université PARIS-SACLAY

NNT : 2018SACLX026

Thèse de doctorat

Human Habits Investigation: from Mobility Reconstruction to Mobile Traffic Prediction

Thèse de doctorat de l'Université Paris-Saclay
préparée à École Polytechnique

École doctorale n° 580 : sciences et technologies de l'information et de la communication

Thèse présentée et soutenue à Paris, le 10 avril 2018, par

Guangshuo Chen

Composition du Jury :

M. Marcelo DIAS DE AMORIM Directeur de Recherche, CNRS et UPMC Sorbonne Universités	Président, Rapporteur
M. M. André-Luc BEYLOT Professeur, CNRS	Rapporteur
M. Jérôme HÄRRI Professeur, EURECOM	Examinateur
Mme Lila BOUKHATEM Maître de Conférences, Université Paris-Sud	Examinateur
Mme Ana AGUIAR Maître de Conférences, Université de Porto	Examinateur
Mme Aline Carneiro VIANA Chargé de Recherche, Inria	Directeur de thèse
M. Marco Fiore Chargé de Recherche, CNR-IEIIT	Invité

ABSTRACT

Characterizing and predicting human mobility and mobile data traffic serves as an important component in cellular networks for network optimization, network management, and application design. Although a large and growing body of the literature on human mobility has characterized the mobility of individuals and proposed human mobility predictive models, only a few studies have focused on the mobile data traffic consumption of individuals, especially from the spatiotemporal viewpoint. Also, such analyses requires the human behavior data mining on the datasets of Charging Data Records (CDR) collected from cellular operators. To this end, this thesis studies the per-user mobile data traffic prediction and the conjoint problems on the utilization of operator-collected datasets.

First, the literature review of the two topic — mobile data traffic prediction and human mobility data utilization — is presented in the thesis, in terms of the existing achievements, techniques, and open research issues.

Second, we mine the operator-collected datasets. We address the challenges brought by the heterogeneity and sparsity of CDR, with respect to mobility measurement, data incompleteness, and data processing. Particularly, we evaluate the mobility data incompleteness in voice call CDR, reveal the geographical bias caused by the mapping of user locations to cell towers, and confirmm the state-of-the-art findings regarding the capability of CDR to model individual movement patterns.

Third, we propose novel approaches to recover incomplete mobility information by formalizing the CDR completion tasks and defining the instant and slotted CDR-based trajectories. For the completion of instant CDR-based trajectories, our proposed techniques achieve an increased temporal completion of CDR data and retain significant spatial accuracy and have 50% of better completion and 5% of higher accuracy compared with the most common proposal in the literature. For the completion of slotted CDR-based trajectories, our technique can precisely recover 55% of the missing one-hour time slots from only 10% of known locations. These techniques serve as a technical foundation for our prediction analysis.

Finally, we study the prediction of mobile data traffic generated by individual mobile network subscribers. To begin with, we propose a novel practice to observe the spatiotemporal correlation of per-user mobile data traffic. We reveal that both the temporal and spatial issues are critical in achieving an accurate prediction. After that, we construct the time series of data traffic volumes and visited locations for large-scale user sets and evaluate the theoretical and practical predictabilities. Our data-driven results show that it is theoretically possible to anticipate the individual demand with a typical accuracy of 81% despite user heterogeneity. Real-world prediction can achieve an average accuracy of 70% and can be further enhanced up to 10% by knowing visited locations. Besides, our results also show that using the contextual information can enhance the performance of human mobility predictive models.

Consequently, this thesis contributes to addressing key problems of operator-collected dataset utilization and providing efficient solutions for forecasting per-user mobile data traffic. Future work lies on using more contextual information to achieve better mobility completion and data traffic prediction.

RÉSUMÉ

Caractériser et prédire la mobilité humaine et le trafic de données mobiles est important dans les réseaux cellulaires pour l'optimisation du réseau, la gestion du réseau et la conception des applications. Bien que de nombreuses études sur la mobilité humaine aient caractérisé la mobilité des individus et proposé des modèles prédictifs de mobilité humaine, seules quelques études se sont intéressées à la consommation de trafic de données mobiles des individus, notamment du point de vue spatio-temporel. En outre, ces analyses requièrent l'exploration de données sur le comportement humain sur les ensembles de données des enregistrements de données de charge (CDR) collectés auprès des opérateurs cellulaires. Cette thèse étudie la prédiction du trafic de données mobiles par utilisateur et les problèmes liés à l'utilisation des ensembles de données collectés par l'opérateur.

Premièrement, la revue de la littérature sur les deux sujets - la prédiction du trafic de données mobiles et l'utilisation des données de mobilité humaine - est présentée dans la thèse, en termes de réalisations, de techniques et de questions de recherche ouvertes. Nous présentons la collection de CDR et les problèmes existants et des solutions sur le traitement de CDR. Nous passons également en revue les études de pointe sur le trafic de données mobiles à partir de points de vue agrégés et individuels.

Deuxièmement, nous exploitons les ensembles de données collectés par l'opérateur. Nous abordons les défis apportés par l'hétérogénéité et la rareté de la CDR en termes de mesure de la mobilité, d'incomplétude des données et de traitement des données. En particulier, nous évaluons l'incomplétude des données de mobilité dans l'appel vocal CDR, révélons le biais géographique causé par la cartographie des emplacements des utilisateurs aux tours cellulaires, et confirmons les découvertes de pointe concernant la capacité de la CDR à modéliser les mouvements individuels.

Troisièmement, nous proposons de nouvelles approches pour récupérer les données de mobilité incomplètes de la CDR. Nous formalisons le problème d'achèvement de CDR et définissons les trajectoires CDR instantanées et ficelées. Pour l'achèvement des trajectoires CDR instantanées, nos techniques proposées réalisent une complétion temporelle accrue des données CDR et conservent une précision spatiale significative et ont un meilleur achèvement de 50% et une précision plus élevée de 5% et plus par rapport à l'approche la plus courante. Pour l'achèvement de trajectoires CDR fendues, notre technique peut précisément récupérer 55% des créneaux horaires d'une heure manquants à partir de seulement 10% des emplacements connus. Ces techniques servent de fondement technique à notre analyse de prédiction.

Quatrièmement, nous étudions la prédiction du trafic de données mobile généré par les abonnés individuels du réseau mobile. Pour commencer, nous proposons une nouvelle pratique pour observer la corrélation spatio-temporelle du trafic de données mobile par utilisateur. Nous révélons que les problèmes temporels et spatiaux sont essentiels pour obtenir une prédiction précise. Après cela, nous construisons les séries temporelles de volumes de trafic de données et les emplacements visités pour les ensembles d'utilisateurs à grande échelle et évaluons les prédictions théoriques et pratiques. Nos résultats basés sur les données montrent qu'il est théoriquement possible d'anticiper la demande individuelle avec une précision typique de 81% malgré l'hétérogénéité de l'utilisateur. La prédiction dans le monde réel peut atteindre une précision moyenne de 70% et peut être améliorée jusqu'à 10% en connaissant les lieux visités.

En outre, nos résultats montrent également que l'utilisation de l'information contextuelle peut améliorer la performance des modèles prédictifs de la mobilité humaine.

Par conséquent, cette thèse contribue à résoudre les principaux problèmes de l'utilisation des ensembles de données collectés par l'opérateur et à fournir des solutions efficaces pour la prévision du trafic de données mobiles par utilisateur. Les travaux futurs reposent sur l'utilisation de plus d'informations contextuelles pour améliorer la mobilité et la prédiction du trafic de données.

Acknowledgement

First of all, I would like to thank my supervisor, Aline Carneiro Viana, for giving me the opportunity to conduct the research presented in this thesis. I also would like to express my gratitude to Marco Fiore, who can be regarded as my co-supervisor without question. Without patient guidance and insightful suggestions of Aline and Marco along these years, this thesis would not have been possible.

I would like to thank those who I have collaborated with, including Sahar Hoteit, Carlos Sarraute, Yota Katsikouli, and researchers who participate the MACACO project. I also would extend my appreciation to Rafael Costa, who reviewed a part of my thesis. I would like to thank all my colleagues. I quite enjoy the time sharing with you.

At last, I own my deepest gratitude to my parents for their encouragement and understanding in the past four years.

PUBLICATIONS

This thesis resulted in the following list of publications:

Published

- G. Chen and A.C. Viana, "On the Quest for Representative Mobile Datasets", ACM S3 Workshop with MobiCom 2015, September 2015, Paris, France.
- S. Hoteit, G. Chen, A.C. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis", ACM CHANTS Workshop with MobiCom 2016, October 2016, New York City, USA.
- G. Chen, A.C. Viana, and C. Sarraute, "Towards an Adaptive Completion of Sparse Call Detail Records for Mobility Analysis", IEEE DAMN! Workshop with PerCom 2017, March 2017, Hawaii, USA.
- S. Hoteit, G. Chen, A.C. Viana, and M. Fiore, "Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis", CoRes, June 2017, Quiberon, France.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "The Spatiotemporal Interplay of Regularity and Randomness in Cellular Data Traffic", IEEE LCN, October, 2017, Singapore.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "Enriching Sparse Mobility Information in Call Detail Records", Elsevier Computer Communications, 2018.
- G. Chen, A.C. Viana, and M. Fiore, "Human Movement Pattern Recovery via Mobile Network Data", CoRes.
- G. Chen, A.C. Viana, and M. Fiore, "Forecasting Individual Mobile Data Traffic Usage", CoRes.

Under Submission

- G. Chen, A.C. Viana, M. Fiore, and C. Sarraute, "Individual Trajectory Reconstruction from Mobile Network Data", EPJ Data Science.
- G. Chen, A.C. Viana, M. Fiore, and C. Sarraute, "The Theoretical and Practical Predictability of Individual Mobile Data Traffic".

Contents

Chapter 1

Introduction

Since the first mobile phone was invented in 1973 [1], the past half-century has seen increasingly rapid advances in the innovation of mobile devices and mobile communication technologies. Nowadays mobile devices are almost necessary for business and personal lives. For instance, mobile devices and connections have reached 8 billion in 2016; they will be 11.6 billion in 2021 [2]. Meanwhile, the tremendous spreading of mobile devices leads to the rapid increment of mobile data traffic: 18-fold from 2011 to 2016 and 7-fold from 2016 to 2021 [2].

The growth of mobile devices, along with the explosion of mobile applications, keeps pressing on the radio access capacity of existing cellular network infrastructures. Besides, the growing diversity of mobile applications brings various needs on the means of mobile connectivity and bandwidth management. To cope with these problems, it is necessary to have a continuous development of techniques for managing cellular networks.

Central to the design of network management techniques is the thorough understanding of human behaviors, which leads to a proliferation of studies, including [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], in the field of wireless and cellular communications. Besides, the quantitative understanding of human behaviors (*e.g.*, individual whereabouts) has recently emerged as a central question in multi-disciplinary research, such as human nature characterization [16, 17], city planning [18], traffic engineering [19], marketing [20, 21] and etc. Individual actions are the root cause of the dynamics that impact technological and economical interests for many research communities.

The ability to foresee future human activities has important implications in network design, management, control, and optimization in cellular networks. For instance, recent developments in data offloading [22] have highlighted the need of the prediction of human mobility [6, 7], data traffic [8, 9], and user communications [10]. The accurate prediction of human behaviors also contributes to the improvement of QoS (quality of service) [5, 11], routing [12], energy saving of mobile devices [13], and application design [14, 15]. At the heart of anticipating a human behavior are the theoretical predictability (that evaluates to what degree the behavior can be foreseen maximumly) and the practical predictability (that measures the performance of actual prediction approaches). *The primary goal of this thesis is to provide technical foundations for the investigation the theoretical and practical predictability of mobile data traffic.* For this, we provide a detailed introduction in Section 1.1.

The investigation of mobile data traffic relies on the availability of corresponding behavioral data. In particular, it requires real-world datasets, which describe the cellular communication activities of mobile network subscribers, to study the variation in individual's mobile data traffic over time and space. In the literature, the primary choice is to use datasets collected by cellular operators [23], which consists of charging data records (CDR) tracking a variety of telecommunication events. Such datasets often cover the entire consumption and generation of mobile data traffic but capture thew limited mobility of mobile subscribers whose locations are logged from the occurrence of telecommunication events. The challenge caused by the latter

often appears in mobility-related analyses [16, 17, 24], including ours in this thesis. Hence, *the secondary goal of this thesis is to provide solid solutions for addressing the problem of limited mobility information in operator-collected datasets.* For this, we provide a detailed introduction in Section 1.2.

Later, we give a brief overview of the contributions in this thesis in Section 1.3.

1.1 Predicting Per-user Mobile Data Traffic

Motivation One of the most critical assumptions for system management is that future conditions of the system are obtainable or predictable. For instance, the high availability of mobility prediction can enable various application scenarios such as location-based recommendation [25], home automation [26], and location-related data dissemination [27], while also help to improve QoS [5, 7]. The better prediction of future mobile data traffic demand can help to improve the design of solutions for network load balancing [8, 12]. A large and growing body of literature has investigated the topic of predicting human mobility [23], while several critical aspects on the prediction of mobile data traffic are still missing.

Existing Studies and Limitations The theoretical predictability is determined by the interplay between regularity and randomness. For instance, the temporal order of a user's visited locations limits the predictability of his mobility [17]; the statistical characteristics such as long-range dependence [28] and self-similarity [29] restricts the predictability of Internet traffic in wired networks [30] or cellular base stations [31, 32]. In particular, for mobile phone subscribers, the regularity of their Internet traffic is not only reflected by the statistics of data but also related to the user's whereabouts [27, 33, 34, 35].

The practical predictability is bounded with actual prediction approaches. In terms of the aggregation level of data traffic, we mass the studies of understanding and predicting mobile data traffic into two classes, *i.e.*, the (1) aggregated and (2) per-user studies. The existing literature mainly focuses on the first class. Several studies have been carried out on the prediction of mobile data traffic of clusters (hotspot) [36], base stations [37, 38], and applications [33, 39]. Particularly, the characterization of aggregated data traffic of base stations has been widely investigated, such as temporal repetitive patterns [27, 37] and spatiotemporal correlation [27, 40]. The characterization of per-user mobile data traffic is investigated in terms of data volumes [27, 34, 41] and applications [42], which provide valuable findings such as (1) a small user group usually takes accounts for most of the traffic and (2) per-user mobile data traffic is also spatiotemporally correlated. Nevertheless, there is still no work using the characterization into the practical prediction. Besides, several approaches are proposed on the prediction of radio channel throughput [43, 44] and latency [11, 45].

To the best of our knowledge, there is no analysis of *(i)* the theoretical and practical predictabilities of per-user mobile data traffic or *(ii)* the associated impacts to the former's prediction brought by users' visited locations.

Objective In this thesis, we fill the taps above and provide a first investigation of the predictability of mobile data traffic generated by individual users. We focus on data traffic volumes, *i.e.*, the amount of kilobytes generated by the mobile services consumed by a mobile

network subscriber. Overall, as our primary goal, we address two central research questions, *i.e.*,

- How predictable is the individual consumption of mobile data traffic?
- Whether how does the information of individual mobility contribute to the prediction of mobile data traffic?

1.2 Utilizing Operator-collected Datasets

Motivation Real human footprints are essential to the human mobility investigation by providing the information of movements. In this context, using specialized spatiotemporal datasets such as GPS surveys seems to be a direct solution, but there is a huge overhead of collecting such a detailed dataset at scale. Hence, charging data records (CDR) have been lately considered as a primary data source for large-scale mobility studies [23]. CDR contain information about *when*, *where* and *how* mobile network subscribers have issued and received voice calls, sent and received text messages, or established data traffic sessions. The operator-collected mobile phone datasets, consisting of CDR, usually cover large populations [23], which makes them a practical choice for performing large-scale mobility analyses.

Nevertheless, a cellular network is far from being a perfect mobility sensory platform. CDR-based trajectories are coarse in space and sparse in time. The accuracy of positioning a mobile device is limited by the antenna coverage, which can be as large as several km^2 in urban areas; the temporal distribution of CDR is highly heterogeneous and sparse, *i.e.*, having typical inter-event times of hours. We refer to this spatiotemporal flaw of CDR data as the *CDR sparsity*. The CDR sparsity along with the indiscriminate use of CDR data may question the validity of research conclusions, such as inferring visited locations [46], learning recurrent movement patterns [46], and measuring mobility-related features [16, 17]. This sparsity also makes the vast majority of the trajectory data unusable: for instance, 100K, 50K and 700 individuals are retained for analysis from populations of 6M, 10M, and 1M in [16, 17, 24], respectively. This maps to percentages between 0.07% and 1.67% of the total user base, whereas the bulk of CDR information is dropped due to the insufficient sampling frequency of a user's movements. Understanding whether and to what extent such sparsity affects mobility studies is a critical issue.

On the purpose of mitigating the CDR sparsity, we propose a task named *CDR completion* for reconstructing functional trajectories of individual users from the original incomplete CDR data. A successful CDR completion would fill the gaps in the individual positioning information conveyed by CDR with the spatiotemporal points that closely match the real-world movement patterns of each subscriber. CDR completion has the potential to increase the size of user populations considered in CDR-based mobility analyses by orders of magnitude, with straightforward benefits to the confidence and generality of results.

Existing Studies and Limitations A few previous works have investigated the validity of mobility studies based on CDR. An influential analysis [46] observed that using CDR allows to correctly identify popular locations that account for 90% of each subscriber's activity; however, biases may arise when measuring individual human mobility features. Works such as [46] or the later [47] discussed biases introduced by the incompleteness of positioning information, *i.e.*, the

fact that CDR do not capture every location that a user has traveled through. Nevertheless, another important bias of CDR is caused by the use of cell tower locations of mobile network subscribers in their footprints instead of their actual positions, while it is still overlooked in the literature.

The literature on CDR completion is fairly thin. The most intuitive solution is to consider that the location in an entry of CDR stays representative for a time interval period (*e.g.*, one hour) centered on the actual event timestamp [17, 35]. Approaches based on interpolation only work in the presence of trajectories composed of thousands of locations per day [24], which is hardly the case with CDR. It is also worth noting that the GPS trace reconstruction (as in [48]) is a very different problem from CDR completion. GPS data are collected with a fixed periodicity and at high frequency (*e.g.*, at every minute), hence do not need to be completed in time.

Objective We address three central research questions as our secondary goal:

- Whether and to what extent does the use of CDR data affect human mobility studies?
- What degree of (in)completeness can one expect in CDR-based trajectories inferred from real-world large-scale mobile phone datasets?
- How efficiently can the CDR completion fill spatiotemporal gaps in CDR-based trajectories?

It is worth noting that the solutions of the CDR completion contribute reversely to the data processing for analyzing mobile data traffic.

1.3 Contributions and Thesis Outline

Building on the importance of mobile data traffic prediction and the presence of CDR sparsity, this thesis focuses on *(i)* per-user mobile data traffic prediction (as our primary goal) and *(ii)* mobility information completion of operator-collected datasets (as our secondary goal). The major results of our work has been developed in the framework of EU CHIST-ERA MACACO project [49].

In this thesis, we first deal with and discuss the problems regarding our secondary goal, because achieving our primary goal needs individual locations and mobile data traffic volumes, and thus, how to obtain such data from our operator-collected datasets has to be addressed in advance. In particular, the dissertation consists of two parts. After introducing the background (*cf.* Chapter 2), in the first part (*cf.* Chapter 3 and 4), we study our datasets' characteristics and biases and develop the CDR completion solutions, so as to achieve the secondary goal. Then, in the second part (*cf.* Chapter 5), we construct the behavioral data based on the CDR completion solutions that we have developed, and investigate the prediction of mobile data traffic generated by individual users, so as to achieve the primary goal. We provide the thesis outline in the following, which also summarizes briefly the contributions of each chapter.

Chapter 2 We begin the thesis by introducing the background in Chapter 2. We analyze the literature and come out with a review of the two topic (*i.e.*, *mobile data traffic prediction* and *human mobility data utilization*) consisting of existing achievements, techniques, and open

research issues with respect to our goals. This chapter serves as the foundation of the following ones.

Chapter 3 This chapter focuses on the datasets that we leverage in the thesis. We explore our datasets' basic characteristics and access the common challenges in dealing with operator-collected datasets in terms of data completeness, mobility information, and data processing. Regarding the impact on results brought by these challenges, we provide a confirmation of previous known issues and validate new issues in the spatial and temporal perspectives. Our results shed light on the use of similar datasets. The work related to this chapter is:

- G. Chen and A.C. Viana, "On the Quest for Representative Mobile Datasets", ACM S3 Workshop with MobiCom 2015, September 2015, Paris, France.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "Enriching Sparse Mobility Information in Call Detail Records", Elsevier Computer Communications, June 2018.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "Individual Trajectory Reconstruction from Mobile Network Data", EPJ Data Science, under submission.

Chapter 4 This chapter deals with a specific challenge brought by CDR temporal sparsity. It aims at reconstructing reliable CDR-based trajectories. For that, we define the CDR-based trajectories and CDR completion tasks, and propose novel effective techniques. We reveal the advantage of our techniques via data-driven simulations: they could achieve a far increased completion and retain essential spatial accuracy compared with the typical proposals in the literature. Our study, on one hand, provides necessary techniques for the next chapter, and on the other hand, fills the literature gaps of the mobility reconstruction. The work related to this chapter is:

- S. Hoteit, G. Chen, A.C. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis", ACM CHANTS Workshop with MobiCom 2016, October 2016, New York City, USA.
- G. Chen, A.C. Viana, and C. Sarraute, "Towards an Adaptive Completion of Sparse Call Detail Records for Mobility Analysis", IEEE DAMN! Workshop with PerCom 2017, March 2017, Hawaii, USA.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "Enriching Sparse Mobility Information in Call Detail Records", Elsevier Computer Communications, June 2018.
- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "Individual Trajectory Reconstruction from Mobile Network Data", EPJ Data Science, under submission.

