Doctoral School of Computer Science
Department of Computer Science

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science
(XXIX Cycle)

Human Mobility in Urban Space

DOCTORAL DISSERTATION OF:
Karim Keramat Jahromi

ADVISORS:
Prof. Gian Paolo Rossi
Dr. Sabrina Gaito

DIRECTOR OF DOCTORAL SCHOOL:
Prof. Paolo Boldi

Academic Year 2016/'17

Contents

Abstract

Nowadays we witness a rapid increase of people mobility as the world population has become more interconnected and is relying on faster transportation methods, simplified connections and shorter commuting times. Unveiling and understanding human mobility patterns have become a crucial issue to support decisions and prediction activities when managing the complexity of the today's social organization. The strict connections between human mobility patterns, the planning, deployment and management of a variety of public and commercial services have fueled the rise of a vast research activity. Throughout this work, we are more interested and mainly focusing on urban mobility because here most of the human interactions take place and mobility has the greatest impact on management and optimization of public and commercial services. In this thesis, we provided a general framework for dealing with the modeling importance of locations from a per-user perspective and identified a few novel properties of human mobility. Also through characterizing the transition patterns driving user movement among visited places, we pave the way to propose a new mobility model in urban spaces. Meanwhile relying on the relevance of visited places, we propose a new algorithm for detecting and distinguishing Home and Workplaces. And finally, we suggest a framework for predicting the different aspects of Encounter/Colocation events. By exploiting the weighted Bayesian predictor we could enhance the accuracy of prediction w.r.t. the standard naïve Bayesian and also to some other state-of-the-art predictors.

Chapter 1

Introduction

In recent years we witnessed a rapid increase of people mobility as the world population has become more interconnected and is relying on faster transportation methods, simplified connections and shorter commuting times. Unveiling and understanding human mobility patterns have become a crucial issue to support decisions and prediction activities when managing the complexity of the today's social organization. In fact, the today's mobility of individuals is mainly driven by social and professional needs and it is easy to find strict connections between human mobility patterns and the planning, deployment and management of a variety of public and commercial services. These arguments have fueled the rise of a vast research activity aimed to capture and model human mobility, and to apply the results to a variety of domains, ranging from disease spreading [2], urban planning, and smart/green transportation to network infrastructure [84, 21], economy and marketing [54], and mobile network services [20]. At the same time, to provide network access to a growing mass of mobile individuals, ad hoc and opportunistic wireless network infrastructures have been deployed alongside cellular networks. The design and performance evaluation of these wireless infrastructures together with their protocols require testbed and testing conditions in which the large amount of involved parameters, such as sensible transmission range, limited buffer space for the storage of messages, delivery latency, are estimated and defined under representative data traffic models and realistic movements of the mobile users (i.e. a mobility model).

Although a lot of work has been performed on mobility understanding and modeling, the described interconnection between human mobility, human sociality, network and service access makes this research domain highly active. In fact, the dynamics of our daily life, mainly in urban spaces, the changing mobility

behaviors and a continuously evolving technology bring together a continuous adaptation of mobility models and a deeper understanding of the properties governing our mobility patterns.
These arguments are at the heart of this thesis in which we identify and define a few novel properties of human mobility, we improve the expressiveness of mobility models by introducing a new model accordingly, and we present preliminary results of a novel approach to predict encounter and colocation events among individuals in urban spaces. Throughout this work, we are more interested and mainly focusing on urban mobility because here most of the human interactions take place and mobility has the greatest impact on management and optimization of public and commercial services. Moreover, although mobility has long been studied, most of the current studies and models are proposed and validated on the base of small datasets collecting the interactions with a nearby radio access network (typically, a WiFi network), covering limited geographical areas, such as campuses, locations of conferences, large events and exhibitions, and involving a small number of individuals.
By contrast, this thesis, in addition to other datasets, based on WiFi and GPS, benefits of a large anonymized dataset of Call Detail Records (CDR) provided by one of the 4 Italian mobile operators. The dataset contains a detailed description of mobile phone activities (voice call, text messages, and data traffic) performed in the metropolitan area of Milano in different time periods by almost 1 million mobile subscribers. This dataset enabled us to stretch out our focus to an urban area and is allowing us to extend the validity of our results and models to such a large and highly populated area where most of the daily life of individuals is performed.

1.1 On Properties of Human Mobility

Nowadays the growing popularity of wireless networks, combined with a wide availability of smartphones and personal wireless devices, are enabling a large population of mobile users to access a huge amount of services through the Internet. Most portable devices are carried by people and kept in proximity of their users; as a consequence, tracking them becomes a powerful indicator of the mobility behavior of their owners. To some extent, we can capture mobility patterns of roaming users in statistical terms. This is relevant in attempting to reveal the mobility patterns related to human behavior, so as to achieve a more realistic mobility model and thereby predict movements of users that can be exploited in different domains and applications (e.g. optimizing and improving

network operation).

In practice, the mobility pattern of each individual consists of the sequence of locations s/he has visited. These locations and their correlations represent the core block of any modeling research and any activity aimed at understanding human mobility. Even though visited locations underpin almost all of the works in this field, their features remain largely unknown. This is due mostly to the fact that they have been considered primarily as points in specific areas or in places of social aggregation, without anchoring spatial features to the behavior of each single user. The main objective of our research is to fill this gap by providing a general framework for dealing with modeling locations from a per-user perspective. Also, it paves the way toward letting the semantic interpretation of locations be overlaid on their spatial distribution.

First we introduce the notion of user's Points of Interest (PoIs), along with the methodology to extract them from different types of mobility datasets. Then we provide a metric to measure the importance of PoIs for a person, along with a methodology to classify them in terms of *(i)* Most Visited Points (MVPs), the places that a person visits on a more regular basis, such as home and work locations; *(ii)* Occasionally Visited Points (OVPs), locations of interest for the user but visited just occasionally; and *(iii)* Exceptionally [1] Visited Points (EVPs), which correspond to seldom visited locations. This classification allows us to define a human mobility profile where the number of PoIs per each class, and even the amount of time spent there, are the characterizing attributes. We further study how people move across PoIs and PoI classes, enriching the knowledge derived from classification of mobility. The proposed classification, the PoIs and user's features provide the basis for understanding human behavior and allow to extract the semantics of visited PoIs, [10, 51, 22, 68].

This work supports its findings by extensively validating results on four datasets with different characteristics in terms of spatial and temporal distribution of the visited places. By showing the validity of our approach throughout datasets with sometimes different characteristics, we demonstrate the independence of our results w.r.t. a specific setting.

Some interesting properties about human mobility emerge from our work. In fact, it turns out that people visit many places in their life, but they have a *very small number of preferred places (MVPs)* which are visited daily (e.g., home, workplace), and *a higher, but still limited, number of places of interest (OVPs)* which are visited with a lower frequency. MVPs are PoIs where people spend most of their time, so best representing and characterizing their lives.

[1] We use the adverb 'exceptionally' as a synonym for rarely, seldom.

On this basis, we propose an algorithm to identify home and workplaces which leverage the relevance of a PoI for a specific person and outperforms other algorithms in terms of semantic accuracy.

By analyzing the transition rules between PoIs, we find that, in contrast with commonly accepted assumptions, the decision to move between two places is not taken on the sole basis of the geographical distance, but according to the relevance individuals ascribe to them. Also, we show that the transition rule based on relevance follows somehow the same distribution law independently of the mobility scenario.

Summarizing the essential contributions of our mobility framework in chapter 3, we can say that it consists of:

- a novel per-user mobility analysis that highlights the following key properties:
  - people visit regularly just a few places where they spend most of their time;
  - HOME and WORKPLACE are in the set of the few places most visited, and, as such, the relevance ratio is a fundamental feature for their semantic identification;
- a classification of visited locations (PoIs) that enables the above-mentioned analysis;
- a classification of users, based on how people move across PoIs and PoIs classes, derived from our mobility analysis;
- a semantic understanding of human behavior based on our mobility analysis;
- a thorough experimental validation on datasets with different properties.

We argue that to produce more realistic mobility traces, a mobility model needs to consider *i)* the new classifications introduced herein, and *ii)* their different classes, their relationships and transition laws among them. Our results could impact a number of areas, for example:
individual mobility characterization; human mobility modeling (since mobility can be described in terms of regular movement among MVPs and OVPs and extemporarily EVPs); localization, (for purposes of predicting the probability that people are in MVPs); social interaction studies and data offloading (as people tend to meet more frequently with people who share the MVPs).

Chapter 3 is based on the following publication:
• M. Papandrea, K. Keramat Jahromi, M. Zignani, S. Gaito, and G.P.Rossi. On properties of human mobility. *Computer Communications, Elsevier 2016.*

1.2 Mobility Modeling in Urban Space

Understanding the rules that govern human mobility is a crux of many studies in multidisciplinary fields, such as urban planning, traffic forecasting and the spreading of biological and computer viruses [28, 71, 81]. Human mobility also determines the formation of social aggregations, so that its awareness is prominent for understanding how specific social networks form and grow [71]. It has also come to the forefront of studies in mobile networks because it is at the core of the decision about the next hop (by predicting the next opportunity) in the design of routing protocols for opportunistic networks [38].

Radio technologies, such as WiFi, Bluetooth, and GPS have been widely adopted as position sources because they offer a simple solution for precisely detecting the location of an individual. Consequently, mobility patterns and features extracted from these datasets have been widely leveraged to design several data-driven models (e.g., [73, 49, 30]). The main limitations of most of these studies and models revolve around the facts that they have been proposed and validated in limited geographical areas, such as university campuses and major event venues, and that they are built upon experiments involving a small number of individuals within a specific area. The recent and massive growth of studies and business models brought together by the quest for new applications and services for smart cities puts the research community in the urgent need to scale up current mobility understanding and modeling so as to involve a much larger number of individuals on an urban, metropolitan scale. This sparks concerns about the feasibility of achieving the mobility requirements posed by smart cities by simply exploiting mobility models derived from proximity detectors in small areas (such as WiFi and Bluetooth).

On the other hand, although most available mobility models share the fact that people move from place to place, what is still unexplored is the characterization of places on a per-user basis. Places are usually modeled as a collection of indistinguishable geographical points [68, 49, 55, 44, 59, 35, 14], having the same importance for all persons. A few notable exceptions are provided by works where home and workplaces are considered [44, 34, 70].

In this thesis, starting from the result of a study about the relevance that each place has for a particular person, we investigate how to simulate human

mobility in a metropolitan area. We do so by exploiting the notion of the relevance that a PoI plays in the mobility patterns of an individual. To this end, we leverage large datasets of Call Detail Records (CDR) containing voice-call, SMS(text messages) and the Internet activities of nearly one million users of a mobile operator in the city of Milan. The datasets have a few important properties about people and places which enable us to study mobility from an urban mobility perspective. Firstly, the observed population represents the real mass and variety of people living in modern cities. Secondly, and unlike other similar datasets, the considered cell towers are regularly spread over the whole metropolitan area and cover a smaller area, less than 200 meters of radius, w.r.t. other datasets in the literature, [67]. For instance in [28] the average service area of each tower was approximately 3 km^2, and just around 30% of the towers covered an area of 1 km^2 or less. The latter feature allows us to reproduce movements of people within a city and to fully characterize places due to the high spatial resolution of the data.

According to our view, PoIs are at the heart of all mobility patterns in an urban space and their relevance attribute is assigned on a per-user basis. To validate whether or not the relevance approach can be scaled down to small environments, we perform the same analysis on a widely used GPS and a WiFi dataset.

In chapter 5 of this thesis we take a first step in the direction of the designing a mobility model that meets the behavioral and scale requirements of modern smart cities. We envision a smart city as a collection of places, each representing a Point of Interest (PoI) with a specific value for single individuals and for a set of them. PoIs are places where the individuals' home and work locations, or the city's resources and services, are located. People are used to going back and forth between different PoIs during their urban life experience on the basis their habits, commitments and social behavior. Consequently, each individual has his/her own mobility footprint, while a few of them share similar mobility patterns. By simulating the mobility of individuals across metropolitan locations, we will be able to properly describe human mobility and social behavior in urban spaces and to extract all necessary information about how a city's resources and services are accessed.

Relying on extensive analysis of heterogeneous datasets, we show in chapter 5 that the proposed mobility model properly reproduces spatio-temporal and regularity mobility patterns at different geographic scales and for different population settings. Synthetic traces allow to correctly capture real-life features and behaviors, and provide good performances when compared with some other

state-of-the-art mobility models. Besides, through the model we can simulate how people might build relationships loosely by sharing a PoI. By leveraging the concept of colocation, we will show that our system can simulate how often and how long people share one of their PoIs.
Chapter 5 is based on the following publication:
• K. Keramat Jahromi, M. Zignani, S. Gaito, and G.P.Rossi. Simulating human mobility patterns in urban areas. *Simulation Modeling Practice and* Theory, Vol. 62, pp. 137-156, Elsevier 2016.

1.3 Prediction of Encounter and Colocation Events

The previously described activities on mobility modeling and analysis of human mobility patterns are at the heart of a research activity aiming to predict encounter and colocation events when people are moving in an urban area. Forecasting the occurrence of this kind of events among mobile careers can be utilized in delay tolerant and opportunistic networks and may lead to achieve high efficiency in performing routing and data forwarding activities [38, 16, 15]. In a scenario of high dynamics and intermittent radio connectivity the awareness about the approximate location, the time duration of an encounter or colocation event and people involved goes further system implications, paving the way for novel applications in a variety of fields, including commercial ADs, recommendation systems, and mobile social networks. A few prediction algorithms have recently achieved accurate results [88] by combining information about individuals' mobility patterns and social ties and behavior. However, data about human sociality are hard to gather; and due to a tightening of privacy restrictions it will become even harder to obtain them in the future. In our research, described in chapter 6, we use only spatio-temporal mobility information to design a novel algorithm able to predict with high accuracy the next encounter or colocation event. Relative features include place, time duration and people to be met. Specifically, the algorithm learns the patterns of the people's mobility on the basis of respective histories; then it predicts the next encounter or colocation event by performing a weighted feature Bayesian predictor. The approach has been extensively evaluated by means of two large datasets, namely the WiFi and the CDR datasets.

We could summarize the key contributions in chapter 6 as follows:
Firstly, describing the implication and different aspects of predicting encounter and colocation events means predicting the places (PoIs) of occurrence of these events along with relative durations; it also entails forecasting who will be the

users who involved in these events.

Secondly, applying the weighted features Bayesian classifier and predictor significantly enhances the accuracy of prediction w.r.t. the standard naïve Bayesian and also to some other state-of-the-art predictors.

Finally, our approach was validated on large-scale datasets involving several hundred mobile users for the duration of several months. These longitudinal datasets using tracking data of smartphones and portable mobile devices in the real daily life conditions allowed us to study encounter and colocation prediction on a large scale.

Chapter 6 is based on the following publications:

- K.Keramat Jahromi, F.Meneses, and A. Moreira, Impact of ping-pong events on connectivity properties of node encounters. In *WMNC 2014, IEEE.*
- K.Keramat Jahromi, F.Meneses, and A. Moreira, On impact of overlapping access points in detecting node encounters. In *Med-Hoc-Net 2015, IEEE.*
- K. Keramat Jahromi, M. Zignani, S. Gaito, and G.P.Rossi. Encounter events prediction in urban space. UNIMI-Internal Report.

Chapter 2

Datasets and Preprocessing

Nowadays smartphones play a crucial role in capturing various behavioral aspects of the users and their interactions with the variety of the today's radio infrastructures are source of data that may reveal a lot about the spatial, temporal and social dimensions of our everyday life. Logs and tracking records from diverse sources, such as WiFi, Bluetooth, GPS, phone voice and text activities, Internet accesses, have become widely used in many areas, from urban planning, prediction and controlling epidemic infection diseases, design and optimization of wireless and infrastructure-less communication systems [81]. Interestingly, these technologies have different characteristics and the interactions of individuals with them enable the extraction of different behavioral information and provide different levels of accuracy and granularity in both time and space. Each dataset we obtain from keeping track of these interactions contains specific clues for better understanding the various facets of human behavior.

The recent growing interest generated by recreating human mobility patterns gave rise to many experiments mainly relying on a variety of mobility capturing technologies to generate traces (see, for instance, surveys [74], mobile phone traces [28], global positioning system (GPS) [68], WLAN associations and AP logging [30], Bluetooth connections [13], etc).
The work on mobility described in this thesis took advantage of the opportunity of using different datasets from different technologies, namely phone cellular network, WiFi and GPS, each playing an important role in the extraction, understanding and evaluation of human mobility.

In this chapter, we describe the features and characteristics of different datasets we used for mobility analysis, validation and prediction. Also, we describe preprocessing of these datasets for extracting Point of Interest (PoI) and discuss general statistics of large-scale CDR (Call Detail Records) datasets

in details.

2.1 Datasets

Since smartphones are carried by people, they can capture movement patterns and behavioral aspects of their human carriers [47]. These mobile devices enable the development of data collection tools to record various behavioral aspects of users, ranging from how the device is used across different contexts to the analysis of spatial, temporal and social dimensions of users' everyday lives, through sources such as GPS, WiFi, call and SMS logs and Internet accesses.

In this work, we exploit all these data in order to highlight mobility features common to different scenarios and geographical areas. Specifically, we performed our studies over four different datasets. The first two datasets are Call Data Records of smartphones collected by a mobile operator. The third ones is WiFi dataset collected by the log of association/disassociation to the Access Points (APs) at Dartmouth campus. The fourth dataset is mainly composed of trajectories collected by means of GPS technology. The first two datasets have different characteristics in terms of spatial and temporal distribution of the visited places w.r.t the other two databases. We will discuss each dataset in more details in next sections. By showing the validity of our approach under different mobility scenarios, we demonstrate the independence of our results from dataset characteristics.

Despite the widespread diffusion of mobility tracking technologies and mobile services, public datasets capturing the mobility of large population of individuals in an urban environment are scarcely available. Actually, most of the mobility datasets (such as GPS, WiFi, and Bluetooth) at our disposal capture the mobility behavior of a specified category of people, mainly students living on a university campus, or of a number of tracked individuals statistically meaningless. These weaknesses reduce the generalizability of the results achievable from their analysis and consequently they impact on the validity of the mobility models built on this analysis. Roughly speaking, how can we be sure that models based and evaluated on campus datasets only, still hold in a wider scenario as a metropolitan area?

To overcome the aforementioned issues, beyond a WiFi campus-based widely used dataset, we also analyzed two cellular network datasets and a GPS dataset. These collected datasets capture the movement of a heterogeneous population and cover a metropolitan area. Moreover, the amount of involved individuals is consistently higher and, in the case of cellular network datasets, cover mo-

bility behavior of nearly 1 million people. These features make these datasets suitable resources to be investigated for proposing our mobility model in the metropolitan area.

2.1.1 Call Detail Records datasets

In our research, we used two smartphone datasets collected in the metropolitan area of Milan, Italy. This type of dataset, known as Call Detail Records, is collected automatically by the cellular wireless operators for billing purposes. The first dataset includes 17 sampling days (May 1st to 17th, 2013) and covers the whole metropolitan area, i.e. the city of Milan and surrounding districts; the second includes 67 days (March 26th to May 31st, 2012) and is limited to the city area. Both datasets contain records about activities of nearly 1 million operator's subscribers. When a user makes a voice call, sends a text message or accesses the Internet, the user id, the cell id of the handling towers, and also the date and time of established contacts are all recorded. In Figure 2.2 we report a small sample for each kind of recorder activity accompanied by a mobility trace that comes from combining the CDR entries. The location is expressed in terms of cell tower ID and its location-name attribute, e.g. street/square name or city's zone, that represents a coarse grain division of the city region. For 67 days dataset the entire dataset contains more than 69 million phone-call records and 20 million text message records. The advantages of the datasets in use, with respect to other datasets [28, 10, 34, 3, 19], are twofold: first, they cover the highly populated area of a big city and, second, they offer the chance to leverage the Internet access data for purposes of mobility pattern analysis [4]. In particular, the billing system records an Internet CDR that reports the position of the user every 10 Mb of traffic data (upload and download) and also every day at midnight. In general, in fact, CDRs may suffer of some spatial and temporal limitations. Spatially, in CDRs the location accuracy depends on the coverage area of cell towers, which varies from a few hundred meters in urban areas to a few kilometers in rural areas. Moreover, as a CDR is generated only when a phone activity is issued, the location of individuals is recorded at the pace given by their phone activities. Both these general limitations are highly mitigated by the datasets we adopted throughout this work. In fact, the analysis of the cell towers involved in our dataset show that each tower has a 200 m coverage radius on average.

Figure 2.1 depicts that the median radius value is of some 120 m in the inner circle of the city (within 3 km from the center)[67].

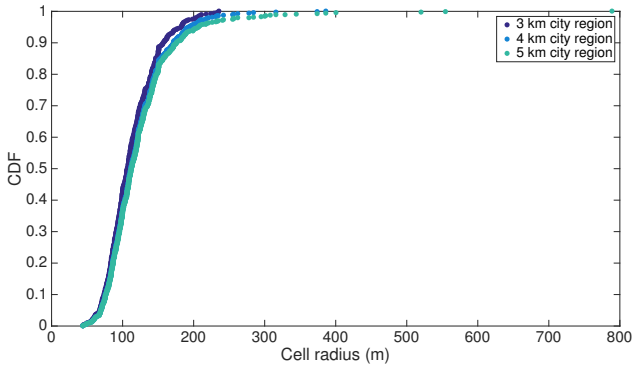


Figure 2.1. *The empirical distribution of the radius of the cell towers in the Milan area.*

Temporally, our analysis benefits of the CDRs associated to the Internet traffic. Indeed, this traffic, with continuous up- and download of data for apps' updates and notifications, provide a fine-grained method to track human mobility. Gonzalez *et al.* [28] showed that trajectories from CDRs are not far from movement recorded by the cellular network infrastructure every two hours.
From here on, we denote the 17-days dataset as CDR-17 and the 67-days one as CDR-67.

2.1.2 WiFi Dataset

We also used a widely adopted WiFi dataset collected through Access Points (APs) at Dartmouth university campus [45].

The Dartmouth WiFi dataset has been used for analyses and sometimes for validation of proposed mobility models, [57, 43, 44]. Whenever a mobile device associates or disassociates with an AP, a log message is recorded. Each record contains a timestamp in seconds, the MAC addresses of the AP and of the mobile device, the Access Session Time in seconds, and the Access Session Status (Start - attach, or Stop - detach). The Dartmouth WiFi dataset with for 4 months duration, from January 3rd to April 30th, 2004, and contains mobility patterns of 17414 anonymized mobile users and 1292 APs. The WiFi dataset has a high temporal continuity and resolution campared with CDR datasets, and w.r.t. GPS, it is more localized in indoor places, although geographically confined to limited areas such as campus or university environment. This WiFi dataset does not include information on the geographical coordinates of the APs

CALL RECORDS
"source","destination","date","time","start_cell","end_cell","dir","duration"
574864,574865,"2012-03-27","13:36:54",47615,47615,"O",0
574864,574867,"2012-03-27","13:55:59",15824,15825,"O",46
574870,574864,"2012-04-02","22:37:41",16677,16677,"I",14

SMS RECORDS
"source","destination","date","time","cell","dir"
1916062,574864,"2012-03-27","21:48:53",16676,"I"
2267867,574864,"2012-03-30","21:59:05",16676,"I"

INTERNET RECORDS
"source","date","time","cell","upload","download"
574864,"2012-03-27","21:35:32",16676,15258,13721
574864,"2012-03-27","21:48:53",16679,76105,78993
574864,"2012-04-02","23:55:45",16677,84589,191681

MOBILITY TRACE
"source","date","time","cell"
574864,"2012-03-27","13:36:54",47615
574864,"2012-03-27","13:55:59",15824
574864,"2012-03-27","21:35:32",16676
574864","2012-03-27","21:48:53",16679
574864,"2012-03-27","21:48:53",16676
574864,"2012-03-30","21:59:05",16676
574864,"2012-04-02","22:37:41",16677
574864,"2012-04-02","23:55:45",16677

Figure 2.2. *The format and a small sample of the call, SMS, and Internet records. The last sample reports a mobility trace that combines the locations given by call, SMS and Internet records associated with a random user. Bold and green entries highlight the problems related to the temporal sparsity of CDR datasets.*

and their spatial distribution.

