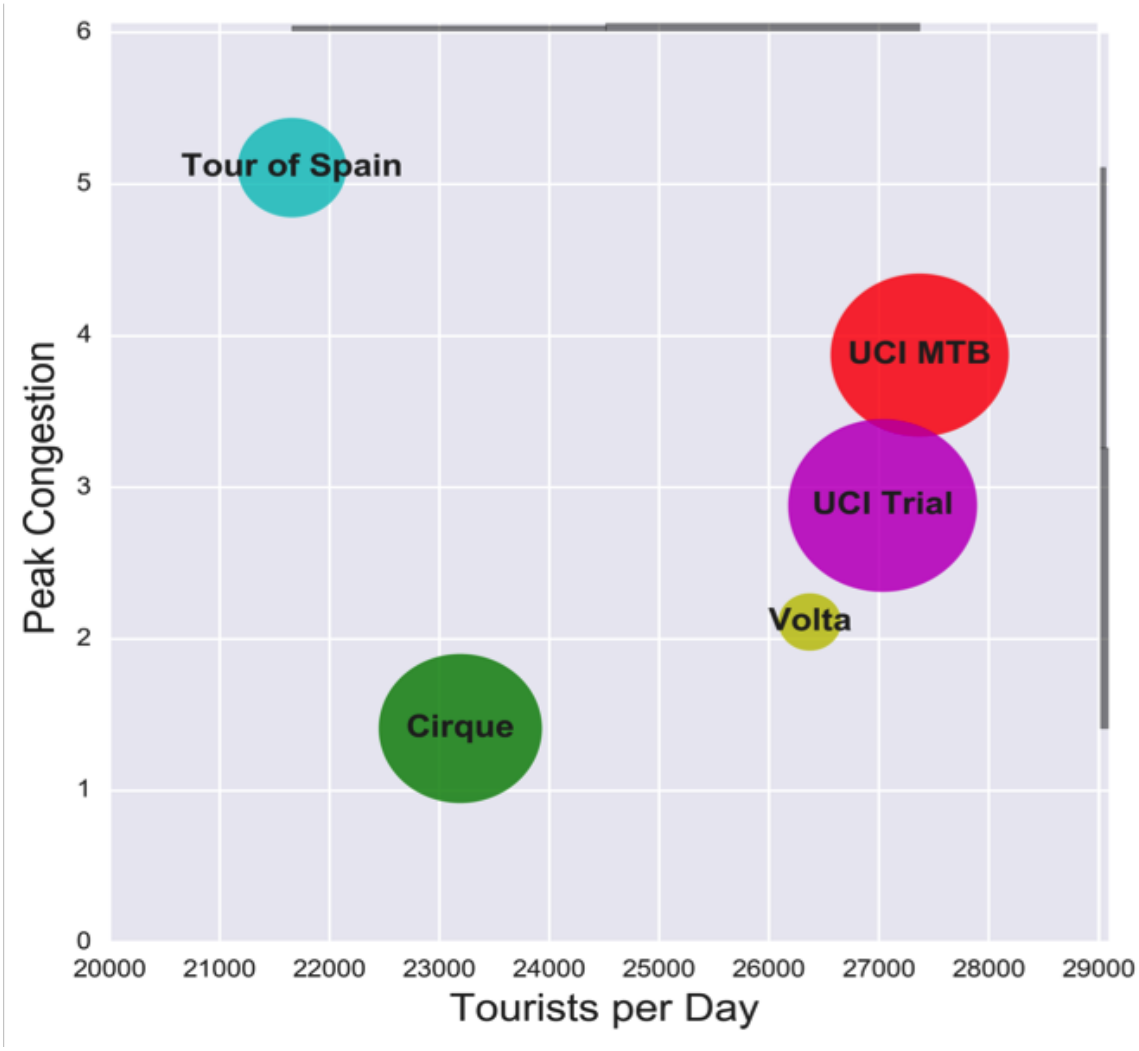

Figure A-4: Summer events analysis: tourists per day vs. peak congestion

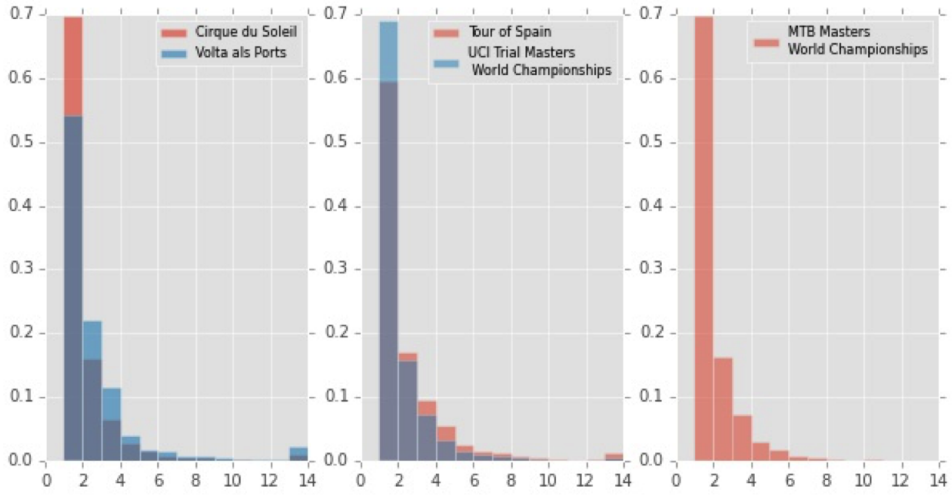

Figure A-5: Histogram of stay length

A.1.2 Winter events

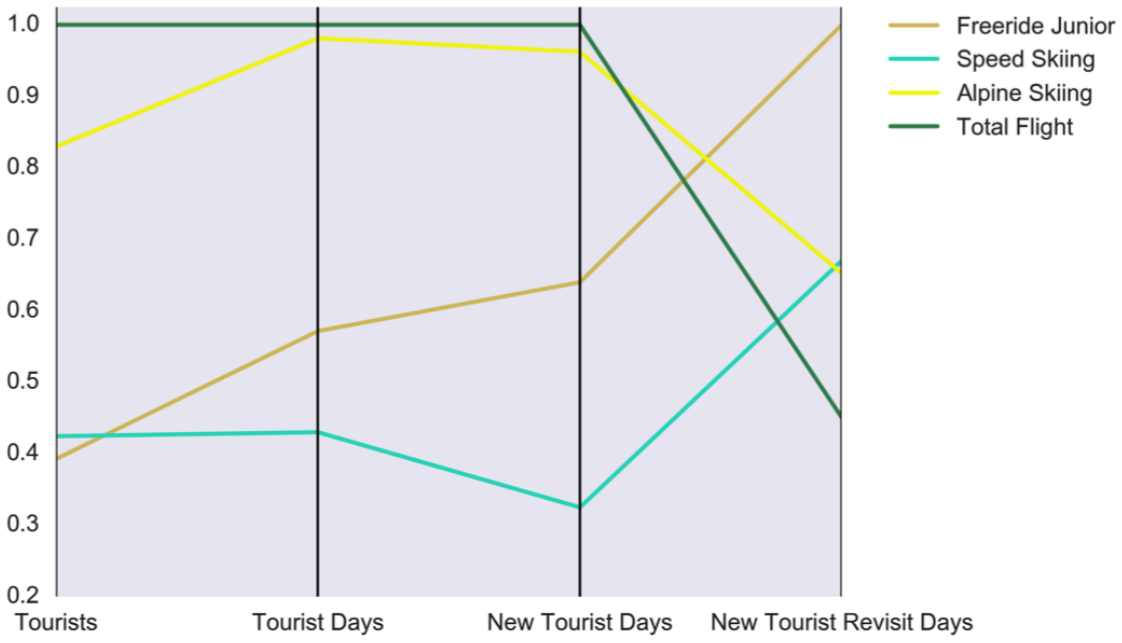

Figure A-6: Winter events analysis

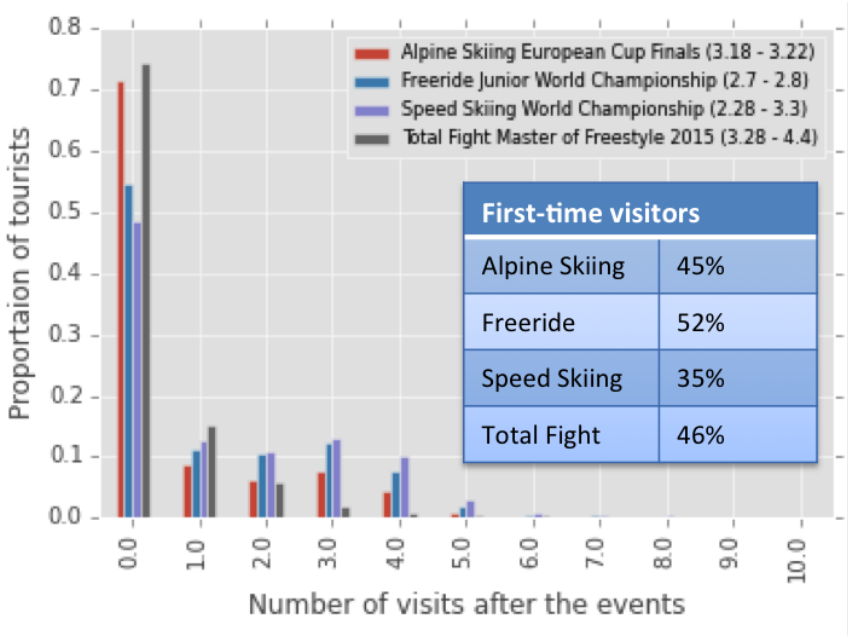

Figure A-7: Winter events analysis: re-visits and first-time visitors

A.2 Tourists analysis

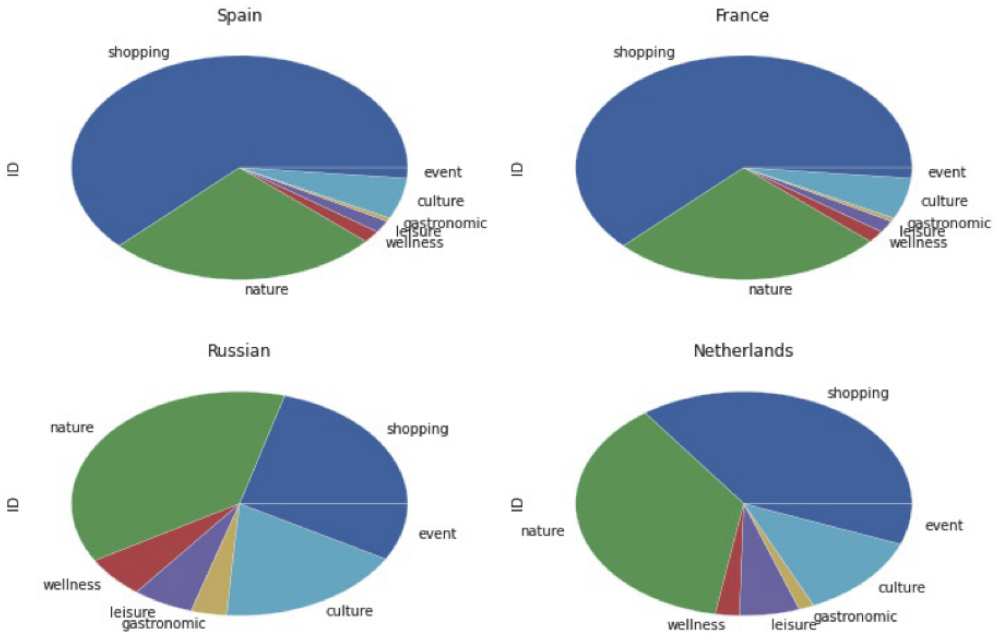

Figure A-8: Tourists intersts vs. nation

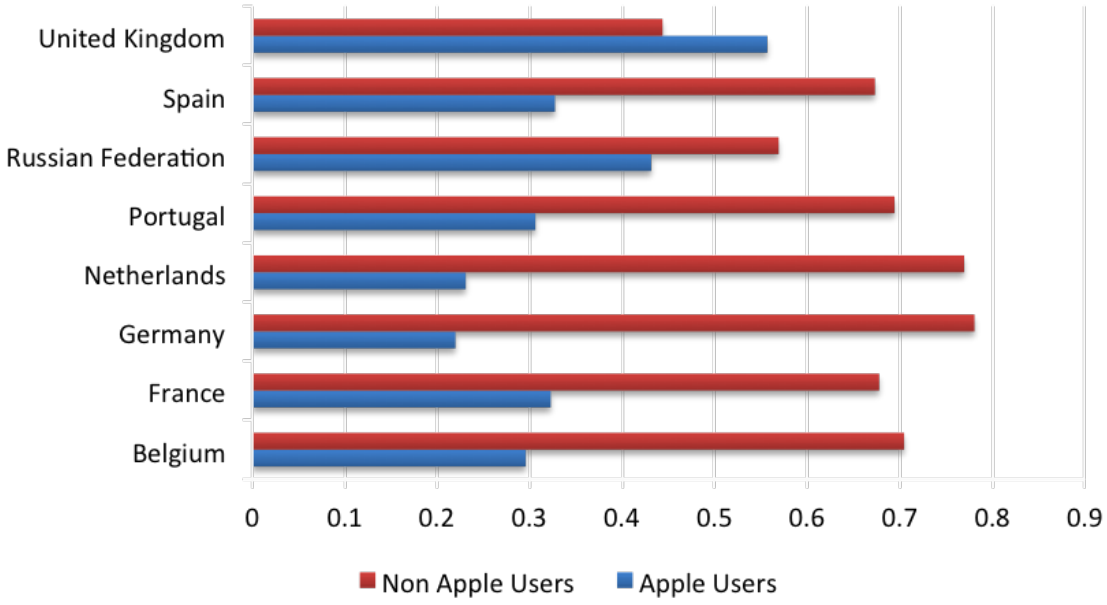

Figure A-9: Nation vs. phone type

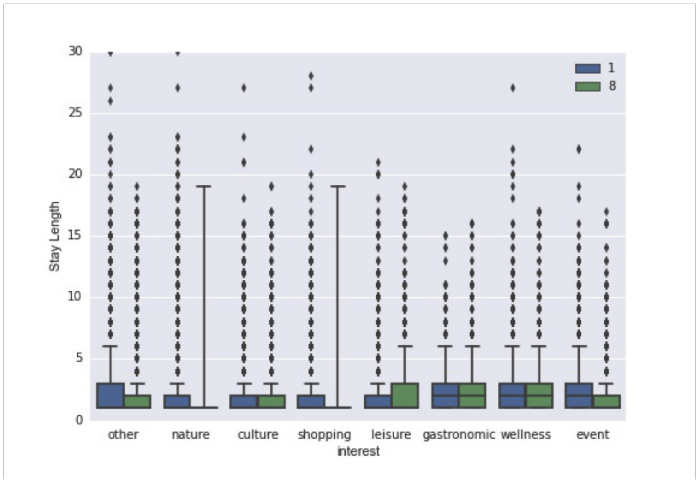

Figure A-10: Tourists interests vs. stay length

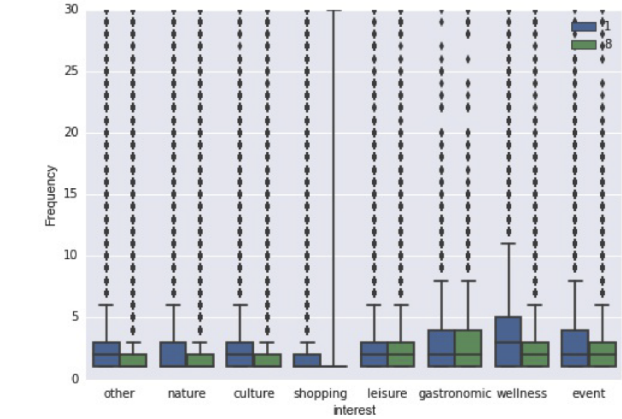

Figure A-11: Tourists interests vs. frequency

Bibliography

[1] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[2] Tim Kindberg, Matthew Chalmers, and Eric Paulos. Guest editors' introduction: Urban computing. *IEEE Pervasive Computing*, (3):18–20, 2007.