Chapter 5 In this chapter, we address the problems in terms of our primary goal, and add solid contributions to the topic of mobile traffic prediction. We perform a detailed spatio-temporal analysis on the theoretical and actual performance of the prediction of per-user mobile data traffic demand. Our analysis suggests the feasibility of anticipating *(i)* how much mobile data traffic will be consumed by a given subscriber and *(ii)* where this will occur in a very effective manner. The work related to this chapter is:

- G. Chen, S. Hoteit, A.C. Viana, M. Fiore, and C. Sarraute, "The Spatiotemporal Interplay of Regularity and Randomness in Cellular Data Traffic", IEEE LCN, October, 2017, Singapore.
- G. Chen, A.C. Viana, M. Fiore, and C. Sarraute, "The Theoretical and Practical Predictability of Individual Mobile Data Traffic", under preparation.

Chapter 6 Finally, this chapter summarizes and concludes the thesis.

Chapter 2

Background

This chapter introduces the background of our studies. It conducts literature reviews of the two major topics in our studies, *i.e.*, *the prediction of mobile data traffic* and *the utilization of operator-collected datasets*.

2.1 Mobile Data Traffic Prediction

To what degree is the Internet traffic predictable? It is a question that has led to a number of attractive issues and has been continuously investigated since the invention of the Internet [50]. In this section, we review the state-of-the-art on the prediction of mobile data traffic. Our discussion is organized from two perspectives:

- **Aggregated mobile data traffic.** In this perspective, we consider the mobile data traffic from the viewpoint of a mobile network operator. Such data traffic is aggregated over many mobile devices within the same cell, the same close geographical area, or the same service/application.
- **Individual mobile data traffic.** Here we discuss in an individual viewpoint, *i.e.*, the mobile data traffic that is generated by a single mobile device.

For each perspective, we briefly introduce the data traffic characterization and particularly present the practical prediction techniques. It is worth noting that, in this section, we focus on the studies on the Internet traffic and exclude those on other traffic (*e.g.*, voice calls).

2.1.1 Literature on Aggregated Mobile Data Traffic

The investigation on aggregated mobile data traffic is mainly driven by the analyses on worldwide large-scale operator-collected datasets. For instance, such datasets that have nationwide populations are deeply mined in the relevant studies by Paul *et al.* [27] (USA), Hoteit *et al.* [51] (France), and Xu *et al.* [37, 38] (China).

2.1.1.1 Characterization

There are two major aspects with respect to the characterization, *i.e.*, temporal dynamics and spatiotemporal correlation.

There is a general agreement on the regularity of the temporal variation of aggregated mobile data traffic [23]. Almost at the same time, Paul *et al.* [27] and Shafiq *et al.* [33] separately investigate the temporal evolution of aggregated mobile data traffic of cell towers and popular applications. They both find that such traffic follows a daily repetitive pattern over weekdays: in general, the traffic has low demand during nighttime and high demand during daytime. The same repetitive pattern is also observed by Xu *et al.* [37, 38]. Is is also

remarked in [27, 33, 52] that the traffic over weekdays and weekends have different repetitive patterns and demands; a larger data traffic demand exists on weekdays than weekends. An interesting fact is that the temporal variations observed by Paul *et al.* [27] and Shafiq *et al.* [33] have different peak hours, which is also observed from other network traffic [23]. For this, a possible explanation is that such temporal variation under a higher temporal resolution partially depends on the area of study.

The spatiotemporal correlation exists among the data traffic generated by cell towers over many users in the same area. In the pure spatial perspective, the distribution of the data traffic is spatially heterogeneous: it varies over different regions as revealed by Paul *et al.* [27] and Xu *et al.* [37, 38]. Further, the latter authors find that the cell towers have similar data traffic profiles regarding their regions (*i.e.*, resident, transport, office, and entertainment) and such profiles of adjacent cell towers are correlated. In the spatiotemporal perspective, the two papers show that the spatial heterogeneity above also varies over time: pick hours depend on regions. The former authors leverage a quantitive measure (*i.e.*, the Moran's I statistic) to evaluate spatiotemporal diversity of the data traffic. They find that in general, the imminent loads of adjacent cell towers are more correlated when these loads are high, but the correlation is relatively weak and almost disappears around midnights. Recently, [53] further investigates the spatiotemporal correlation and propose an approach to infer the hidden spatial and temporal structures of aggregated mobile data traffic.

Also, several studies reveal the spatial heterogeneity aggregated over applications. The earlier work by Trestian *et al.* [40] already shows that the Internet traffic over services and applications is consumed differently at home and work locations. Hoteit *et al.* [51] find that the data traffic loads of cell towers have different inner diversities among TCP- and UDP-based services. Later, the extended analysis by Shafiq *et al.* [39] finds that the data traffic aggregated by popular applications is strongly heterogenous over regions. This provides the capability of categorizing cell towers into four classes (web browsing, email, audio, and mixed traffic) with respect to the major applications in their data traffic loads.

2.1.1.2 Prediction

Some efforts have been put on the prediction of aggregated mobile data traffic. They aim at converting the observed dynamics and correlations above to practical prediction techniques. In the following, we review the proposed prediction techniques according to the level of the aggregation.

- **Cell-level data traffic**. There is a common observation on the fact that the data traffic of cell towers has a high degree of both theoretical and practical predictabilities. Regarding the theoretical viewpoint, Zhang *et al.* [31, 32] investigate the limits of the theoretical predictability by observing the traffic of $7,000$ cell towers in China. They find that under the temporal resolution of 30 minutes, aggregated traffic (voice, text, and data) can be well predicted from the historical demand of the preceding 15 hours; the theoretical predictability of the data traffic is lower than that of the data flow of voice calls or text messages. They also find that the knowledge of the traffic demands of adjacent cells towers can enhance the theoretical predictability, but in a less degree on the data traffic than the others, which supports the quantitative evaluation on the spatiotemporal correlation by Paul *et al.* [27]. Their results ensure the capability of time series prediction techniques on the prediction of such traffic.

Regarding the practical prediction techniques, Xu *et al.* [37, 38] show that the cell-level data traffic is predictable via a linear combination of four primary components corresponding to human activities. Zang *et al.* [54] propose a mixed machine learning approach composed of K-means clustering, Elman Neural Network, and wavelet decomposition. An alternative prediction approach is proposed by Yi *et al.* [55]; it builds a complex network among cell towers, measures the traffic on the very important ones, and predicts the others' traffic using Support Vector Regression – another machine learning method. It can recover the whole picture of the traffic demand from only 8% of the total cell towers. In the opposite viewpoint, Nika *et al.* [36] perform an empirical study on data hotspots using a large-scale operator-collected dataset of $5,327$ cell towers, and show the availability of standard machine learning methods on the prediction of future hotspots (cells towers) of the traffic demand from the past history.

- **Application-level data traffic**. The early paper by Keralapura *et al.*[56] proposes a technique to cluster users and their browsing profiles. The authors find that user behavior in terms of Internet surfing can be captured using a small number of clusters. Such heterogeneity of aggregated mobile data traffic is also investigated by Ying *et al.* [57]. Later, Shafiq *et al.* [33] uses a Zipf-like model to capture the distribution of application-level mobile data traffic and finds that the regularity makes the temporal variation of the traffic highly predictable from the history of the past demand using a simple Markovian method. Recently, Zhang *et al.* [58] design a mixed application-level traffic prediction framework that leverages the α-stable modeled property and dictionary learning to separately deal with the temporal variation and the spatial sparsity of the traffic. Marquez *et al.* [59] extend the analysis in [33] and reveal a strong heterogeneity in difference mobile service demands using correlation and clustering. They show that the temporal usage patterns are quite different from service to service. Besides, several works focus on the traffic generated by special services, such as chatting (*e.g.*, WhatsApp [60] and WeChat [61]), video streaming [62], and mobile cloud [63].

In summary, the proposed techniques extend the technical bound on the prediction of mobile data traffic: they not only leverage the legacy tools that used for analyzing wired network traffic (*e.g.*, the entropy, Markov property, α-stable modeled property) to capture the temporal variation but also import several state-of-the-art machine learning tools to utilize the spatiotemporal correlation.

2.1.2 Literature on Individual Mobile Data Traffic

A relatively small body of literature is on the investigation of individual mobile data traffic, which is also driven by data mining. Differently, the relevant studies utilize both large-scale operator-collected datasets, *e.g.*, by Paul *et al.* [27] and Oliveira *et al.* [34, 52], and small-scale mobile crowdsensing datasets, *e.g.*, by Jo *et al.* [35].

2.1.2.1 Characterization

The characterization from the individual viewpoint is performed by Paul *et al.* [27], Jo *et al.* [35], Li *et al.* [42], Oliveira *et al.* [34, 52], among others.

There is a general agreement on the heterogeneity of the data traffic, with respect to the user population and the time. It is shared by Paul *et al.* [27] and Oliveira *et al.* [34, 52]. They show that most of the total data traffic is generated from a small group of "heavy" users.

Regarding the temporal variation, both the authors above find that, in general, each user is highly active only in a few hours per day, and similarly, the temporal variation is different on weekdays and weekends, as in aggregate mobile data traffic. The latter authors [34, 52] find that individual mobile data traffic also follows daily repetitive patterns and the users also have peak and non-peak hours in terms of data traffic. In particular, they find that the variation of different hours within the same day is stronger than that of the same hours overs different days.

As to the spatiotemporal correlation, Paul *et al.* [27] point out that a user is usually active at only a few of his common locations. Jo *et al.* [35] mine a small dataset of locations and services of 124 users over 16 months and they identify the spatiotemporal correlations of service usage patterns.

Other dynamics with respect to social features are also revealed. For instance, Oliveira *et al.* [34, 52] find that the distribution of individual mobile data traffic is slightly heterogeneous over the age and gender; Li *et al.* [42] focus on the major smartphone operating systems and discuss the traffic dynamics and major application in each system.

2.1.2.2 Prediction

Yet, fewer studies have addressed the prediction of individual mobile data traffic. Regarding the bandwidth of mobile devices around a cell tower, a theoretical analysis is performed by Bui *et al.* [43, 44]. Based on a theoretical LTE radio model, the authors propose a model to predict the bandwidth of mobile devices over a wide range of time scales [43]. Their model considers both the user location and the statistic of bandwidth availability. They also design a refined model aiming at the prediction of short-term bandwidth using Gaussian Random Walks [44]. Regarding the latency of each data session, Zhao *et al.* [11, 45] address the static and dynamic latency estimation problems and propose a distance-feature decomposition algorithm based on the Matrix Factorization technique to predict the latency.

In summary, the current literature has already shown the temporal dynamics and spatio-temporal correlations of both aggregated and individual mobile data traffic, while only a few practical prediction techniques are proposed aiming at the latter's prediction. Still, a large amount of the investigation on individual mobile traffic is needed. For instance, to perform data-driven prediction analysis on the theoretical and practical predictabilities of individual mobile data traffic.

2.1.3 Taxonomy of Prediction Techniques

In this thesis, we study the prediction of individual mobile data traffic by mining individual time series of locations and data sessions. For that, here we review the state-of-the-art techniques for the time series prediction. Some of them have been already used in the prediction of the Internet traffic.

2.1.3.1 Theoretical Predictability Measurements

The theoretical predictability of a time series describes its inherent forecasting capacity and is independent of any actual prediction technique. It evaluates to what degree the time series can be foreseen and bounds to the maximum forecasting performance. The theoretical predictability is correlated with the uncertainty that is usually measured by the entropy in information theory [64]. Feder *et al.* [65] build the quantitive correlation between the predictability and entropy, which is later applied on the study of an actual behavior (*i.e.*, human mobility) by Song *et al.* [17]. They are then followed by similar predictability studies on human mobility [66, 67], vehicular traffic [68], radio spectrum state [69], and aggregated mobile data traffic [31]. Nevertheless, the theoretical predictability of per-user mobile data traffic remains untouched.

Analyzing the theoretical predictability helps the design and implementation of practical prediction techniques *e.g.*, human mobility prediction [66]. In this thesis, we study the theoretical predictability of mobile data traffic on the foundation of Feder *et al.* [65]. Yet, the practical predictability of time series still depends on the prediction techniques, *e.g.*, probabilistic or regression models. We summarize these practical techniques in the following.

2.1.3.2 ARIMA Prediction Models

This class contains the typical ARIMA, AR, MA, and ARMA models. They are designed for the prediction of time series of continuous scalar values and leverage moving average and autoregressive filtering. They often appear in the time series analyses of statistics and economic studies [70]. The application of the models assumes that the target time series or its difference is *stationary* [70]. Besides the typical models, the others in this class, such as MEAN, LAST, BM, FARIMA, and GARCH, are used to predict network resources in wired networks [71, 72, 73, 74] with smoothing solutions (*e.g.*, wavelet approximation) [43]. Nevertheless, the designs of the typical models cannot combine discrete values (*e.g.*, visited cell tower locations) in their prediction, and none of these works has considered the spatial dynamics of their targets.

2.1.3.3 Markovian Prediction Methods

The methods in this class, *i.e.*, MC, PPM, SPM, and ALZ, leverage the Markov property. They are mainly designed for the prediction of time series of discrete observations. Their application assumes that the target time series has the Markov property, *i.e.*, its current value is always determined by a limited number of its previous values. In this class, a prediction method predicts the current value X_t of a time series by building a probabilistic model from its full history and solving the following maximization problem: $\hat{x}_t = \arg\max_x P(X_t = x | x_{t-1}, \cdots, x_{t-k})$ where $x_{t-1}, \cdots, x_{t-k}$ are the newest k previous values observed in the time series.

MC (Markov Chain) This is almost the simplest Markovian method [75]. A k-th order Markov chain, represented as MC(k), makes a prediction of the state X_t solely based on the fixed previous k states. It builds a transition matrix consisting of the probabilities of transitions from the past k states to the current one. There are several common practices to compute the probabilities, such as MLE (maximum likelihood estimation) [76] and MCMC (Markov Chain Monte Charlo) [77]. However, it needs a large number of samples to compute probabilities, which grows quickly with respect to the order k and the alphabet of discrete values.

PPM (Prediction by Partial Matching) This is an improved method of the Markov chain, used massively in the lossless text compression [78]. A k-th order PPM model, represented as PPM(k), is a combination of MC(m), $\forall m \leq k$. It computes the so-called *escape* probabilities of the state X_t as the weighted sums of the probabilities of all the Markov models in the combination. We use the implementation of Moffat *et al.* [78, Method C] in this thesis.

SPM (Sampled Pattern Matching) This is another improved method of the Markov chain designed by Jacquet *et al.* [79]; for predicting the current state X_t, it considers much larger immediately preceding states than the MC or PPM does. In a SPM predictor, instead of using a fixed order k, the considered length of immediately preceding context is determined as a fixed fraction (represented as the parameter α) of the longest context which has previously appeared. An SPM model with the parameter α is represented as SPM(α).

ALZ (Active LeZi) This is an improved online prediction algorithm based on the classical LZ78 data compression scheme proposed by Gopalratnam *et al.* [80]. It also employs the power of the Markov property and is able to incrementally learn the sequence and to deliver real time predictions. The ALZ algorithm also makes a prediction of the state X_t in the time series given the preceding context, while a variable window of immediately preceding symbols is maintained, of which the length is the longest phrase previously observed in a classical LZ78 parsing. With this window, the algorithm can compute statistics on all possible preceding contexts. For the pseudo code of this algorithm, we refer the reader to [80, Figure 3].

2.1.3.4 Supervised Machine Learning Methods

This class contains a group of state-of-the-art techniques that are categorized to the field of *supervised machine learning* in practice [81]. These techniques can solve problems in the shape of $\mathbf{y} = f(\mathbf{x})$ where $\mathbf{x}$ and $\mathbf{y}$ are the input and output vectors. Each of them builds a model (which is composed of kernel functions, decision/regression tress, or neuron networks) upon a training set that consists of known instances of $\mathbf{x}$ and its corresponding $\mathbf{y}$ as the output classes (in a classification problem) or values (a regression problem). Then, the trained model can predict $\mathbf{y}$ from $\mathbf{x}$ in a new instance. They are capable of forecasting time series of continuous and discrete values.

SVM (Support Vector Machine) This is actually a classification algorithm [75]. It assigns a new input vector $\mathbf{x}$ to one of the already known output classes $\mathbf{y}$, just as other classification algorithms do. SVM is adaptive to both linear and non-linear discriminant functions during the training process. It applies a kernel function to map the training data to a high-dimensional feature space, which ensures the ability to produce a more accurate classification. Further, SVM is insensitive to the number of input features.

GBRT (Gradient Boosted Regression Trees) This algorithm is on the basis of the ensemble learning technique [82]. It build a prediction model as an ensemble of weak prediction models, *i.e.*, regression trees, in a state-wise fashion and allows optimization of an arbitrary differentiable loss function. The GBRT algorithm has advantages in terms of flexibility for heterogeneous features, good predictive power, and training speed, and has well-tested implementations.

MLP (Multilayer Perceptron) This algorithm is a typical supervised learning algorithm using artificial neural networks. It is designed for both regression and classification problems. A MLP network is a feedforward artificial neural network that is fully connected. The MLP algorithm accepts different activation functions, layers and neurons [83].

To the best of our knowledge, there is still no study about these prediction techniques' performance on per-user mobile data traffic prediction. We are not able to judge which is superior so far, and thus, will perform a comparative study in this thesis. First, ARIMA models are not considered because they do not support discrete values, which is necessary for our analysis. Second, all Markovian methods will be utilized, since they show impressive performance on similar time series prediction scenarios, such as human mobility prediction [76] and aggregated mobile data traffic prediction [33]. Besides, supervised machine learning techniques are shown to be effective in multiple application scenarios [75, 83], including aggregated Internet traffic [84]. We will also consider them in our analysis.

2.2 Operator-collected Mobility Data Utilization

A large amount of operator-collected datasets has been utilized in the literature [23], due to the spreading of mobile devices and the enlarging openness of mobile network operators. In the wide scope of research disciplines driven by mining the operator-collected datasets, the analytic data investigation of human mobility is one of the most significant topics [23]. In this context, the mobile network works like a gigantic and multi-modal sensing platform that produces time-stamped *human footprints*. In this section, we introduce the background on the utilization of the operator-collected datasets. Our focus is on the state-of-the-art means of processing human footprints with the limited spatiotemporal granularity [23] in order to provide an introductory guide.

2.2.1 The Events that Produce Human Footprints

Human footprints in most of the operator-collected datasets come from CDR – once referred to *Call Detail Records* and later to *Charging Data Records* in the 3GPP specification [85, TS 32.298]. The collection of CDR is triggered by the so-called *charging events* [85, TS 32.250/32.251]. In the following, we list the types of charging events that lead to human footprints.

- **Voice call**. This is the most common trigger to the collection of operator-collected datasets in the literature [23]. Each voice call CDR marks *(i)* when the call originates and terminates and *(ii) only* the cell tower where the call originates. Significant effort has been put on the characterization of voice calls [46, 86, 87, 88, 89, 90, 91]; it is well known that the temporal heterogeneity and sparsity of voice calls limit the quality of mobility information and cause a heavy data loss of locations.

- **Text message**. This is also a common trigger and produces text message CDR that are similar to voice call CDR. Yet, each text message triggers two CDR separately, which correspond to the origination and delivery of the message and mark the cell towers from/to which the message is created/delivered. Usually, text message CDR appear together with voice call CDR in the operator-collected datasets [23]. According to the

characterization [46], text messages also only provide limited mobility information due to their temporal heterogeneity and sparsity.

- **Internet visit**. This event triggers multiple CDR types such as S-CDR, SGW-CDR, and PGW-CDR [85, TS 32.251]. Only the S-CDR has a mandatory parameter describing the location of devices; it points to the cell tower where a mobile device originated a session for Internet visits. This event appears as a new source of human footprints in recent years, and is used in several studies [27, 40, 46, 92, 93, 94, 95]. Internet visits provide far more human footprints than voice calls or text messages; they still suffer the temporal heterogeneity but are far less sparse than the other two events [52]. Therefore, a operator-collected dataset of Internet visits usually appears as the ground-truth data as in [46, 96].

- **Mobility update**. Every time a mobile device enters a roaming process or has an inter-system handover, a CDR is triggered and contains the identifier of the new cell tower. Several operator-collected datasets contain human footprints collected by this events along with Internet visits, including [96, 97, 98].

- **Location request**. The 3GPP specification [85, TS 23.271] describes the capability for both the network and mobile devices to initiate location requests. Once a location request is created, the network use its built-in LCS (location services) method [99] to measure the position of the target mobile device, and two CDR are recorded: one for the origination of the location request and the other for the termination [85, TS 32.251]. However, to the best of our knowledge, none of the datasets collects the location request CDR.

- **Others**. Events such as power on/off and network switch (LTE to 2G/3G) may also trigger the generation of CDR of their own types and help the collection of human footprints, while they rarely appear in the literature: we only find one dataset [96] uses these events together with Internet visits.

Consequently, voice calls and text messages are still the major contributors to human footprints in the literature heretofore [23], while the mobility information among them has limited temporal granularity due to their high temporal heterogeneity and sparsity. Nevertheless, due to the fact that there are more geo-referenced CDR types while having more sensitive user behavior data, we may have mobile network operators with better openness and operator-collected datasets with a higher temporal resolution in future.

2.2.2 Extracting Geographical Coordinates from CDR

The aforementioned charging events help the collection of human footprints in the literature. Particularly, in the CDR of these charging events, there are specific parameters which provide those human footprints, named and defined in the 3GPP specification [85, TS 29.002]. With respect to the provided information, we categorize these parameters into three classes, *i.e.*, *who* (Subscriber Equipment Number, Served IMEI, or Served IMSI), *when* (Event time stamps), and *where* (Cell Identifier or Location Estimate) [85, TS 29.002].

The *where* parameters need some explanation. The *cell identifier* parameter is the digit identifier of the serving sector, antenna, or cell tower that a mobile device communicates with.

It is included as a *mandatory* parameter in the CDR of *all* the charging events mentioned above, while it is not a geographical coordinate. The *location estimate* parameter represents the geographical coordinate of a mobile device, but is *only* included in the CDR of a location request which is not a common trigger of the collection of human footprints. Therefore, in most of the operator-collected datasets, the dataset users or the data providers have to convert those cell identifiers to actual geographical coordinates. In the following, we summarize the major means of the conversion.