2.1.3 GPS mobility

Alongside the cellular network datasets, we used a GPS dataset that collects the movement of 178 people in a period of over 4 years (from April 2007 to October 2011). It was released by Microsoft Research Asia under the GeoLife Project. The project provides GPS data taken from a heterogeneous group of people[1] equipped with GPS loggers. This localization technology leads to a spatio-temporal fine granularity of the dataset. More precisely 91% of the GPS trajectories are recorded in a dense representation, i.e. every 1-5 seconds. This allows locating people more precisely. The dataset covers a large portion of the Earth from Europe to the USA to Asia. However, in this work, we limit our analysis to GPS data collected in the Beijing metropolitan area, as our main goal is mostly to characterize and validating our mobility model in an urban

[1] students, government staff, employees from Microsoft and several other companies

area. Overall the dataset contains about 17,000 trajectories which cover about one million kilometers and a total duration of 48,203 hours. Despite their fine spatio-temporal granularity, GPS dataset exhibits a high level of fragmentation, especially regarding features as the effective duration of the trajectories, the data collection period and the number of trajectories per user. Loss of GPS signal inside buildings, energy consumption and voluntary shutdown of the devices are the main reasons of the above issues. Indicatively, more than half of the trajectories span less than one hour, while about 60% of users collected data for less than a month.
In next section, we unify the representation of the mobility datasets captured by the cellular network, GPS and WiFi by expressing users' mobility as a sequence of Points of Interest.

2.2 Preprocessing and general statistics

In this subsection, we describe how we prepare our data to obtain a homogeneous description of people mobility. Each dataset needed to be preprocessed firstly in order to get the required information and secondly to re-conduct all the datasets to a unique representation, i.e. a sequence of temporally annotated Points of Interest (PoIs).

Given the different nature of the employed datasets, the characteristics of a PoI change slightly with respect to the analyzed data. Yet, its main meaning remains the same: namely, it is a place or area which is visited by a user. For the CDR datasets, a PoI is identified by a cell where a user is performing an on-phone activity (e.g., call, SMS, Internet access).

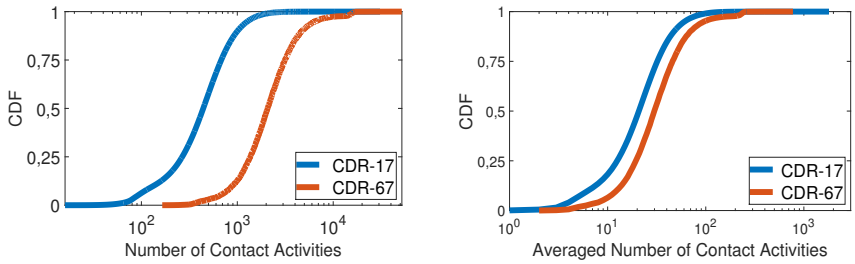
To extract mobility characteristics of individuals we need to have enough CDR samples to study the movement of users. Therefore we select users with at least one activity per day in each dataset and we restrict our analysis to this subset of users. Also, we combine call/SMS and Internet traffic records to get more data about users' positions. Internet access reports the position of user every 10 Mb of traffic (upload and download) and also at midnight. This way, we can consider as Points of Interest for a user, the cells he/she visits, *i.e.* where he/she performs an on-the-phone activity.

The number of users and the number of visited cells covered by each dataset have been indicated in Table 2.1. The results indicate the portion of active users w.r.t. the total number of users by increasing the geographic area.

Figure 2.3a reports the empirical cumulative distribution function (CDF) of the aggregated number of activities (SMS or call). To fit the empirical distribu-

Table 2.1. *The number of users and network cells in the CDR datasets. The last column reports the number of users on which our analysis is based.*

Dataset	Users	Cells	Users with at least one contact activity per day
CDR-17	1,291,416	12,898	543,085
CDR-67	734,149	5,398	17,400



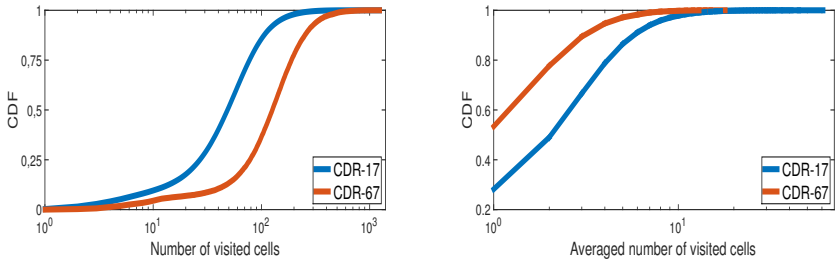
(a) *(b)*

Figure 2.3. *a The empirical distribution of the number of contact activities per user. b The empirical distribution of the averaged number of contact activities per user per day.*

tions, we compare different distributions, whose parameters have been estimated by MLE; and from those that pass the Kolmogorov-Smirnov (KS) goodness-of-fit test[2], we select the model which gets the highest KS statistic. The evaluated distributions are Log-Logistic (3P), Log-Logistic, Pearson, Log-Pearson, Log-Normal, Log-Normal (3P), Weibull (3P), Weibull, Gamma, Log-Gamma, Exponential, Pareto, Levy, Chi-Squared. According to the above method the Log-Logistic (3P) distribution with parameters $\alpha = 2.4584$, $\beta = 1979.1$ and $\gamma = 83.6$ (p-value ≈ 0.2632) obtained the best result for CDR-67 dataset. The Log-logistic (3P) can be considered as heavy-tail distributions family, that imply far more small thing than larger ones. The number of contact activities distribution indicates the majority of people perform an almost small number of contact activities and a minority of people perform a high number of contact activities. Here log-logistic conformed with heavier tail in distribution.

For CDR-17, none of the mentioned distributions passed the test. The average and standard deviation of the number of activities per user in CDR-17

[2]'Data follow the distribution X ' is the null hypothesis. A p-value greater than 0.05 usually indicates that the null hypothesis has not been rejected.



(a) *(b)*

Figure 2.4. *a The empirical distributions of number of distinct visited cells per user. b The empirical distribution of averaged number of distinct visited cells per user per day.*

dataset are circa 532 and 412 contacts; in CDR-67 dataset these values are higher, 2,722 and 2,578 respectively, as the observation period is much longer.

Figure 2.3b shows the CDF of the number of activities per user, averaged over the time frame of a day. We observe that the distribution related to CDR-17 is located above the one related to CDR-67.

We applied the average over the day in order to have comparable values: the measured average corresponds to 25 ($\sigma = 20$) in CDR-17 and 40 ($\sigma = 38$) in CDR-67. In general, by combining the information of the above distributions, the set of users captured by the CDR datasets are quite active and some of them are very active. That represents a good advantage since active users result in more mobility data.

In Figure 2.4a we report the distributions of the number of distinct visited cells per user for each dataset. First of all, almost 90 percent of users have visited fewer than 100 and 260 distinct cells, respectively in CDR-17 and CDR-67 datasets. This implies that most of the people visit a limited number of cells (places), while only a few of them visit a huge number of cells [79]. The CDF of CDR-67 lies under the CDR-17, implying that over a longer period people are more likely to discover and visit new places [28]. The best fitting distributions (from those on the already mentioned list) of the number of distinct visited cells are Log-Normal (3P) with parameters $\sigma = 0.6108$, $\mu = 4.125$, $\gamma = -14.693$ and p-value ≈ 0.646 for CDR-17 and Log-Logistic (3P) with parameters $\alpha = 3.6538$, $\beta = 183.1$ and $\gamma = -57.57$ (p-value ≈ 0.6455) for the CDR-67 dataset. In broader terms, the number of distinct visited cells follows a heavy-tailed distribution.

Figure 2.4b reports the CDF of the number of distinct visited cells per day

Table 2.2. *Summary about the four datasets: cardinality of the datasets before and after the preprocessing, the number of days each dataset spans and the number of distinct visited PoIs.*

Datasets	Number of Users		Number of Days	Number of PoIs
	Before Preprocessing	After Preprocessing		
CDR-17	1,291,416	543,085	17	12,898
CDR-67	734,149	17,400	67	5,398
GPS Trajectories	178	21	20	672
WiFi	17,404	14,082	120	907

and per user. Most people visit on a daily basis a very low number of cells, but there is a long tail accounting for people who visit many cells every day. As the considered mobility area is larger in the 17-day CDR dataset, this dataset captures a higher number of locations visited per day by users. Although our CDR datasets have a higher number of users than the GPS dataset, we should note that CDR datasets are more sporadic in the temporal dimension and coarse in the spatial one w.r.t GPS and WiFi datasets. Even though the coarse spatial granularity of the cellular towers better fits the definition of the region of interest, in this work we mainly use the term Point of Interest to make the notation and the presentation more uniform.

For WiFi dataset, the PoIs correspond to WiFi APs and their coverage areas. In order to extract significant PoIs from the WiFi dataset, we filtered out APs where the user has just passed from by considering only APs with Pause Time > 15 min.

However, for the trajectory dataset, a PoI is identified by a place where the user is either standing still (data gap between consecutive trajectories) or an area within which the user is moving very slowly. To infer user's PoIs from GPS dataset, the clustering-based method is applied as presented in [90]. A PoI extracted from the GPS trajectory is approximated as circle with a radius of 60 m. For more details about the approach and analyzing results refer to [62].

The characteristics of the four datasets after the different preprocessing phases have been summarized in Table 2.2. The following analysis of the mobility behaviors is going to be based on the preprocessed datasets.

The detection of the PoIs allows us to compare the mobility habits in terms of visited places with different type of datasets. After extracting PoIs we can unify the representation of the mobility datasets captured by the cellular network, GPS and WiFi by expressing users' mobility as a temporally annotated sequence of PoIs, i.e. each PoI P for a mobile user u is characterized by the arrival time a and the departure time d. This way the entire history of the movements of

mobile user u is denoted by the temporally annotated sequence $S =< P_i^{(a_i,d_i)} >$, for $i = 1, \ldots, n$ where n is the last visited PoI. egion

Chapter 3

Properties of Human Mobility

The current age of increased people mobility calls for a better understanding of how people move: how many places does an individual commonly visit, what are the semantics of these places, and how do people go from one place to another. In this chapter, we show that the number of places visited by each person (Points of Interest - PoIs) is regulated by some properties that are statistically similar among individuals. Subsequently, we present PoIs classification in terms of their relevance on a per-user basis. In addition to the PoIs relevance, we also investigate the role of the relevance of place that describes the travel rules among PoIs. Most of the existing works on mobility are mainly based on spatial distance. Here we argue that for human mobility the PoIs relevance are rather the major driving factors. With the support of different datasets, this chapter provides an in-depth analysis of PoIs distribution; it also shows that our results hold independently of the nature of the dataset in use. We illustrate that our approach is able to effectively extract a rich set of features describing human mobility and we argue that this can be seminal to novel mobility research.

3.1 Related Work

Spatial mobility patterns have been analyzed in different disciplines, from physics to pervasive computing. Works from the physicists' community focus on concepts from statistical mechanics and thermodynamics. Their main goal is to identify what kind of diffusion process is able to best reproduce human mobility. For these reasons they analyze the displacement and the length of movements, searching for evidence of sub- or super-diffusive processes [89, 71]. On the contrary, works from computer science focus more on human mobility properties,

which can be exploited in the deployment of different services (from opportunistic networks to link prediction in location-based social networks).

In their seminal work Brockmann *et al.* [9] investigated human traveling statistics by analyzing the circulation of banknotes in the United States. Based on a huge dataset of over a million individual displacements, they found that the distribution of the traveling distances decays as a power law, indicating that trajectories of banknotes are similar to Lévy flights. Secondly, they showed that the probability of staying in a confined region (pause time distribution) is characterized by a long tail leading to a sub-diffusive process.

Gonzalez *et al.* [28] also focused on distances covered by people. In particular, they analyzed mobile phone users for a six-month period in a large area. They found that the distribution of the distance between two consecutive calls is well approximated by a truncated power-law. Moreover, each individual tends to return to a few frequented locations with high probability.

Rhee *et al.* [68] were the first to deal with the statistical properties of human mobility using GPS dataset. By analyzing GPS dataset collected on a campus they reported that bursty hot spot sizes play an important role in causing the heavy-tail distribution of distances in the human walk. They show that visit points are clustered and that pause time distribution in hot spots follows a truncated Pareto.

Common properties observed in WLAN datasets in daily life according to [37] are skewed location visiting preferences and time-dependent mobility behavior. Location visiting preference refers to the percentage of time a mobile user spends at a given AP with skewed distribution. The time-dependent mobility refers to the observation that users visit different locations depending on the time of the day. It means that users tend to visit just a few locations, where they spend the majority of their time (Skewed Location Preference) with time dependence periodical re-appearance.

Kim *et al.* [41] used Access Point (AP) log data to extract information about users' movements and pause times. They found that pause time and speed distributions follow a log-normal distribution and that the directions of movement follow the direction of popular roads and walkways on the campus showing asymmetry across 180 degrees.

According to the experiments with small and large datasets [77, 89], the distribution of Pause Time, i.e., the time interval that a user spends at one location, follow Truncated Power Law (TPL) with an exponent and a cut-off. The linear decaying power law head part is likely to reflect common temporal scaling pattern of human mobility such as daily life activity while the exponential

part is more likely relevant to the uncommon pattern of human mobility such as long time travel to overseas. This Truncated Power Law feature of Pause Time is co-existed with temporal heterogeneity in human mobility. It means that people pause short time in the majority of visiting locations while they spend a long time in few of locations.

Authors in [77] have shown that f_k frequency at which a user visits its k^{th} most visited location, follow Zipf's law $f_k \sim k^{-\varepsilon}$ with parameter $\varepsilon \approx 1.2 \pm 0.1$ implied the preference for visiting different areas is skewed. They also have proposed the probability of a user to visit a given location, f times (i.e., visitation frequency) follows $p(f) \sim f^{-(1+\frac{1}{\varepsilon})}$. On the other hand authors in [7] claim that the Zipf's law observed in visitation frequencies distribution is influenced by a recency bias expressed as a tendency to return to recently visited locations.

The Return Time is another important temporal metric, which also has the impact on the property of Pause time. Return Time Probability is the probability of returning to one of the previous visited location at time t. According to [28, 37], the estimated Probability Density Function (PDF) of Return Time indicates peaks at 24 h, 48 h and 72 h that confirms the diurnal and periodical human activities. Also, authors in [89] discuss power law distribution of return time and pause time as a universal property of temporal human mobility because humans visit locations periodically and especially following the diurnal cycle pattern. Most of above works have shown that individuals follow reproducible patterns despite the diversity of their travel history. As the users always return to several of their highly frequented locations such as home and workplace, significant regularity can be identified in their trajectories. Daily and weekly spatio-temporal regularity in the human movement have been studied [32, 29, 8] through analyzing WiFi and CDR datasets, respectively. In this section, we plan to characterize this daily mobility regularity in visiting places from the perspective of an individual user and mine the importance of the visited places per each user.

3.2 Daily Mobility regularity

We know that people do not move randomly; by contrast, their movements are influenced by their needs, commitments, and their social ties. As the users always return to several of their highly frequented visited locations such as home or work, significant regularity can be identified in their trajectories [32, 29, 8, 81]. As a result, mobility patterns show daily regularity and periodicity. These regularities in mobility patterns could be characterized by defining the *Relevance*

Ratio (RR) of the PoI (place) P for a user u as:

$$RR(P,u) = \frac{d_{visit}(P,u)}{d_{total}(u)} \tag{3.1}$$

where $d_{visit}(P,u)$ is the number of days a given place P has been visited by user u and $d_{total}(u)$ is the number of days recorded in the user's dataset [62]. We adopt the day as temporal window since it represents the fundamental period when considering life routine of individuals. The relevance ratio captures how likely an individual will be moving towards a place or return back to it according to his/her tracking history.

The empirical relevance ratio distributions obtained from all datasets in Table 2.2 are shown in Figure 3.1. In CDR and GPS datasets a high number of PoIs are sporadically visited, while a few PoIs are almost daily visited (high value of RR). In the most trustable dataset, the 67 days cellular dataset with 3 months recorded, this pattern is highlighted as there was the time to record the high number of PoIs that users visit once, or very few times by chance for special events.

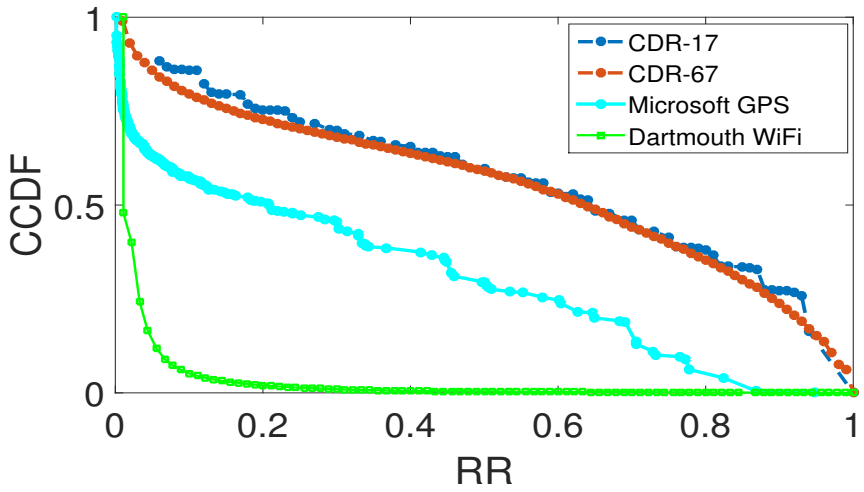


Figure 3.1. *The empirical Relevance Ratio (RR) distributions in CDR-17, CDR-67, GPS and WiFi datasets, aggregated on all users and PoIs.*

Although the nature of datasets and the way they have been collected are different, but conforming a few preferred and recurrent locations characterize the mobility of each user.

3.2.1 PoI Classification by relevance

The above results show how many days people spend at their PoIs, but places play different roles in their lives.

We adopt a classification of PoIs based on the relevance. It is organized into three classes, where each class accounts for places with different importance and semantic values in the user's daily life experience. The proportionality between the place importance for a user and the frequency characterizing the place visit represents the key assumption on which the classification relies. This way relevance ratio enables us to identify:

- **Mostly Visited PoIs (MVP)**: locations most frequently visited by the user. Often they can be associated with the home location and workplace.
- **Occasionally Visited PoIs (OVP)**: locations of interest for the user, but visited just occasionally. Often they correspond to favorite places or meeting points weekly visited.
- **Exceptionally Visited PoIs (EVP)**: rarely visited PoIs. They represent outliers w.r.t. the regular mobility behavior of the user.

The evaluation of the PoIs' relevance allows us a straightforward per-user identification of these three classes, as they will be described in the following section. But simply by examining the aggregated relevance distribution shown in Figure 3.1 we can assign most of the probability distribution to the multitude of EVPs with very low relevance. Meanwhile, the first set of points expresses the few albeit highly relevant MVPs. The central part of the distribution contains OVPs.

3.2.2 Relevance class detection algorithm

Although the described classes of PoIs and their meanings are shared among all users, the relevance class bounds we use to identify them could be different on a per-user basis and cannot be fixed *a priori*. This argument advocates a clustering algorithm that adaptively adjusts according to the single user's mobility pattern. In particular, we adopt an unsupervised approach which groups the PoIs of a single user based on the PoI relevance and maximizes their separability. To this end, we have chosen the k-means algorithm. To avoid the problem related to the initial choice of the centroids, we run 10 replicas of k-means with different initial seeds and choose the partition that minimizes the within-cluster sums of point-to-centroid distances, thus maximizing the separability. We run

k-means with $k = 1, 2, 3$, then we assign to the user the number of relevance classes corresponding to the value of k with the best clustering performance, by choosing the value k which maximizes the silhouette separability. In Figure 3.2, as an example we show the result of the k-means, with $k = 3$, clustering on a sampled user. The EVP class (first box on the left) covers the range from 0.01 to 0.12, the OVP (central box) spans the range from 0.16 to 0.46 and the MVP class (first box on the right) contains only one PoI with relevance 0.82.

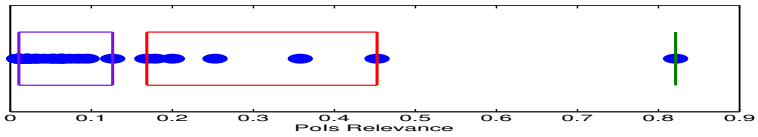


Figure 3.2. *Three classes of relevance in a sampled user*

The adoption of a clustering algorithm for detecting the three relevance classes allows us to adaptively select their bounds and avoid the choice of fixed thresholds. In fact, the application of a clustering algorithm best suits the diverse human mobility patterns and mitigates the spatio-temporal heterogeneity which characterizes the different datasets. However, the clustering of the relevance for each single user could generate overlapping among the classes of different users. For instance, relevance values which belong to the OVP class for a user could correspond to the MVP class for another user.

In this section, we apply the class detection algorithm described above on the PoIs derived from the different datasets and analyze the obtained classes to extract their features.

In GPS dataset the best separability is achieved by $k = 3$ for nearly all users [62]; however, the mobility captured by the CDR and WiFi datasets are more varied and not every user satisfies the above classification. CDR datasets differ from GPS dataset in many respects, as discussed in section 2.1, both meaning and characteristics of PoIs extracted from these datasets are radically different, especially with reference to the relevance classes. First of all, the spatial granularity of PoIs is wider in CDR data than in GPS data. In the former case, an urban PoI coincides with a cell tower and approximates a hexagon with a few hundred meters side. While a PoI is extracted from the GPS trajectory (see section 2.2) it approximates a circle with a radius of 60 m. Consequently, a PoI extracted from a CDR dataset could actually aggregate other PoIs. This would require the finer grain of the GPS to emerge. For instance, a cell-based PoI could aggregate workplace and coffee shop or home and nearby stores. Moreover, the

Table 3.1. *Users' distribution among groups identified by the number of mined relevance classes.*

Group	Percentage of Users			Distinct visited cells/APs		
	CDR-17	CDR-67	WiFi	CDR-17	CDR-67	WiFi
1	25.16%	18.42%	3.82%	11,534	2,509	33
2	46.37%	47.6%	12.41%	11,689	2,845	65
3	26.94%	33.97%	27.91%	11,425	2,643	106

CDR datasets only record the cell where the user is performing a phone activity. As a result, the number of visited PoIs that can be extracted from a CDR dataset is smaller than the one obtained from GPS dataset.