[3] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM, 2013.

[4] Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012.

[5] Fahad Alhasoun, May Alhazzani, and Marta C González. City scale next place prediction from sparse data through similar strangers.

[6] João Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. Where will you go? mobile data mining for next place prediction. In *Data Warehousing and Knowledge Discovery*, pages 146–158. Springer, 2013.

[7] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3, 2013.

[8] Jan Petzold, Faruk Bagci, Wolfgang Trumler, and Theo Ungerer. Comparison of different methods for next location prediction. In *Euro-Par 2006 Parallel Processing*, pages 909–918. Springer, 2006.

[9] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

[10] Akinori Asahara, Kishiko Maruyama, Akiko Sato, and Kouichi Seto. Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of*

the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, pages 25–33. ACM, 2011.

[11] 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.

[12] 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.

[13] Wesley Mathew, Ruben Raposo, and Bruno Martins. Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 911–918. ACM, 2012.

[14] Manlio De Domenico, Antonio Lima, Marta C González, and Alex Arenas. Personalized routing for multitudes in smart cities. *EPJ Data Science*, 4(1):1–11, 2015.

[15] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 2015.

[16] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1–55, 2015.

[17] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. 2014.

[18] 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.

[19] Mi Diao, Yi Zhu, Joseph Ferreira, and Carlo Ratti. Inferring individual daily activities from mobile phone traces: A boston example. *Environment and Planning B: Planning and Design*, 2015.