- The most common means is to use the LCS methods [99], which can only be done by the data providers or the mobile network operators. In most cases, the Pure CID (Cell Identity) method is used in the operator-collected datasets to obtain cell tower locations from identifiers. It returns the coordinates having the location resolution of cell towers [99]. In a very few cases, other LCS methods (*e.g.*, Enhanced CID method [99]) are used in mobile networks and lead to more accurate human footprints, *e.g.*, in the datasets used by Sahar *et al.* [24] and Jiang *et al.* [97].

- When CDR are directly exposed without geographical coordinates but only cell identifiers (*e.g.*, in the dataset used by Schlaich *et al.* [100]), the dataset users are able to obtain geographical coordinates via a series of third-party databases and services, which are listed in the following.

  - OpenCelliD [101] is a community database that documents the locations of cell towers and WiFi access points all over the world. The positioning leverages multiple Android smartphone applications (OpenCelliD apps [101, Data Source]) that are capable of reading identifiers of the cell towers and GPS coordinates of the mobile devices. Due to this methodology, the precision is limited, *i.e.*, usually in a degree of kilometers.
  - France OpenData [102] is a government database that accompanies the opening data of the States and the administrations. It provides a dataset of the cell tower locations, categorized by the generation (2G, 3G, and LTE) and the mobile network operator (Orange, SFR, Free, and Bouygues) in France. The dataset is considered precise because it is the mobile network operators who publish the geographical coordinates of their cell tower locations.
  - Google Geolocation [103], Unwired Labs [104], OpenSignal [105] and Mozilla Location Service [106] provide the services letting mobile devices determine their locations by sending the identifier of a cell tower or WiFi access point. Their precision is determined by the community of data contributors and thus varies widely.

- Rarely, when cell identifiers are partially exposed in CDR, *i.e.*, only their *location area codes* [85, TS 32.298], the users can still obtain "locations" in a very low spatial resolution, such as Zip Codes used in the dataset of Issacman *et al.* [107].

Consequently, in the operator-collected datasets, human footprints are converted from cell identifiers of the CDR and have the location resolution of cell towers. Their spatial granularity depends on the density of the cell tower deployment. This issue needs to be carefully considered in the utilization, while the evaluation of the positioning error in these human footprints is still neglected in the literature. On the positive side, there are many other more accurate LCS

methods, such as A-GPS that improves the accuracy to under 5 meters [99]. It is reasonable to expect a operator-collected dataset with a higher location resolution in future.

2.2.3 Reliability of Human Footprints

As introduced above, the operator-collected datasets have the limited spatiotemporal granularity. This may question the reliability of the results and cause some biased conclusions. A large amount of effort has been put in assessing the reliability [23]. For such studies, the ground-truth datasets with higher granularities are necessary, in order to capture a more complete picture of the mobility. In the following, we summarize the ground-truth used in the literature to validate the operator-collected datasets.

In the validation, human footprints as ground-truth are either those of the same users captured from their events that appear more frequently, or those captured from another user population. Ranjan *et al.* [46] employ a dataset (500K users; 1 month) captured from San Francisco as the ground-truth. In the one-month observing period, each of their users has voice calls or text messages in a median magnitude of $10^{1.5}$; the ground-truth data has human footprints in a median magnitude of $10^{2.5}$ [46, Fig 2(a)]. Yet, this amount of human footprints as ground-truth is not always sufficient to capture a user's entire mobility in a month. The same situation occurs in the study of Zhao *et al.* [96], their ground-truth (1loc/6hrs; 686K users; 1 day) has almost the same temporal granularity as the dataset of voice calls. Gonzalez *et al.* [16] employ a small dataset (0.5loc/hr; 206 users; 1 week) to show the reliability of a far larger voice call dataset (100K users; 6 months). Similarly, Song *et al.* [17] use a small number of trajectories (1loc/hr; 100 users; 8 days) to validate their methodology to estiamte the entropy of locations that is later applied on a large-scale CDR dataset (0.5loc/hr; 50K users; 14 weeks). Hoteit *et al.* [24] make a ground-truth dataset of highly active subscribers (10^3loc/day; 707 users; 1 day) to validate their trajectory completion approach. Even the locations manually reported in a web survey appear as ground-truth in [107]. In all, none of these ground-truth datasets that have been ever used, has full advantage in terms of the population, the length of the observing period, the accuracy of positioning, and the granularity, compared with the operator-collected datasets captured by voice calls or text messages.

Besides, the voice calls or text messages made by the very users in the ground-truth are sometimes needed. In an ideal case, they are included in the ground-truth, as in [46, 96]. Otherwise, equivalent CDR-like data needs to be artificially generated. Heretofore, there is not any best practice to such generation. Song *et al.* [108] perform such CDR-like data by uniformly removing locations from the ground-truth, while the generated data does not fit the appearance pattern of voice calls. Hoteit *et al.* [24] proceed more reasonably; they remove locations according to the distribution of voice calls.

The existing analyses on the reliability only aim at the human footprints captured by voice calls or text messages. Those captured by Internet visits are in default regarded as the ground-truth. However, since the mobile data traffic is also temporally heterogeneous and sometimes sparse [40, 46], the reliability of this type of human footprints remains an open question and needs for attention. Based on the discussions above, we conclude that *(i)* CDR are hitherto an appropriate choice to obtain human footprints; *(ii)* more accurate ground-truth data is needed to have more convincible reliability analyses.

2.2.4 Biased Mobility Feature Measurement

A major use of human footprints is to measure mobility features, such as uncertainty of locations [17], location displacement [16], and significant locations [26, 109, 110]. These features describe per-user mobility in certain levels and are of importance in mobility-related studies [23]. Due to the limited spatiotemporal granularity in human footprints, the biased results may be obtained. We separate the validation of the results by each feature in the following.

- **Uncertainty of locations**. This feature is usually measured by the entropy or entropy rate of locations [17], as a common prepositive step of the prediction analyses [31, 66], and is also used in network management (*e.g.*, cellular paging [111]). Due to the nature of the entropy, the loss of locations impacts the accuracy of the entropy management. The early paper by Song *et al.* [17] shows that the entropy rate of time series of locations, measured by a operator-collected dataset, is smaller than the actual entropy rate; they develop a technique to estimate the unbiased entropy rate in [108]. Later, Ranjan *et al.* [46] confirmed this bias.

- **Location displacement**. This is a significant feature measuring the individual span of movement. A common metric of the location displacement is the radius of gyration. It was firstly measured in a large-scale operator-collected dataset by Gonzalez *et al.* [16], but they did not consider the impact of the data loss. The later paper by Ranjan *et al.* [46] shows that the distribution of the radius of gyration of highly-frequent voice call users is close to the actual distribution. Another common metric is the appearance frequency of each location. Ranjan *et al.* [46] show that, when the frequencies are computed from voice calls, the less frequent locations may disappear and the frequent locations may have their frequencies less than the actual ones [46]. This incompleteness of less frequent locations is later discussed by Zhang *et al.* [47].

- **Significant locations**. Locations such as individual's *home* and *work* places are important in various applications, and are measured by several studies [16, 26, 110]. Ranjan *et al.* [46] find that their measurement is merely suffered from the loss of locations: using operator-collected datasets allows to correctly identify popular locations that account for 90% of each subscriber's activity. Besides, Zhang *et al.* [47] find that CDR-based individual trajectories can match the reference information from public transport data, *i.e.*, GPS logs of taxis and buses, as well as subway transit records.

Recall that another critical bias caused by the use of cell tower locations instead of users individual actual positions has been overlooked in the literature. The early paper by Isaacman [107] unveils that using CDR as positioning information may lead to a distance error within 1 km compared to ground-truth collected from 5 users. Still, we do not have a global view of how the use of cell tower locations impacts the user positioning. We address this issue with the presentation of our datasets (*cf.* Section 3.5).

2.2.5 Practices to Overcome the Limited Granularity

We have introduced the collection of human footprints and the biased results of mobility features. In mobility studies, the data preliminary of the operator-collected datasets is mandatory and in most cases, aims at addressing the challenges brought by the limited spatiotemporal

granularity. In the following, we summarize the representative data preliminary practices on human footprints.

- **Filter out "bad" users**. This is a common procedure in the literature. It is to set a threshold based on the demand and then eliminate from the analysis the users who are below the threshold. There is hardly a standard on this procedure. One representative threshold is to keep the users that have at least one call every two hours on average and have at least two unique locations during the period of study, which is used by Gonzalez *et al.* [16], Song *et al.* [17][112], and Ranjan *et al.* [46]. This threshold guarantees the collection of a highly active group of users and keeps the significant locations [46]. Note that such filtering, whether or not effective, may drop a large number of users and thus should be applied on a large population to ensure the scale of the remaining users, *e.g.*, keeping thousands of the users as in [16, 17, 24].

- **Reduce the resolution**. This relates to the reduction of the resolution of time and space and the merging of human footprints. It can reduce the requirement on the number of human footprints. Usually, the temporal resolution depends on the type of events and varies according to the quality of the data, *e.g.*, 15 minutes [46], 1 hour [16, 17, 112], 2 hours [113], and 6 hours [96]. Although it is not very common, a few studies use reduced spatial resolutions. For instance, De Montjoye *et al.* [113] replace the cell tower identifiers with the clusters that are merged by one or more cells.

- **Segment the observing period**. This concerns the separation of the observing period into segments and make different user groups for the segments. For instance, divide into daytime and nighttime hours [90, 96], weekdays and weekends [110, 114], weeks or months [17]. This practice may loose some of the long-term behaviors but can keep more users than using the whole observing period.

- **Correlate with the mobility loss**. This is to build a function between results and inherent features of human footprints (*e.g.*, the loss rate of locations) by leveraging a ground-truth dataset. The function is then used to fix the biased result obtained by the incomplete mobility information. For instance, Song *et al.* [17] find a linear correlation between the loss rate of voice calls and the logarithm of the entropy rate of time-ordered locations [108, Fig. S3]. Iovan *et al.* [86] find a quadratic fit between the number of charging events and the median daily travelled distance of a user.

- **Perform controlled experiments**. This is to repeat the methodology or the mobility measurement on an alternative controlled dataset, used in [16, 17, 112, 115]. The controlled dataset usually has a higher resolution than the counterpart used in the study. The conclusion is more convincible provided that the same results can be obtained from both datasets.

- **Infer missing locations**. This is to infer a user's position during a close time when a time-stamped location is recorded, so as to enlarge the availability of human footprints. Regarding voice calls, Ficek *et al.* [87] propose a probabilistic inter-call mobility model, using a finite Gaussian mixture model to determine users' positions between their consecutive voice calls. Their work contributes to the location inference and particularly, supports the common intuitive solution that is to assume that a human footprint stays

representative for a time interval period (*e.g.*, one hour) centered on the actual event timestamp, as used in [17, 35]. Approaches based on interpolation only work in the presence of trajectories composed of thousands of locations per day [24], which are not very adaptive to common operator-collected datasets of voice calls or text messages.

- **Reconstruct trajectories**. This is to use a user's existing human footprints to reconstruct a stable trajectory. For instance, Zhang *et al.* [116] leverage the tensor factorization technique to generate a spatiotemporal graph from a operator-collected dataset for their analysis; they do not validate the accuracy of their spatiotemporal graph. The mobility reconstruction techniques designed for GPS location samples (*e.g.*, [48]) may be adaptive on the datasets of Internet visits but not on those of voice calls or text messages because they are originally designed for highly frequent location samples.

Consequently, the literature on the reliable practices dealing with the limited granularity is fairly thin. In this thesis, we contribute to the location inference and the trajectory reconstruction on the human footprints of voice calls or text messages, or in other words, on the mobility data having a low sampling frequency. To the best of our knowledge, there is no validated investigation on the trajectory reconstruction.

2.3 Summary

In this chapter, we review the literature on the prediction of mobile data traffic, the theoretical and practical prediction techniques, and the utilization of operator-collected datasets. In the following, we present our datasets (*cf.* Chapter 3) and propose our techniques of the location inference and trajectory reconstruction (*cf.* Chapter 4), so as to fill the gap in the literature. Then, we leverage the prediction techniques introduced in this chapter to investigate the prediction of individual mobile data traffic (*cf.* Chapter 5).

Chapter 3

Datasets: Characteristics and Challenges

Our analysis of human mobility and data traffic demand relies on the availability of datasets. It is highly desirable that such datasets cover a large (*e.g.*, citywide or nationwide) population and a proper time period (*e.g.*, weeks to months), so as to ensure a fairly good diversity of human behaviors and to support the generality of conclusions. It is also highly favored to leverage multiple datasets, so as to ensure the validity of conclusions and to reduce the risk of obtaining biased conclusions.

In this context, multiple datasets are used in this thesis. This chapter focuses on the datasets and is summarized as follows:

1. The common practice of collecting human behavioral data in the literature is firstly introduced in Section 3.1. Then, two sets of the utilized datasets categorized by their means of the collection are presented in Section 3.2 and Section 3.3, respectively.

2. The incompleteness of mobility information in the operator-collected datasets and the challenge brought by such incompleteness are explored and discussed in Section 3.4.

3. The incompleteness above may lead to biased conclusions. Regarding this issue, the particular challenge of the measurement of mobility features is discussed in Section 3.5.

4. The challenge of processing the operator-collected datasets is discussed Section 3.6 in terms of the mixed data quality, the joint data complexity, and the large data amount.

Finally, Section 3.7 summarizes this chapter.

3.1 Human Behavior Collection

For the investigation of human behaviors, mobile devices have become almost the most popular data source in the past decade [23]. In the literature, there are two major means of collecting the data of mobile devices. They have contributed to the collection of the datasets used in this thesis. We introduce the means in this section and the datasets later on.

3.1.1 Mobile Traffic Gathering

The primary means is to use logs of mobile network traffic. We name those collected by this means as the *operator-collected* datasets. The ever higher penetration rate of mobile devices and their continuous interaction with the cellular network infrastructure give mobile network operators the possibility to easily record time-stamped and geo-referenced events of a very large population at a small cost for billing or network management purposes [23]. Recall that, in

this context, mobile network operators collect CDR including a variety of telecommunication events (*cf.* Section 2.2.1). Each record is associated to a specific mobile device and is also time-stamped and sometimes geo-referenced with the location of the antenna that communicates with the device.

3.1.2 Mobile Crowdsensing

The secondary means is to directly extract the sensory data from smartphones or other mobile devices, which is also called the mobile crowdsensing [117]. The sensory data comes from volunteers who have agreed with the installation of special applications mainly for the purpose of research. We name those collected by this means as the *application-based* datasets. Thanks to the openness of mobile operating systems, smartphone applications can obtain various information about a mobile device from its equipped sensors, *e.g.*, rotation, acceleration, GPS location, WiFi and Bluetooth access points and their signal strength. The sensory data provides precise sensitive information but usually includes a small number of mobile devices due to the difficulty of enlarging participants in the mobile crowdsensing.

3.2 Operator-collected Large-scale Datasets

Both human footprints and their mobile data traffic demands can be extracted from CDR. Thus, we leverage four large-scale operator-collected datasets in the thesis. The first three, *i.e.*, `CDR`, `Session`, and `Flow` datasets, come from the same mobile network operator in Latin America, contain a nationwide amount of mobile network subscribers, share common users between each other, and are anonymized by the same hashing function on the users' identifiers. The last, *i.e.*, `Shanghai` dataset, is collected by a major network operator in China from a central area in Shanghai and has a citywide population.

All the datasets introduced in this section are extracted from CDR, while we only name the first one using the abbreviation `CDR` due to the fact that its collection is triggered by voice calls, *i.e.*, the most common event that produces CDR in the literature (*cf.* Section 2.2.1). We present each dataset, along with the basic characterization on the spatial and temporal perspectives, in the following.

3.2.1 CDR Dataset

This is the most significant dataset in this thesis. It consists of time-stamped and geo-referenced records of voice phone calls of mobile network subscribers (also called Call Detail Records). Each record contains the hashed identifiers of the caller and the callee, the call duration in seconds, the timestamp for the call time, and the location of the cell tower to which the caller's device is connected to when the call originated. This dataset consists of approximately 3.8 billion of records generated by 3.6M prepaid mobile subscribers during a 15-month period. Due to a non-disclosure agreement with the data owner, we cannot reveal the geographical area or the exact collecting period of this dataset.

Spatial Characterization Each location in a CDR entry corresponds to a cell tower. In total, 4.2K cell towers have appeared in `CDR` dataset as latitude and longitude pairs. In the metropolitan area, their deployment is fairly dense as shown in Figure 3.1, where (red) dots

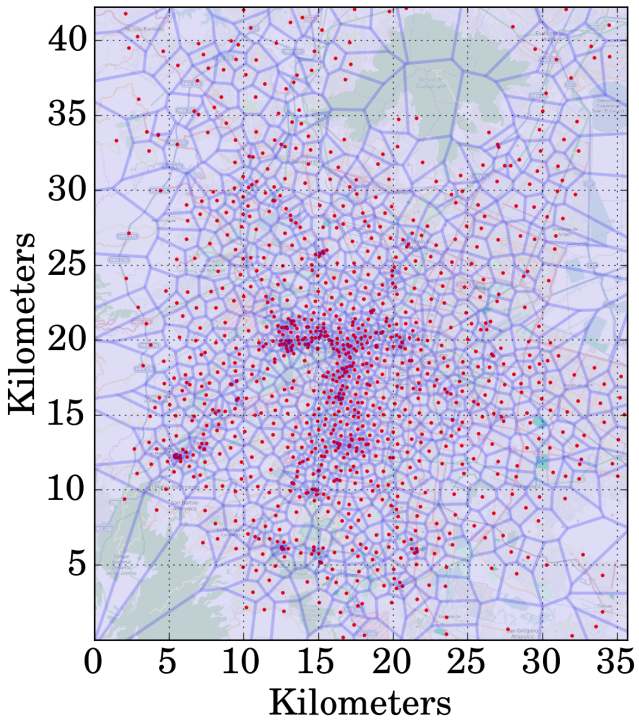


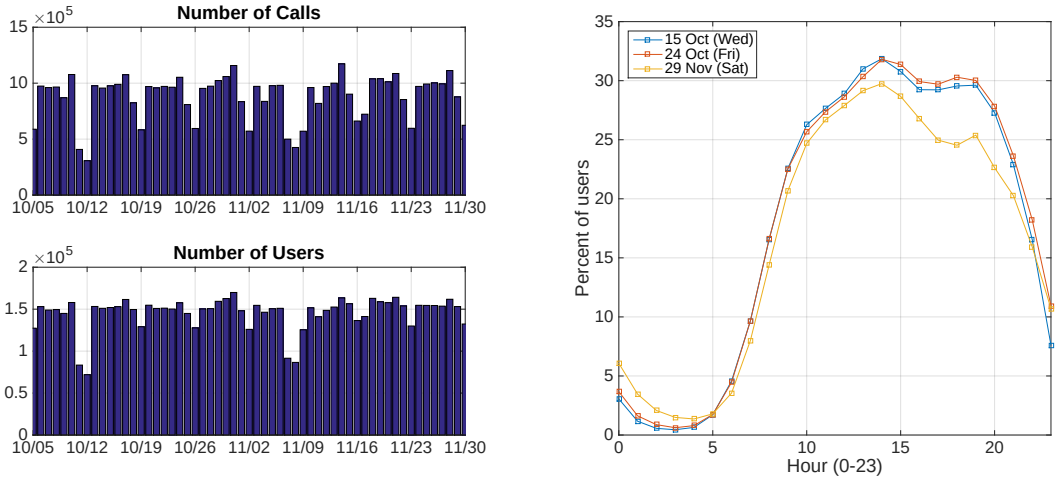
Figure 3.1: Deployment of cell towers in a metropolitan area from CDR dataset. Red dots represent the base stations and the Voronoi tessellation approximates the coverage of each cell.

represent the cell towers and the Voronoi tessellation approximates the coverage of each cell: on average, a cell tower covers an area of 2 km^2. This grants a fair spatial granularity in the positioning of mobile network subscribers.

Temporal Characterization The telecommunication events of mobile subscribers are driven by endeavors and habits that are far from deterministic. In particular, the numbers of both mobile phone subscribers and their CDR are not uniform over time. We plot in Figure 3.2(a) the numbers of the CDR and users on each day of a 2-month period and see that the numbers follow a weekly periodic pattern. In particular, the number of calling active users increases on weekdays, reaches a weekly maximum on Friday, and declines quickly at weekends. On several days there has been an abnormal wave, which partially corresponds to the public holidays (*e.g.*, high values on October 31, Halloween, and lower values on November 17, Revolution Day Memorial). On each day, the number of calls also varies hourly. As shown in Figure 3.2(b), the majority of users have phone calls from $8am$ to $11pm$, while less than 10% from $0am$ to $7am$: there are more devices having calls in working hours and evenings than early hours in the morning. We will present a more detailed analysis on the temporal perspective of this dataset in Section 3.4.

3.2.2 Session Dataset

This dataset serves as our source of mobile data traffic in this thesis. It consists of *data session records* that are extracted from a part of the parameters of S-CDR and describe Internet



(a) Number of CDR and users

(b) Number of users in hours

Figure 3.2: (a) Number of CDR and users during a 2-month period in `CDR` dataset. (b) Percentage of users appeared in `CDR` dataset in three particular days.

visits (*cf.* Section 2.2.1). Each record describes a data session and contains the hashed device identifier, the volume of upload and download data exchanged in KiloBytes, and the timestamp denoting the starting time of the session.

This term requires some explanation. In the 3G/LTE standards, a data session is created when a radio channel is allocated to a subscriber whenever it has data to send. The allocated radio channel is revoked when the subscriber is dormant for certain period known as the *dormancy period* [27] (typically from 5 to 30 seconds). According to the relevant study [27] that mines the same type of records, a data session stays alive around 100 seconds on average; over 90% of the sessions are less than 1000 seconds. Thus, we say that this `Session` dataset provides entire data of the individual mobile data traffic volumes in a degree of minutes.

Spatial Characterization Data session records do not have the location parameter of the original S-CDR and thus do not contain any geographical information. Yet, since `CDR` and `Session` datasets share the common users and the observing period, we can obtain a fairly large amount of data sessions' locations by linking the two types of records according to their identifiers.