Users with fewer than 3 PoIs have been discarded: nevertheless, they represent only 1.53% and 0.01% of the users in the 67- and 17-day CDR datasets, respectively. For all of the other users, we apply the k-means algorithm, as explained in section 3.2.2. While in the GPS dataset for nearly all users the best separability was achieved by $k = 3$, in the CDR datasets the aggregation of PoIs in broader cells led to different results. For many users, PoIs clusterization according to their relevance achieves better performance when two (k-means with $k = 2$) or one (k-means with $k = 1$) classes are considered. Thus we consider three groups of users, each characterized by the number of relevance classes achieving the best performance in PoIs k-means clustering.

For CDR and WiFi datasets we categorized users into different groups: Group 1 includes users whose PoIs have been grouped into a single class, while Group 2 and Group 3 include users whose visited PoIs can be separated into two and three classes, respectively.

The distribution of users among these groups is reported in Table 3.1. Only for about one-third of users, those belonging to Group 3, it is possible to identify all three classes of PoIs: MVP, OVP, EVP.

The distributions of the relevance in different groups for CDR-17 and CDR-67 datasets have been depicted in Figure 3.3.

In contrary to CDR datasets, around 55% of users in WiFi dataset (with 3 months duration) have visited fewer than 3 PoIs that will be discarded from our analysis. The distributions of the relevance in different groups for WiFi dataset have been depicted in Figure 3.4.

The difference of k-mean algorithm output is mainly due to the spatio-temporal nature of CDR and WiFi datasets. To make the following analysis more uniform across the mobility datasets, we focus on the Group 3, ignoring

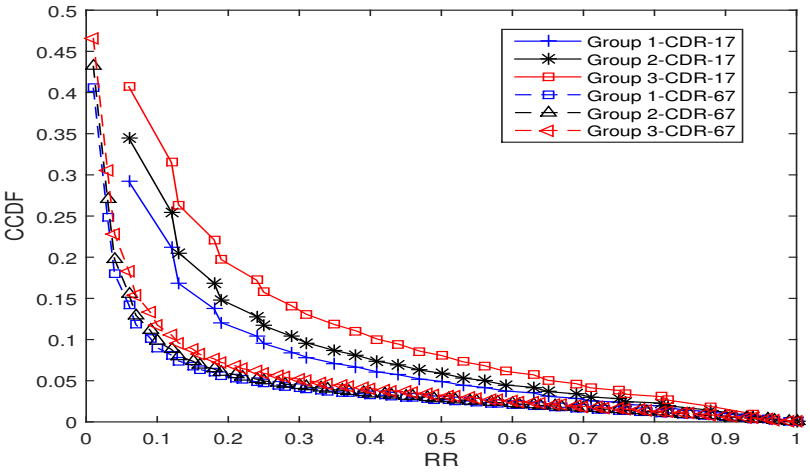


Figure 3.3. *The empirical CCDF distributions of the relevance ratio in different groups for the two CDR datasets.*

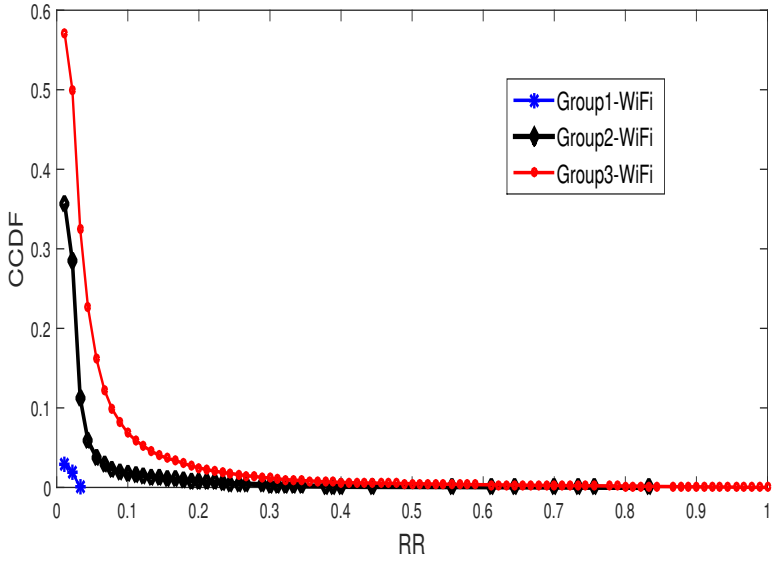


Figure 3.4. *The empirical CCDF distributions of the relevance ratio in different groups for the WiFi dataset.*

users who belong to Groups 1 and 2 that POIs visited by them clustered in One class (Group 1) or Two classes (Group 2), due to lack of enough on-phone activities or sedentary. The percentages of users belonging to Group 3 in the CDR-67, CDR-17, and WiFi Dartmouth datasets are 27%, 34% and 27.91%, respectively.

As we can see in Figure 3.5, the distribution of CDR-67 is located under CDR-17. That implies users in CDR-17 are more regular in visiting PoIs w.r.t. CDR-67. This observation is a consequence of the longer duration of the dataset in CDR-67. According to [28], a longer duration results into a higher number of distinct visited PoIs as also reported in Figure 3.7. So mobile users in the CDR-17 move among a more limited number of PoIs compared to the CDR-67, resulting into a higher regularity. Finally, in the Dartmouth WiFi dataset, although the number of distinct visited PoIs per user is less than in the CDR datasets, we observe that regularity is fallen under the CDR-17 distribution, due to the longer duration w.r.t. CDR-67. In WiFi dataset (with three months duration), we observe a higher number of PoIs regularly visited. In fact regularity in visiting PoIs in a campus environment is higher as a campus is a very limited area where PoIs correspond to precise activities (classes, offices, library) performed regularly in people daily life.

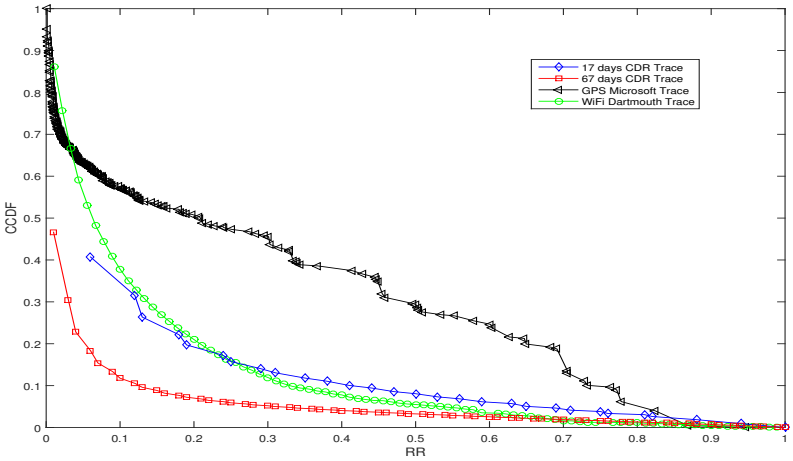
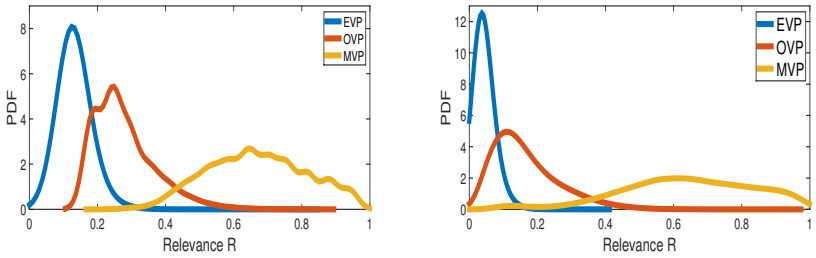


Figure 3.5. *The empirical CCDF Relevance Ratio (RR) distributions in Group 3 of CDR-17, CDR-67, GPS and WiFi datasets.*

In Figures 3.6a and 3.6b we show the distributions of the relevance characterizing MVPs, OVPs and EVPs in Group 3 of CDR-17 and CDR-67 datasets,

respectively. In both CDR datasets, the relevance distributions reveal the high level of separability of the relevance classes. Besides, MVPs relevance is much higher than EVP and OVP ones, accounting for PoIs actually visited very frequently and regularly, versus the two other classes which are visited occasionally and exceptionally.



(a) CDR-17 *(b) CDR-67*

Figure 3.6. *The empirical probability density function estimated through KDE (kernel density estimation) of the relevance in each class. EVP and MVP functions have been resized for a better visualization. Classes are separable.*

Once extracted the relevance classes, we focus on two main features: *a)* the number of distinct PoIs in each class of relevance, *b)* how people move across relevance classes. As it will be detailed in the next section, the last characteristic is fundamental in the model description, since it allows us to model the human mobility as the movement among PoIs belonging to different relevance classes.

As far as *a)* is concerned, in Figures 3.7, 3.8, 3.9 and 3.10, we show the per-user number of distinct PoIs associated to each class of relevance in CDR, GPS and WiFi datasets, respectively. We observe in Figure 3.7 that the per-user number of distinct visited PoIs increases when moving from 17- to 67-day CDR datasets, with the consequence that the number of visited PoIs grows over time.

For all datasets, we observe a remarkable difference between the number of EVPs and the PoIs into the other relevance classes (OVP, MVP). This points out the general users' habit to visit many new locations, but also that they regularly move towards very few of them. If we limit our attention to OVP and MVP classes it turns out that the number of visited OVPs is limited and the average number of this kind of visited places per user in CDR-67 is around 5; also for the MVPs the number of visited places per user is limited, and its average value is around 2. As expected, the above analysis underlines the fact that each user has a very small number of favorite PoIs (MVP) which are visited daily (e.g., home, workplace), and a higher, but still limited number of location

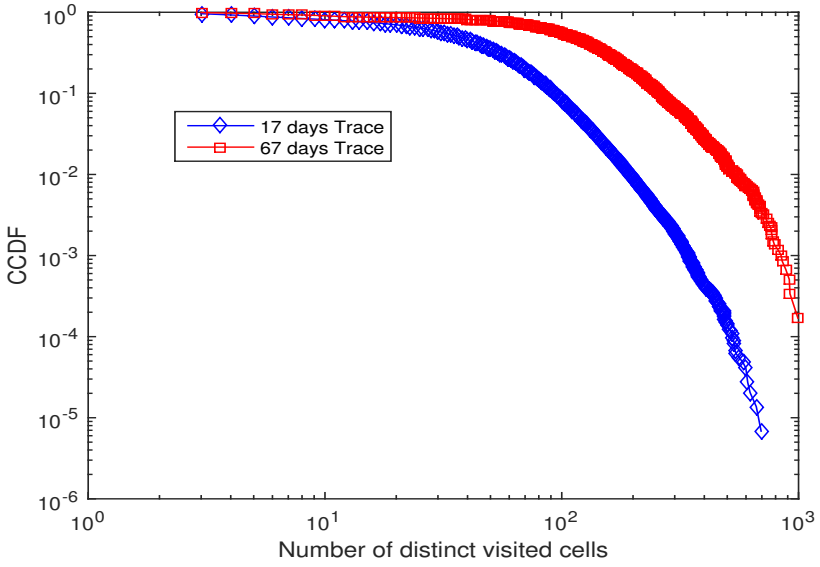


Figure 3.7. *The empirical CCDF distributions of the number of distinct visited PoIs per user, aggregated over the three relevance classes in the both cellular network datasets.*

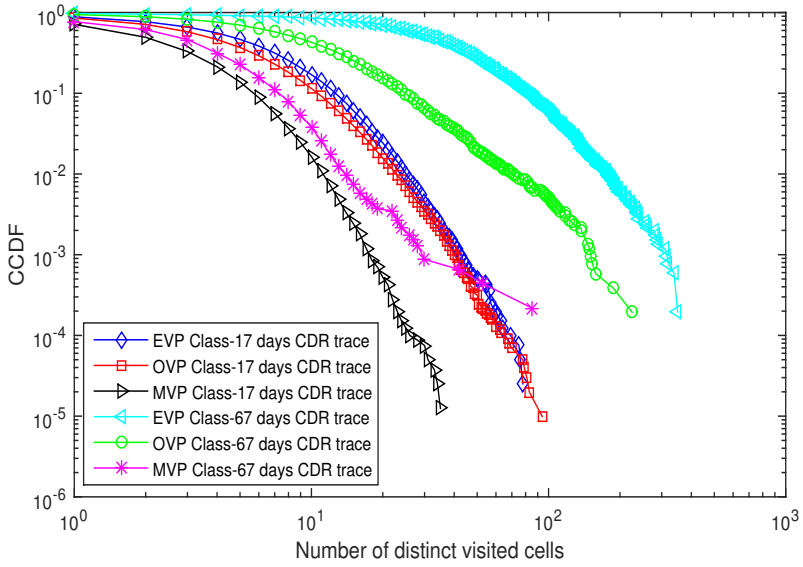


Figure 3.8. *The empirical CCDF distributions of the number of distinct visited PoIs per user in each relevance class in the two cellular network datasets.*

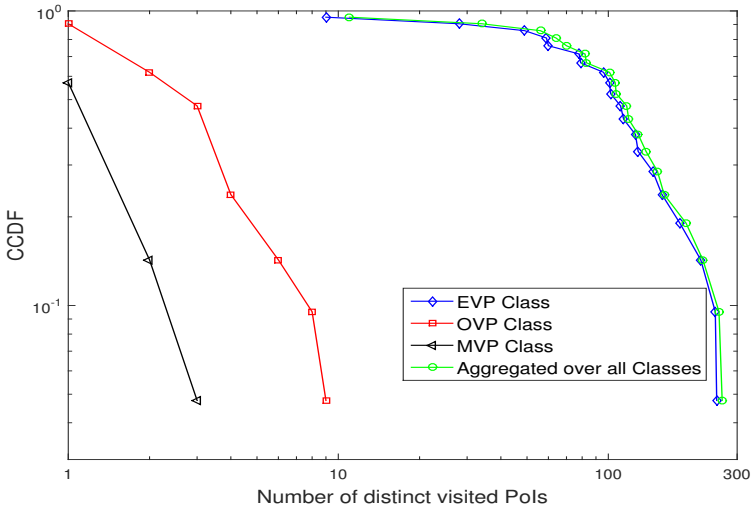


Figure 3.9. *The empirical CCDF distributions of the number of distinct visited PoIs per user in GPS Microsoft dataset.*

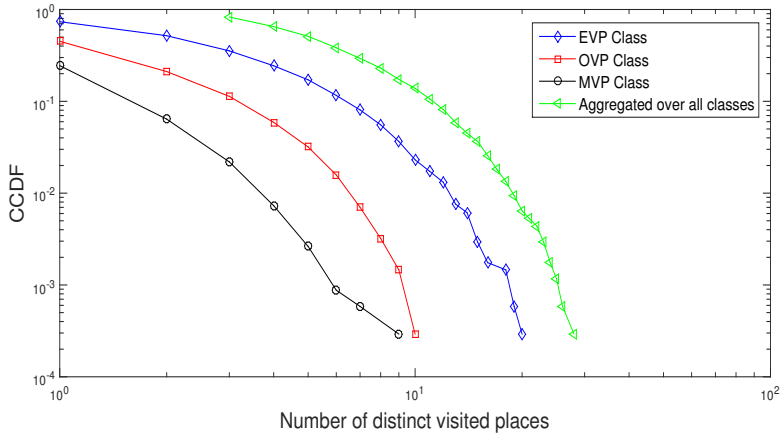


Figure 3.10. *The empirical CCDF distributions of the number of distinct visited PoIs per user in WiFi Dartmouth dataset.*

Table 3.2. *Average percentage of visited PoIs in each class in Group 3*

Datasets	P_{evp} %	P_{ovp} %	P_{mvp} %
CDR-17 days	77.10	14.60	8.30
CDR-67 days	62.69	23.38	13.93
GPS	96.13	2.69	1.18
WiFi	47	29.50	23.50

of interest (OVPs) which are visited with a lower frequency (e.g., gym, favorite pub, parent's house).

The observed characteristic in MVP class, with a heavy tail distribution of the number of visited PoIs, implies that the majority of users visit just a few PoIs more frequently and regularly which is well aligned with location preference property in human mobility already reported in previous works such as [28, 33].

In WiFi dataset, the concept of EVP, OVP, and MVP places should be adapted to the campus environment. For instance, the most frequently visited PoIs might be the laboratory or the office which would correspond to home or workplaces in CDR and GPS datasets.

Finally, we enhance the generalizability of the feature of relevance class throughout different datasets by analyzing the percentage of PoIs relying on the 3 classes, as reported in Table 3.2. The behavior is quite similar for all datasets. Most points belong to the EVP class; there are very few MVPs, while OVPs account for a number of PoIs similar to the MVPs class. The average percentage of the number of distinct PoIs in each relevance class for different datasets is indicated in Table 3.2. We can, therefore, conclude that the classification we identified in terms of relevance at the beginning of this section (MVPs, OVPs, EVPs) is generally significant since the distribution of the per-user number of PoIs associated to each class of relevance is similar across datasets with very different characteristics. We have shown that, independently of the dataset characteristics, the PoIs visited by people fall mainly in the EVP class. However, most of the people spend most of their time in MVPs or OVPs; many of them can be found more than half of the time in MVPs. Although the Relevance Ratio (RR) was defined based on the frequency that each PoI is visited on daily base, but the extracted Pause Time distributions in three classes of PoIs (MVP, OVP and EVP) indicate that there is a positive correlation between relevance ratio and pause time in the PoIs (Refer to the Figure 5.4 for WiFi and [62]

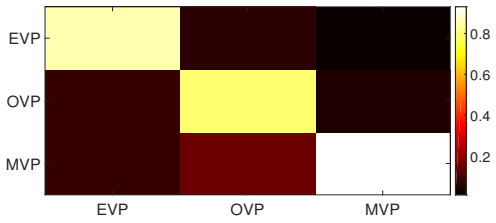
for GPS). It implies that in the classified PoIs as MVP and OVP (with higher value of RR) the probability of staying for longer duration is higher.

3.2.3 Transition rules

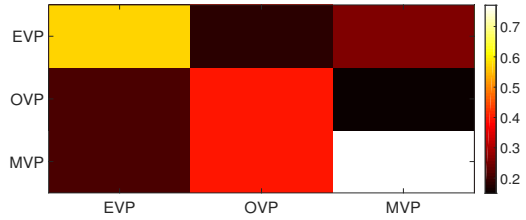
The human decision to move from one place to another emerges from a decision-making process that is influenced by a variety of human and contextual behaviors. To improve the understanding of this process, we want to measure the impact of relevance on the chance to get to a given arrival PoI A.

The impact of the class of relevance of the departure PoI is independent of the scale of the scenario when we analyze the conditional probability to move from a PoI in a class c_1 to a PoI in a class c_2. In Figures 3.11a and 3.11b the conditional probabilities of moving among the relevance classes in CDR-17 and CDR-67, respectively, are depicted. As shown in Figure 3.11a, we observe that the most probable movements occur between the same classes, *i.e.* the relevance class of the destination will likely be the same class as the departure location. Otherwise, movements among different classes are less probable. The scenario and the mobility habits change a little in the CDR-67 dataset. In this case (see Figure 3.11b), people mainly commute between MVPs classes and with lower probability from OVPs classes to OVPs /MVPs classes.

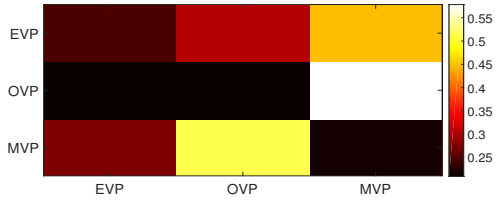
In Figure 3.11c; WiFi dataset, the most probable movements occur from/to MVPs to/from OVPs classes. Even if the conditional probabilities are heavily affected by the great number of EVPs, people commute to/from OVPs from/to MVPs, *i.e.* occasionally visited locations such as pub or free time spaces are related to home/workplaces (most visited PoIs). The transition patterns among the relevance classes for GPS dataset has been reported in [62].



(a) Class Transition in CDR-17



(b) Class Transition in CDR-67



(c) Class Transition in WiFi Dataset

Figure 3.11. *Transition probability among relevance classes. Each square represents the conditional probability to move from a PoI in a class c_1 to a PoI in a class c_2,* i.e. *$P(C_{new} = c_2 | C_{old} = c_1)$. On the x-axis the conditioning variable C_{old} and on the y-axis the conditioned variable C_{new}.*

Chapter 4

Home and Work detection

We have established that the locations visited by people can be classified in terms of their relevance as well as the rules that characterize the mobility between them. However, it is also important to understand the semantic value of such locations so as to better define human mobility. In particular, home and workplace are the most meaningful locations in human life. They are both characterized by a set of features, not shared with other places visited by a user. First of all, they are the places people visit more frequently and regularly than others. This characteristic is fully measured by the relevance ratio (RR) described in the previous chapter. In this chapter we introduce a novel algorithm for home and workplace detection based on the concept of relevance and evaluate it on the CDR datasets.

4.1 Related Work

A great effort has been devoted to the assessment of the visited locations, trying to detect (learn) the significant places (PoI) and assign a particular meaning to each of them.

Authors in [34] proposed an algorithm for identifying important places in people's lives from Cellular Network Data. In first stage, the cell towers that appear in a user's trace are spatially clustered (Assuming be aware of the geographical position of cell towers) and in next stage the important cluster will be identified using a model derived from a logistic regression.

Laasonen et al. [46] proposed a conceptual framework for identifying significant locations of users from GSM dataset collected through their personal mobile phone. Without reference to any physical locations, cell graph serves as

the representation of cell topology. By exploiting the cell clustering approach (considering the issue of cells overlapping), significant locations are identified as locations user spends a large portion of his/her time.

Kang et al. [36] discuss the definition of user's significant places as places where the user spends a substantial amount of time and or/ visit frequently. Authors proposed a Time-Based Clustering approach that exploit the clustering locations obtained through WiFi beacons for identifying places people visit. Authors in [72] propose an approach for extracting significant places from WiFi logs. In this approach the most frequently visited APs are natural candidates to be represented as significant places. Providing that a sufficient number of visits has been recorded to a given AP by a user, that AP could be considered as a significant place for that user if number of visits is higher than the threshold **n**. In [65] authors beside explaining the concepts for movement presentation such as **Location** and **Place**, to infer places (place learning); integrate observations collected from three sensors means GSM, GPS and WiFi through simultaneous data fusion.

Among the different problems in the evaluation of the location semantic, we focus on the detection of home and workplaces from cellular network data, based on the frequency of daily visits, a.k.a. relevance. To solve the aforementioned issue, Isaacman *et al.* [34] have proposed a technique based on clustering and regression to identify important places then assign them a semantic such as home and work. By contrast, Csaji *et al.* [19] have combined principal component analysis with clustering to robustly identify home and workplaces. Arai and Shibasaki [5] have proposed a methodology for the estimation of home and work locations based on time windows. After recognizing important places according to the length of stay and frequency of visits, they base the home/work identification on core hours at home/work. Authors in [54] propose a technique to automatically predict the approximation residential location of anonymized cell phone users based on their calling behavior, having access to the geographical position of contacts handling cell towers and also having a small set of users for whom their approximate residential location is known. All of the above approaches require knowledge of the tower position (GPS coordinates), but this information is not always available. So the strategies and methodologies proposed in above literature are not applicable in our case since our datasets don't include geographical data or tower positions.