[20] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20, November 2014.

[21] Shan Jiang, Joseph Ferreira Jr, and Marta C González. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore.

[22] Jameson L Toole, Serdar Colak, Fahad Alhasoun, Alexandre Evsukoff, and Marta C Gonzalez. The path most travelled: mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636*, 2014.

[23] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. Towards a comparative science of cities: using mobile traffic records in new york, london, and hong kong. In *Computational approaches for urban environments*, pages 363–387. Springer, 2015.

[24] Iva Bojic, Emanuele Massaro, Alexander Belyi, Stanislav Sobolevsky, and Carlo Ratti. Choosing the right home location definition method for the given dataset. *CoRR*, abs/1510.03715, 2015.

[25] Rein Ahas, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[26] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-time urban monitoring using cell phones: A case study in rome. *Intelligent Transportation Systems, IEEE Transactions on*, 12(1):141–151, 2011.

[27] Lishan Sun, Liya Yao, Shuwei Wang, Jing Qiao, and Jian Rong. Properties analysis on travel intensity of land use patterns. *Mathematical Problems in Engineering*, 2014, 2014.

[28] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[29] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[30] Francesco Calabrese, Jonathan Reades, and Carlo Ratti. Eigenplaces: segmenting space through digital signatures. *Pervasive Computing, IEEE*, 9(1):78–84, 2010.

[31] Shan Jiang, Joseph Ferreira, and Marta C González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.

[32] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[33] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[34] Ying Zhang. User mobility from the view of cellular data networks. In *INFOCOM, 2014 Proceedings IEEE*, pages 1348–1356. IEEE, 2014.