Temporal Characterization Quite different from voice calls in `CDR` dataset, Internet visits or data sessions have a relatively uniform pattern, mostly due to automatically-generated background traffic and push notifications that are periodic and fairly independent of human activity [46]. Regarding the per-user mobile data traffic, we plot in Figure 3.3 the CDF of the mean daily volume per user. We see that the per-user daily data traffic varies from 1 KB to 2 GB, and only 10% of the users generate more than 10 MB of mobile data traffic every day. This indicates the heterogeneity in mobile data traffic generation and supports the similar finding in [27, 118]: only a minority group of mobile network subscribers take account for most of the data traffic in the network. Consequently, in our analysis, we need to focus on those

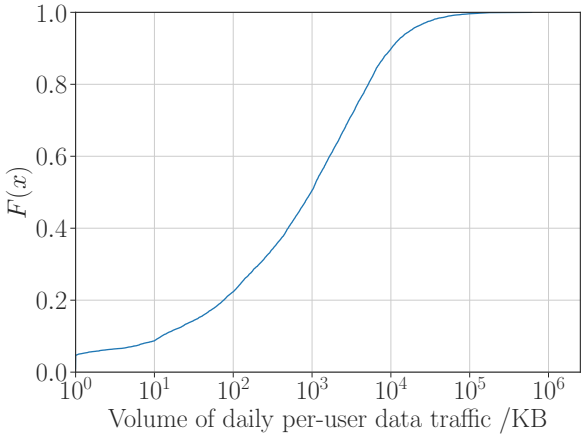


Figure 3.3: Distribution of volumes of the mean daily per-user mobile data traffic in `Session` dataset.

who actually leverage the cellular network for Internet visits, *i.e.*, to select the heavy users.

3.2.3 Flow Dataset

This dataset is also triggered by Internet visits as `Session` dataset, while its records that we call *data flow records* have more parameters of the S-CDR and importantly, have the location parameter. Compared with `Session` dataset, this dataset has only a 5-day observing period and a far smaller population of 20K mobile network subscribers, and describes data sessions of certain services (*i.e.*, Facebook, Google Services, P2P applications, and WhatsApp). Thus, we ignore its incomplete data traffic information and only use its mobility information as a complement to `CDR` dataset. Each data flow record contains the hashed device identifier, the type of service, the volume of exchanged upload and download data, the timestamps denoting the starting and ending time of the session, and the location of the cell tower when the session started.

Spatial Characterization Each data flow record has a location corresponding to the cell tower where the traffic was generated as in `CDR` dataset. `Flow` and `CDR` datasets have locations collected from the same observing area and the same cell tower deployment. Thus, they have the same spatial granularity.

Temporal Characterization Thank to the high frequency of allocating and revoking of radio channels, each user has considerably more records in `Flow` than in `CDR`, which also can provide more human footprints. The temporal heterogeneity of records still exists but in a far less degree than in `CDR`. Due to the fact that the data is missing in some nighttime hours, we mainly use this dataset during daytime hours. For that, we plot in Figure 3.4(a)(b) the distributions of the average inter-event time and the average number of records per user during the time period (10am, 6pm). We note that, in 98% of cases, the inter-event time is less than 5 minutes, and only in less than 1% of cases, the inter-event time is higher than 10 minutes. Besides, most of the users have their mean daily numbers of records ≥ 100. Consequently, `Flow` dataset can provide enough mobility information to construct a fine-grained mobility

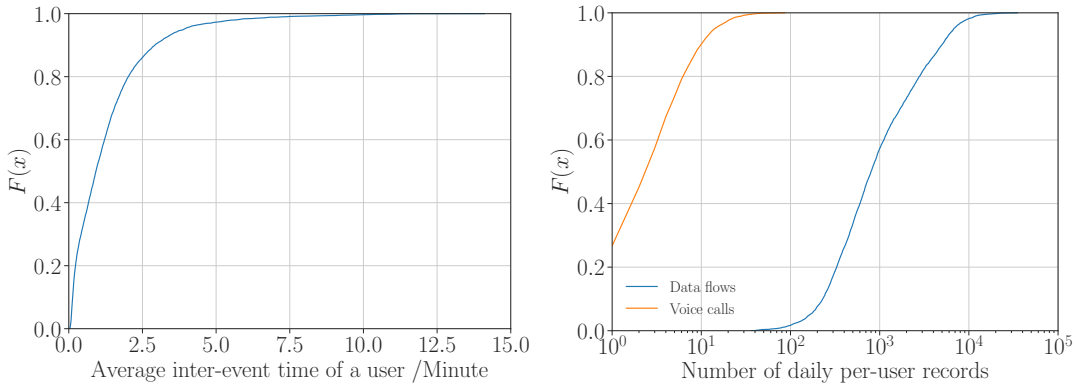


Figure 3.4: Distributions of (a) the inter-event time and (2) the number of daily data records per user during (10am, 6pm) in `Flow` dataset.

dataset of its users. The coarse-grained version of the same users' mobility data can also be constructed by extracting the records in `CDR` dataset.

3.2.4 Shanghai Dataset

This dataset is collected from 642K mobile phone subscribers. It provides aggregated human footprints in the frequency of one location per hour during a period of 2 weeks. The locations in this dataset are gathered by merging the locations of CDR in each one-hour interval, and thus do not appear all the time due to the temporal heterogeneity of charging events. Each user has a trajectory with 336 one-hour time slots in the 14-day observing period. Each time slot has either a geographical coordinate or a *unknown* mark that represents a missing location.

Spatial Characterization Each location of an hour represents the user's centroid of the hour with the precision of 200 meters according to the instruction of the data provider. This accuracy of positioning is higher than that of `CDR` or `Flow` datasets. Yet, the data provider neither explains which LCS method is used in positioning nor provides the cell tower deployment. This limits the use of this dataset.

Temporal Characterization We plot in Figure 3.5 the CDF of the ratio of known locations (*i.e.*, the completeness) of each trajectory. We see that, although most of the trajectories are highly incomplete, 7.6% of the trajectories, *i.e.*, approximately 48K users, have their completeness larger than 0.8, which means that each trajectory has locations in 20 hours per day on average. These highly completed trajectories are used in this thesis: considering the high precision of the locations, we use them as ground-truth of the mobility data in the validation of our CDR completion approach (*cf.* Chapter 4).

3.3 Application-based Mobility Datasets

We use the `Geolife` and `MACACO` application-based datasets to have more fine-grained mobility data. They were collected from two separate mobile phone applications, and in the thesis, are

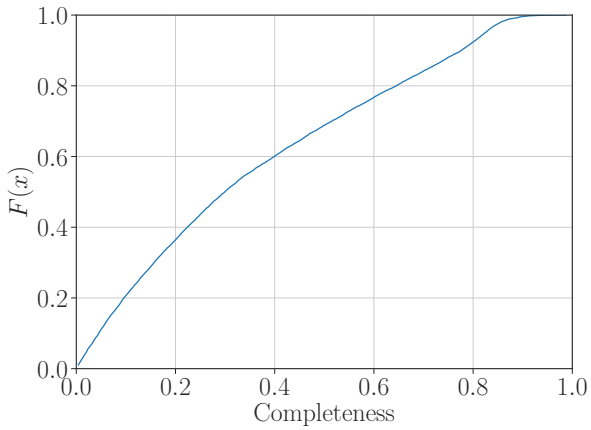


Figure 3.5: Distribution of the completeness of the 642K trajectories in `Shanghai` dataset with the temporal resolution of one hour.

processed collaboratively with the `CDR` dataset. The two datasets are introduced along with the generation of their coarse-grained equivalents in the following.

3.3.1 Geolife Fine-grained Dataset

This is one of the first publicly released datasets containing mobility data. In this thesis, we use its latest version [25]. It provides time-stamped GPS locations of 182 individuals, mostly in Beijing [25]. The dataset spans a three-year time period, from April 2007 to August 2012. Unfortunately, the `Geolife` dataset is often characterized by large temporal gaps between subsequent data records. As a result, not all users present a number of locations or mobility level sufficient to our analysis. Hence we select users given the criteria that the entropy rate of each individual's data points falls below the theoretical maximum entropy rate, which are used in [67] to select the Geolife users for analyzing individual human mobility.

3.3.2 MACACO Fine-grained Dataset

This dataset is obtained through an Android mobile phone application, MACACOApp, developed in the context of the EU CHIST-ERA MACACO project [49]. The application collects the data related to the user's digital activities such as used mobile services, generated uplink/-downlink traffic, available network connectivity, and visited GPS locations. These activities are logged with a fixed periodicity of 5 minutes. We remark that this sampling approach differs from those employed by popular GPS tracking projects, such as MIT Reality Mining [119] or GeoLife [25]. With respect to such previous efforts, the regular sampling in the data grants a neater and more comprehensive overview of a user's movement patterns. The data covers 84 users who live in 6 different countries and travel worldwide. The data collection spans 18 months approximately, from July 10, 2014, to February 4, 2016.

3.3.3 Generating Coarse-Grained Equivalents

We do not have access to any operator-collected dataset that shares common users with `MACACO` or `Geolife` datasets, while our analysis requires the CDR entries of the users in the two

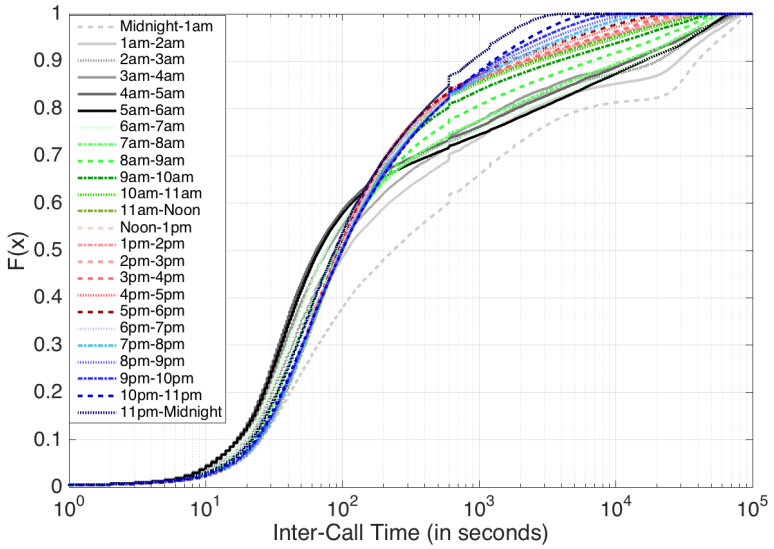


Figure 3.6: CDF of the inter-event (voice call) time in `CDR` dataset collected on an hourly basis.

datasets. For that, we generate CDR-equivalent coarse-grained datasets by leveraging the experimental distributions of the inter-event time in the `CDR` dataset, which refers to CDR from voice calls, as shown in Fig. 3.6. Specifically, we downsample `MACACO` and `Geolife` datasets so that the inter-event times match those in the experimental distributions. Therefore, we first randomly choose one GPS record of the user as the seed CDR entry. We then randomly choose an inter-event time value from the distribution for the corresponding hour of the day, and use such interval to sample the second GPS record for the same user, mimicking a new CDR entry. We repeat this operation through the whole fine-grained trajectories of all users, and obtain datasets of downsampled GPS records that follow the actual inter-event time distributions of CDR.

Note that tailoring the inter-event distribution on a specific hour of the day allows taking into account the daily variability of CDR sampling. Also, upon downsampling, we filter out users who have an insufficient number of records, *i.e.*, users with less than 30 records per day on average or less than 3 days of activity. The final CDR-like coarse-grained versions of the `MACACO` and `Geolife` datasets contain 32 and 42 users, respectively.

Preparing Combinations of Coarse- and Find-grained Dataset With the availability of our datasets, we can pair some datasets into the combination of a *coarse-grained* one and a *fine-grained* one that have the same users. Such combinations of the datasets are necessary in our analysis. In particular, *coarse-grained* datasets provide CDR entries and feature significant spatiotemporal sparsity as well as user locations mapped to cell tower positions; *fine-grained* datasets describe the mobility of the same user populations in the coarse-grained datasets with a much higher level of details and spatial accuracy, used as ground-truth to validate the results.

In this context, our `CDR` and `Flow` datasets share the same set of subscribers, and thus represent a readily usable pair of coarse- and fine-grained datasets. Besides, the coarse-grained counterparts of `MACACO` and `Geolife` datasets are instead artificially generated above by downsampling the original fine-grained data. As a result, we have three combinations of fine- and coarse-grained datasets, as shown in Figure 3.7.

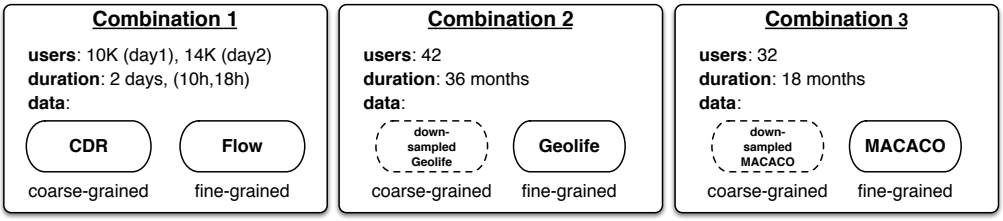


Figure 3.7: Combinations of corresponding coarse- and fine-grained datasets.

The combination of the `CDR` and `Flow` datasets needs some explanation. The number of common users shared by the two datasets varies significantly every day. In order to have a good number of users, we only select two days (a Sunday and a Monday) from `Flow` dataset and focus on the (10 am, 6 pm) time interval. In particular, over 10K and 14K subscribers recorded on Sunday and Monday (as shown in Table 3.1) and are separated into two similarly sized classes based on the number of their voice calls in `CDR` dataset as follows:

- *Rare CDR users* are not very active in placing or receiving voice calls and thus have limited records in the `CDR` dataset. As in [120], we use the threshold of 0.5 event/hour, a user having CDR below this threshold is considered to belong to this category.
- *Frequent CDR users* are more active callers or callees and have more than 0.5 event/hour in the `CDR` dataset.

These selected users have actual CDR entries from `CDR` dataset and ground-truth of their mobility from `Flow` dataset.

Day of the week	Users	Rare CDR users	Frequent CDR users
Sunday	10,856	6,154	4,702
Monday	14,353	7,215	7,138

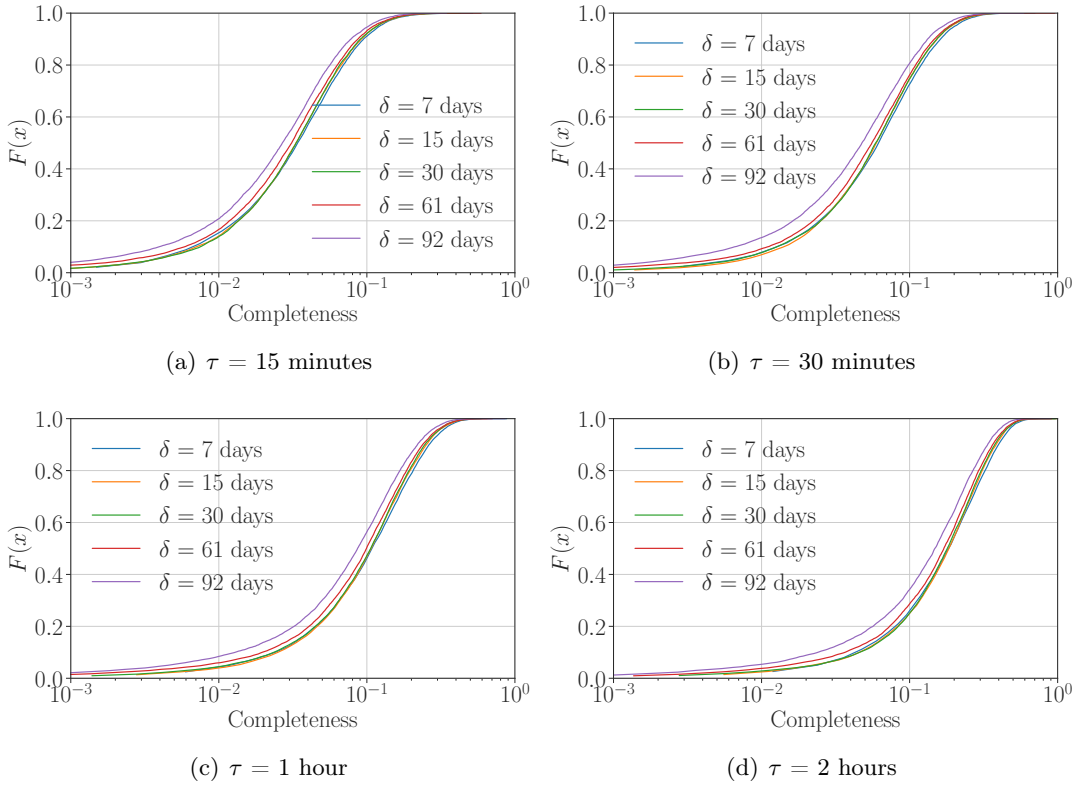
Table 3.1: Number of users in the combination of `CDR` and `Flow` datasets

3.4 Challenge of Completeness

In this section, we present the challenge in the use of operator-collected datasets with respect to the completeness of mobility information. The `CDR` dataset collected from voice calls is our major source of mobility information. The sequence of time-stamped CDR associated to a mobile device provides an implicit and approximated sampling of the device's trajectory. We refer to such a sampling as a *CDR-based trajectory*. Due to temporally varying human activities on making and receiving voice calls, CDR-based trajectories are typically fairly incomplete. In the light of this situation, a relevant question arises, *i.e.*, *what degree of (in)completeness can one expect in CDR-based trajectories inferred from real-world operator-collected datasets?*

3.4.1 Limited Completeness in `CDR` dataset

To answer this question in a structured manner, we first define the notion of *completeness* of a CDR-based trajectory as the fraction of time intervals during which the location of a



(a) $\tau = 15$ minutes

(b) $\tau = 30$ minutes

(c) $\tau = 1$ hour

(d) $\tau = 2$ hours

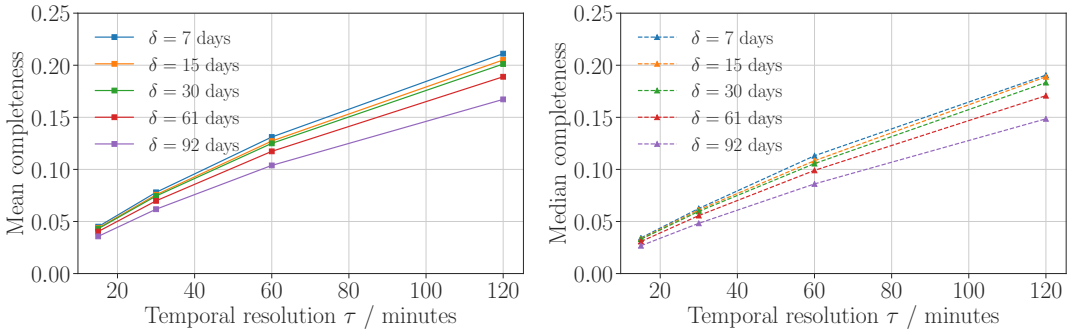
Figure 3.8: Distributions of the completeness of the CDR-based trajectories of 1.8M users for combinations of the observing period δ and resolution τ.

mobile subscriber is recorded at least once. This definition implies that the overall observing period covered by the dataset, referred to as δ, is discretized into time intervals of which the resolution we denote as τ. For example, given a dataset with the observing period $\delta = 7$ days and the resolution $\tau = 1$ hour, a CDR-based trajectory with locations in 80 different hours has a completeness ratio as $80/(7 \times 24) = 0.476$.

Intuitively, the completeness of CDR-based trajectories depends on both the observing period δ and resolution τ. We investigate how the two parameters affect the completeness. We consider all combinations of the observing period $\delta \in \{7, 15, 30, 61, 92\}$ days and the resolution $\tau \in \{15, 30, 60, 120\}$ minutes. Given each combination of an observing period δ and a resolution τ we generate CDR-based trajectories and compute their completeness from a consecutive 3-month period of `CDR` dataset.

Figure 3.8 portrays the CDF of the resulting CDR-based trajectory completeness. Each plot refers to a different resolution and curves map to diverse observing periods. We make the following observations.

- Whichever combination is chosen, fully complete trajectories are very hard to obtain from CDR directly. Not a single complete trajectory is inferred from our reference dataset despite the million-scale user population covered, even when combining the shortest observing period $\delta = 7$ days with the lowest resolution $\tau = 2$ hours.
- Interestingly, the completeness is only very slightly affected by the period covered by the dataset. As one could expect, CDR-based trajectories that span a shorter time period



(a) (b)

Figure 3.9: The (a) mean and (b) median completeness versus the resolution τ with respect to the observing period δ.

tend to have a higher completeness while the difference of the completeness remains fairly small and never exceeds on each observing period δ (from 7 to 92 days). In fact, the completeness hardly varies at all when the observing period $\delta \leq 30$ days.

- Instead, the completeness is significantly affected by the resolution τ. For instance, considering a dataset with the observing period $\delta = 30$ days, only 10% of the CDR-based trajectories display their completeness above 0.1 when the resolution τ is 15 minutes; this percentage grows to 75% when the resolution is 120 minutes. To better exploit this observation, we plot in Figure 3.9 the mean and median completeness versus the resolution τ over all the observing periods, which show that the completeness and the resolution are almost linearly correlated. In fact, ignoring the observing period, the Pearson correlation coefficient between the two parameters is still 0.983.

Overall, our results highlight that, although CDR never capture fully complete trajectories, the vast majority of CDR-based trajectories features a non-negligible level of the completeness (*e.g.*, $0.1 - 0.8$) given an appropriate resolution (*e.g.*, in the order of hours). These trajectories may be not all adaptive to the mobility studies, but, as we will show in Chapter 4, they retain sufficient information to be salvaged by our CDR completion techniques.

3.4.2 Statistical Regularity of the Completeness

We also provide a data-driven model of the completeness of CDR-based trajectories. To this end, we derive the theoretical distributions that best fit the empirical CDF shown in Figure 3.8 under all the combinations of the observing period δ and resolution τ.