An identification method not founded on knowledge of the tower positions has been presented by Alhasoun *et al.* [4]. In their work they identify the places where each user is more active (call) by dividing a day into daytime and

night. Home is the most active place during the night window, while work is the most active location during the day. Apart from being time window dependent, the method does not consider regularity in visiting places as the main feature defining home and work. However, it is commonly accepted that most users regularly visit and commute between home and workplace on workdays. Thus, solely the number of activities is not a good indicator for home and workplace since users may make a burst of on-phone activities in places which are not frequently and regularly visited.

In [22] the authors analyzed call and Bluetooth logs of approximately a hundred users for a duration of nine months in order to identify a structure in the daily life routine of mobile users. They attempted to quantify the amount of predictable structure in an individual's life using an information entropy metric. They expected people with low-entropy lives to be more predictable across all time scales. By using the discovered patterns and contextualized proximity information extracted from Bluetooth logs, they proposed a model for identifying locations such as home, workplace and also activities.

4.2 Home vs. Work discrimination from cellular network data.

We exploit the relevance ratio (RR) to identify home and workplace among all visited PoIs. Home and Workplaces are the places that people visit more frequently and regularly than other places, so we expect home and workplaces have the highest RR. Specifically, PoIs belonging to the class of most visited places (MVP) are the natural candidates for home and workplace identification as they have the highest relevance ratio, as shown in Figures 3.6a and 3.6b. Beyond this main measure, a set of other features can help identifying home and workplace. Considering that these are the places where people spend the bulk of their lives, it is also reasonable to assume that they are the places where people perform the highest number of contact activities. Thus, we introduce a feature to quantify this aspect. Finally, to distinguish between home and work, we argue that, on average, people rarely spend most of the night at their workplace; therefore, we take into account the initial time of on-phone activities. The overview of the recognition strategy is presented in Figure 4.1, and it is mainly based on the relevance of a PoI and represent only the values of the relevance which identify the MVP class for a given user.

We then apply this strategy to the two CDR datasets, as the other dataset

such as Microsoft GPS presents small number of users and also we do not have any ground truth for Home and workplace locations for them.

```
Data: L = list of the locations visited by the user u
H, W = null;
H = heapify(L);
while H.size > 0 do
    L ← H.extract_max();
    switch R(L, u) do
        case R(L, u) ≥ HighRR
            if H = null then H ← L;
        end
        case R(L, u) ∈ [MediumRR, HighRR)
            if Start time of contact activities during the NIGHT then
                if H = null then H ← L;
            else
                if W = null then W ← L;
            end
        end
        case R(L, u) ∈ [LowRR, MediumRR)
            if Start time of contact activities during the DAY then
                if W = null then W ← L;
            end
        end
    endsw
end
```

Algorithm 1: Home/Workplace Recognition

As evident in Figure 4.1, we identify three relevance intervals where we can look for the home and workplace candidate locations. If a PoI belongs to the red interval (High RR- on the right), it becomes the HOME. If more than one PoI has the same highest relevance due to the ping-pong effect[69], we recognize as HOME the PoI where most of the user's activities occur, discarding the other PoIs in High RR from the candidates set for workplace recognition. But as aforementioned, CDR datasets are not continuous, so potentially the HOME location may not appear in the High RR interval. In this case, we can have a situation where HOME and WORK both have medium relevance (Medium RR- orange middle interval). Consequently, we need to introduce a further feature:

the starting time of contact activities. We distinguish between night and day time. With this new feature, identifying contacts starting at nighttime, we again classify the highly ranked PoI as the HOME location. Otherwise, if it starts during the day, we identify it as the WORK location. For low relevance (Low RR - on the left) home identification becomes less stringent since these users are very likely to live outside the city and come into city only for work purposes, so we identify only the WORK location. This is further detailed in algorithm 1. The algorithm receives a list of PoIs and builds the heap $\mathcal{H}$. In the heap, PoIs are primarily ordered by their relevance and by the number of activities on the part of user u in the case of relevance equality. At each iteration the algorithm extracts and removes from the heap the maximum element and assigns it to the right relevance interval depicted in Figure 4.1. In the end, the variables H and W contain the home and workplace whereas they are detectable.

The CDR datasets we analyze are related to the urban area of Milan, which is why we consider the time interval 8 A.M. to 8 P.M. as day time. Similarly, from the relevance distribution, we can classify a PoI as a location with high relevance when $RR >= 0.9$, *i.e.* being at home for at least 90% of the days. Medium relevance corresponds to $0.8 <= RR <= 0.9$, which means visiting a PoI at least $5 - 6$ days per week. We classify the relevance of a PoI as low if $0.65 <= RR <= 0.8$, which corresponds to 5 working days and also possible holidays. Otherwise, the information is not significant. Also, the start time of the activities provides a semantic for distinguishing between home and workplace in the case of medium relevance: home if it is between 8 P.M. and 8 A.M. (when people are expected to be at home), workplace in all other instances.

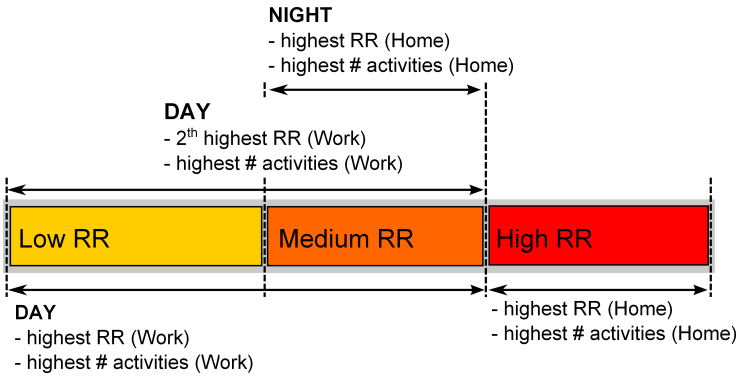


Figure 4.1. *Home/Workplace recognition process*

In Table 4.1 we report the number of users for whom the algorithm is able

Table 4.1. *Percentage of recognized home/work locations.*

Dataset	HOME	WORK
CDR-17	37093/80143=46.28%	62258/80143=77.68%
CDR-67	2577/4578=56.29%	3383/4578=73.90%

to recognize the home and work locations. Overall we analyze 80,143 and 4,578 users belonging, respectively, to CDR-17 and CDR-67. Our methodology assigns a home location to 37,093 (46.28 %) and 2,577 (56.29 %) users, a workplace to 62,258 (77.68%) and 3,383 (73.90%) ones, respectively. For users with low relevance in visiting MVP places, it is not possible to recognize their home/workplaces. Since a ground truth for the home/work detection does not exist, the goodness of the recognition algorithm is only partially verifiable. As already mentioned in section 2.1, we exploit the billing mechanism to get an approximation of the ground truth. In particular, the billing system records an Internet CDR every day at midnight indicating the position of the user. The most visited PoI on weekdays at midnight can be reasonably expected to correspond to the home location. Since the billing system is operator-dependent and undocumented in most cases, we have decided not to include this heuristic in the detection algorithm. Rather, we employ it in the evaluation. Keeping this setting, we measure an accuracy rate equal to 78.3% in CDR-67, which is a good performance for the home detection task.

Table 4.2. *Conformity percentage of recognized Home/Work places between Alhanson and Relevance based approaches.*

Dataset	HOME		WORK	
	Cell Level	Area Level	Cell Level	Area Level
CDR-17	83%	91.20%	69.73%	76.60%
CDR-67	83%	91%	56.50%	77.74%

In addition, we want to show that the relevance is of paramount importance and that our approach, where the main criteria are relevance, has some advantages compared to the similar approaches that use different criteria. For that reason, we compare our algorithm to the one proposed in Alhansoun *et al.* [4] which uses only the highest number of total contact activities in day and night windows, to recognize home and work locations. The Accuracy rate of Alhan-

Table 4.3. *Differences in the results among relevance-based and Alhanson approaches.*

Approach	Dataset	Relevance Range		Number of recognized	
		Home Places	Work Places	Home Places	Work Places
Relevance Based	**CDR-17**	0.80-1	0.65-0.90	37093	62258
	CDR-67	0.80-1	0.65-0.90	2577	3383
Alhasoun	**CDR-17**	0.42-0.88	0.27-0.93	80143	80143
	CDR-67	0.47-0.97	0.31-1	4578	4578

soun's algorithm for the home detection task is 63% in CDR-67, lower than the rate obtained by our algorithm (78.30%). In Table 4.2, we observe that there is 83% match of recognized home places between the two approaches. For workplaces, the percentage drops to 69.73% and 56.50%, respectively, in CDR-17 and CDR-67 datasets. If we consider the spatial granularity of a tracking area (which covers several nearby cell towers) instead of a single cell tower, the percentage of conformity between home places increases to 91.20 and 91, and the percentage of workplaces increases to 76.60 and 77.74 in CDR-17 and CDR-67. The differences in the recognized home and workplaces between our approach and the one presented by Alhasoun *et al.*[4] are due to the poor correlation between the number of contact activities in a PoI and its relevance.

Figure 4.2 depicts the distributions of the relevance of PoIs recognized as workplaces by Alhasoun's approach [4], which are different from the PoIs we recognize as workplaces. We observe that at least around 40% of the workplaces recognized by the approach described in [4] have low relevance, as shown in Table 4.3, although they have the highest total number of contact activities (since they get recognized). This means that most of these workplaces are not visited regularly by users; they do have, however, the highest number of on-the-phone activities. Also, PoIs that have relevance higher than 0.9 can rarely be workplaces, since it is very unlikely that people went to work almost every day throughout the duration of the collected datasets. Therefore, we can conclude that our approach based on relevance allows reducing the number of errors induced by counting the number of contact activities only.

Table 4.3 indicates the differences among the results obtained by the two approaches and highlights the relevance bounds which characterize home and workplaces extracted by Alhasoun's approach.

In the case of using GPS or WiFi datasets (high temporal continuity) the approach would be similar to what is discussed above; almost the same, just pause time duration would be used instead of the number of contact activities.

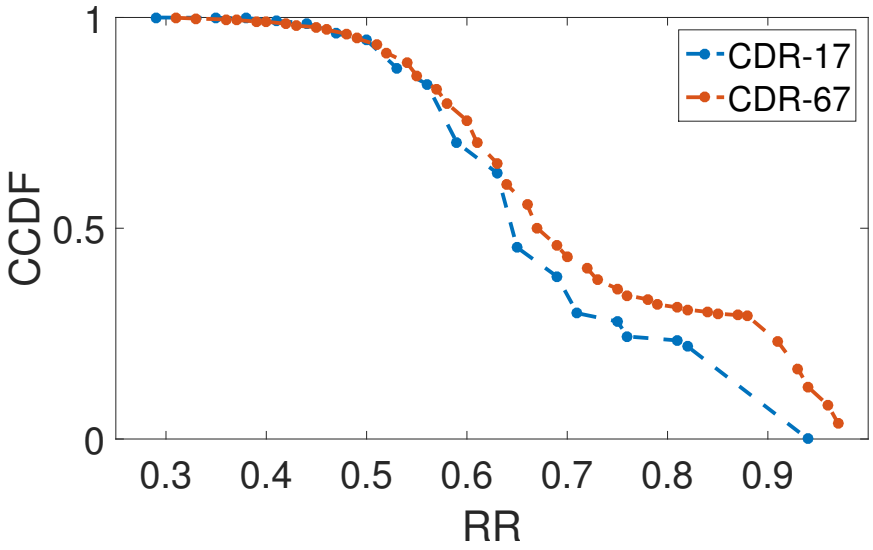


Figure 4.2. *The empirical CCDF distributions of the relevance of the PoIs recognized as workplaces by Alhasoun's approach but not identified as workplaces in our approach.*

Anyway, we don't have any ground truth of Home or Workplaces for WiFi and GPS datasets.

Chapter 5

Simulating Human Mobility in Urban Space

With the rise of smart cities people are moving within urban spaces and still be able to pervasively interact with information, services, city's resources and other people. In such a highly connected scenario, smartphones, and other wireless portable devices are carried by humans, exhibit the same mobility behavior of their human carriers and their movements strongly impact on the underlying network operation and performance. The understanding of human mobility in an urban space has become crucial to optimize the network management, to plan the adaptive allocation of critical resources and ensure constant quality of the user experience. This chapter takes a first step in the direction of the design of a mobility model meeting behavioral and scale requirements of modern smart cities. We envision a smart city as a collection of places, each representing a Point of Interest (PoI) with a specific value for single individuals and for a set of them. As a consequence, each individual has his/her own mobility footprint, while few of them share similar mobility patterns. By simulating the mobility of each individual across city's places, we will be able to properly describe human mobility and social behavior in urban spaces, and to extract all needed information about how city's resources and services are accessed. The extensive use of CDR, GPS, and WiFi datasets enables us to analyze the characteristics of city PoIs, classify them for each individual according to their importance and study how the individuals move across them. The common features observed are the key points to build a metropolitan mobility model able to reproduce the regularity in the spatio-temporal behavior of mobile users and also the way city sociality is built around PoIs of the city. The model exhibits high flexibility and can be applied in wide geographical and population scales.
The described analysis in chapter 3 has helped to build the key elements driving human mobility in metropolitan and campus environments. These elements con-

stitute the foundational pillars of our mobility model. Conceptually, individuals move according to the rules driving the movement among classified PoIs and following temporal behavior in accordance with the literature. More precisely, movements among PoIs are chosen by a first-order Markov chain generating synthetic relevance-based traces.

5.1 Related Work

Prior mobility models were mainly Random models such as Random Walk, Random Waypoint, and Gauss-Markov model [12]. In these models, waypoints, velocity of mobile users and direction of movement are chosen randomly in the simulation area. By choosing purely random parameters, mobile users are allowed to make rapid and sudden changes in their movement. However, unexpected changes rarely occur in human mobility patterns. As a consequence, random models poorly reflect realistic aspects of human mobility and it is ineffective to rely on them for an accurate understanding of protocol performance. By emerging detailed and long-term human mobility datasets that capture actual human movement, researchers discarded random models and focused on new trace-based mobility models able to realize real properties of human mobility. On the other hand, Synthetic models aim to create driving forces of individual mobility such as social attitude, location preferences, and regular schedules. Synthetic models are mathematical models, such as sets of equations, which try to capture the movement of the mobile node. Datasets, in this case, might be explored for validation. Synthetic approaches range from the basic random walk (e.g., Brownian motion) to more sophisticated models, which could use a detailed map of the region, individual schedules, etc. Authors in [75] proposed an analytical mobility model which is an alternative to the gravity law [83] that estimates commuting and mobility fluxes in the daily activity of population at the macroscopic level. Although this model realizes mobility and transport patterns observed in a wide range of population and geographical scales, it does not retrieve any mobility information at the microscopic level. Also, we believe such kind of model is more appropriate for studying movement across countries and across cities instead of intra-city mobility. Most of the recent trace based mobility models are distinguished based on which of the three driving forces of human mobility (social relationships, preference for particular locations as movement waypoints and human activities schedule) has a dominant role in creating the model [38].

Orbit model is one of the primary models in this category [26]. This model

creates a set of location preferences (spots) in the simulation area, randomly, for all mobile users (nodes). Each node chooses a random subset of these locations from a Hub List and pauses in each one according to predefined Hub Stay Time. In Orbit the mobility is divided into two main scenarios: inside spots (Intra-spots) and between spots (Inter-spots). For both components, any known mobility model may be chosen.

SLAW [49] captures the truncated power law jump sizes by considering waypoints as distributed according to a power-law distribution of distance in the simulation area, and by selecting the waypoint of the next movement according to the minimized distance from the current location (least action principle). This model enables mobility through locations that are more popular than others and also satisfies location preference and short distance traveling properties of human mobility.

SWIM [55, 44] is based on home locations. For each mobile node, a point over the network area is chosen randomly and uniformly, which is called home. Then, the mobile node itself assigns to each possible destination a weight that grows with the popularity of the place and decreases with the distance from home in a way that captures the power law distribution of jump size. The node then selects a destination for the next moves with probability proportional to the calculated weights. The popularity of a place is calculated as the number of other people encountered when the node visited the place recently. The basic characteristic of this model is that it represents a trade-off proximity and popularity. In SWIM, the speed of nodes is proportional to the distance covered.

The Exploration and Preferential Return (EPR) model [76], in contrast to the above-mentioned models, does not fix prior the number of visited places but let them emerge spontaneously. In this model, an individual has two choices: exploration and preferential return. With a given probability, the individual mobile node returns to one of the previously visited places (Preferential return phase) and with complementary probability the mobile node moves to a new location (Exploration phase). The probability to explore new places decreases as the number of visited places increases, as a result, the model has a warm up period of greedy exploration, while after long run mobile node mainly moves around a set of previously visited places. This model recently has been modified by [7] through adding the recency of visited places during the preferential return phase, or adding a preferential exploration step to account for the collective preference for places and also the returners and explorers dichotomy.

Authors in [73] use datasets collected from realistic and human-like virtual world for analyzing and characterizing properties and patterns of human mo-

bility. According to observed characteristics, the proposed SAMOVAR human mobility model captures some realistic characteristics of human mobility such as population heterogeneity, popular location preferences, truncated power law of jump size, pause time and inter-contact time.

In [70] data collected from smart subway fare card transactions are used to model urban mobility patterns. In this model, authors utilize the idea of preferential selection to popular places in the city for modeling the spatial distribution of individual's movements for selecting destinations. In Samiul et al., the popularity of places in the city is the factor driving the interactions among different individuals. In fact, in contrast to our work, the popularity at a population-level, i.e. the most preferred places by the entire population, impacts on the choice of the individuals' destinations.

Our metropolitan mobility model takes the first step in modeling human mobility patterns in urban scenarios as sequences of the point of interests, each with associated the per-user semantic value, and with a data-driven validation reproducing real metropolitan settings in terms of geographic size and city's population.

From an agent-based perspective, a mobile user (agent) chooses the next destination according to the category of the current PoI and the transition matrix which characterizes the probability of movements from one PoI in a class to another PoI in the same or other classes. A similar agent-based model, base on a hierarchical structure of the PoIs (aimed at setting the spatial structure of the mobility of the agents) has been presented in [35] where agents choose the next destination according to a uniform or preferential criterion.

5.2 Movement model

The described analysis in chapter 3 has helped to build the key elements driving human mobility in metropolitan and campus environments. These elements constitute the foundational pillars of our mobility model. Conceptually, individuals move according to the rules driving the movement among classified PoIs and following temporal behavior in accordance with the literature. More precisely, movements among PoIs are chosen by a first-order Markov chain generating relevance-based synthetic traces.

5.2.1 Environment setting

The first step in the definition of the environment setting is the geographical placement of the visitable places. In our metropolitan mobility model, PoIs are uniformly distributed all over the area of the city, as other recent works do [44, 70, 59].

After this preliminary setting of PoIs positioning, the model adopts a per-user strategy where each user performs the activities described below. The model requires the setting of the user's visitation set, i.e. the set of visitable PoIs grouped by relevance class. The cardinality of the overall visitation set (K), the MVP set (NP_{mvp}) and the OVP (NP_{ovp}) is generated by adopting the same corresponding distributions reported in Fig. 3.7 and Fig. 3.8 for CDR datasets, Fig. 3.9 for GPS dataset and Fig. 3.10 for WiFi dataset, while the remaining $NP_{evp} = K - (NP_{mvp} + NP_{ovp})$ denotes the cardinality of the EVP set [1]. Once these cardinalities have been set-up, the relative sets are created by randomly placing without replacement the obtained amount of PoIs for each class. The union of drawn PoIs in all three classes forms the visitation set of each user. Each PoI of the visitation set is labeled with the value of the relevance ratio the PoI has for that user. At the end of this initial setup, we obtain for each user three random vectors of size NP_{mvp}, NP_{ovp} and NP_{evp} describing the PoIs the user can visit, subdivided per relevance class and weighted with its relevance ratio value extracted from the relevant relevance distribution in each class. Each user starts to move from his/her MVP with the highest relevance value as it represents his/her favorite PoI.

5.2.2 Where to go next?

Once a mobile node is inside the PoI P_i, it has to decide the next PoI be visited. To this purpose, the model provides a 2-steps algorithm. The first step provides the selection of the next relevance class according to a first-order Markov chain characterized by three states: 'M', 'O' and 'E' associated to MVP, OVP and EVP classes, respectively. Each element of the matrix $\mathbb{T}$ indicates the probability that a node in a place with relevance class c_i has to move towards a generic PoI inside relevance class c_j, i.e. $\mathbb{T}_{ij} = P(c_j|c_i) = P(c_i, c_j)/\sum_j P(c_i, c_j)$.

[1] If $NP_{evp} <= 0$, the extraction process is repeated until the previous condition is true

$$\mathbb{T} = \begin{bmatrix} P_t(M \mid M) & P_t(O \mid M) & P_t(E \mid M) \\ P_t(M \mid O) & P_t(O \mid O) & P_t(E \mid O) \\ P_t(M \mid E) & P_t(O \mid E) & P_t(E \mid E) \end{bmatrix} \tag{5.1}$$

In the following we calculate and report the transition matrices characterizing the movements among classes in CDR-17 ($\mathbb{T}_{CDR-17}$), CDR-67 ($\mathbb{T}_{CDR-67}$), GPS ($\mathbb{T}_{GPS}$) and WiFi ($\mathbb{T}_{WiFi}$) datasets, respectively.

$$\mathbb{T}_{CDR-17} = \begin{bmatrix} 0.9384 & 0.0462 & 0.0154 \\ 0.1428 & 0.8095 & 0.0477 \\ 0.0714 & 0.0714 & 0.8572 \end{bmatrix} \tag{5.2}$$

$$\mathbb{T}_{CDR-67} = \begin{bmatrix} 0.7797 & 0.1525 & 0.0678 \\ 0.4091 & 0.4091 & 0.1818 \\ 0.2105 & 0.2105 & 0.5790 \end{bmatrix} \tag{5.3}$$

and

$$\mathbb{T}_{GPS} = \begin{bmatrix} 0.3901 & 0.5722 & 0.2572 \\ 0.2809 & 0.1870 & 0.0951 \\ 0.3270 & 0.2408 & 0.6477 \end{bmatrix} \tag{5.4}$$

and

$$\mathbb{T}_{WiFi} = \begin{bmatrix} 0.2179 & 0.5102 & 0.2719 \\ 0.5808 & 0.2103 & 0.2089 \\ 0.4504 & 0.3067 & 0.2429 \end{bmatrix} \tag{5.5}$$

Once the new relevance class has been selected, the second step enables the node to extract the next PoI to be visited among the PoIs assigned to the class according to a probability that relies on the relevance ratio.

5.2.3 Transition time

The user moves toward the next PoI when the pause time in the current PoI runs out. The time required to move between two PoIs in a metropolitan area cannot be modeled with a uniform distribution of the speed. The structure of GPS and WiFi datasets enable us to extract the distribution of the transition time, while in CDR datasets this is not possible because of temporal sparsity. In the latter case, we consider that people have several options to move in urban

space, but it has been shown that the speed of the movement is proportional to the distance [59]. Thus, we define a step-function which returns the transfer time given the distance d to be covered:

$$T_t(d) = \begin{cases} t_0 & d \leq 100m \\ t_1 + \frac{d}{v_1} & 100m < d \leq 1km \\ t_2 + \frac{d}{v2} & d > 1km \end{cases} \tag{5.6}$$

where t_0, t_1, t_2 are offsets that could be set according to the scenario (t_2 might resemble the average waiting time for public transportation) and v_1, v_2 are average speed of pedestrian, cars or buses; respectively. For instance, authors in [68, 59] derive functions to estimate the transition time from travel distances. The simulation set-up we have chosen is: $t_0 = 4$ min, $t_1 = 2.5$ min; $v_1 = 4$ km/h; $t_2 = 5$ min; $v_2 = 12$ km/h accordingly to data of public transportation in Milan city.