[35] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in peopleâĂŹs lives from cellular network data. In *Pervasive computing*, pages 133–151. Springer, 2011.

[36] David L Davies and Donald W Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[37] M Kathleen Kerr and Gary A Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16):8961–8965, 2001.

[38] Anil K Jain and JV Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987.

[39] JE McKenna. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. *Environmental Modelling & Software*, 18(3):205–220, 2003.

[40] C Lee Giles, Steve Lawrence, and Ah Chung Tsoi. Noisy time series prediction using recurrent neural networks and grammatical inference. *Machine learning*, 44(1-2):161–183, 2001.

[41] John A Bullinaria. Recurrent neural networks.

[42] Wim De Mulder, Steven Bethard, and Marie-Francine Moens. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1):61–98, 2015.

[43] Tomáš Mikolov. Recurrent neural network based language model.

[44] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Q Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 957–966, 2015.

[45] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013.

[46] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[47] FranÃğois Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[48] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[49] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. Using recurrent neural networks for slot filling in spoken language understanding. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(3):530–539, 2015.

[50] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[51] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994.

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[53] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. 1996.

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

[55] Daniel Austin, Robin M Cross, Tamara Hayes, and Jeffrey Kaye. Regularity and predictability of human mobility in personal space. *PloS one*, 9(2):e90256, 2014.

[56] Hsun-Ping Hsieh, Cheng-Te Li, and Xiaoqing Gao. T-gram: A time-aware language model to predict human mobility. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[57] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

[58] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 285–290. IEEE, 2014.

[59] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[60] Xie Chen, Yongqiang Wang, Xunying Liu, Mark JF Gales, and Philip C Woodland. Efficient gpu-based training of recurrent neural network language models using spliced sentence bunch. 2014.

[61] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[62] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20, 2014.

[63] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[64] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[65] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[66] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter Collaborative Filtering Recommender Systems, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[67] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[68] Sheetal Girase, Debajyoti Mukhopadhyay, et al. Role of matrix factorization model in collaborative filtering algorithm: A survey. *arXiv preprint arXiv:1503.07475*, 2015.

[69] Betim Berjani and Thorsten Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, page 4. ACM, 2011.

[70] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 374–383. ACM, 2013.

[71] Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 433–442, New York, NY, USA, 2015. ACM.

[72] Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee. Clr: a collaborative location recommendation framework based on co-clustering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 305–314. ACM, 2011.

[73] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 831–840. ACM, 2014.

[74] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 325–334. ACM, 2011.

[75] Jia-Dong Zhang and Chi-Yin Chow. igslr: Personalized geo-social location recommendation: A kernel density estimation approach. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 334–343, New York, NY, USA, 2013. ACM.

[76] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 363–372. ACM, 2013.

[77] Chen Cheng, Haiqin Yang, Irwin King, and Michael R Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[78] Manh Cuong Pham, Yiwei Cao, Ralf Klamma, and Matthias Jarke. A clustering approach for collaborative filtering recommendation using social network analysis. *J. UCS*, 2011.

[79] Huiji Gao, Jiliang Tang, and Huan Liu. Personalized location recommendation on location-based social networks. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 399–400, New York, NY, USA, 2014. ACM.

[80] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. Recommendations in location-based social networks: a survey. *GeoInformatica*, 19(3):525–565, 2015.

[81] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 971–976, Dec 2010.

[82] 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.

[83] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[84] 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.

[85] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.

[86] Shan Jiang, Joseph Ferreira Jr, and Marta C González. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. In *Int. Workshop on Urban Computing*, 2015.

[87] Hoteleria i turisme. *original*, 2015.

[88] Moshe Ben-Akiva and Michel Bierlaire. Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pages 5–33. Springer, 1999.

[89] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Informational braess' paradox: The effect of information on traffic congestion. *arXiv preprint arXiv:1601.02039*, 2016.