Earlier studies have shown that the distribution of charging events (*i.e.*, voice calls and text messages) across mobile subscribers tends to be long-tailed [121]. Since the completeness of CDR-based trajectories is determined by the user's level of events, it makes sense to look at heavy-tailed distributions. We run a maximum likelihood estimation of the parameters of six standard long-tailed distributions (namely, Generalized Pareto, Lévy, Power-law, Lognormal, Gamma, and Weibull) on the empirical CDF of the completeness. We evaluate the quality of the fitting by both computing the coefficient of determination (denoted as R^2) and running the Kolmogorov-Smirnov test (whose output, *i.e.*, the Kolmogorov-Smirnov statistic, is indicated as D_{KS}) on the empirical and estimated distributions.

Table 3.2: R^2 and D_{KS} of the best fits of the six standard long-railed distributions

		Weibull		Lognormal		Gamma		GenPareto		Levy		Powerlaw	
δ	τ	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2	D_{KS}	R^2
$7d$	$15min$	0.0334	**0.9990**	**0.0318**	0.9973	0.3710	0.4679	0.0495	0.9969	0.2509	0.8046	0.3538	0.5031
	$30min$	**0.0302**	**0.9993**	0.0345	0.9967	0.0548	0.9962	0.1716	0.8973	0.2649	0.7848	0.3587	0.4894
	$60min$	**0.0278**	**0.9990**	0.0372	0.9958	0.0475	0.9977	0.1198	0.9516	0.2888	0.7506	0.4162	0.2041
	$120min$	**0.0375**	**0.9972**	0.0443	0.9938	0.0629	0.9853	0.1043	0.9768	0.3238	0.6977	0.2456	0.7450
$15d$	$15min$	0.0264	**0.9986**	**0.0233**	0.9984	0.3594	0.5007	0.0627	0.9893	0.2620	0.7853	0.3949	0.4120
	$30min$	**0.0208**	**0.9995**	0.0264	0.9980	0.0271	0.9991	0.0724	0.9851	0.2765	0.7647	0.3576	0.4975
	$60min$	**0.0188**	**0.9997**	0.0279	0.9974	0.0257	0.9987	0.0935	0.9719	0.2997	0.7315	0.3176	0.6076
	$120min$	**0.0254**	**0.9987**	0.0332	0.9961	0.0913	0.9572	0.1880	0.8780	0.3349	0.6792	0.2721	0.7116
$30d$	$15min$	0.0239	**0.9985**	**0.0207**	**0.9985**	0.3514	0.5261	0.0700	0.9835	0.2619	0.7872	0.3829	0.4365
	$30min$	0.0205	**0.9992**	0.0216	0.9983	**0.0203**	0.9991	0.1528	0.8976	0.2763	0.7661	0.3912	0.4149
	$60min$	**0.0149**	**0.9996**	0.0263	0.9975	0.0289	0.9982	0.0895	0.9702	0.3003	0.7346	0.3131	0.6212
	$120min$	**0.0239**	**0.9984**	0.0315	0.9962	0.0995	0.9479	0.1069	0.9538	0.3337	0.6893	0.3154	0.6176
$61d$	$15min$	0.0266	**0.9980**	**0.0239**	0.9977	0.1990	0.8284	0.0458	0.9934	0.2527	0.8076	0.3850	0.4156
	$30min$	**0.0233**	**0.9985**	0.0234	0.9976	0.0286	0.9974	0.1944	0.8669	0.2649	0.7903	0.3897	0.4212
	$60min$	**0.0207**	**0.9985**	0.0264	0.9970	0.0310	0.9980	0.0772	0.9769	0.2883	0.7622	0.3451	0.5635
	$120min$	**0.0245**	**0.9975**	0.0316	0.9958	0.0754	0.9750	0.0946	0.9617	0.3209	0.7196	0.3491	0.5070
$92d$	$15min$	**0.0298**	**0.9954**	0.0329	**0.9954**	0.3619	0.5248	0.0322	0.9954	0.2371	0.8390	0.3866	0.4528
	$30min$	0.0336	**0.9958**	**0.0335**	0.9950	0.0583	0.9864	0.0398	0.9942	0.2487	0.8264	0.3819	0.4774
	$60min$	**0.0305**	**0.9960**	0.0355	0.9944	0.0454	0.9913	0.0611	0.9843	0.2705	0.8020	0.3231	0.6123
	$120min$	**0.0369**	0.9940	0.0379	0.9932	0.0486	0.9927	0.0760	0.9743	0.2998	0.7687	0.3359	0.5885

Table 3.2 summarizes the results. Two distributions fit far better than the others as they occupy most of the maximum values of R^2 and the minimum values of D_{KS} among all the combinations of the observing period and resolution: they are Lognormal with the PDF as follows:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{\ln(x/\mu)^2}{2\sigma^2}}, \tag{3.1}$$

and Weibull with the PDF as follows:

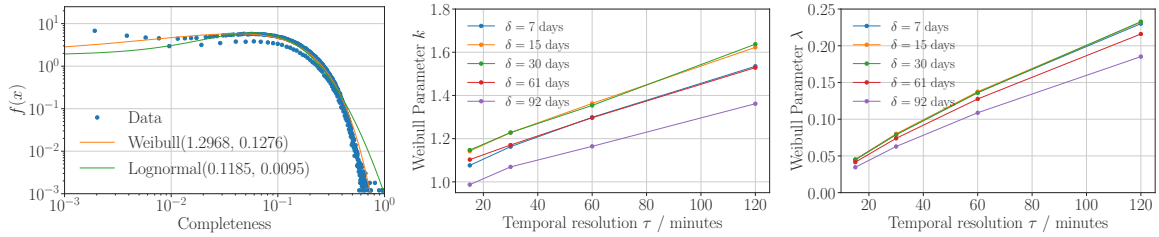
$$f(x) = \begin{cases} \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(x/\lambda)^k} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}. \tag{3.2}$$

An illustrative example for the observing period $\delta = 61$ days and resolution $\tau = 60$ minutes is shown in Figure 3.10(a). We also remark that, quite as expected, the models are well aligned with those of charging events (*e.g.*, voice calls) found in the literature [121].

The theoretical modeling of the completeness distributions paves the way to an in-depth analysis of how the system settings affect the model parameters. We focus on the Weibull model, of which the fittings are in general better than those of the LogNormal model[1]. Figure 3.10(b) and Figure 3.10(c) show an interesting phenomenon together: both parameters of the Weibull model, k and λ, scale linearly with the resolution τ for any observing period δ. More precisely, the Pearson correlation coefficient with respect to the resolution τ of k and λ are 0.89 and 0.98, respectively.

Although the generality of these relationships needs more confirmation on other CDR datasets, they open interesting perspectives on the possibility to characterize the level of completeness of CDR-based trajectories directly from the time resolution τ.

[1] A possible reason is that the bursting of events pushes multiple human footprints into a time slot as a single location in the calculation of the completeness so that the distribution is shifted from Lognormal to Weibull.



(a) (b) (c)

Figure 3.10: (a) Empirical PDF and Weibull, Lognormal theoretical fittings of the completeness of CDR-based trajectories with the observing period $\delta = 61$ days and resolution $\tau = 60$ minutes. (b)(c) Linear correlation of the Weibull k and λ parameters versus resolution τ for the varying observing period δ.

3.5 Challenge of Mobility Measurement

In this section, we present the challenge in the use of operator-collected datasets with respect to the measurement of mobility features. Recall that human footprints such datasets are usually coarse in space and sparse in time, which may affect the validity of the results, which has been partly investigated (*cf.* Section 2.2.4). In the following, we present an updated analysis of the suitability of common CDR datasets, *i.e.*, those captured by voice calls. Our analysis in this section is based on the dataset combinations that we have constructed (*cf.* Section 3.3.3).

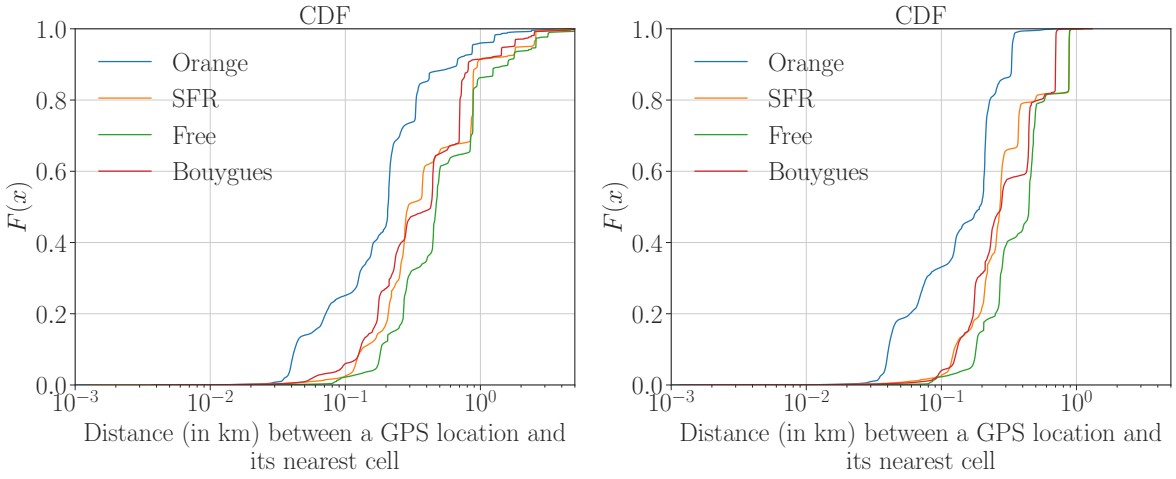
3.5.1 User Mapping of Cell Tower Locations

Recall that the position information in CDR is represented by the cell tower location handling the charging events (*cf.* Section 2.2.1). Hence, a shift from the user's actual location to the cell tower location always exists in every CDR entry. Such a shift may impact the accuracy of individual mobility measurements. Usually, CDR are collected in metropolitan areas. In this case, the precision of human locations provided by CDR is related to the local deployment of base stations. Fig. 3.1 shows the deployment of cell towers in the metropolitan area where our `CDR` dataset was collected. The presence of cell towers is far from uniform, with a higher density in downtown areas where a cell tower covers an approximately 2 km^2 area on average: in these cases, the cell coverage grants a fair granularity in the localization of mobile network subscribers. The same may not be true for cells in the city outskirts, which cover areas of several tens of km^2.

We evaluate how the cell deployment can bias human mobility studies. To this end, we perform a quantitative analysis of spatial shifts introduced by CDR positioning information by leveraging GPS logs in the `MACACO` dataset. Our focus on the `MACACO` dataset is due to two reasons: *(i)* the `Flow` and `CDR` datasets lack GPS information of visited locations or only provide cellular-level information of visited locations of their users; *(ii)* no available reliable source allows extraction of the cell tower information (*i.e.*, coordinates or covered area of deployed cell towers) in the area of Beijing that the `Geolife` users are mainly from.

We first extract 718, 987 GPS locations in the mainland of France[2] from the `MACACO` dataset.

[2]The study focuses on the area in the latitude and longitude ranges of $(43.005, 49.554)$ and $(-1.318, 5.999)$, respectively.



(a) Entire country

(b) Metropolitan areas

Figure 3.11: Distributions of the distances to the nearest cell tower (shifts), for 718,987 GPS locations in the MACACO data of users in (a) the whole area and (b) major metropolitan areas (Paris Region, Lyon, Toulouse) in France.

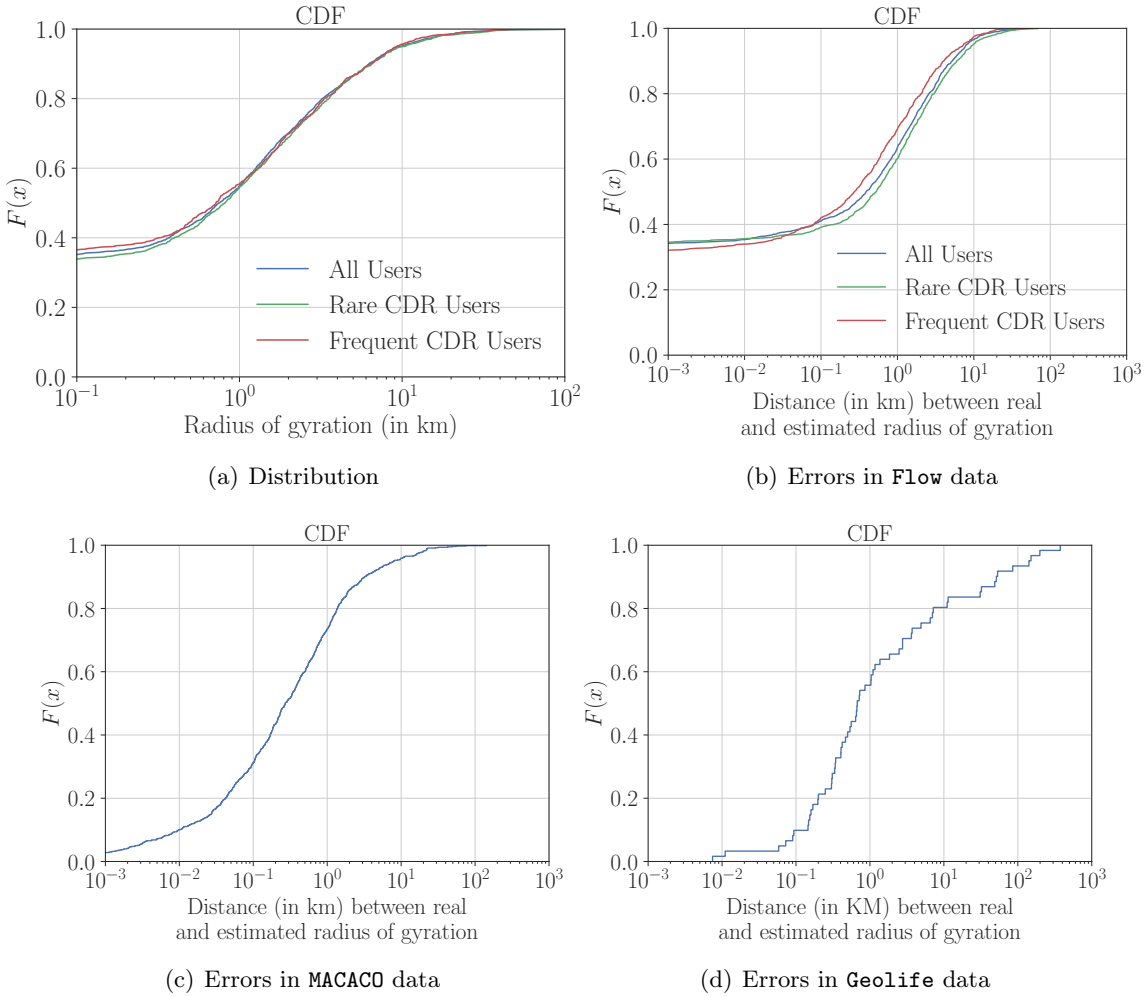
Among these locations, 74% are collected from the major metropolitan areas in France, including Paris Region, Lyon, and Toulouse. We then extract cell tower locations of the four major mobile network operators in France (*i.e.*, Orange, SFR, Free, and Bouygues) from the open government data [102].

Fig. 3.11(a) is the CDF of the distance between each GPS location in the MACACO dataset and its nearest cell tower. We observe that most of the locations have distances below 1 km when shifting to their nearest cells (*i.e.*, 95% for Orange, 91% for SFR, 86% for Free, and 91% for Bouygues). Nevertheless, when we focus on the metropolitan areas as shown in Fig. 3.11(b), almost all the shifts (*i.e.*, over 99%) are below 1 km and all the operators have their median shifts around $200 - 500$ meters. This indicates that the shifts above 1 km are all observed in rural areas. Still, most of the shifts are higher than 100 meters, indicating the presence of some biases of using cell tower locations. We stress that these values provide an lower bound to the positioning error incurred by CDR, as mobile devices may be associated to antennas that are not the nearest ones due to the signal propagation phenomena or load balancing policies enacted by the mobile network operator.

Still, the level of accuracy in Fig. 3.11, although far from that obtained from GPS logs, is largely sufficient for a variety of metropolitan-level or inter-city mobility analyses. For instance, it was shown that a spatial resolution of $2 - 7$ km is sufficient to track the vast majority of mobility flows in a large dual-pole metropolitan region [122].

3.5.2 Human Movement Span

We then examine whether mining CDR data is a suitable means for measuring the geographical span of movement of individuals. For that, we compute for each user u in the set of study $\mathcal{U}$ the *radius of gyration*, *i.e.*, the deviation of the user's positions to their centroid. Formally, $R_g^u = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{r}_i^u - \mathbf{r}_{\text{centroid}}^u||_{\text{geo}}^2}$, where $\mathbf{r}_{\text{centroid}}^u$ is the center of mass of locations of the



(a) Distribution

(b) Errors in `Flow` data

(c) Errors in `MACACO` data

(d) Errors in `Geolife` data

Figure 3.12: (a) CDF of the radius of gyration of two categories (Rare and Frequent) of CDR users in the `Flow` dataset. (b)(c)(d) CDF of the distance between the real and the estimated radius of gyration from CDR over the users of the (b) `Flow`, (c) `MACACO`, and (d) `Geolife` datasets.

user u, *i.e.*, $\mathbf{r}^u_{\text{centroid}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{r}^u_i$. This metric reflects how widely the subscribers move and is a popular measure used in human mobility studies [16, 27, 120, 123]. An individual who repeatedly moves among several fixed nearby locations still yields a small radius of gyration, even if he may total a large traveled distance.

We are able to compute both *estimated* (due to the temporal sparsity of the actual or the equivalent CDR data) and *real* (due to the finer granularity in the ground-truth provided by the `Flow`, `MACACO`, and `Geolife` datasets) radius of gyration for each user. Fig. 3.12 summarizes the results.

Let us first consider the users of the `Flow` dataset and their radii of gyration. Three curves denote different cases: *all*, *rare*, and *frequent* CDR users. The associated radius of gyration CDF are portrayed in Fig. 3.12(a). The three distributions are quite similar, indicating that one can get a reliable distribution of R^u_g from a certain number of users even if they are rare CDR users, *i.e.*, have a limited number of voice calls.

When considering the error between real and estimated radius of gyration, in Fig. 3.12(b) for the `CDR` and `Flow` datasets, and in Fig. 3.12(c) and 3.12(d) for the `MACACO` or `Geolife` datasets, we observe the following:

- The distribution of large errors is similar in all cases and outlines a decent accuracy of the coarse-grained CDR or CDR-like datasets. For approximately 90% of the `Flow` users, 95% of the `MACACO` users and 70% of the `Geolife` users, the errors between the real and the estimated radius of gyration are less than 5 km. The higher errors obtained from the `Geolife` dataset may be interpreted by the irregular sampling in the original data and the presence of very large gaps between consecutive logs.
- A more accurate radius of gyration can be obtained for the CDR users who are especially active: 92% of the frequent CDR users have their errors lower than 5 km, while the percentage decreases to 86% for the rare CDR users.
- When considering small errors, the distributions tend to differ, with far lower errors in the `CDR` than in the `MACACO` or `Geolife` dataset. This is in fact an artifact of considering cell tower locations as the ground-truth user positions in the fine-grained `Flow` dataset. In the more accurate GPS data of `MACACO` and `Geolife`, around 30% and 10% of the users enjoy their errors lower than 100 meters, while 35% of the users in the `CDR` dataset have errors below 1 meter.

Overall, these results confirm the previous findings on the limited suitability of CDR for the assessment of the spread of human mobility [46]. They also unveil how different datasets can affect the data reliability at diverse scales.

3.5.3 Missing Locations

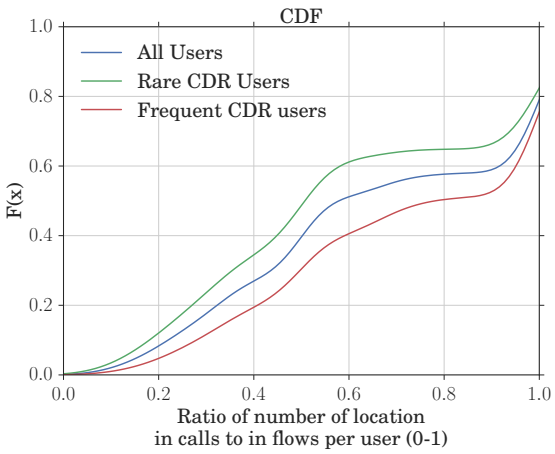
Due to spatiotemporal sparsity, the mobility information provided by CDR is usually incomplete. We investigate the phenomenon in the case of users in the `CDR` dataset, and plot in Fig. 3.13(a) the ratio r_{N_L} of unique locations detected from CDR (N_L^{CDR}) to those from the ground-truth (N_L^{Flow}), *i.e.*, `Flow` data, as follows:

$$r_{N_L} = N_L^{\text{CDR}} / N_L^{\text{Flow}}. \tag{3.3}$$

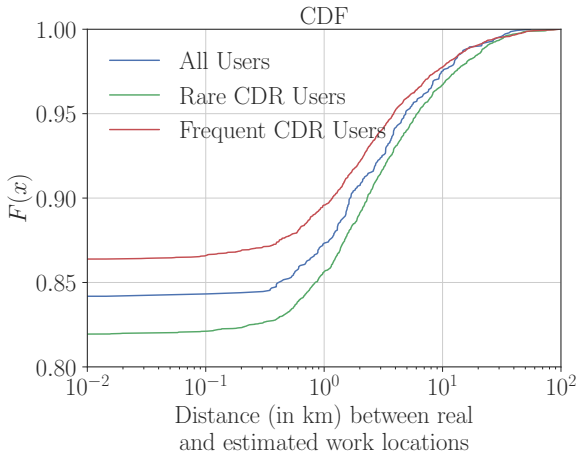
We notice that 42% in the population of study (*i.e.*, all users) have their r_{N_L} higher than 0.8. For these users, 80% of their unique visited locations appear in their CDR. The percentage of all users having this criterion is slightly higher for the frequent CDR users (50%) and lower for the rare CDR users (37%). These results confirm that using CDR to study very short-term mobility patterns is not a good idea due to the high temporal sparsity and the lack of human footprints captured by CDR.