In this model, despite most of the existing mobility models like the gravity model[83], SLAW, SWIM or STEP [86], the distance decay effect has not been taken into account directly. In fact, although distances among places influence people mobility, the transition time (Time of Travel) has gained a more important role especially in intra-urban environment [61, 70]. Thinking in terms of temporal distance should be more informative and closer to the criteria people choose to minimize the time spent to travel or commute. According to this point of view, we use the transition time among PoIs instead of the usual geographical distance.

5.2.4 Pause time distribution

After reaching the chosen PoI, each person remains inside it for a time period drawn from a truncated power law distribution whose parameters depend on the class of the PoI. The choice of a heavy-tailed distribution is driven by the analysis of pause time reported in [28, 68, 63]. In fact, people are likely to spend a long period of time in MVP places (for instance home, work office or school) whereas they stay less in OVP and EVP places (post office, bank, cafeteria). In Figure. 5.3 and 5.4 the distributions of pause time of EVP, OVP and MVP places from GPS and WiFi datasets, have been depicted, respectively. So, when a user arrives at one of (MVP/OVP/EVP) PoI, pauses there according to the relevant pause time distribution. The distribution in all classes follows a truncated power law but with a different slope. Due to the high temporal sparsity of CDR datasets, extracting pause time distribution isn't straightforward task for them.

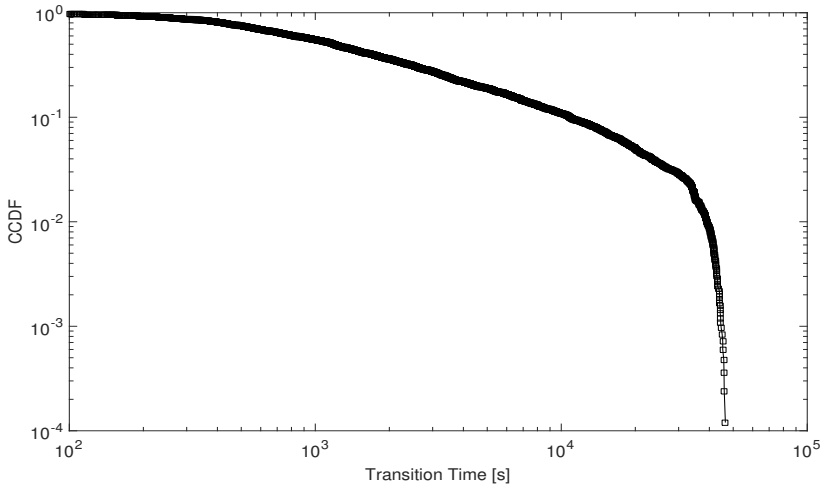


Figure 5.1. *The empirical Transition Time distribution for Microsoft GPS dataset.*

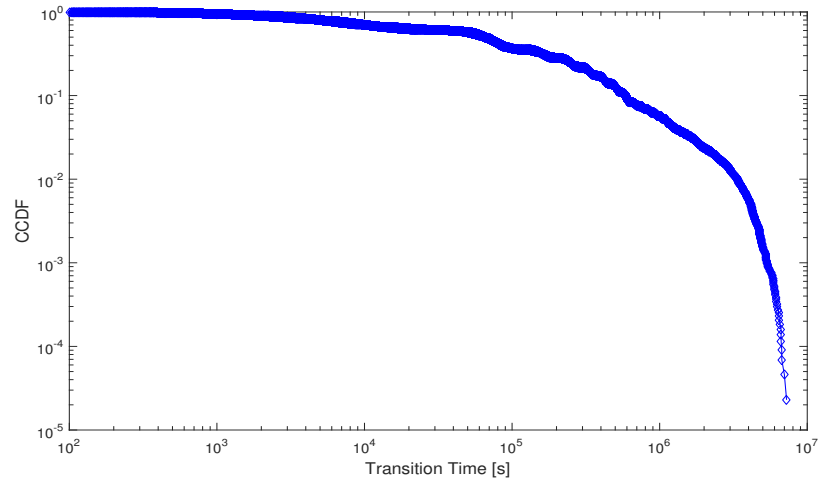


Figure 5.2. *The empirical Transition Time distribution for Dartmouth WiFi dataset.*

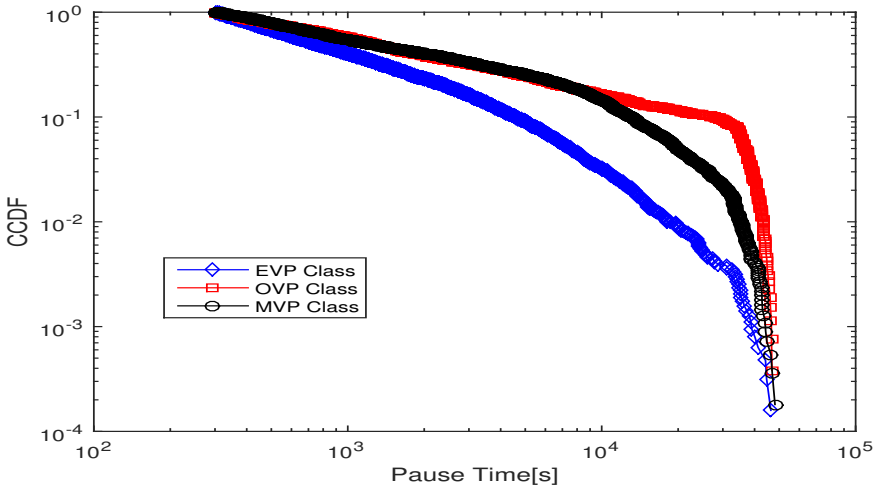


Figure 5.3. *The empirical Pause Time distributions in different classes of Microsoft GPS dataset.*

In Figure 5.3 is counter-intuitive that the OVP has longer pauses than the MVP. Here the Pause time might be affected by gaps in the data collection process. So we observe pause time in OVP PoIs is longer than MVP PoIs.

The entire procedure driving the mobility of each user is expressed in the listing 2).

5.3 Evaluation

In this section, we show that our mobility model reproduces realistic spatio-temporal patterns and connectivity characteristics of human movements.

5.3.1 Simulation setup

In the following, we compare the synthetic traces produced by our model against the real traces from the CDR-67, the GPS, and WiFi datasets. Consequently, the input parameters and distributions correspond to the following mobility scenarios:

- *CDR-67 scenario.* In this scenario, the model handles about 5,000 mobile nodes. The distributions, required as input to the model, have been

Input: relevance distribution $\mathcal{R}$; visitable place distribution $\mathcal{K}$, $\mathcal{K}_M$ and $\mathcal{K}_O$; $\mathbb{T}$; $\mathcal{P}$ = set of the visitation places; pause time distributions $\mathcal{PT}$s

repeat
| $K \sim \mathcal{K}$, $K_M \sim \mathcal{K}_M$, $K_O \sim \mathcal{K}_O$, $K_E = K - (K_M + K_O)$
until $K_E \leq 0$;
foreach $c \in \{M, O, E\}$ **do**
| Pick NP_c elements without replacement from $\mathcal{P}$ into the set c set.
end
$\mathbf{w} \sim K$
Assign weights $\mathbf{w}$ to visitable places
$S \leftarrow$ empty sequence
$P_0 = P \in M$ with the highest weight
$pause \sim \mathcal{PT}_{mvp}$
$c_{now} = P_0^{(0,pause)}$, S.append(c_{now})
$time \leftarrow pause$
$i = 1$
while *Stop condition* **do**
| $c \leftarrow$ extract the relevance class from the probability distribution $\mathbb{T}(.|rc(c_{now}))$
| $pause \sim \mathcal{PT}_c$
| $P_i \leftarrow$ extract from c a place with probability proportional to its weight
| $d \leftarrow$ compute distance between c_{now} and P_i
| $transfer \leftarrow T_t(d)$, $time \leftarrow time + transfer$
| $c_{now} = P_i^{(time,time+pause)}$, S.append(c_{now})
| $i \leftarrow i + 1$
| $time \leftarrow time + pause$
end

Algorithm 2: Mobility model

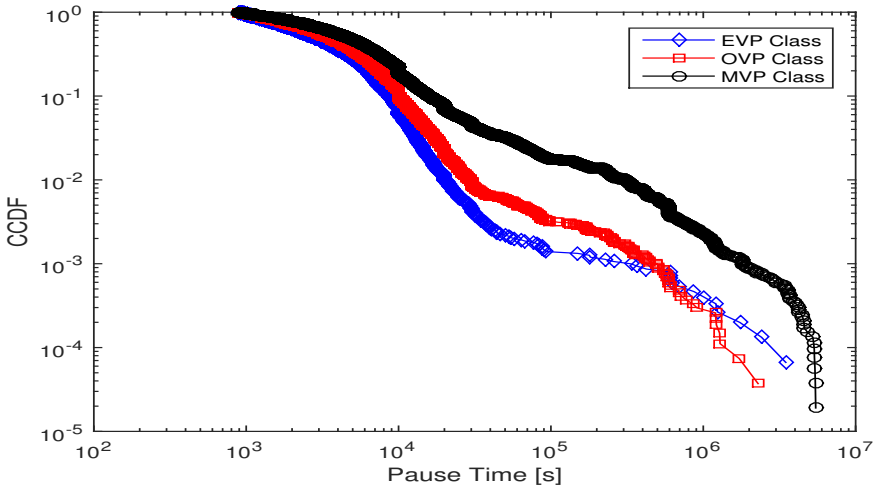


Figure 5.4. *The empirical Pause Time distributions in different classes of Dartmouth WiFi dataset*

extracted from the Group 3 of CDR-67 (see Section 3.2) as well as the transition matrix $\mathbb{T}_{CDR-67}$.

- *GPS scenario.* The number of mobile nodes in this simulation setting is 5000 and the input distributions are extracted from Microsoft GPS dataset distributions (see Section 3.2) as well as the transition matrix $\mathbb{T}_{GPS}$.

- *WiFi scenario.* The number of mobile nodes and PoIs in this simulation setting are 3224 and 596 (extracted from Group 3), respectively. The input distributions are extracted from Dartmouth WiFi dataset with three months duration (see Section 3.2) as well as the transition matrix $\mathbb{T}_{WiFi}$.

In the first scenario(CDR-67), the simulation area is a disk with the radius of 6 km (almost corresponding to the whole Milan's area). PoIs correspond to the 2634 cells (belong to the Group 3) and are uniformly distributed in the city area according to Disk Point Picking [1]. In the second scenario there are 672 PoIs (extracted from GPS dataset) and in the third scenario 596 (extracted from 3 months WiFi dataset-Group 3).
The transition times among visited PoIs in the first scenario are calculated according to equation (5.6) while in second and third scenarios could be directly extracted from datasets, as indicated in Figure. 5.1 and Figure. 5.2 for GPS and WiFi datasets.

For GPS and WiFi scenarios, pause times can be drawn from the corresponding empirical distributions depicted in Figure. 5.3 and 5.4. As pause times are not available in CDR dataset, we use the GPS pause time distribution to model them. In all scenarios, all users start moving from their home places (the PoI with the highest relevance for the user).

We evaluate the goodness of the mobility model by comparing empirical and synthetic traces onto the two main metrics able to account for metropolitan mobility. The first one is the relevance ratio as it describes the relationship of each individual with the Point of Interest of the city. The second one is the colocation which accounts for the relationship between people sociality and city places.

We compare our model with a gravity model adapted for an urban scenario [61] and a modified version of the SLAW mobility model which only implements the least action trip planning algorithm, i.e. it prefers closer PoIs. The gravity model was implemented by setting the probability of transition between two PoIs i and j, $p_{ij}(u)$, proportional to $\frac{m_i * m_j}{d_{ij}^{\beta}}$, where m_i is the mass of the PoI i and d_{ij} is the geographical distance between PoIs i and j. We define the mass m_i of a PoI i as the number of PoIs falling within a circle of radius 2 km centered on i.

5.3.2 Relevance

In Figures 5.5, 5.6 and 5.7 the relevance ratio distributions in different relevance classes, extracted from CDR-67, GPS Microsoft, Dartmouth WiFi datasets are compared with the corresponding distributions extracted from our model in the three scenarios, respectively. We observe that the relevance ratio distributions of the synthetic traces well follow the distributions of the collected datasets in each relevance class for all scenarios, especially in MVP class. It implies that the proposed mobility model realizes the spatio-temporal regularity of visiting PoIs that has been observed in collected datasets.

5.3.3 Colocation duration

The colocation duration is a good index to validate the generated synthetic traces against real ones. It represents the time period in which a pair of mobile nodes is colocated, i.e. they are both in the same PoI. In order to get an estimation of colocation duration in CDR datasets, we assume that mobile nodes are steady under the coverage area of the same cell for a time period lasting T_h second before and after each contact activity time stamp. For instance, if t_j^k

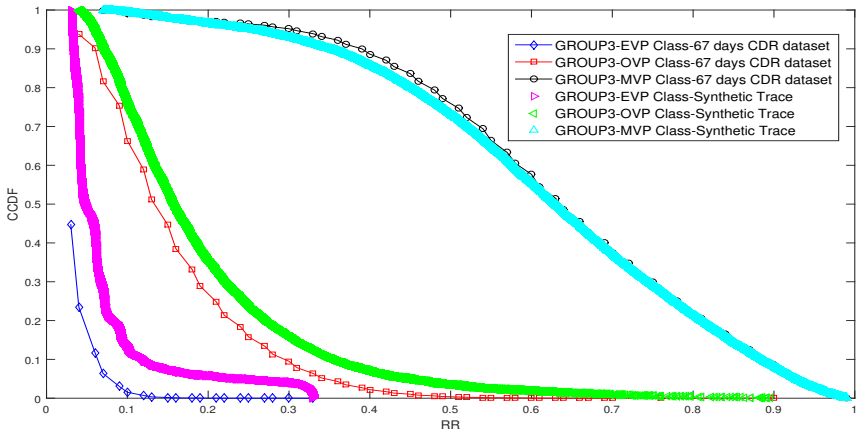


Figure 5.5. *Compare the empirical CCDF distributions of relevance ratio in different classes among 67 days CDR dataset and Metropolitan Synthetic trace in the first scenario.*

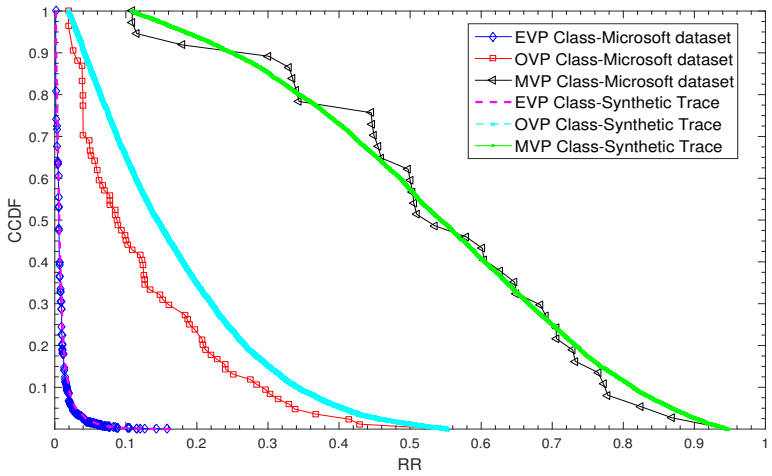


Figure 5.6. *Compare the empirical CCDF distributions of relevance ratio in different classes of Microsoft GPS dataset and Synthetic trace in the second scenario.*

and Td_j^k are the initial time stamp and the call duration of the k^{th} activity of user j in cell c, then we assume that the mobile node j is available under cell c, at least within the time interval $[t_j^k\text{-}T_h/2\ ,\ T_h/2{+}t_j^k{+}Td_j^k]$. For messaging and Internet activities $Td_j^k = 0$ holds and in general we set $T_h = 1800$ seconds.

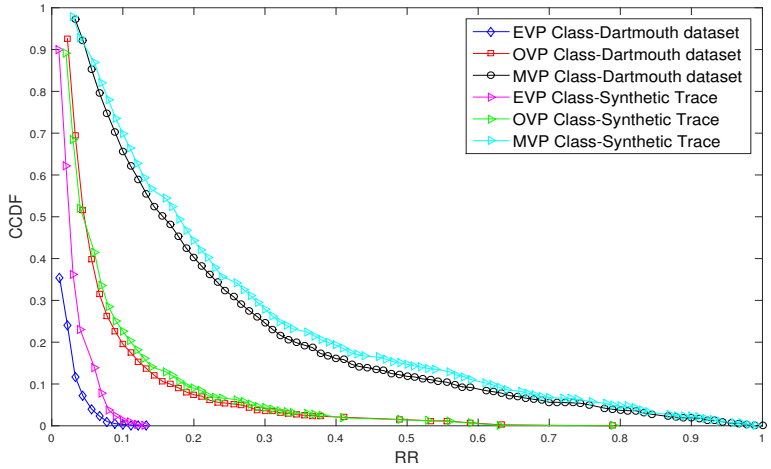


Figure 5.7. *Compare the empirical CCDF distributions of relevance ratio in different classes of Dartmouth WiFi dataset and Synthetic trace in the third scenario.*

The comparison between empirical and synthetic colocation duration distributions are depicted in Figures 5.8, 5.9 and 5.10. In all scenarios, despite their relative diversity, the colocation patterns are well reproduce.

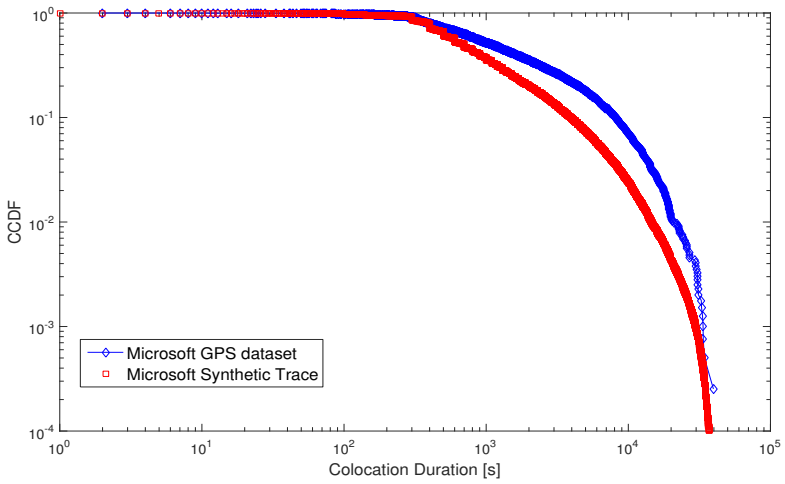


Figure 5.8. *Compare the empirical CCDF distributions of colocation duration of Microsoft GPS dataset and Synthetic trace in second scenario.*

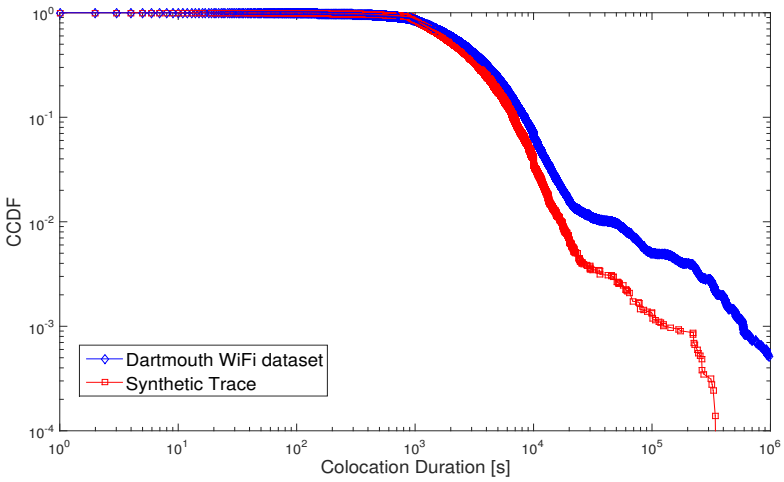


Figure 5.9. *Compare the empirical CCDF distributions of colocation duration of WiFi Dartmouth dataset and Synthetic trace in third scenario.*

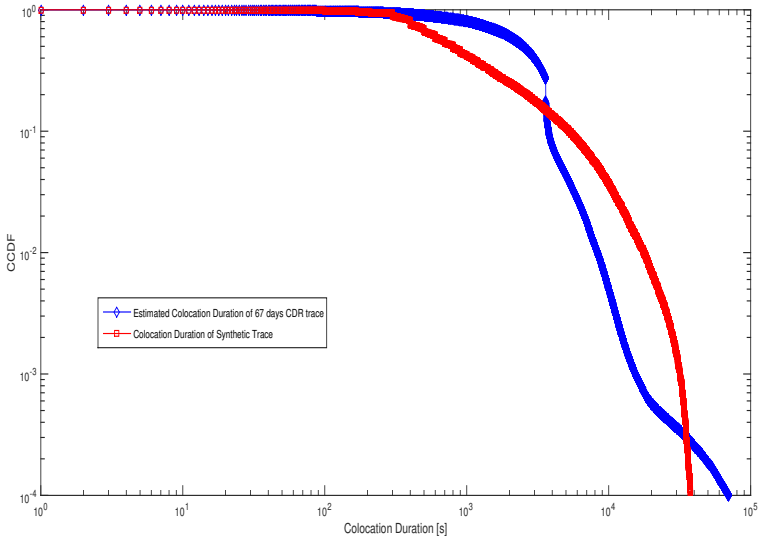


Figure 5.10. *Estimated colocation duration distribution of CDR dataset compared with the distribution extracted from Synthetic trace in first scenario.*

In order to quantify the similarity between the empirical and the synthetic distributions, we adopt the Hellinger distance [85, 52], which measures the dis-

tance between two probability distributions. If $P = (p_1, p_2, ..., p_k)$ and $Q = (q_1, q_2, ..., q_k)$ are two probability distributions, the log scale weighted Hellinger distance is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\frac{1}{\log(i+1)} \sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5.7}$$

The calculated weighted Hellinger distances between the distributions reported in Figures 5.8, 5.9 and 5.10 are 0.1083, 0.0554 and 0.1837 (Table 5.1), respectively. We observe that the fitness in the second (WiFi dataset) and third (GPS dataset) scenario is better than in the first scenario (CDR dataset) due to the high temporal sparsity of CDR dataset.

In Figure 5.11 the colocation duration distributions extracted from dataset and synthetic trace are compared with the gravity ($\beta = 1$) and SLAW (with distance decay coefficient $\alpha = 0.75$) mobility models in third scenario (Microsoft GPS). We observed that the colocation duration distribution reproduced by our model is closer to the empirical one w.r.t. the Gravity model and Slaw. In fact, the Hellinger distance between the colocation duration distributions extracted from dataset and the Gravity and Slaw mobility models is 0.217.

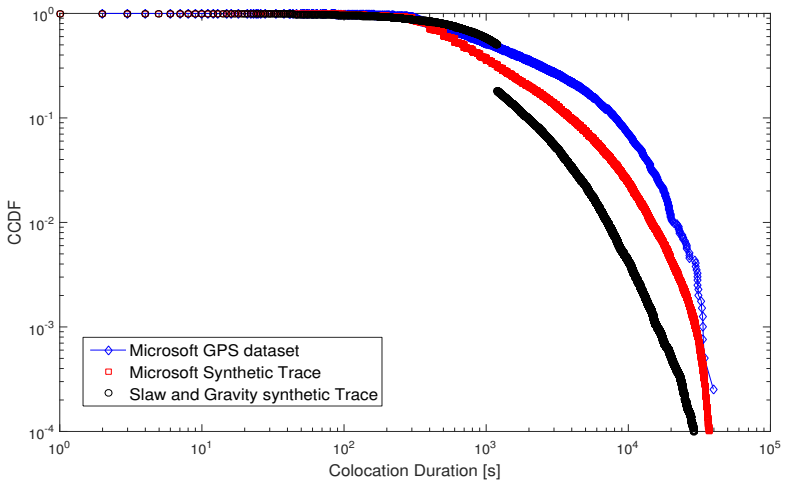


Figure 5.11. *Comparison between the distributions of the colocation duration of Microsoft GPS dataset and those reproduced by the Gravity and Slaw mobility models in the third scenario. The distributions from the Gravity and Slaw models almost perfectly overlap and present a discontinuity at about 1000 sec.*

5.3.4 Inter colocation time

Inter colocation time is the elapsed time between two consecutive colocation events of a pair of mobile nodes. Fig. 5.12 depicts Inter colocation time distribution extracted from GPS Microsoft dataset and synthetic Metropolitan traces in the second scenario. In Figure 5.12 and 5.14 the comparison is shown for the GPS and CDR datasets, respectively. In both GPS and CDR scenarios the model very well reproduces inter colocation times, except for the very upper tail of the distribution while the accordance is limited when the WiFi scenario is considered.

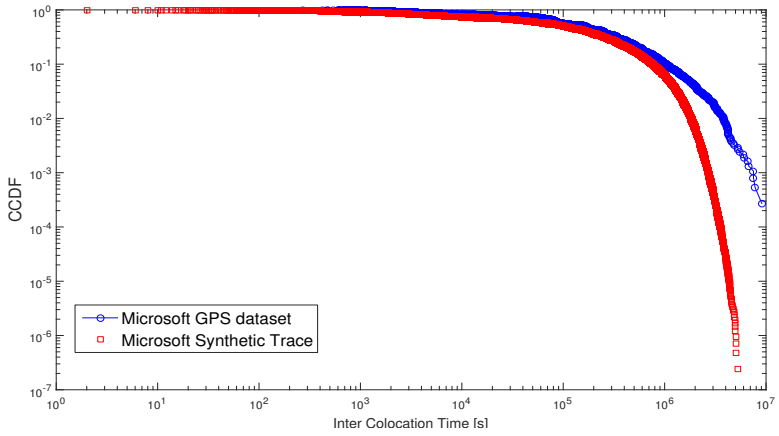


Figure 5.12. *Distributions of Inter colocation time of Microsoft GPS dataset and Synthetic trace.*

The Hellinger distances between the model and the empirical inter colocation distributions (see Figures 5.12, 5.13 and 5.14) are 0.1252, 0.074 and 0.0952 (refer to Table 5.1), respectively.

The above evaluation shows that the synthetic traces well emulate critical connectivity properties, such as colocation duration and inter colocation time of human behavior while moving in a variety of scenarios ranging from campus to urban areas.

Figure 5.15 depicts the inter colocation time distributions from the real, synthetic trace, gravity ($\beta = 1$) and SLAW ($\alpha = 0.75$) traces for the third scenario setting. The Hellinger distances between the distributions from the real traces and the Gravity and Slaw models are 0.1667 and 0.2082, respectively.

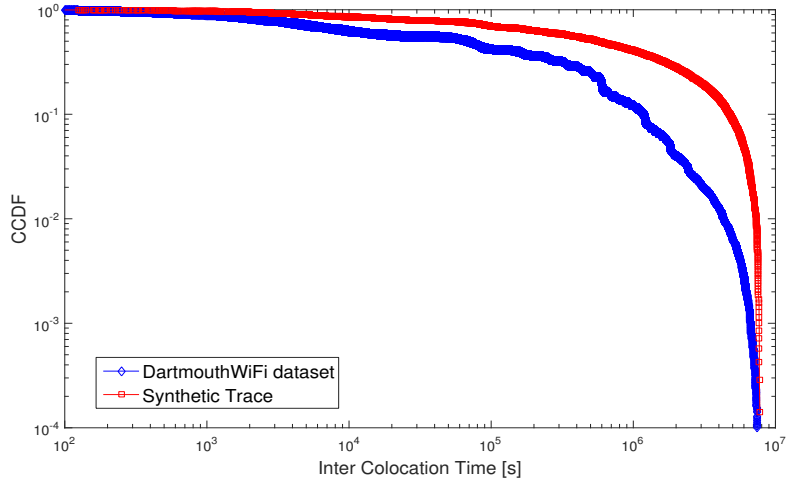


Figure 5.13. *Distributions of Inter colocation time of Dartmouth WiFi dataset and Synthetic trace.*

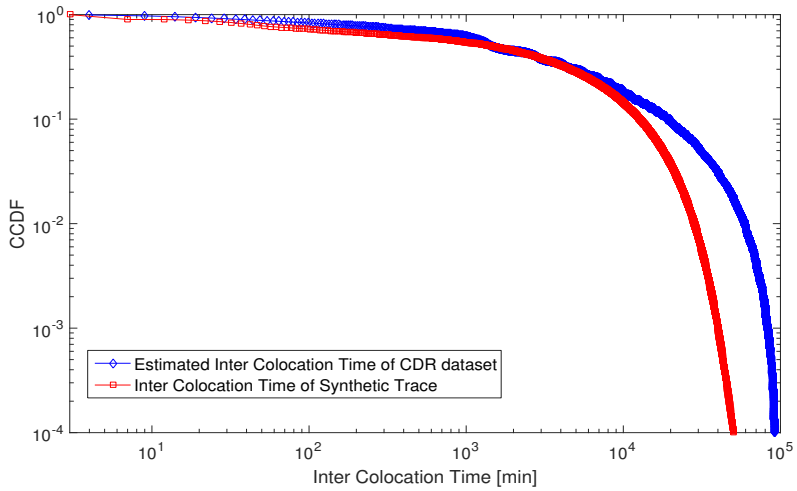


Figure 5.14. *Estimated Inter colocation time distribution of CDR dataset compared with the distribution extracted from the Synthetic trace in the first scenario.*

Table 5.1 indicates the summery of the calculated Hellinger distances between real datasets and synthetic traces in different scenarios.

In general, we observed that the distribution reproduced by our synthetic model is closer, in terms of Hellinger distance, to the empirical one; extracted

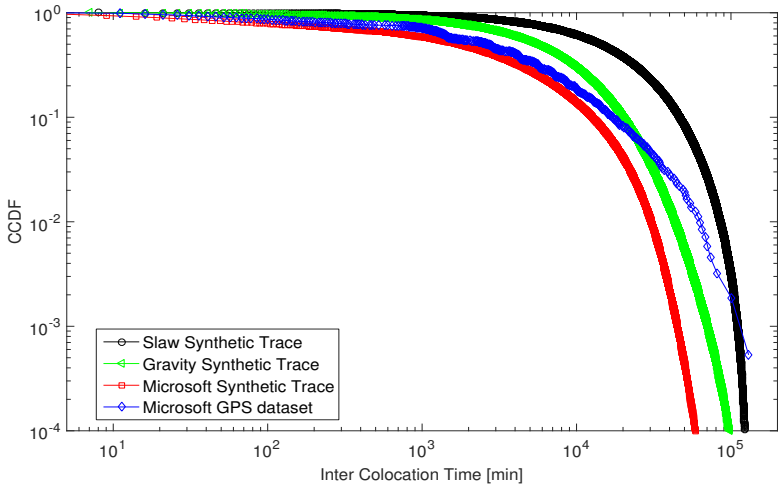


Figure 5.15. *Comparison of the distributions of Inter colocation time of Microsoft GPS dataset with synthetic Microsoft, Gravity and Slaw traces in the third scenario.*

Table 5.1. *Hellinger distance between Real datasets and Synthetic traces*

Model	**Datasets**	**Hellinger Distance**	
		Colocation Duration	**Inter colocation Time**
Relevance Base	GPS	0.108	0.095
	WiFi	0.055	0.074
	CDR-67	0.184	0.125
Slaw	GPS	0.217	0.208
Gravity	GPS	0.217	0.167

form dataset, than the distributions returned by the gravity and the SLAW mobility models.

Chapter 6

Encounter and Colocation Prediction

The previously described activities on mobility modeling and analysis of human mobility patterns are at the heart of a research activity aiming to predict encounter and colocation events when people are moving in the urban area.

In situations where continuous and direct observation of human mobility is difficult or even against privacy rights, knowledge about the usage of WiFi networks collected through Access Points (APs) and also CDR datasets of cellular networks can be used to perform an analysis of the encounter and colocation events among users. Taking into account the coarser temporal and spatial granularity of CDR w.r.t. WiFi datasets, we therefore find a significant difference in the spatial granularity for these two cases. So we use the encounter and colocation events in case of WiFi and CDR datasets, respectively.

The forecasting of the occurrence of this kind of event has great impact on the design of delay tolerant and opportunistic networks and may lead to achieving high efficiency in performing routing and data forwarding activities [38, 16, 15]. In a scenario of high dynamics and intermittent radio connectivity the awareness about the approximate location, the time duration of an encounter or colocation event and people who involved in these events goes further system implications paving the way for novel applications in a variety of fields, including commercial ADs, recommendation systems, and mobile social networks.

6.1 Related Work

Pattern recognition and prediction are closely related tasks since human movement cues are usually periodic and/or repetitive [18, 29, 42]. Therefore repetitive encounter and colocation events can be learned and predicted reliably as long

as we could collect enough observation data from smartphones and other mobile devices.

One challenge for using the WiFi and Cellular datasets described in chapter 2 for prediction (especially PoI prediction) is that PoIs are represented by ID (symbolic place) without any coordinates. Therefore some existing prediction scheme such as [56, 72, 80] would not be applicable. So our used datasets do not support the arithmetic or logic operation, which is usually are used to process GPS coordinates for location prediction.

Peddemors et al. [64] propose an approach based on the prediction of the time of next occurrence of an event of interest, such as arrival time to a certain place with a focus on the prediction of network visibility events as observed through the wireless network interfaces of mobile devices. Their approach is based on a predictor that analyses the events stream for forecasting context changes. The authors found that including predictors of infrequently occurring events can improve the prediction performance.

Gao et al. in [25] proposed a location predictor model that captures the spatio-temporal contexts of the visited places. They exploited a smoothing technique in the training of spatio-temporal model, to avoid the over-fitting problem due to a large number of spatio-temporal trajectory patterns. They assume that temporal features (day of the week and hour of the visit of a place) to be independent, and estimate the distributions of the day of the week and the visiting hour by Gaussian distributions.

In contrast to the wide range of future place prediction works [78, 6, 24, 53, 27] relying on Markov chain, needed to keep track n previous visited places, our proposed approach just need temporal context as an inquiry for predicting PoIs.

The Next Place predictor method in [72] captures the concurrent temporal periodicity of mobile users when they visit their most important places. This spatio-temporal predictor relies on a non-linear time series analysis of the arrival time and on pause time durations of users in their most relevant places. This predictor, besides predicting the arrival time to the next place and its stay duration, is also able to predict the interval time between two subsequent visits to the predicted place. This approach has been only applied to the most important visited places and needs a large amount of data to constitute time series. Due to these requirements, its application is limited just to the most frequently visited places (home and workplaces).

In [82] a probabilistic kernel method for visited places prediction using spatio-temporal information via multiple kernel functions are presented. The kernel density estimation is a smoothing technique for sparse data collected

by smartphones. Even though this approach exploits i.i.d assumption among spatio-temporal context in Bayesian predictor, it has obtained good accuracy in predicting the next visiting places just for the next few hours.

Authors in [88], by analyzing MIT Reality Mining CDR dataset, have observed a strong correlation between calling pattern and colocation patterns of mobile users. By exploiting this social interplay on top of user periodic behaviors, they proposed a self-adjuster symbolic predictor which combines the output of social interplay and periodicity predictors to estimate the next cell to visit. Although authors only used calls pattern, they achieve higher prediction accuracy than the other state of the art schemes at cell tower level. Considering that MIT Reality Mining CDR has been collected in 2004 and during that period definitely calls were dominated contact activities among people, so it makes sense if authors exploited call activities for capturing social interplay among participants in their experiment while nowadays the majority of contact actives among people and friends have been oriented towards the wide variety of Internet-based applications. However, such Internet-based contact activities somehow will be hidden from CDR datasets, since when a user accesses to the Internet just the Internet traffic data will be recorded in CDR. Therefore nowadays we should be conservative about extracting social interplay among mobile users just relying on calls and even text SMS.

Authors in [16, 17] according to their observations of the time of encounters (nodes encounter patterns tend to have centers in the middle of the day), proposed the Gaussian process for predicting the time of encounters and also predicting the number of encounters in the predicted time frame.

In other works [60, 17] authors predicted the user future contacts to be used in a routing protocol to improve the efficiency of message delivery in opportunistic networks.

6.2 Encounter Events

An encounter event in the real world means meeting face to face, which implies physical proximity among people. The extent of this physical proximity is not always exactly clear and may be vary on different scenarios, applications, and domains. For instance, in the biological field and in disease spreading, physical proximity or distances between involving objects are short and even can be considered as direct touching, while in wireless networks it depends on the coverage areas of mobile devices or wireless network infrastructures. Nowadays smartphones and the majority of short-range wireless devices are carried by humans

that can be used to observe mobility and extract physical proximity information. In the communication network literature, an encounter between two mobile devices occurs when they are in communication range of each other or within the same coverage area of WLAN infrastructures that devices are associated with; the latter also called Indirect encounter [89, 50]. Most researchers define an encounter event occurrence in a WLAN when two or more mobile nodes are associated to the same AP during an overlap time interval. This definition may not always reflect proper and exact realistic physical encounters among mobile nodes due to some challenges. For instance, mobile nodes might be physically close to each other but associated to different APs, or they are out of coverage area of APs, or they are involved in ping-pong events or in overlap areas among radio coverage areas of nearby APs [39, 40]. Despite some challenges and limitations, if collected WiFi datasets are used carefully (i.e. accounting for the effects of ping-pong events, overlap in coverage areas and missed encounters) it would appear to be a good source of empirically-derived data on human encounters since large amount of data can be gathered easily at low cost, allowing large-scale analysis of encounter patterns.

Here for WiFi dataset after smoothing the ping-pong events according to [39], we extract encounter events. The resulting record for an encounter event is:

```
UserA,UserB,PoI Id,Encounter Start Time,Encounter End Time.
```

6.3 Colocation Events

Colocation event has been defined in cellular networks among mobile users while they are connected to the same cell tower for an overlap time interval. Taking into account the coarser temporal and spatial granularity of CDR datasets, therefore, there is a significant difference in the spatial range of encounter and colocation events.

Colocation event among a number of people interestingly give rise to a significant amount of in-proximity voice/data traffic on the cellular network and advocate the provisioning of a new class of services supporting it. The analysis and understanding of colocation interactions can inspire the design of new mobile service that can detect physical proximity among people and deploys the mobile social network supporting proximity interaction. In fact, they are supposed to operate as close as possible to their users to ensure better user's experience; by placing the interaction services at the edge of operator's network,

has a payoff in terms of traffic off loading from the core network and decreasing latency such as Direct LTE structure and the next generation mobile radio system(5G). Such structures allow devices to exchange data over short range and possibly on more reliable channels. Such paradigm shift allows not only to save the base station's resources but also to increase the data rate and QoS requirements.

Each colocation event is characterized by a specific time interval and place (PoI). Authors in [88, 11] have characterized the spatio-temporal features of colocation events and observed a reasonable subset of actual face-to-face meetings between users. In order to get an estimation of colocation events and their durations in CDR dataset (note that due to temporal sparsity of CDR dataset, extracting colocation event is not straightforward), we assume that mobile users are steady under the coverage area of the same cell for a time period lasting T_h second before and after each on-phone activity time stamp. For instance, if t_j^k and Td_j^k are the initial time stamp and the call duration of the k^{th} activity of user j in cell c, then we assume that the mobile user j is available under cell c, at least within the time interval [t_j^k-T_h , T_h+t_j^k+Td_j^k]. For messaging activities $Td_j^k = 0$ holds and in general we set $T_h = 900$ seconds.

The resulting record for an colocation event is:

```
UserA,UserB,PoI Id,Colocation Start Time,Colocation End Time.
```

6.4 Prediction Methodology

While predicting the user's events, we seek the answers to three fundamental questions [66]: (1) where will the encounter or colocation event occur for a fixed user at a future time (i.e., PoI)? (2) how long will he/she be encountered or colocated with other users at that PoI (i.e., event duration)? and (3) who will he/she meet (i.e., encounter/colocation contacts)?

In this work, our goal is to predict the places (PoIs) where a user will experience an encounter or colocation event, give an estimate of its duration and people are involved. We assume no a prior knowledge on the temporal relation between encounter and colocation events.

Focusing on the temporal and the spatial information of encounter and colocation traces, we try to learn the dependencies between these contextual variables with next encounter or colocation events. The temporal context captures regular patterns in the occurrence of these events from the weekly calendar, such as an event occurring at given time of day and day of the week. On the

Table 6.1. *Encounter/Colocation features.*

ϕ	τ		UserID	PoIID	Encountered/Colocated User ID
	τ_s	τ_e			

other hand, the dynamics of these events can be explored through the spatial information. Since daily schedule of people usually is different on weekdays and weekend, to constructing the encounter or colocation predictors, we consider several parameters to efficiency capture multiple aspects of temporal contexts. The temporal context features are: *i*) day time slot, and *ii*) day of the week. The "day time slot" τ is an integer feature and depends on the length H of the time slot, i.e $\lfloor t/H \rfloor$ where $t = 0, \ldots, 23$. We set $H = 2$ hours since it represents a trade-off which offers a good daily resolution and a robustness against small changes in the daily movement routine, e.g. being late for work due to an exceptional traffic jam or little delays in the agenda. The "day of the week" ϕ maps a day of the week to an integer, where Monday is 1 and Sunday is 7. We compute the above features on the encounter and colocation start and end times, so that each record is defined by the user-ids of the mobile users, the place IDs (PoI IDs), ϕ, τ_s for the start time, τ_e for the end time (see Table 6.1).

We adopt a per-user perspective, i.e. for each mobile u we will make a prediction relying on her/is context history, only. The predictor, trained on the event records having User ID equal to u, will accept as input the tuple $T = (\phi, \tau)$ and will return the place (PoI ID) where the encounter or colocation event will occur, its duration and the users who u will meet during the temporal context T.

6.5 Predictive Model

The naïve Bayesian classifier is one of the most common classification techniques. Naïve Bayesian classifiers are based on the Bayes' theorem with naïve independence assumption between the features and apply a decision rule, known as Max a Posteriori or MAP decision rule, which selects the hypothesis with the highest probability. In this work, similarly to other recent works on mobility prediction [82, 27, 23, 15], we use Bayesian classifiers for encounter and colocation prediction. Besides its simplicity and being fast compared to other classifiers, it can be trained with a few observation records; especially in our case that encounter and colocation traces are coarse and sporadic, and still achieve reliable

results.

6.5.1 Encounter and Colocation Place(PoI ID) prediction

Most of the recent location-based services are based on the knowledge of the current or future place of the mobile user. For instance, by exploiting the future visiting places, we can access to the information such as nearby PoIs or available services. For prediction of the encounter or colocation PoI, we consider the conditional probabilistic of a PoI with ID, $L = l$ given the temporal context $T = (t_1 = \phi, t_2 = \tau)$.

Under independence assumption:

$$P\left(L = l \mid T\right) \propto P\left(L = l\right) \prod_{k=1}^{2} P(t_k \mid l) \tag{6.1}$$

By exploiting the MAP decision rule

$$l_k = argmax_j(P(\phi = \phi_k \mid l_j)P(\tau = \tau_k \mid l_j)P(L = l_j)) \tag{6.2}$$

Applying the standard formulation of a naïve Bayesian classifier poses some problems due to the conditional independence assumption. In human mobility context, for example, people have a different schedule on weekdays and weekend, i.e a person may visit different places on weekdays and weekends during the same time slot. In this case, the independent assumption for ϕ and τ would be violated. So we used a weighted naïve Bayesian technique to alleviate the independence assumption issue, based on the Kullback-Leibler divergence [48]. The Kullback-Liebler measure for feature a and class label c is defined as:

$$\mathfrak{KL}(C \mid a) = \textstyle\sum_c P\left(c \mid a\right) log(\frac{P(c|a)}{P(c)}) \tag{6.3}$$

Where $\mathfrak{KL}(C \mid a)$ is the average mutual information between the class event c and the feature value a with expectation taken with respect to a posteriori probability distribution of C. This can be considered as an asymmetric information theoretic similarity between two probability distributions, which measures how dissimilar is a priori and a posteriori. This distance measure corresponds to the amount of divergence between a priori distribution and a posteriori distribution. The weight of a feature can be defined as a weighted average of the $\mathfrak{KL}$ across the feature values. The features weighting improves the performances of the naïve Bayesian classifier since it relaxes the independence among the context features.

The introduction of the features weights, results in the following formulation of the predictor:

$$l_k = argmax_j(P(\phi = \phi_k \mid l_j)^{w_{\phi,L}} P(\tau = \tau_k \mid l_j)^{w_{\tau,L}} P(L = l_j)) \tag{6.4}$$

Where $w_{\phi,L}$ and $w_{\tau,L}$ are average feature weights of ϕ and τ calculated for PoIs label in training set according to the [48]. The weights are shared over all users in the training set. Finally, since the naïve Bayesian model return the probability $P(L = l|T)$, we can retrieve the p-most likely places (PoI IDs) l_k given the temporal context $T = (\phi = \phi_k, \tau = \tau_k)$.

6.5.2 Encounter Duration Predictor

The encounter duration predictor estimates how long the encounter event at the predicted PoI will last. Indeed, the predictor depends not only on the temporal context T but also on the outcome of the PoI predictor, i.e. l_p. In this setting we aim at finding the duration d_j which maximizes $P(D|T, L = l_p)$, i.e.

$$d_k = argmax_j \left(P(\phi = \phi_k, \tau = \tau_k, l = l_k \mid d_j) P(D = d_j)\right) \tag{6.5}$$

In above equation addition to the temporal features, the predicted PoI also considered as a spatial feature. By applying the feature weighting for naïve Bayesian classifiers we obtain:

$$\begin{aligned} d_k = argmax_j(&P(\phi = \phi_k \mid d_j)^{w_{\phi,D}} P(\tau = \tau_k \mid d_j)^{w_{\tau,D}} P(L = l_k \mid d_j)^{w_{l,D}} \\ &P(D = d_j)) \end{aligned} \tag{6.6}$$

Where $w_{\phi,D}$ and $w_{\tau,D}$ are average features weights of ϕ and τ, and $w_{l,D}$ is the spatial weight related to the predicted PoI. So the event duration predictor will learn a function whose input is the tuple $(\phi = \phi_k, \tau = \tau_k, l = l_k)$.