3.5.4 Important Locations

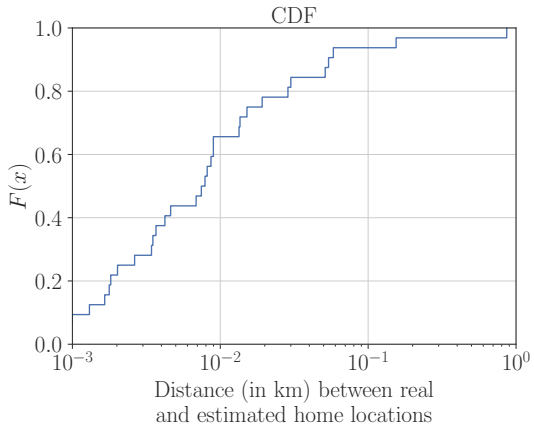
The identification of significant places where people live and work is generally regarded as an important step in the characterization of human mobility. Here, we focus on home and work locations: we separate the period of study into two time windows, mapping to work time (9 am to 5 pm) and night time (10 pm to 7 am) for both CDR-like and ground-truth datasets. For each user, the places where the majority of work time records occur are considered a proxy for



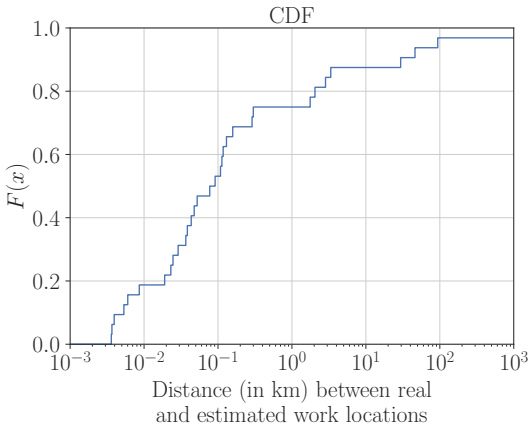
(a) Missing Ratio



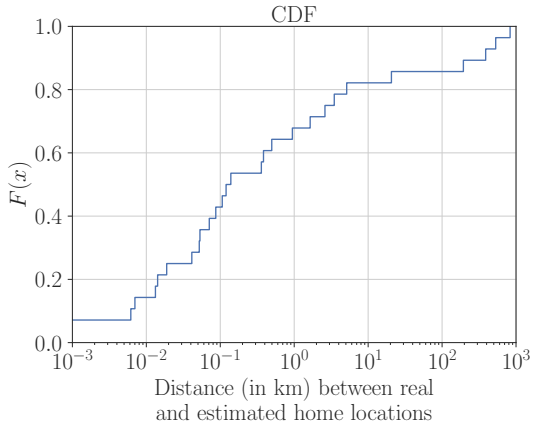
(b) Errors in Flow data: Work



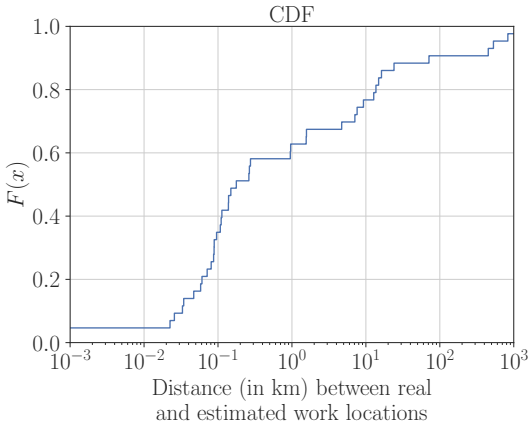
(c) Errors in `MACACO` data: Home



(d) Errors in `MACACO` data: Work



(e) Errors in `Geolife` data: Home



(f) Errors in `Geolife` data: Work

Figure 3.13: (a) CDF of the radio r_{N_L} of the number of locations in each user's coarse-grained trajectory to the one in his fine-grained trajectory. (b)(c)(d)(e)(f) CDF of the distances between each user's real and estimated important locations located by his CDR and ground-truth: (b) work locations over the `Flow` users; (c) home and (d) work locations over the `MACACO` users; (e) home and (f) work locations over the `Geolife` users.

work locations; the equivalent records at night time are considered a proxy for home locations as in [124]. It is worth noting that, as the `Flow` dataset covers only $(10am, 6pm)$, we only infer work locations for this dataset.

Formally, let us consider a user u from the user set. The visiting pattern of the user u is a sequence of samples $\{(\ell_u^1, t_u^1), \ldots, (\ell_u^n, t_u^n)\}$, where the i-th sample (ℓ_u^i, t_u^i) denotes the location ℓ_u^i where the user u is recorded at time t_u^i. The home location ℓ_u^H of the user u is then defined as the most frequent location during night time:

$$\ell_u^H = mode(\ell_u^i \mid t_u^i \in t^H), \tag{3.4}$$

where t^H is the night time interval. The definition is equivalent for the work location ℓ_u^W of the user u, computed as

$$\ell_u^W = mode(\ell_u^i \mid t_u^i \in t^W), \tag{3.5}$$

where t^W is the work time interval.

We use the definitions in (3.4) and (3.5) to determine home and work locations and then evaluate the accuracy of the CDR-based significant locations by measuring the geographical distance that separates them from the equivalent locations estimated via the corresponding fine-grained ground-truth datasets.

The results are shown in Fig. 3.13(b)-(f) as the CDF of the spatial error in the position of home and work places for different user groups for the three datasets. We observe the following:

- The errors related to home locations are fairly small in the `MACACO` dataset, but are relatively higher in the `Geolife` dataset. For the `MACACO` users, the errors are always below 1 km and 94% are within 100 meters. For the `Geolife` users, we observe that 17% of the errors are higher than 10 km. A possible interpretation is that some `Geolife` users are highly active and do not stay within a stable location during nighttime.

- For both `MACACO` and `Geolife` users, the errors associated with work locations are sensibly higher than those measured for home locations. For instance, as shown in Fig. 3.13(d), while 75% of the `MACACO` users have an error of less than 300 meters, the work places of a significant portion of individuals (around 12%) are identified at a distance higher than 10 km from the positions extracted from the GPS data. A close behavior can be noticed from the `Flow` and `Geolife` users, as shown in Fig. 3.13(b) and Fig. 3.13(f). These large errors typically occur for users who do not seem to have a stable work location and may be working in different places depending on, *e.g.*, the time of day.

- The errors are significantly reduced when using cell tower locations as in the `Flow` dataset instead of actual GPS positions as in the `MACACO` or `Geolife` datasets. For the `Flow` users in Fig. 3.13(b), the errors between the real and the estimated significant locations are null for approximately 85% of all users, indicating that the use of the coarse-grained dataset is fairly sufficient for inferring these significant locations.

- The errors are non-null for the remaining `Flow` users (15%). Among them, 10% have relatively small errors (less than 5 km), while 5% have errors larger than 5 km.

- There is only a slight difference in the distribution of the errors associated with work locations between the rare and the frequent CDR users as shown in Fig. 3.13(b). The reason is that, most of CDR are generated in significant locations, and hence the most

frequent location obtained from CDR of a user is likely to be his actual work location during daytime. Still, it is relatively difficult to capture actual location frequencies if a user has only a few of CDR. Hence the rare CDR users have higher errors.

Overall, these results confirm previous findings [46], and further prove that CDR yield enough details to detect significant locations in users' visiting patterns. Besides, the results reveal a small possibility of incorrect estimation in the ranking among such locations.

3.6 Challenge of Data Processing

Besides the challenges discussed in the two previous sections, particular features also bring important processing issues, such as the millions of records, the extremely large population, the long observing period. In the next, we discuss the challenges that we have met and addressed when processing our datasets.

3.6.1 High Shrinkable Population of Study

Our spatiotemporal analysis of mobile data traffic demand requires operator-collected datasets describing time-stamped visited locations of mobile devices as well as their generated data traffic at those locations. For this, the merge of records of different datasets (*e.g.*, `CDR` and `Session` datasets) with the common users is sometimes necessary, while it can significantly reduces the availability of the user population. The population shrinking is mainly due to the incompleteness of mobility information: a large amount of users do not have enough visited locations captured in the datasets and have to be filtered out as in relevant studies [16, 17, 24]. What is worse, other filters regarding mobile data traffic are applied on the users in our case, which further threaten the scale of the user population. To address this challenge, we propose techniques for the `CDR` completion in order to fill the spatiotemporal gaps in user trajectories and to reduce the number of filtered users from the datasets, which is later presented in Chapter 4.

3.6.2 High Computational Cost

Analyzing individual behaviors by mining `CDR` and `Session` datasets is a time-consuming task. For each user, we apply the completion techniques (*cf.* Chapter 4) and the prediction models on his trajectory of locations and time series of data traffic volumes (*cf.* Chapter 5). An individual task like this costs a small CPU power and thus takes a short time. For an application-based dataset, such an analysis of individual behaviors is highly feasible thanks to the small population. Nevertheless, when we deal with thousands or even millions of users in `CDR` and `Session` datasets, its large population makes such an analysis spend a large amount of time. Although in our case the analysis can be naturally paralleled in the level of users, it still takes ten of days or even weeks to perform on a personal workstation or a standalone server due to their small numbers of CPU cores.

We address this high computational cost from two aspects. On one side, we carefully design and implement the program by leveraging the processor-level enhancement. For instance, we utilize the Intel Math Kernel Library to have highly optimized popular mathematical functions, which provides a 5-10x speed-up in terms of the execution time of each individual. On the other side, we specifically implement a version of the analyzing program to perform our analysis

on a high-performance machine cluster using the MPI (Message Passing Interface) along with the other optimization schemes on the data storage and transmission among machines. The cluster has 20 servers and 240 CPU cores in total, providing enough capability for parallelization. Using the cluster reduces the execution time from weeks to less than 5 days. Overall, our experiences have shown that, human behavior analyses on large-scale operator-collected datasets need a careful design of the analyzing program and a powerful hardware platform.

3.7 Summary

In this chapter, we have introduced our datasets and discussed the challenges brought by the utilization of these datasets. The latter leads to several major findings which we summarize in the following.

- We assess the completeness of the mobility information in `CDR` dataset and provide empirical distribution estimates of relevant metrics including the period of coverage, the sampling frequency, and the loss patterns. Our results show that legacy data preliminary techniques risk filtering out users with substantial although irregular location information, thus rejecting useful information.
- We show that the geographical shifts caused by the mapping of user locations to cell tower are less than 1 kilometer in the most of cases (*i.e.*, $85\% - 95\%$ in the entire country or over 99% in the metropolitan areas in France), and the median shift is around $200 - 500$ meters (varying across mobile network operators). This result substantiates the validity of many large-scale analyses of human mobility that employ CDR.
- We provide a confirmation of previous findings in the literature regarding the capability of CDR to model individual movement patterns: (1) CDR provides the limited suitability for the assessment of the spread of human mobility and the study of short-term mobility patterns; (2) CDR yield enough details to detect significant locations in users' visiting patterns and to estimate the ranking among such locations.

Chapter 4

CDR-based Trajectory Completion

Human footprints in operator-collected datasets serve a large and growing body of literature on human mobility analyses [23]. As introduced in Section 2.2, such data usually covers a large amount of users but provide incomplete mobility information; hence one of the common practices of human mobility analyses is to recover the information (*e.g.*, infer missing locations or reconstruct trajectories), while the literature on this aspect is fairly thin. In this context, we address this mobility data recovery problem with respect to the utilization of operator-collected datasets in this chapter, to achieve our secondary goal proposed in Chapter 1.

Particularly, we study this problem to perform our further joint analysis of individual mobility and mobile data traffic demand. From our datasets, we need to extract for each of the users of study a mixed time series of locations and data traffic volumes in a certain temporal granularity. Although our `Session` dataset can offer the complete information regarding data traffic volumes, our `CDR` dataset cannot provide all the locations (*cf.* Section 3.4). Therefore, its incomplete mobility information has to be addressed before we proceed to the joint analysis.

This chapter focuses on the recovery of mobility information in terms of human footprints in operator-collected datasets: a task we name as *CDR completion*. We contribute to this topic in a number of ways as follows.

1. By summarizing the utilization of CDR, we formulate two types of CDR-based trajectories (instant and slotted) and explain how we proceed to enhance their completeness by defining two corresponding CDR completion tasks. Details are given in Section 4.1.

2. For the completion of instant CDR-based trajectories, we implement different existing proposals and propose novel approaches that build on them. We assess the quality of all the techniques in presence of ground-truth data with high spatiotemporal granularity, and show that our proposed techniques outperform the others: on average, we achieve an increased temporal completion of CDR data (75% of daytime hours) and retain significant spatial accuracy (having errors below 1 km in 95% of completed time); compared with the most common proposal in the literature, our best adaptive approach grants 50% increase in completion, and, at the same time, grants 5% higher accuracy. Details are given in Section 4.2.

3. For completing slotted CDR-based trajectories, we propose a hierarchical completion approach powered by mobility characteristics and state-of-the-art techniques of missing value inference. We also validate its quality by comparing with the state-of-the-art proposals using the real-world ground-truth data. Generally, our technique can precisely recover 55% of the missing one-hour time slots from only 10% of known locations in the slotted CDR-based trajectories of our ground-truth, and have fairly low distance errors on the inference of the remaining missing locations. Details are given in Section 4.3.

Finally, Section 4.4 summarizes the chapter.

4.1 Terminology

4.1.1 Instant and Slotted CDR-based Trajectories

Extracted from any typical operator-collected dataset (*e.g.*, our `CDR` dataset), each user of the dataset has an ordered sequence of time-stamped and geo-referenced footprints collected by the user's CDR, which we name as *a CDR-based trajectory* as firstly introduced in Section 3.4. Based on the experience on the utilization of such CDR-based trajectories in the literature [16, 17, 23, 27, 35, 113, 116, 125, 126, 127], we categorize those trajectories into two types, *i.e.*, the *instant* and *slotted* CDR-based trajectories, which are defined in the following.

Definition 4.1: Instant CDR-based trajectory

A user's instant CDR-based trajectory is defined as a finite set of time-stamped locations, formulated as $\{(\mathbf{l}_1, t_1), (\mathbf{l}_2, t_2), \cdots, (\mathbf{l}_n, t_n)\}$, where n is the number of locations and $(\mathbf{l}_k, t_k)$ represents that the user is at the k-th location $\mathbf{l}_k$ at the instant time t_k. This type of trajectories are seen in [16, 27, 35, 125].

Definition 4.2: Slotted CDR-based trajectory

A user's slotted CDR-based trajectory is defined as a N-slot time series of locations: $\{\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_N\}$, where $\mathbf{l}_k$ is the *representative* location (*i.e.*, the centroid or the most occupied location) in the k-th time slot. The observing period is known and is divided into N equivalent slots. If no location sample is observed in a time slot, the location is marked as *unknown*. This type of trajectories are seen in [17, 113, 116, 126, 127].

4.1.2 CDR Completion Tasks

Given a dataset consisting of CDR, for each user, an instant CDR-based trajectory can be naturally extracted from his CDR, and a slotted CDR-based trajectory can be also extracted by choosing a segmenting duration and a representative location of each time slot. The temporal heterogeneity of the primary charging events (*i.e.*, voice calls and text messages) used in the CDR data collection usually leads to incomplete mobility information (*cf.* Section 2.2.1). Therefore, each CDR-based trajectory, whichever type it is, is likely to be highly incomplete. In particular, an instant one may endure the existence of many time periods without any locations/CDR captured; a slotted one may have many locations marked as *unknown*. For instance, we have shown a high degree of incompleteness of the CDR-based trajectories in our `CDR` dataset in Section 3.4. In both the cases, we need *CDR completion* as a rescue. Intuitively, the objective of the CDR completion is to fill spatiotemporal gaps in CDR-based trajectories. With respect to the type of CDR-based trajectories, we define two separate CDR completion tasks in the following.

- **Instant CDR completion task**. For an instant CDR-based trajectory, this is to expand each instant time-stamped cell tower location on both sides of time to a period (*i.e.*, *temporal cell boundary*) that the user dwells in the same location, so as to enlarge the availability of mobility information and to fill the spatiotemporal gaps as much as possible.

- **Slotted CDR completion task**. This is to infer missing values of a slotted CDR-based trajectory for all the time slots having *unknown* locations, so as to fully reconstruct the user's mobility in a certain spatiotemporal resolution and to obtain a fully complete version of the trajectory.

In the remaining sections of this chapter, we evaluate the state-of-the-art techniques and develop our novel techniques for the two tasks. Note that we assume all the locations of a CDR-based trajectory represent the cell towers so that the geographical information remains available not only at instant time points but in durations when the cell towers cover the user. Note that this assumption is in line with the actual CDR data collection (*cf.* Section 2.2.2).

4.2 Completing Instant CDR-based Trajectories

In this section, we address the instant CDR completion task. Particularly, we develop novel techniques that expand an instant cell tower location sample to a dwelling period as a temporal cell boundary. To achieve this goal, we first introduce briefly the preliminary of our datasets in Section 4.2.1. Second, we discuss the existing approaches to this task in Section 4.2.2. Finally, we propose our techniques that achieve improved accuracy and completion, and their data-driven evaluations during nighttime and daytime, in Section 4.2.3 and Section 4.2.4, respectively.

4.2.1 Data Preliminary

For this study, we use the three combinations of our datasets that we introduced in Section 3.3.3. Recall that each combination has two mobility datasets sharing the same users: a *coarse-grained* one consisting of CDR used as our target, and a *fine-grained* one with a higher resolution used in the validation of the results as ground-truth. Considering the observing period and the user population, we use each combination in the following way. For the coarse-grained `CDR` and fine-grained `Flow` datasets, this combination has the largest population and covers the daytime hours, *i.e.*, (10am, 6pm). Therefore, it is used into the CDR completion task of daytime. As to the other two combinations, *i.e.*, the coarse-grained and fine-grained versions of `MACACO` and `Geolife` datasets, they have small user populations and are used into the CDR completion tasks of nighttime.

4.2.2 Current Approaches to Instant CDR Completion

Recall that the CDR completion aims at filling time gaps among CDR by estimating users' locations in between their charging events. Several strategies for the instant CDR completion task have been proposed to date. In this section, we introduce and discuss the two most popular solutions adopted in the literature.

4.2.2.1 Basic Solution: `Static`

A simple solution is to hypothesize that a user remains static at the same location where he is last seen in his CDR. This methodology is adopted, *e.g.*, by Khodabandelou *et al.* [128] to compute subscriber's presence in mobile traffic meta-data used for population density estimation. We will refer to this approach as the `static` solution and will use it as a basic benchmark

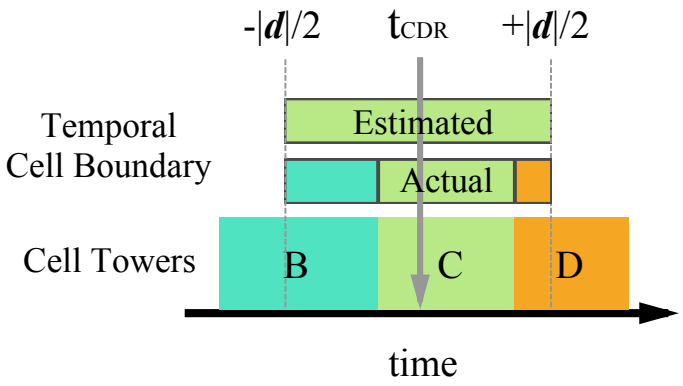


Figure 4.1: An example of a temporal cell boundary in the `stop-by` approach: A period $(t_{\text{CDR}} - |\mathbf{d}|/2, t_{\text{CDR}} + |\mathbf{d}|/2)$ is given as a temporal cell boundary at the cell C attached with a CDR entry at time t_{CDR}. In this temporal cell boundary, the user is assumed to be at the cell C, while actually he moves from the cell B to D: this leads to a spatial error.

for more advanced techniques. It is worth noting that this solution has no spatiotemporal flexibility; its performance only depends on the number of CDR a user generates in the period of study: *i.e.*, the higher is the number of CDR, the lower will be the spatial error in the completed data by the `static` solution. In other words, there is no space (configurable setting or initial parameter) for customizing this solution to obtain better accuracy.

4.2.2.2 Basic Solution: `Stop-by`

Building on in-depth studies proving individuals to stay most of the time in the vicinity of their voice call places [87], Jo *et al.* [35] assume that users can be found at the locations where they generate some digital activities for an hour-long interval centered at the time when each activity is recorded. If the time between consecutive activities is shorter than one-hour, the inter-event interval is equally split between the two locations where the bounding events occur. This solution will be denoted as `stop-by` in the remaining sections.

The drawback of the `stop-by` is that it uses a constant hour-long interval for all calls as well as users in CDR, which may be not always suitable. This solution lacks flexibility in dealing with various human mobility behaviors. As exemplified in Figure 4.1, a single CDR is observed at time t_{CDR} at cell C. Following the `stop-by` solution, the user is considered to be stable at this cell C during the period $\mathbf{d} = (t_{\text{CDR}} - |\mathbf{d}|/2, t_{\text{CDR}} + |\mathbf{d}|/2)$, while in fact the user has moved to two other cell towers during this period. We call the period estimated from an instant CDR entry, a *temporal cell boundary*. In the example of Figure 4.1, this temporal cell boundary is overestimated.

Nevertheless, this solution has more flexibility than the `static` solution does, *i.e.*, the time interval $|\mathbf{d}|$ affects its performance and is configurable. Although a one-hour interval ($|\mathbf{d}| = 60$ minutes) is usually adopted in the literature, we are interested in evaluating the performance of the `stop-by` solution over different intervals, which has never been explored before.