Encounter duration among mobile users in a specific PoI, due to dynamic nature of their movements, varies in time. As a consequence, the prediction of the duration is not straightforward. On the other hand since most of the people follow daily schedule tasks, their encounter durations in a specific PoI are not always same, but we expect that the variation of duration lies in a limited range. These observations reflect on how we evaluate the accuracy for the encounter duration event. We extract from the test set P_k, the set of encounter durations occurring in PoI l_k for the temporal context $\{\phi = \phi_k, \tau = \tau_k\}$. After removing outlier durations by using skewness [31]; we obtain $P_k = \{pt_1, pt_2, \cdots, pt_{|P_k|}\}$,

where $|P_k|$ is the size of P_k. Then we compute the average μ_k and the standard deviation σ_k on the set P_k. If the predicted duration for the temporal context $\{\phi = \phi_k, \tau = \tau_k, l = l_k\}$ lies in the interval $[\mu_k - \sigma_k, \mu_k + \sigma_k]$, we assume the event duration prediction to be correct.

6.5.3 Encounter and Colocation Contacts Predictor

Because of the critical role of predicting future encounter and colocation events in content delivery and routing protocols in opportunistic and delay tolerant network [38, 87], in this section we focus on predicting event contacts (whom a user will meet) in specific period $\{\phi = \phi_k, \tau = \tau_k\}$.

$$c_k = argmax_j \left(P(\phi = \phi_k, \tau = \tau_k \mid c_j) P(C = c_j)\right) \tag{6.7}$$

Since the set of people met by a mobile user may change between weekdays and weekend even during the same time slot, we alleviate the conditional independence assumption between temporal contexts of $\phi = \phi_k$ and $\tau = \tau_k$ by features weighting:

$$c_k = argmax_j (P(\phi = \phi_k \mid c_j)^{w_{\phi,C}} P(\tau = \tau_k \mid c_j)^{w_{\tau,C}} P(C = c_j)) \tag{6.8}$$

Where $w_{\phi,C}$ and $w_{\tau,C}$ are average feature weights of ϕ and τ, are calculated for contacts label; respectively, according to [48].

The above-calculated weights are shared over all users in the training set. Above equation retrieves most likely User IDs, who have been encountered or colocated with under study mobile user in day ϕ_k and time slot τ_k.

6.6 Evaluating Prediction Accuracy

In this section, we evaluate the goodness of the Bayesian predictors with weighted features on the encounter and colocation traces separately due to the significant difference in the spatio-temporal granularity of WiFi and CDR datasets which results in the prominent difference in the spatial granularity range of encounter and colocation events.

In each bellow subsection we train and evaluate the Bayesian classifier with weighted features and also the standard naïve Bayesian classifier for each user, separately. This way we obtain a set of accuracy values, whose distribution captures the performance of the approach for a specific task. Moreover, since the Bayesian classifier can return the most p likely items, we report the results of

the evaluation for $p = 1, 2, 3$. We conducted the evaluation by using 4-fold cross validation and the average accuracy as a performance metric. The accuracy prediction rate per user is defined as the number of correct prediction over the total observation records for that user in the test set.

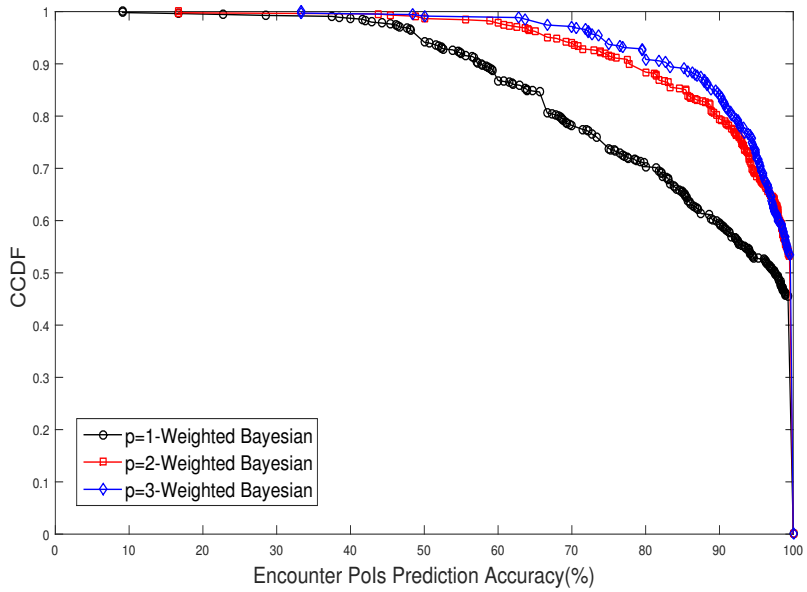


Figure 6.1. *The empirical CCDF distributions of encounter PoI prediction accuracy of weighted features Bayesian predictor for* $p = 1, 2, 3$.

6.6.1 Evaluating Encounter Prediction Accuracy

For total four months duration of encounter trace (extracted from Dartmouth WiFi dataset, collected from Jan 3rd to April 30rd, 2004, includes 17414 mobile users and 1292 APs) and also only choosing users with at least 75 encounter event records to have enough records for training classifier, we run predictor four times, each time by considering 3 months as training set and one month as a test set.

In this section, the accuracies of predicted PoIs, durations and also contacts of encounter events per user will be analyzed.

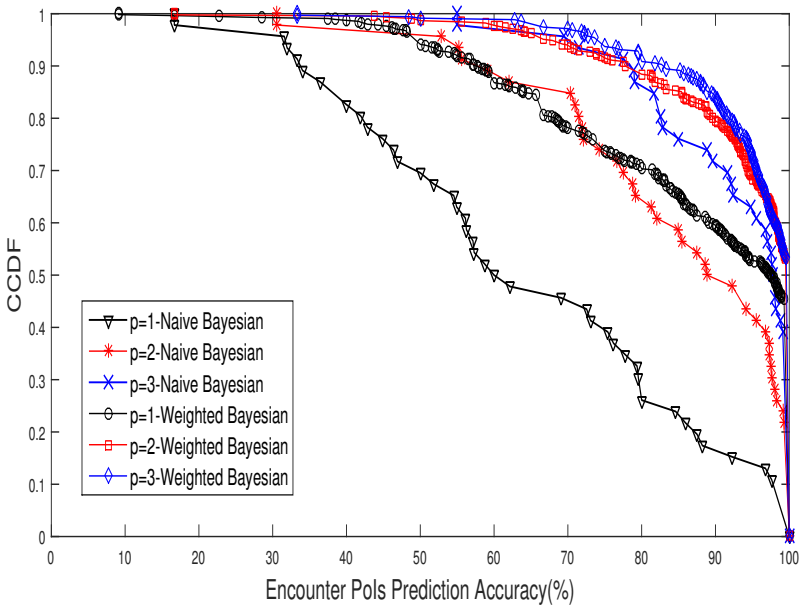


Figure 6.2. *Comparing the empirical CCDF distributions of encounter PoI prediction accuracy for weighted features and standard naïve Bayesian predictors.*

Encounter PoIs Predictor

In Figure 6.1, predicting the first, two and also three most probable PoIs, we observe sensible accuracy rate for case $p = 3$ that is higher than the case $p = 2$ and also the case $p = 1$, since by increasing p, we enlarge the prediction set of PoIs and the probability that the prediction set will contain the correct PoI. When $p = 3$, more than 90% of mobile users have more than 80% accuracy in predicting PoIs of encounter events, compare to the case $p = 1$ that around 70% of users have more than 80% accuracy in the predicted PoIs.

In Figure 6.2, the distributions of weighted and standard naïve Bayesian encounter PoI prediction (per user) are compared. We observe a pronounced improvement in the PoIs accuracy prediction especially for $p = 1$ and $p = 2$ by exploiting the weighted Bayesian predictor because of relaxing the conditional independence assumption between temporal contexts.
Comparing to the location prediction accuracy have been reported in [25] (with best prediction accuracy rate 50 %) and also the overall precision results of Next Cell in [88], our PoI predicted accuracy rates even for p=1 are significantly higher.

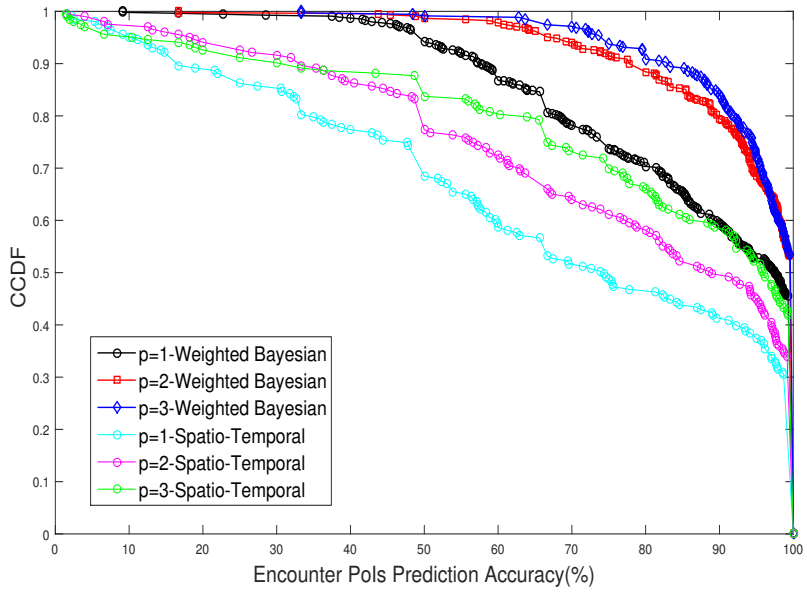


Figure 6.3. *Comparing the empirical CCDF distributions of encounter PoIs prediction accuracy for weighted features Bayesian and spatio-temporal based predictors.*

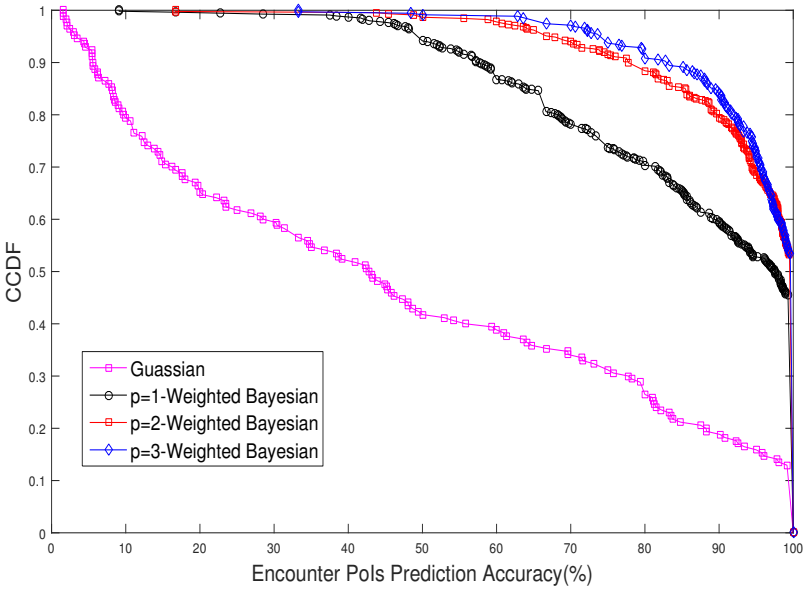


Figure 6.4. *Comparing the empirical CCDF distributions of encounter PoIs prediction accuracy for weighted features Bayesian and Gaussian process predictors.*

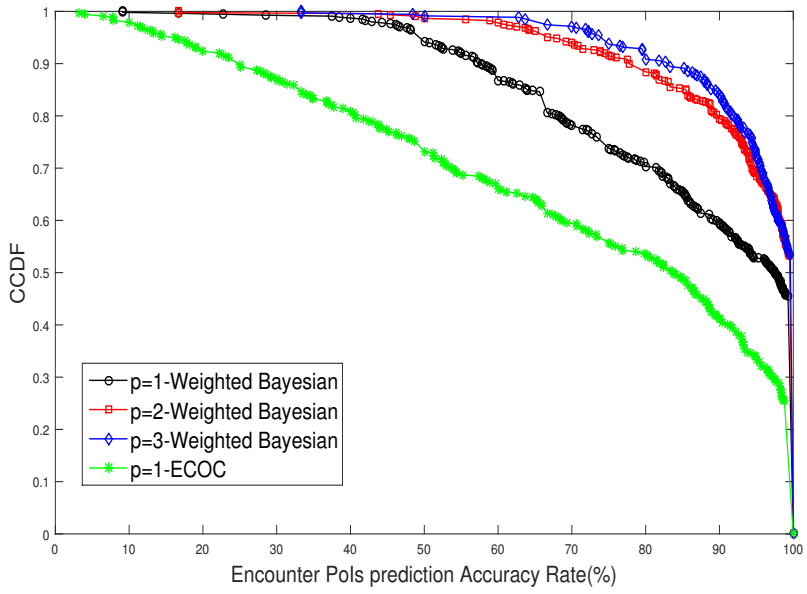


Figure 6.5. *Comparing the empirical CCDF distributions of encounter PoIs prediction accuracy for weighted Bayesian and ECOC predictors.*

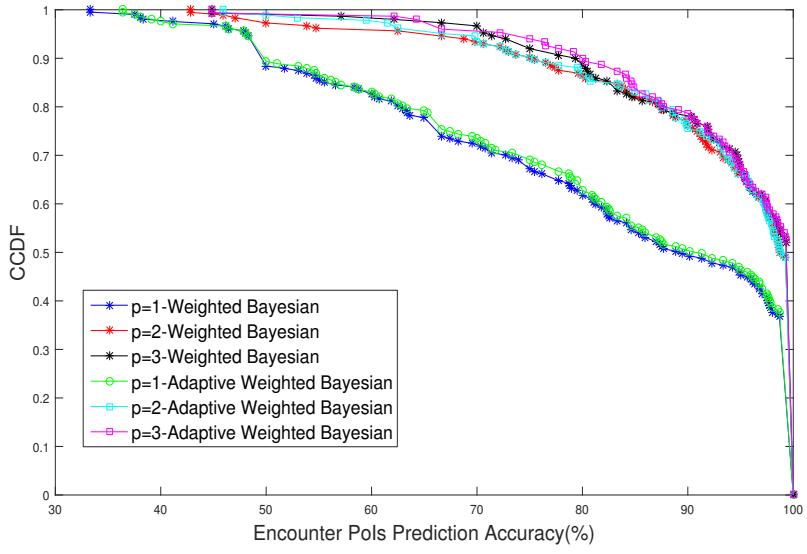


Figure 6.6. *Comparing the empirical CCDF distributions of PoIs prediction accuracy for weighted features and Adaptive weighted features Bayesian predictors.*

We also compared our approach with other algorithms: Gao et al. [25] and Ciobanu et al.[16]. The distributions of spatio-temporal location prediction in [25] which is based on i.i.d (independent and identical distributions) assumption about spatio-temporal contexts and also relying on daily and hourly Gaussian distributions of temporal context are compared with our approach and depicted in Figure 6.3 which confirms that our approach outperforms this predictor in terms of accuracy for $p = 1, 2, 3$, significantly.

In another experiment, our approach has been compared with the work of [16] that Gaussian process exploited for prediction of encounters. The comparison among the distributions of Gaussian process (with constant mean function and covariance ARD) and weighted features Bayesian predictor in terms of accuracy have been indicated in Figure 6.4. We can observe that even for $p = 1$, our approach outperforms the Gaussian approach in terms of accuracy.

We also compared our weighted features Bayesian predictor with ECOC (Error-Correcting Output Code) predictor that is the multi-class version of SVM classifier. Again in Figure 6.5, we can observe that even for $p = 1$, our approach outperforms the ECOC predictor.

Finally, as mentioned in section 6.5, the calculated features weights are shared and averaged over all users in the training set. In case that for each user the features weights are calculated independently according to those records in the training set that just belong to that individual user, we will have Adaptive weighted Bayesian predictor; means that features weights will be different for each user. Figure 6.6 depicts the comparison among distributions of PoIs prediction accuracy for weighted and Adaptive weighted Bayesian predictors. We observe some minor improvements in prediction accuracy for the case of adaptive features weights.

Encounter Duration Predictor

For encounter duration as discussed in section 6.5.2, if the predicted duration for the temporal context $\{\phi = \phi_k, \tau = \tau_k, l = l_k\}$ lies in the interval $[\mu_k - \sigma_k, \mu_k + \sigma_k]$, we assume the event duration prediction to be correct. The encounter duration accuracy prediction is defined as the number of correct prediction over the total observation records for that user in the test set. By choosing the average prediction accuracy rate for users in all runs, we have achieved the accuracy distributions for encounter duration for $p = 1, 2, 3$, depicted in Figure 6.7.

In Figure 6.7 the distributions of encounter duration for weighted and naïve Bayesian predictors are compared. We observe almost a pronounce enhancement in predicting encounter duration in terms of accuracy for $p = 1, 2, 3$ in the case

of weighted Bayesian predictor rather than standard Bayesian naïve although in general the accuracy rate is not high.

Encounter Contacts Predictor

For encounter contacts if one of the predicted contacts (User IDs) be match with the User ID in test set for a given temporal context, the contact prediction will be correct. The encounter contact accuracy prediction is defined as the number of correct prediction over the total observation records for that user in the test set.

The distributions of encounter contacts accuracy (for weighted features Bayesian) for $p = 1, 2, 3$ per user are depicted in Figure 6.8. We observe pronounce differences of accuracy (per user) for $p = 1$ and $p = 3$. For $p = 1$ just almost 20% of mobile users, have accuracy rate more than 80% for predicted encounter contacts while this percentage for $p = 3$ is more than 35%.

In Figure 6.9 the distributions of encounter contacts accuracy for $p = 1, 2, 3$ for both cases of weighted features and standard naïve Bayesian predictors are depicted. The enhancements of the prediction accuracies are almost significant in the weighted features cases for $p = 1, 2$.

6.6.2 Evaluating Colocation Prediction Accuracy

For total 67 days duration of colocation trace (extracted from CDR dataset) and also only choosing users with at least 350 colocation events records to have enough records for training set, we run predictor four times, each time by considering 50 days as training set and remaining as a test set. The accuracy predictor rate per user is defined as the number of correct prediction over the total observation records for that user in the test set. Through choosing average prediction accuracy rate for users in all runs we have achieved the accuracy distributions for $p = 1, 2, 3$ in different aspects. In this section, the accuracies of predicted PoIs, and also contacts of colocation events per user will be analyzed.

Colocation PoIs Predictor

In Fig 6.10 predicting first, two and also three most probable colocation PoIs, similar to the encounter PoI prediction in the previous section, we observe sensible accuracy rate for case $p = 3$ that is higher than the case $p = 2$ and also the case $p = 1$. When $p = 3$, around 50% of mobile users have more than 50% correctness in the predicting PoIs of colocation events, compare to the cases $p = 1$ and $p = 2$ that around 15% and 20% of users, respectively, have more than 50%

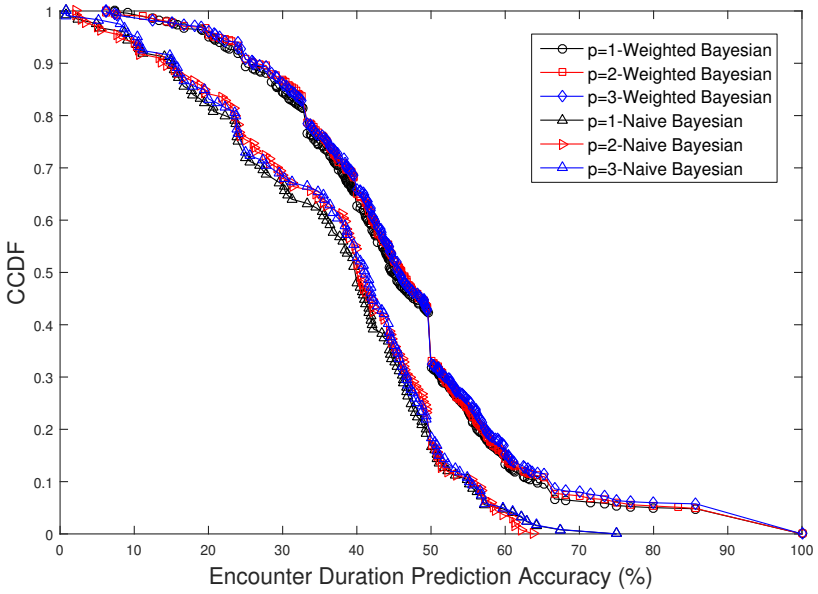


Figure 6.7. *Comparing the empirical CCDF distributions of encounter Duration prediction accuracy for weighted features and standard naïve predictors for* $p = 1, 2, 3$.

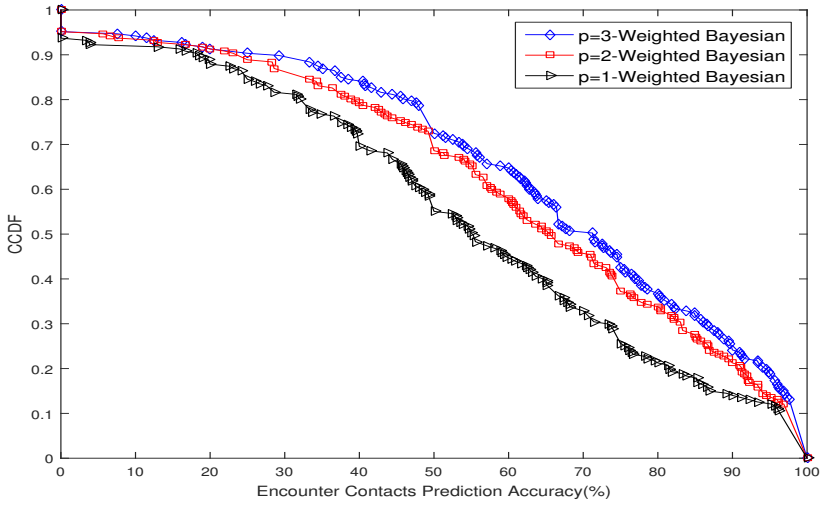


Figure 6.8. *The empirical CCDF distributions of encounter Contacts prediction accuracy of weighted features Bayesian predictor for* $p = 1, 2, 3$.

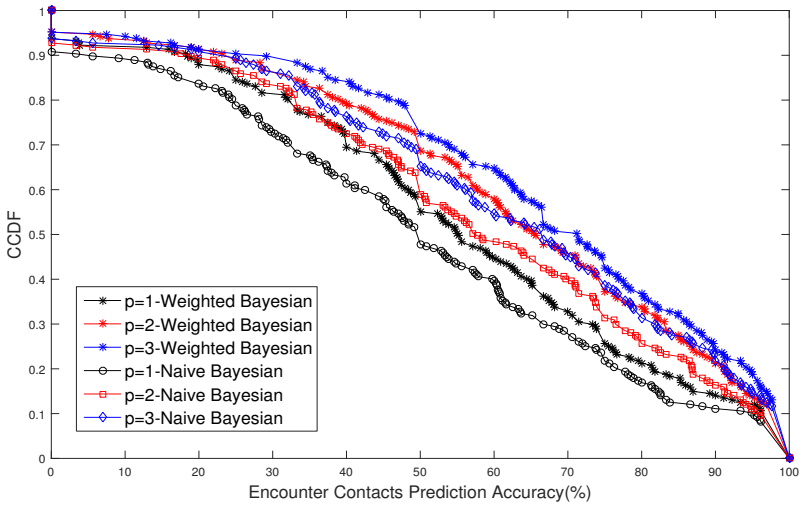


Figure 6.9. *Comparing the empirical CCDF distributions of encounter Contacts prediction accuracy for weighted features and naïve Bayesian predictors.*

correctness in predicted PoIs. We can observe that the percentage of the PoIs accuracy prediction in colocation events is less than the encounter events, most probably due to the spatio-temporal sparsity of CDR dataset rather to the WiFi dataset.

In Figure 6.11, the distributions of weighted features and standard naïve Bayesian PoI colocation prediction (per user) are compared. We can observe a pronounced improvement in the accuracy of PoI prediction especially for $p = 1$ and $p = 2$ through using weighted features Bayesian predictor.

Colocation Contacts Predictor

The distributions of colocation contacts accuracy for $p = 1, 2, 3$ per user are depicted in Fig. 6.12. The collocation contacts prediction accuracy are not comparable with the encounter contacts prediction. Most probably since the number of contacts (mobile users) is more limited in the coverage areas of APs rather than cell towers. On the other hand, colocation trace is suffered by high spatial spareness and temporal coarseness. Moreover, the spatio-temporal regularity of encounter events in campus environments is much more than the colocation events in metropolitan areas.

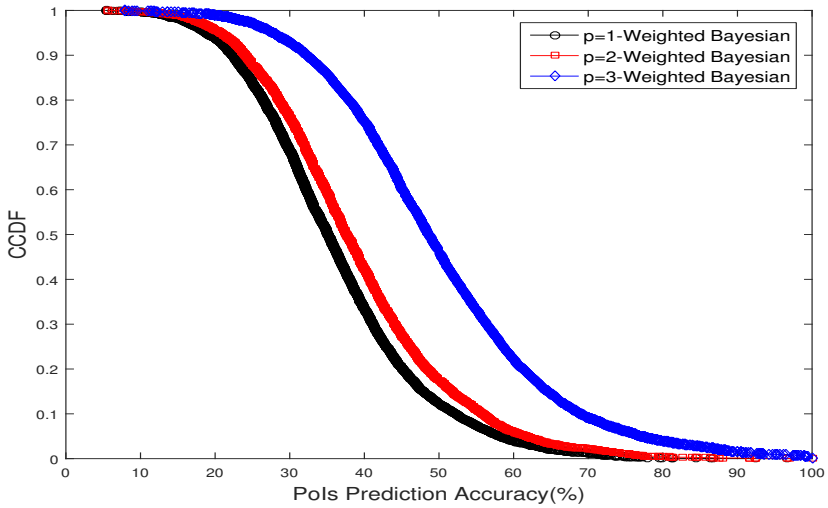


Figure 6.10. *The empirical CCDF distributions of PoIs colocation prediction accuracy of weighted features Bayesian predictor for* $p = 1, 2, 3$.

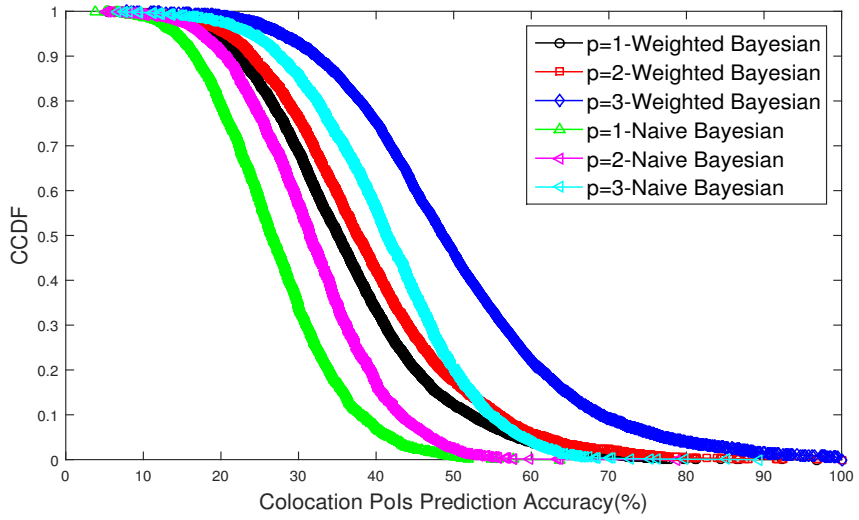


Figure 6.11. *Comparing the empirical CCDF distributions of colocation PoIs prediction accuracy for weighted features and standard naïve Bayesian predictors.*

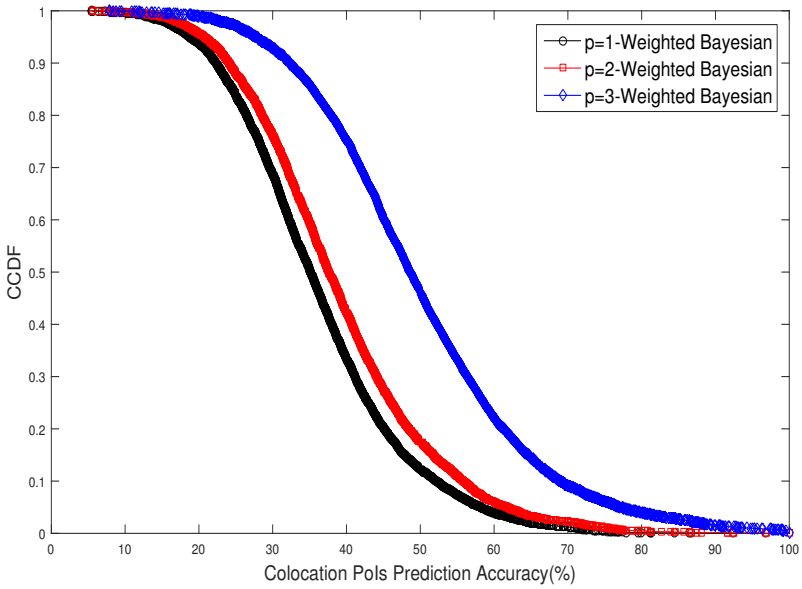


Figure 6.12. *The empirical CCDF distributions of colocation contacts prediction accuracy of weighted features Bayesian predictor for* $p = 1, 2, 3$.

Chapter 7

Conclusion

In this work, we have taken a fresh look at the concept of location. We have proposed a general framework for extracting, characterizing, and classifying the points of interest of each individual according to their relevance for her/him. We have also proposed suitable metrics and algorithms to describe the semantic values of locations and the commuting rules among them. Our key observations are as follows:

- Individuals are regularly drawn to a limited set of locations where they spend most of their time;
- HOME and WORK are among the most frequently visited locations, and, as such, the relevance ratio is a fundamental feature for their semantic identification.

These observations hold true across different datasets with completely different properties.

Based on above observations, we have derived a mobility framework where we are able to classify PoIs, the users and the way they move along PoIs, as well as the semantic meaning of PoIs. We have validated our framework with extensive experimental work.

These novel methods and results can change the way mobility is analyzed and modeled: we argue that, to produce more realistic mobility traces, a mobility model needs to consider (i) the new classifications of PoIs introduced, and (ii) the new features, their relationships and transition patterns and laws among them. Similarly, in localization activity, such laws can enormously simplify the prediction of the next location. Finally, our framework successfully and powerfully combines social and physical characteristics, so it can serve as a basis

for social analysis of mobile complex networks. This can be used, for instance, in Recommendation Systems for Location Based Social Networks [58], where the next location can be recommended based on the class of PoIs that a user has already visited as well as on his/her own social history. According to the observed features, patterns and regularities in different classes, we proposed a metropolitan mobility model whose main axis is the regularity in visiting MVP, OVP, and EVP places. The evaluations depict that spatio-temporal regularity in visiting PoIs and also connectivity properties of human mobility such as colocation duration, and inter colocation time are well realized through this model visually and quantitatively (through Hellinger coefficient) and even outperforms the other state of the art mobility models such as Gravity and Slaw mobility models.

In this work, we also started to cope with the prediction of different aspects of encounter and colocation events, such as PoIs, their durations and also people who involved in the meeting. Being able to forecast these properties by simply leveraging the mobility patterns of the user is one of the crucial aspects in applications directly relying on the human behaviors, e.g. opportunistic networks, online and location-based social networks or epidemiology. Although encounter and colocation events are not always predictable, learning the repetitive patterns and the spatio-temporal regularity in occurrence history of these events help us to improve the prediction accuracy of them from contextual clues collected from smartphones and other portable wireless devices.

To reach our goal, we turn the main task into a multi-class classification problem. Specifically, we use spatio-temporal mobility information to train a multi-class weighted features Bayesian classifier which predicts with high accuracy the next encounter along with its characteristics, and we validate the prediction upon two different datasets (namely, WiFi and CDR datasets).

One the other hand our proposed weighted features Bayesian prediction approach outperforms significantly the standard naive Bayesian predictor and also some other location prediction approaches in previous works.

Bibliography

[1] http://mathworld.wolfram.com/diskpointpicking.html.

[2] R. Agarwal, V. Gauthier, M. Becker, T. Toukabrigunes, and H. Afifi. Large scale model for information dissemination with device to device communication using call details records. *Computer Communications*, 59:1–11, 2015.

[3] R. Ahas, S. Silm, O. Jarv, E. Saluveer, and M. Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Urban Technology, Special Issue: Mobile Positioning and Tracking in Geography and Planning*, 17(1):3–27, 2010.

[4] F. Alhasoun, A. Almaatouq, K. Greco, R. Campari, A. Alfaris, and C. Ratti. The city browser:utilizing massive call data to infer city mobility dynamics. In *ACM International Workshop on urban computing*, ACM Workshop, pages 1–8. IEEE, 2014.

[5] A. Arai and R. Shibasaki. Estimation of human mobility patterns and attributes analyzing anonymized mobile phone cdr: Developing real-time census from crowds of greater dhaka. In *Proceedings 2nd AGILE PhD School 2013*, 2nd Agile PhD school 2013, pages 1–6, 2013.

[6] A. Asahara, K. Maruyama, K. Sato, and K. 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*, GIS'11, pages 25–33. ACM, 2011.

[7] H. Barbosa, F. de Lima-Neto, A. Evsukoff, and R. Menezes. The effect of recency to human mobility. *EPJ Data Science*, 4(1):1–14, 2015.

[8] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific Data*, 2, 2015.

[9] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[10] F. Calabrese, M. Diao, G. D. Lorenzo, J. Ferreira, and C. Ratti. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313, 2013.

[11] F. Calabrese, Z. Smoreda, V. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PlosOne*, 6:1–6, 2011.

[12] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2:483–502, 2002.

[13] A. Chaintreau, H. Pan, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *Transactions on Mobile Computing*, 6(6):606–620, 2007.

[14] Y. Chang and H Liao. Emm: an event-driven mobility model for generating movements of large numbers of mobile nodes. *Simulation Modeling Practice and Theory*, 13(4):335–355, 2005.

[15] C. Chilipirea, A.C. Petre, and C. Dobre. Predicting encounters in opportunistic networks using gaussian process. In *19th International Conference on Control Systems and Computer Science*, pages 99–105. IEEE, 2013.

[16] R. Ciobanu and C. Dobre. Predicting encounters in opportunistic networks. In *Proceedings of the 1st ACM workshop on High performance mobile opportunistic systems*, HP-MOSys'12. ACM, 2012.

[17] R. Ciobanu, C. Dobre, V. Cristea, F. Pop, and F. Xhafa. Sprint-self: Social-based routing and selfish node detection in opportunistic networks. *Mobile Information Systems*, 2015, 2015.

[18] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network. In *Proceedings of the DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, pages 1–5, 2007.

[19] B. Csáji, A. Browet, V.A. Traag, J. Delvenne, E. Huens, P. Dooren, Z. Smoreda, and V. Blondel. Exploring the mobility of mobile phone users. *Physica A:Statistical Mechanics and its Application*, 392:1459–1473, 2013.

[20] M. Daoui, A. M'zoughi, M. Lalam, M. Belkadi, and R. Aoudjit. Mobility prediction based on an ant system. *Computer Communications*, 31(14):3090–3097, 2008.

[21] A. Domenico, E. Strinati, and A. Capone. Enabling green cellular networks: A survey and outlook. *Computer Communications*, 37:5–24, 2014.

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

[23] V. Etter, M. Kafsi, E. Kazemi, M . Grossglauser, and P. Thiran. Where to go from here? mobility prediction from instantaneous information. *Pervasive and Mobile Computing*, 9(6):784–797, 2013.

[24] S. Gambs, M. Killijian, and M. Cortez. Show me how you move and i will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL'10, pages 34–41. ACM, 2010.

[25] H. Gao, J. Tang, and H. Liu. Mobile location prediction in spatio-temporal context. In *Nokia Mobile Data Challenge Workshop,*, pages 1–4, 2012.

[26] J. Ghosh, S.J. Philip, and C. Qiao. Sociological orbit aware location approximation and routing (solar) in manet. In *2nd International Conference on Broadband Networks*, IEEE BroadNets, pages 641–650. IEEE, 2005.

[27] Jo. Gomes, C. Phua, and S. Krishnaswamy. Where will you go? mobile data mining for next place prediction. *Data Warehousing and Knowledge Discovery*, 8057, 2013.

[28] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[29] A. Helmy and S. Moon. Understanding periodicity and regularity of nodal encounters in mobile networks: A spectral analysis. In *SIGMOBILE Mobile Computing and Communications Review*, GLOBECOM'10, pages 1–5. ACM, 2010.

[30] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. *Computer Networks*, 52(14):2690–2712, 2008.

[31] S. Heymann, M. Latapy, and C. Magnien. Outskewer: Using skewness to spot outliners in samples and time series. In *ACM International Conference on Advance in Social Networks Analysis and Mining*, ASONAM'12, pages 527–534. IEEE, 2012.

[32] W. Hsu and A. Helmy. On nodal encounter patterns in wireless lan traces. *TRANSACTIONS on Mobile Computing*, 9(11), 2010.

[33] W.J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy. Modeling time-variant user mobility in wireless mobile networks. In *26th IEEE International Conference on Computer Communications.*, INFOCOM'07, pages 758–766. IEEE, 2007.

[34] S. Isaacman, R. Becker, C. Caceres, and S. Kobourov. Identifying important places in people's lives from cellular network data. *Pervasive computing*, pages 133–151, 2011.

[35] T. Jia, B. Jiang, K. Carling, M. Bolin, and Y. Ban. An empirical study on human mobility and its agent-based modeling. *Statistical Mechanics: Theory and Experiment*, 2012:11–24, 2012.

[36] J. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of locations. In *Proceedings of the 2Nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, pages 110–118. ACM, 2004.

[37] T. Karagiannis, J. Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of inter-contact times between mobile devices. *Transactions on Mobile Computing*, 9(10), 2010.

[38] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella. Human mobility models for opportunistic networks. *Communications Magazine*, 49(12):157–165, 2011.

[39] K. Keramat Jahromi, F. Meneses, and A. Moreira. Impact of ping-pong events on connectivity properties of node encounters. In *7th IFIP Wireless and Mobile Networking Conference*, WMNC'14. IEEE, 2014.

[40] K. Keramat Jahromi, F. Meneses, and A. Moreira. On the impact of overlapping access points in detecting node encounters. In *14th Annual Mediterranean Ad Hoc Networking Workshop*, Med-Hoc-Net'15. IEEE, 2015.

[41] M. Kim and D. Kotz. Extracting a mobility model from real user traces. In *Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies*, INFOCOM'06, pages 1–13. IEEE, 2006.

[42] M. Kim and D. Kotz. Periodic properties of user mobility and access-point popularity. *Personal and Ubiquitous Computing*, 11(6):465–479, 2007.

[43] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *25th IEEE International Conference on Computer Communications*, INFOCOM'06, pages 1–13. IEEE, 2006.

[44] S. Kosta, A. Mei, and J. Stefa. Large-scale synthetic social mobile networks with swim. *IEEE Transactions on Mobile Computing*, 13(1), 2014.

[45] D. Kotz, T. Henderson, and I. Abyzov. Crawdad dataset dartmouth/campus (v. 2007-02-08). 2013.

[46] K. Laasonen, M. Raento, and H. Toivonen. Adaptive on-device location recognition. In *International Conference on Pervasive Computing*, Pervasive 2004, pages 287–304. Springler, 2004.

[47] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150, 2010.

[48] C. Lee, F. Gutierrez, and D. Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In *11th International Conference on data mining*, pages 1146–1151. IEEE, 2011.

[49] K. Lee, S. Hong, S.J. Kim, I. Rhee, and S. Chong. Slaw: A new mobility model for human walks. In *INFOCOM'09*, pages 855–863. IEEE, 2009.

[50] F. Legendre, T. Spyropoulos, and T. Hossmann. A complex network analysis of human mobility. In *IEEE Conference on Computer Communications Workshops*, INFOCOM WKSHPS'11. IEEE, 2011.

[51] M. Lin, W. Hsu, and Z. Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the ACM Conference on Ubiquitous Computing*, UbiComp'12, pages 381–390. ACM, 2012.

[52] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian. Understanding intra-urban trip patterns from taxi trajectory data. *Geographical Systems*, 14(4):463–483, 2012.

[53] X. Lu, E. Wetter, N. Bharti, A. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. pages 1–9, 2013.

[54] V. Martinez, J. Virseda, A. Rubio, and E. Martinez. Toward large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, pages 1–10. IEEE, 2010.

[55] A. Mei and J. Stefa. Swim: A simple model to generate small mobile worlds. In *INFOCOM'09*, pages 2106–2113. IEEE, 2009.

[56] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 637–646. ACM, 2009.

[57] A. Mtibaa and K.A. Harras. Exploring social information in opportunistic mobile communications. *CRC press*, 2015.

[58] S. Mudda and S. Giordano. Regula: Utilizing the regularity of human mobility for location recommendation. In *International Workshop on GeoStreaming*, SIGSPATIAL, 2015.

[59] Z. Narmawala and S. Srivastava. Community aware heterogeneous human mobility (cahm) model and analysis. *Pervasive and Mobile Computing*, pages 119–132, 2015.

[60] H. Nguyen. Context information prediction for social-based routing in opportunistic networks. *Ad Hoc Networks*, 10(8), 2012.

[61] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, 7:1–11, 2012.

[62] M. Papandrea, K. Keramat Jahromi, M. Zignani, S. Gaito, S. Giordano, and P. Rossi. On properties of human mobility. *Computer Communications*, pages 1–10, 2016.

[63] M. Papandrea, M. Zignani, S. Gaito, S. Giordano, and G. P. Rossi. How many places do you visit a day? In *International Conference on Pervasive Computing and Communications Workshops*, PERCOM Workshops, pages 218–223. IEEE, 2013.

[64] A. Peddemors, H. Eertink, and l. Niemegeers. Predicting mobility events on personal devices. *Pervasive and Mobile Computing*, 6(4):401–423, 2010.

[65] J. Peixoto and A. Moreira. Human movement analysis using heterogeneous data sources. *International Journal of Agricaltural and Environmental Information Systems*, 2013.

[66] P. Pirozmand, G. Wu, B. Jedari, and F. Xia. Human mobility in opportunistic networks: Characteristics, models and prediction methods. *Network and Computer Applications*, 42:45–58, 2014.

[67] C. Quadri, S. Gaito, and G. P. Rossi. Big data inspired, proximity-aware 4g/5g service supporting urban social interactions. In *International Conference on Smart Computing*, SMARTCOMP'16. IEEE, 2016.

[68] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking*, 19(3):630–643, 2011.

[69] A. Rodriguez-Carrion, S.K. Das, C. Campo, and C. Garcia-Rubio. Impact of location history collection schemes on observed human mobility features. In *International Conference on Pervasive Computing and Communications Workshops*, PERCOM Workshop, pages 254–259. IEEE, 2014.

[70] H. Samiul, C.M. Schneider, V.U. Satish, and M.C. Gonzalez. Spatio-temporal patterns of urban human mobility. *statistical physics*, 151(1), 2013.

[71] N. Scafetta. Understanding the complexity of the levy –walk nature of human mobility with a multi-scale cost/benefit model. *Chaos, an interdisciplinary Journal of Nonlinear Science*, 21(4), 2011.

[72] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell. Nextplace: A spatio-temporal prediction framework for pervasive systems. *Pervasive Computing*, 6696:152–169, 2011.

[73] S. Shen and A. Iosup. Modeling avatar mobility of networked virtual environment. In *ACM Proceedings of International Workshop on Massively Multiuser Virtual Environments*, pages 1–6, 2014.

[74] j. Silvis, D. Niemeier, and R. D'Souza. Social networks and travel behavior: report from an integrated travel diary. In *in 11th International Conference on Travel Behavior Research*, tokyo, 2006.

[75] F. Simini, M. González, A. Maritan, and A. Barabási. A universal model for mobility and migration patterns. *Nature*, 484:96–100, 2012.

[76] C. Song, T. Koren, P. Wang, and A. Barabasi. Modeling the scaling properties of human mobility. *Nature physics*, 6(10):818–823, 2010.

[77] C. Song, Z. Qu, N. Blumm, and A. Barabási. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

[78] L. Song, D. Kotz, R. Jain, and X. He. Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Mobile Computing*, 5(12):1633–1649, 2006.

[79] A. Thakur, U. Kumar, A. Helmy, and W. Hsu. Analysis of spatio-temporal preferences and encounter statistics for dtn performance. In *7th International Wireless Communications and Mobile Computing Conference*, IWCMC'11. IEEE, 2011.

[80] R. Trasarti, R. Guidotti, A. Monreale, and F. Giannotti. Myway: Location prediction via mobility profiling. *Information Systems*, 2015.

[81] R. Trestian, P. Shah, H. Nguyen, Q. Vien, O. Gemikonakli, and B. Barn. Towards connecting people, locations and real-world events in a cellular network. *Telematics and Informatics*, 34(1):244–271, 2016.

[82] T. Tri Do, O. Dousse, M. Miettinen, and D. Gatica-Perez. A probabilistic kernel method for human mobility prediction with smartphones. *Pervasive and Mobile Computing*, 20, 2015.

[83] J. Truscott and N. Ferguson. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modeling. *Plos*, 8(10), 2012.

[84] H. Tuncer, S. Mishra, and N. Shenoy. A survey of identity and handoff management approaches for the future internet. *Computer Communications*, 36(1):63–79, 2012.

[85] J. Vegelius, S. Janson, and F. Johansson. Measures of similarity between distributions. *Quality and Quantity*, 20(4):437–441, 1986.

[86] R. Vogt, L. Nikolaidis, and P. Gburzynski. A realistic outdoor urban pedestrian mobility model. *Simulation Modelling Practice and Theory*, 26:113–134, 2012.

[87] L. Vu, Q. Do, and K. Nahrstedt. 3r: Fine-grained encounter-based routing in delay tolerant networks. In *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, WoWMoM'11. IEEE, 2011.

[88] D. Zhang, H. Xiong, L.T. Yang, and V. Gauthier. Nextcell: Predicting location using social interplay from cell phone traces. *Transaction on Computers*, 64(2), 2015.

[89] M. Zhao, L. Mason, and W. Wang. Empirical study on human mobility for mobile wireless networks. In *Military Communications Conference*, MILCOM'08. IEEE, 2008.

[90] M. Zignani, S. Gaito, and G.P. Rossi. Extracting human mobility and social behavior from location-aware traces. *Wireless Communications and Mobile Computing*, 13(3):313–327, 2013.