Intuitively, a spatial error occurs if the user moves to other different cells during the temporal cell boundary. To have a quantitative manner of such an error, we define the spatial

error of a temporal cell boundary with a period $\mathbf{d}$ as follows:

$$\text{error}(\mathbf{d}) = \frac{1}{|\mathbf{d}|} \int_{\mathbf{d}} \left\| c^{(\text{CDR})} - c_t^{(\text{real})} \right\|_{\text{geo}} \mathrm{d}t. \tag{4.1}$$

This measure represents the average spatial error between a user's real cell location over time $c_t^{(\text{real})}$ and his estimated cell location $c^{(\text{CDR})}$, during the time period $\mathbf{d}$. The interpretation of the spatial error is straightforward, as follows:

- When $\text{error}(\mathbf{d}) = 0$, it means that the user stays at the cell $c^{(\text{CDR})}$ during the whole temporal cell boundary. Still, the estimation of $\mathbf{d}$ may be conservative, since a larger $|\mathbf{d}|$ could be more adapted in this case.
- When $\text{error}(\mathbf{d}) > 0$, it means that the temporal cell boundary is over-sized: the user in fact, moves to other cells in the corresponding time period. Thus, a smaller $|\mathbf{d}|$ should be used for the cell.

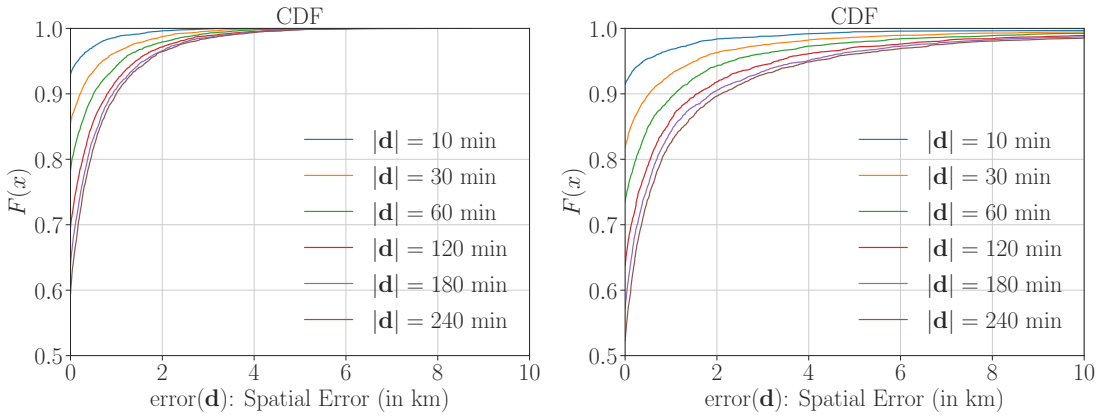
Due to the relevance of this parameter on the model performance, in the following we evaluate the impact of $|\mathbf{d}|$ on the spatial error.

4.2.2.3 Impact of Parametrization on `stop-by` Accuracy

We evaluate the performance of the `stop-by` approach by considering the combination of `CDR` and ground-truth `Flow` datasets (*cf.* Section 3.2 and 3.3.3). CDR are used to generate temporal cell boundaries, while locations in the fine-grained data of flows are adopted as actual locations and are used to compute the spatial errors. We consider a comprehensive range of values $|\mathbf{d}| = \{10, 30, 60, 120, 180, 240\}$ minutes for the `stop-by` parameters.

Figure 4.2(a) and 4.2(b) show the CDF of the spatial errors of temporal cell boundaries on Monday and Sunday, respectively. We observe that $\text{error}(\mathbf{d}) = 0$ for 80% of CDR on Monday (*cf.* 75% on Sunday) when $|\mathbf{d}| = 60$ minutes, and for 60% of CDR on Monday (*cf.* 53% on Sunday) when $|\mathbf{d}| = 240$ minutes. This result is a strong indicator that users tend to remain in cell coverage areas for long intervals around their instant locations recorded by CDR. It is also true that many users are simply static, *i.e.*, only appear at one single location in their Internet flows, and, consequently have an associated radius of gyration $R_g^u = 0$: this behavior accounts for approximately 35% and 40% on Monday and Sunday, respectively. The high percentage of temporal cell boundaries with $\text{error}(\mathbf{d}) = 0$ in Figure 4.2 may be due to these static users, since they will not entail any spatial error, under any $|\mathbf{d}|$. To account for this aspect, we exclude the static users in the following, and only consider the *mobile* users, *i.e.*, ones having $R_g^u > 0$.

An interesting consideration is that the spatial error incurred by the `stop-by` approach is not uniform across cells. Intuitively, a cell tower covering a larger area is expected to determine longer user dwelling times and hence better estimates with `stop-by`. We thus compute for each cell its coverage as the *cell radius*: specifically, we assume a homogeneous propagation environment and an isotropic radiation of power in all directions at each cell tower, and roughly estimate each cell radius as that of the smallest circle encompassing the Voronoi polygon of the cell tower. We remark that this approach yields overlapping coverage at temporal cell boundaries, which reflects what happens in real-world deployments. In the target area under study, shown in Figure 3.1, 70% of the cells have radii within 3 km, and the median radius is approximately 1 km.



(a) CDF of Spatial Error: Monday

(b) CDF of Spatial Error: Sunday

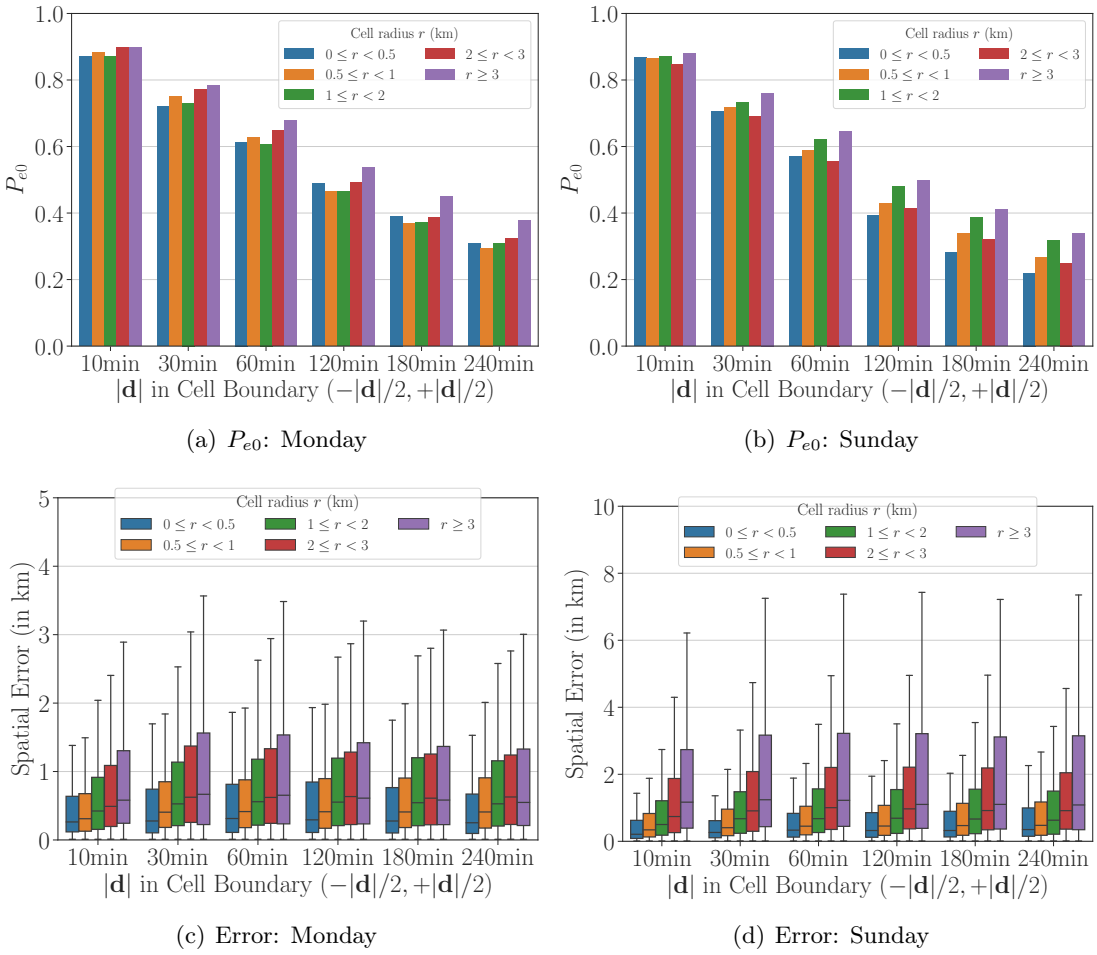
Figure 4.2: CDF of the spatial error of temporal cell boundaries of CDR generated by the `stop-by` solution over two groups of the users in the `Flow` dataset on (a) Monday and (b) Sunday.

We can now evaluate the probability of having a temporal cell boundary with a null spatial error, as $P_{e0} = \mathbf{Pr}\{\text{error}(\mathbf{d}) = 0\}$. Figure 4.3(a) and 4.3(b) present the probabilities P_{e0} grouped by the cell radius, when applying varying sizes of temporal cell boundary on the days of study. We notice the following.

- The probability P_{e0} decreases with the increasing period marked by $|\mathbf{d}|$, indicating that using a large period on the temporal cell boundary increases the chances of generating some spatial errors. For instance, for $|\mathbf{d}| = 30$ minutes, the probability of having a null spatial error is around 0.7 depending on the date and on the cell radius. When a larger $|\mathbf{d}|$ is used, the probability significantly increases (*e.g.*, for $|\mathbf{d}| = 60$ minutes, the probability P_{e0} reduces to around 0.6).
- The probability P_{e0} correlates positively with the cell radius r. This trend is seen on both Monday and Sunday (except some cases), indicating that the cell size has an impact on the time interval during which the user stays within the cell coverage. Intuitively, handovers are frequent for users moving among small cells and less so for users traveling across large cells.

The results support the idea that there is a strong correlation between the temporal cell boundary and the cell coverage. Nevertheless, since CDR are usually sparse in time, using a small temporal cell boundary could only cover an insignificant amount of cell visiting time, while using a big temporal cell boundary increases the risk of having a non-null spatial error. To investigate this trade-off, we plot the variation of the statistical distribution of the spatial errors after excluding the null errors (*i.e.*, keeping only cases with non-null error($\mathbf{d}$)) in Figure 4.3(c) and 4.3(d). We observe the following.

- The spatial error varies widely: it goes from less than 1 km to very huge values (up to 3.6 km on Monday and to 7.5 km on Sunday). Hence, for some users, the `stop-by` solution is unsuitable for reconstructing visiting patterns due to the presence of such high spatial errors.



(a) P_{e0}: Monday

(b) P_{e0}: Sunday

(c) Error: Monday

(d) Error: Sunday

Figure 4.3: Spatial errors of temporal cell boundaries of CDR generated by the `stop-by` solution over users with their $R_g > 0$: (a)(b) the probability (P_{e0}) of having a non-error temporal cell boundary $(-|\mathbf{d}|, |\mathbf{d}|)$, where $|\mathbf{d}| \in \{10, 30, 60, 120, 180, 240\}$ minutes, under several groups of cell radius on (a) Monday and (b) Sunday; (c)(d) Box plot of non-zero spatial errors, grouped by the cell radius and the time period of temporal cell boundary on (c) Monday and (d) Sunday. Each box denotes the median and $25^{th} - 75^{th}$ percentiles and the whiskers denote $5^{th} - 95^{th}$ percentiles.

- The spatial error grows with the cell radius: when the cell size increases, the variation of the error becomes wider, while the mean value also increases. This is reasonable because the higher the cell radius is, the farther the cell is from its cell neighbors. Hence, when a spatial error occurs, it means that the user is actually in a far cell that has a larger distance to $c^{(\text{CDR})}$.

4.2.2.4 Key Insights

Overall, we assert that temporal cell boundary estimates user's locations with a high accuracy when $|\mathbf{d}|$ is small. This validates the previous finding that users usually stay in proximity of call locations for certain time. The accuracy reduces significantly, giving rise to spatial errors,

when increasing $|\mathbf{d}|$. Hence, the trade-off between the completion and the accuracy should be carefully considered when completing CDR using temporal cell boundaries. Using a constant $|\mathbf{d}|$ over all users as in the `stop-by` solution is unlikely to be an appropriate approach.

Building on these considerations, we propose enhancements to the `stop-by` and `static` solutions in the remainder of the paper. Our strategies introduced in the following leverage common trends in human mobility, in terms of (1) attachment to a specific location during night periods, and (2) a tendency to stay for some time in the vicinity of locations where digital activities take place. In particular, we tell apart the instant CDR completion task at nighttime and daytime: Section 4.2.3 presents nighttime completion strategies inferring the home location of users; Section 4.2.4 introduces our adaptive temporal cell boundary strategies leveraging human mobility regularity during daytime.

4.2.3 Identifying Temporal Home Boundaries

The main goal of our strategies for the instant CDR completion during nighttime is to infer temporal boundaries where users are located, with a high probability, at their home locations. We refer to this problem as the identification of the user's *temporal home boundary*. Gaps in CDR occurring within the home boundary of each user are then filled with the identified home location. The rationale for this approach stems from our previous observations that CDR allow identifying the home location of individuals with high accuracy (*cf.* Section 3.5.4).

4.2.3.1 Proposed Solutions

We extend the `stop-by` solution (*cf.* Sec. 4.2.2.2) in the following ways. Note that all techniques below assume that the home location is the user's most active location during some night time interval $\mathbf{h}$, and that CDR not in $\mathbf{h}$ are completed via legacy `stop-by`.

- The `stop-by-home` strategy adds fixed temporal home boundaries to the `stop-by` technique. If a user's location is unknown during $\mathbf{h} = (10pm, 7am)$ due to the absence of CDR in that period, the user will be considered at his home location during $\mathbf{h}$.
- The `stop-by-flexhome` strategy refines the previous approach by exploiting the diversity in the habits of individuals. The fixed night time temporal home boundaries are relaxed and become flexible, which allows adapting them on a per-user basis. Specifically, instead of considering $\mathbf{h} = (10pm, 7am)$ as the fixed home boundaries for all users, we compute for each user u the most probable interval of time $\mathbf{h}^{(u)}_{\text{flex}} \subseteq \mathbf{h}$ during which the user is at his home location. Then, as for `stop-by-home`, the user will be considered at his home location to fill gaps in his CDR data during $\mathbf{h}^{(u)}_{\text{flex}}$.
- The `stop-by-spothome` strategy augments the previous technique by accounting for positioning errors that can derive (1) from users who are far from home during some nights, or (2) from ping-pong effects in the association to base stations when the user is within their overlapping coverage region. In this approach, if a user's location during $\mathbf{h}^{(u)}_{\text{flex}}$ is not identified, and if he is last seen at no more than 1 km from his home location, he is considered to be at his home location.

We compare the above strategies with the `static` and the legacy `stop-by` solution introduced in Section 4.2.2, assuming $|\mathbf{d}| = 60$ min. Our evaluation considers dual perspectives.

The first is *accuracy*, *i.e.*, the spatial error between mobility metrics computed from ground-truth GPS data and from CDR completed with the different techniques above. The second is *completion*, *i.e.*, the percent of the time during which the position of a user is determined. Note that the `static` solution (*cf.* Section 4.2.2) provides user locations at all times, but this is not true for `stop-by` or the derived techniques above. In this case, the CDR is completed only for a portion of the total period of study, and the users' whereabouts remain unknown in the remaining time.

4.2.3.2 Accuracy and Completion Results

We first compute the geographical distance between the positions in the GPS records in `MACACO` and `Geolife` and those in their equivalent CDR-like coarse-grained datasets. These strategies are not designed to provide positioning information at all times expect the `static` solution, hence distances are only measured for GPS samples whose timestamps fall in the time periods for which completed data is available.

Figure 4.4(a) and Figure 4.4(b) summarize the results of our comparative evaluation of accuracy, and allow drawing the following main conclusions:

- The `static` approach provides the worst accuracy in both datasets.
- The `stop-by-flexhome` technique largely improves the data precision, with an error that is lower than 100 meters in $90-92\%$ of cases for the `MACACO` users and with a median error around 250 meters for the `Geolife` users.
- The `stop-by-spothome` technique provides the best performance for both datasets. For instance, about 95% of samples lie within 100 meters of the ground-truth locations in the `MACACO` dataset, while the median error is 250 meters (the lowest result) in the `Geolife` dataset.

These results confirm that a model where the user remains static for a limited temporal interval around each measurement timestamp is fairly reliable when it comes to accuracy of the completed data. They also support previous observations on the quite static behavior of mobile network subscribers [87]. More importantly, the information of home locations can be successfully included in such models, by accounting for the specificity of each user's habits overnight.

The `stop-by` and derived solutions do not provide full completion by design. Figure 4.4(c) and Figure 4.4(d) show the CDF of the hours per day during which a user cannot be localized by such solutions, for individuals in the `MACACO` and `Geolife` CDR-like datasets, respectively. The completion performance is in fact very heterogeneous across users, for all solutions: it can range from one hour per day for some individuals up to 23 hours per day for other subscribers. By comparing the plots, we assert that the more irregular sampling of the `Geolife` dataset translates into larger time gaps and smaller completion. Interestingly, the `stop-by` approach yields the worst result for both datasets, with unknown user positions in 12 and 19 hours per day in the median cases. Our proposed refinements to the `stop-by` solution increase the completion by inferring missing user positions overnight, when the CDR sampling is reduced. The improvement is significant, with a median gain over the basic `stop-by` solution of $4-5$ hours for `MACACO` dataset and $3-7$ hours for the `Geolife` dataset.

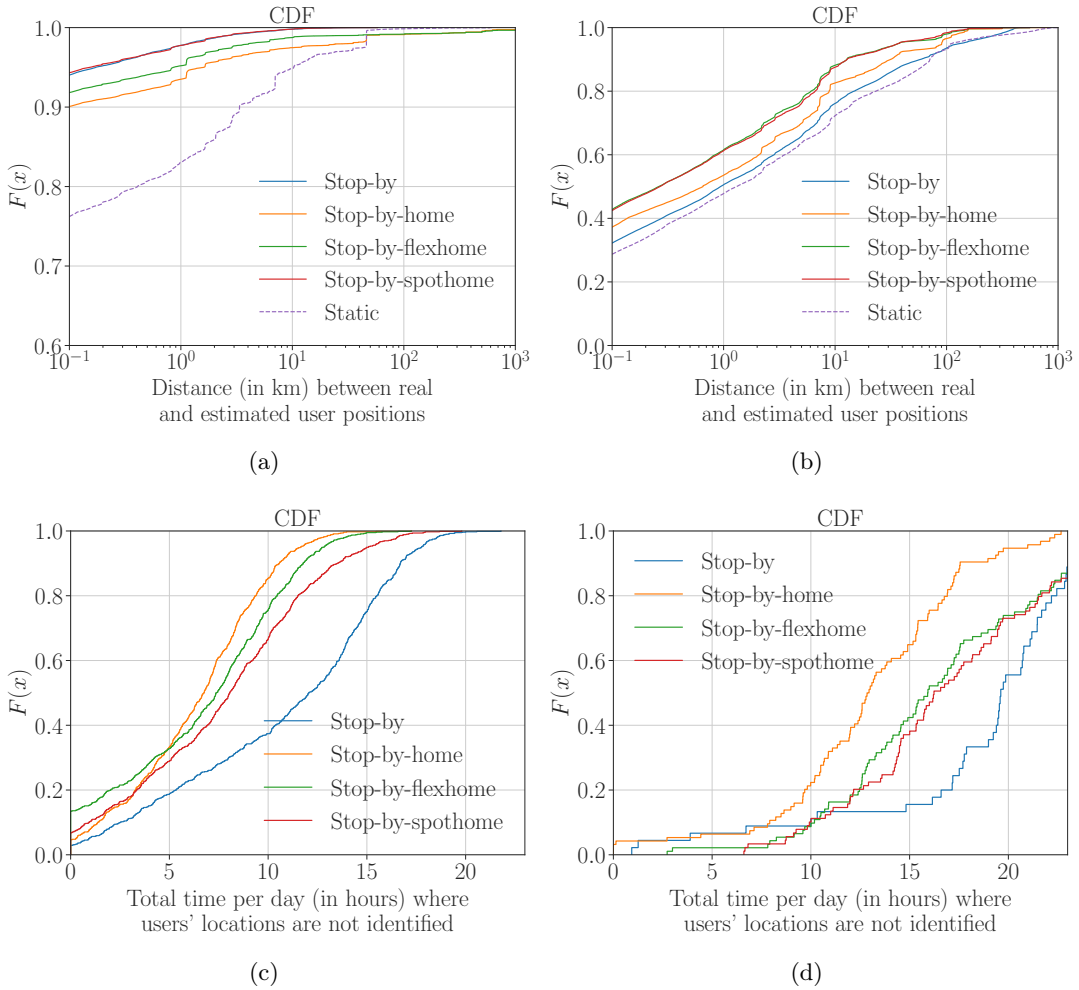


Figure 4.4: CDF of the spatial error (in km) between samples from the GPS and completed data over the (a) `MACACO` and (b) `Geolife` data. CDF of the temporal coverage of completed data over the (c) `MACACO` and (d) `Geolife` data.

Overall, the combination of the results in Figure 4.4 indicates the `stop-by-spothome` solution as that achieving the best combination of high accuracy and fair completion, among the different completion techniques considered.

4.2.4 Identifying Temporal Cell Boundaries

We now consider the possibility of completing CDR during daytime. Our strategy is based again on inferring temporal boundaries of users. However, unlike what has been done with nighttime periods in Sec. 4.2.3, here we leverage the communication context of human mobility habits and extend the time span of the position associated with each communication activity to so-called *temporal cell boundaries*.

4.2.4.1 Factors Impacting Temporal Cell Boundaries

Hereafter, we aim to answer the following question: *how to choose a proper and adaptive period for a temporal cell boundary instead of a static fixed-to-all period?* To answer the question, we need to understand the correlation between the routine behavior of users in terms of mobile communications and their movement patterns. For this, we first study how human behavior factors that can be extracted from CDR may affect daytime temporal cell boundaries. We categorize factors in three classes, *i.e.*, event-related, long-term behavior, and location-related, as detailed next. Then, we leverage them to design novel approaches to estimate temporal cell boundaries.

Event-related Factors We include in this class the meta-data contained in records of common CDR datasets. They include the activity `time`, `type` (*i.e.*, voice call or text message), and `duration`[1]. Intuitively, these factors have direct effects on temporal cell boundaries. For instance, in terms of `time`, a user may stay within a fixed cell during his whole working period. In terms of `type` and `duration`, a long phone call may imply that the user is static, while a single text message may indicate that the user is on the move. Besides, these factors are commonly found in and easily extracted from any common CDR entries.

Long-term Behavior Factors This class includes factors describing users' activities over extended time intervals. They are the radius of gyration (`URg`), the number of unique visited locations (`ULoc`), and the number of active days during which at least one event is recorded (`UDAY`). These factors characterize a user by giving indications of *(i)* his long-term mobility and *(ii)* his habit on generating calls and text messages, which may be indirectly related to his temporal cell boundaries. For each user, these factors are computed from our CDR dataset (*cf.* Chapter 3) by aggregating data during the whole 3-month period of study.

Location-related Factors Factors in this class relate to positioning information. The first factor is the cell radius (`CR`), which we already proved to be affecting the reliability of CDR completion schemes in Section 4.2.2. The other location-related factors take account for the relevance that different places have for each user's activities. The intuition is that individuals spend long time periods at their important places. Specifically, we explore it by applying the algorithm presented by Isaacman *et al.* [110], which determines prominent locations where the user usually spends a large amount of time or visits frequently.

The algorithm applies Hartigan's clustering [129] on visited cell locations of users in CDR and use logistic regression to estimate a location's importance to the user from factors extracted from the cluster that the location belongs to. To start with, the cluster approach chooses the cell tower from the first CDR and makes it the first cluster. Then, it recursively checks all cell towers in the remaining CDR. If a cell tower is within the distance threshold (we use 1 km) to the centroid of a certain cluster, the cell tower is added to the cluster, and the centroid of the cluster is moved to the weighted average of the locations of all the cell towers in the cluster. The weights assigned to locations are the fractions of days in which they are visited over the whole observing period. The clustering process finishes once all cell towers are assigned to clusters.

[1] We set the duration text messages to 0 seconds.

Once clusters are defined, the importance of each cluster is identified according to the following observable factors: *(i)* the number of days on which any cell tower in the cluster was contacted (`CDay`); *(ii)* the number of days that elapse between the first and the last contact with any location in the cluster (`CDur`); *(iii)* the sum of the number of days cell towers in the cluster were contacted (`CTDay`); *(iv)* the number of cell towers inside the cluster (`CTower`); *(v)* the distance from the registered location of the activity to the centroid of the cluster (`CDist`).

These factors derived from a cluster correlate with the time that the user spends in the cluster's locations, as shown by Isaacman *et al.* via their logistic regression model [110]. It is worth noting that we cannot reproduce the exact model in [110], since the used ground-truth is not publicly available. However, we can still use the same factors for our objective, *i.e.*, identifying temporal cell boundaries.

4.2.4.2 Supervised Temporal Cell Boundary Estimation

So far, we have introduced human behavior factors that might be directly or indirectly related to temporal cell boundaries. In order to use them for our purpose, we need a reliable model linking them to actual temporal cell boundaries. In the following we introduce two approaches to do so, both based on supervised machine learning.

Symmetric and Asymmetric Temporal Cell Boundaries We define two kinds of temporal cell boundaries: symmetric and asymmetric. Given a CDR entry at time t, determining its temporal cell boundary means to expand the instantaneous time t to a time interval $\mathbf{d}$, during which the user is assumed to remain within coverage of the same cell. For a symmetric temporal cell boundary, this period is generated from a CDR-based parameter $d^{\pm}$ as $\mathbf{d} = (t - d^{\pm}, t + d^{\pm})$, *i.e.*, it is symmetric with respect to the CDR time t. Instead, the period of an asymmetric temporal cell boundary is generated from two independent parameters d^{+} and d^{-} as $\mathbf{d} = (t - d^{-}, t + d^{+})$.

We design `sym-adaptive` and `asym-adaptive` approaches, both of which receive a CDR entry as input and return an estimate of its associated temporal cell boundary. More precisely, the factors aforementioned are extracted for each user and CDR record, and converted to an input vector $\mathbf{x}$, under the following rules: *(i)* the categorical factor `type` is converted to two binary features by one-hot encoding[2]; *(ii)* the `time` is converted to the distances (in seconds) separating it from $10am$ and from $6pm$[3]; *(iii)* the other factors are used as plain scalar values. Given a CDR entry and its input vector $\mathbf{x}$, we have the following approaches:

- The `sym-adaptive` approach contains one model that accepts the input vector and predicts the parameter $d^{\pm}$ as a symmetric estimation of the corresponding temporal cell boundary, *i.e.*, $d^{\pm} = F_{\text{sym}}(\mathbf{x})$.
- The `asym-adaptive` approach contains two models that separately predict the parameters d^{+} and d^{-} as a joint asymmetric estimation of the corresponding temporal cell boundary, *i.e.*, $d^{+} = F^{+}_{\text{asym}}(\mathbf{x})$ and $d^{-} = F^{-}_{\text{asym}}(\mathbf{x})$.

We use supervised machine learning techniques to build the models. It is worth noting that the user identifier is not in the input vector $\mathbf{x}$ because we do not want to train models that

[2]Used to deal with the unbalanced occurrence of the types.

[3]Daytime interval covered by the used combination (*cf.* Section 3.3.3).

bound themselves to any particular user. This gives our models better flexibility and ensures higher potential for applying the trained model into other mobile phone datasets where the same factors can be derived.

Estimating Temporal Cell Boundaries Via Supervised Learning We detail our methodology and results, by *(i)* formalizing the optimization problems that capture our goal, *(ii)* discussing how they can be addressed via supervised machine learning, and *(iii)* presenting a complete experimental evaluation.

(i) Optimization problems. All the models are generalized from a training set $\mathcal{X}$ consisting of CDR entries (as input vectors) and their real temporal cell boundaries (which are originally asymmetric), *i.e.*, $\mathcal{X} = \{(\mathbf{x}_i, d_i^+, d_i^-)\}$.

To build the `asym-adaptive` approach, the objective is to find two separate approximations, as $F^+_{\text{asym}}(\mathbf{x})$ and $F^-_{\text{asym}}(\mathbf{x})$, to functions $F^+(\mathbf{x})$ and $F^-(\mathbf{x})$ that respectively minimize the expected values of two losses $L(d^+, F^+(\mathbf{x}))$ and $L(d^-, F^-(\mathbf{x}))$, *i.e.*,

$$F^+_{\text{asym}}(\mathbf{x}) = \arg\min_{F^+} \mathbb{E}_{d^+,\mathbf{x}}[L(d^+, F^+(\mathbf{x}))], \tag{4.2}$$

$$F^-_{\text{asym}}(\mathbf{x}) = \arg\min_{F^-} \mathbb{E}_{d^-,\mathbf{x}}[L(d^-, F^-(\mathbf{x}))], \tag{4.3}$$

where L is the squared error loss function, *i.e.*, $L(x, y) = \frac{1}{2}(x - y)^2$.

To build the `sym-adaptive` approach, a modified training set $\mathcal{X}^{\pm} = \{(\mathbf{x}_i, d_i^{\pm})\}$ is firstly generated from the original $\mathcal{X}$ by applying $d_i^{\pm} = \min\{d_i^+, d_i^-\}$ on each real asymmetric temporal cell boundary. Then, as our objective, we need to find an approximation $F_{\text{sym}}(\mathbf{x})$ to a function $F^{\pm}(\mathbf{x})$ that minimizes the expected value of the loss $L(d^{\pm}, F^{\pm}(\mathbf{x}))$, *i.e.*,

$$F_{\text{sym}}(\mathbf{x}) = \arg\min_{F^{\pm}} \mathbb{E}_{d^{\pm},\mathbf{x}}[L(d^{\pm}, F^{\pm}(\mathbf{x}))]. \tag{4.4}$$

(ii) Learning technique. In order to compute the approximations, we utilize a typical supervised machine learning technique, *i.e.*, Gradient Boosted Regression Trees (GBRT) [82, 130]. Although several supervised learning techniques can be adopted, we pick the GBRT technique because *(i)* it is a well-understood approach with thoroughly-tested implementations, *(ii)* it has advantages over alternative techniques, in terms of predicative power, training speed, and flexibility to accommodate heterogeneous input (which is our case) [81], and *(iii)* it returns quantitative measures about the contribution of each factor to the overall approximation [82].

In the GBRT technique, an approximation function is the weighted sum of an ensemble of regression trees. Each tree divides the input space (*i.e.*, the vector $\mathbf{x}$ of factors) into disjoint regions and predicts a constant value in each region. The GBRT technique combines the predictive power of all regression trees having a weak predicting performance by making a joint predictor: it is proved that the performance of such a joint predictor is better than that of each single regression tree [130]. The ensemble is initialized with a single-leaf tree (*i.e.*, a constant value). During each iteration, a new regression tree is added to the ensemble by minimizing the loss function via gradient descent. An algorithm of the GBRT technique for building the approximation of the function F_{sym} in the `sym-adaptive` approach is given in Alg. 1. In the algorithm, the function `FitRegrTree` is used to build a *regression tree* based on the input and the gradients of the function in the last iteration, of which we refer the reader to [130, Chapter 9.2.2] for the detail. Two important tuning parameters are in the algorithm, *i.e.*, the

Algorithm 1: GBRT algorithm [130, Algorithm 10.3] for finding the approximation function $F_{\text{sym}}(\mathbf{x})$ in the `sym-adaptive` approach

function GBRT($\mathcal{X}^{\pm}, M, \nu$);
Input : $\mathcal{X}^{\pm}$ - training set, M - number of iterations, ν - learning rate
Output: $F_{\text{sym}}(\mathbf{x})$ - symmetric temporal cell boundary estimation function
$F_0^{\pm}(\mathbf{x}) = \arg\min_{\gamma} \sum_i L(d_i^{\pm}, \gamma)$;
for $m \leftarrow 1$ **to** M **do**
  for $(d_i^{\pm}, \mathbf{x}_i) \in \mathcal{X}^{\pm}$ **do**
    $g_i = -\frac{\partial L(d_i^{\pm}, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)}$;
  end
  $\mathcal{G} = \{(g_i, \mathbf{x}_i)\}$;
  $h_m(\mathbf{x}) =$ `FitRegrTree`$(\mathcal{G})$;
  $\rho_m = \arg\min_{\rho} \sum_i L(d_i^{\pm}, F_{m-1}(\mathbf{x}_i) + \nu \cdot \rho \cdot h_m(\mathbf{x}_i))$;
  $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \rho_m \cdot h_m(\mathbf{x})$;
end
return $F_M(\mathbf{x})$;

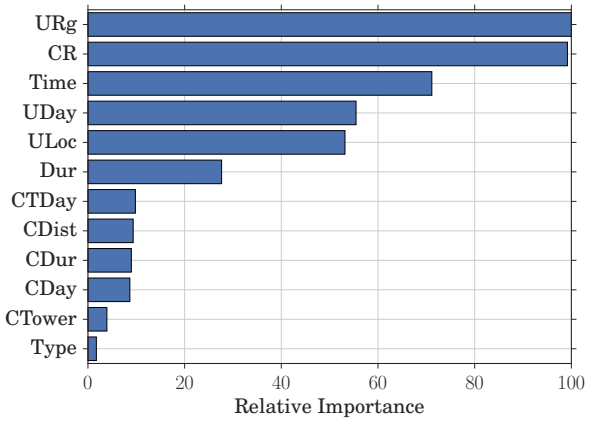


Figure 4.5: Relative Importance of features in determining accurate temporal cell boundaries.

number of iterations M (*i.e.*, the number of regression trees to be added to the ensemble) and the learning rate ν (*i.e.*, the level of contribution expected by a new regression tree), which we determine via cross validation and discuss later. In the `asym-adaptive` approach, the same algorithm is used except that the training set $\mathcal{X}^{\pm}$ is replaced by $\mathcal{X}$.

(iii) Experiments. The first step is to build the training sets. For that, we randomly select 50% of the users from the two available days (*i.e.*, a Monday and a Sunday) in the `Flow` dataset (*cf.* Section 3.2). In particular, from the CDR and `Flow` datasets, we first extract for each CDR entry of these selected users its corresponding input vector $\mathbf{x}$ as well as the parameters d^+, d^- of its real temporal cell boundary. We then build the two training sets $\mathcal{X}$ and $\mathcal{X}^{\pm}$.

The second step is to build the approximation functions (*i.e.*, F^+_{asym}, F^-_{asym}, and F_{sym}) from the training sets. For that, we have to first tune the M and ν parameters of Alg. 1 of each approximation function. To this end, we use a 3-fold cross-validation to select the number

of iterations M from the candidate set $\{100, 500, 1000, 2000, \cdots, 10000\}$ and the learning rate ν from the candidate set $\{0.1, 0.2, \cdots, 1\}$. In particular, we divide the training set $\mathcal{X}$ (or $\mathcal{X}^{\pm}$) into equal-sized three subsets. For each combination of M and ν, we train the model corresponding to each approximation function based on one subset and validate it on the other two subsets. We repeat this operation three times with each of the subsets used as training data. We select as our actual parameters the M and ν values that achieve the lowest loss in the cross-validation. Finally, we use the training sets $\mathcal{X}$ and $\mathcal{X}^{\pm}$ and the tuning parameters that we select to build the functions F^{+}_{asym}, F^{-}_{asym}, and F_{sym} corresponding to the `asym-adaptive` and `sym-adaptive` approaches.

Figure 4.5 shows the relative importance of factors with respect to the estimation of a temporal cell boundary in the training procedure of the GBRT technique. For each factor, its importance is computed as a relative value of the sum of its corresponding importance in all the three approximations. The importance indicates the degree of a feature contributing to the construction of the regression trees. This figure allows us drawing the following main conclusions, valid for both approaches.

- The three most important factors are the timestamp of the activity, the cell radius, and the radius of gyration. This indicates that the time spent by a user within coverage of a same cell mainly depends on the cell size, the precise time when the activity occurred, and the user's long-term mobility.

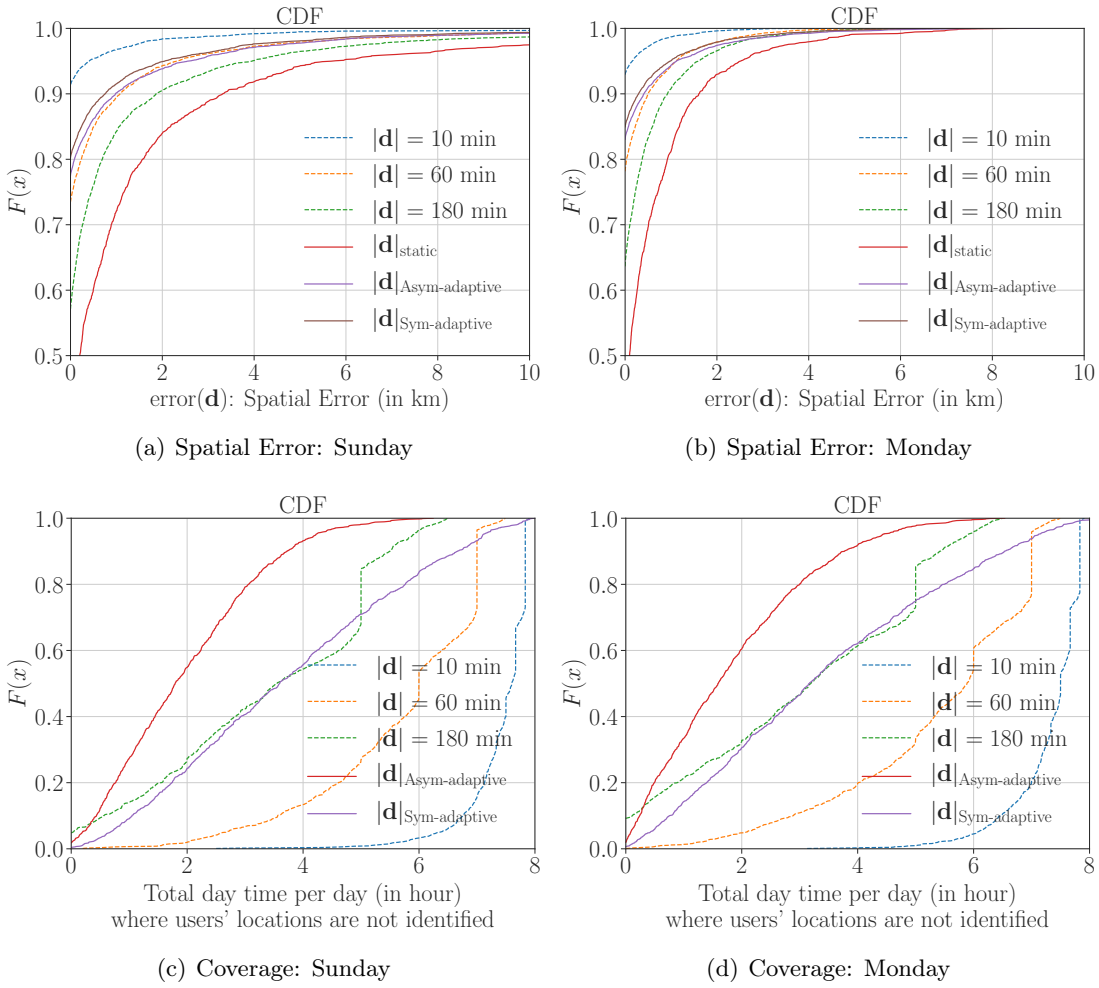
- Surprisingly, the activity's `type` is the least relevant factor, indicating that knowing whether a user generates a call or a message is useless in determining a temporal cell boundary.

4.2.4.3 Accuracy and Completion Results

We compare our two trained approaches with the `stop-by` and `static` approaches using the CDR from the remaining 50% of the randomly-selected users. For the two `sym-adaptive` and `asym-adaptive` approaches, we build two testing sets from the CDR entries of the remaining users. We then let them generate adaptive symmetric and asymmetric temporal cell boundaries using the input vectors in the testing sets. Besides, we let the `stop-by` approach generate temporal cell boundaries using $|\mathbf{d}| = \{10, 60, 180\}$ minutes. As in Section 4.2.3, we make a comparative study by evaluating the solutions regarding *accuracy* and *completion*, where the accuracy is measured by evaluating the *spatial error* in Equation (4.1) (*cf.* Section 4.2.2). Recall that a good data completion approach should cover the observing period as much and precise as possible, *i.e.*, satisfying high accuracy and completion simultaneously.

Figure 4.6(a) and 4.6(b) display the distribution of the spatial errors over all temporal cell boundaries. Our results confirm that the spatial error increases as t_d becomes larger when using the `stop-by` approach. More importantly, the two adaptive approaches perform slightly better than the `stop-by` approach does with its most common setting ($|\mathbf{d}| = 60$ minutes) in terms of the spatial error. As expected, the `static` solution has the worst performance, similarly to what observed in the case of home boundaries using the `MACACO` and `Geolife` datasets.

Figure 4.6(c) and 4.6(d) plot the distribution of the completion per users over all approaches except `static` (of which the completed data always covers the whole period). The x-axis of the figures has 8 hours because the `Flow` dataset only covers an eight-hour day time, *i.e.*, $(10am, 6pm)$. We remark that both our adaptive approaches score a significant performance



(a) Spatial Error: Sunday

(b) Spatial Error: Monday

(c) Coverage: Sunday

(d) Coverage: Monday

Figure 4.6: CDF of the spatial errors of temporal cell boundaries computed on (a) Sunday and (b) Monday; CDF of the completion of completed data on (c) Sunday and (d) Monday, across the `stop-by`, `static`, `sym-adaptive`, and `asym-adaptive` approaches.

improvement in terms of completion: the amount of time during which users' locations stay unidentified is substantially reduced with respect to the legacy `stop-by` approach. On average, only approximately 2 hours (25% of the period of study) of the user's day time remains unidentified after applying the `asym-adaptive` approach, while 3 hours remains unidentified after using the `sym-adaptive` and `stop-by` ($|\mathbf{d}| = 180$ minutes) approaches. The `stop-by` approach with its most common setting ($|\mathbf{d}| = 60$ minutes) has the same degree of accuracy as the adaptive approaches but has a far less degree of completion (*i.e.*, a median of 6 unidentified hours).

Overall, these results highlight a clear advantage provided by adaptive approaches for CDR completion based on supervised learning. Consequently, the adaptive approaches achieve a slightly better performance in terms of accuracy but have a far better performance in terms of completion. The `asym-adaptive` approach has an obvious advantage than the competitors: it completes 75% of the day hours with a fairly good accuracy.

4.3 Completing Slotted CDR-based Trajectories

In this section, we address the slotted CDR completion task. First, we discuss the new challenge and the rationale for this task in Section 4.3.1. Second, we introduce the existing practices for completing slotted CDR-based trajectories in Section 4.3.2. Finally, we present the design and evaluation of our *Context-enhanced Slotted CDR Completion* (CSCC) approach in Section 4.3.3 and Section 4.3.4, respectively.

4.3.1 Rationale

Recall that our target is now the slotted CDR-based trajectories. Before we proceed our discussion, we hereby provide a more self-contained definition of those trajectories. For a slotted CDR-based trajectory, *i.e.*, $\{\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_N\}$ as in Definition 4.2, we use a set $\mathcal{T} = \{1, \cdots, N\}$ to represent the N equivalent time slots and a set $L^u_{\mathcal{T}}$ to represent the trajectory itself as a time series with respect to the time slots $\mathcal{T}$ and the user u as follows: $L^u_{\mathcal{T}} = \{\mathbf{l}^u_i | i \in \mathcal{T}\}$ where $\mathbf{l}^u_i$ represents the representative location of the user u in the i-th time slots, instead of the previous symbol $\mathbf{l}_i$ in Definition 4.2. Due to the loss of locations, slotted CDR-based trajectories are generally incomplete. For the trajectory $L^u_{\mathcal{T}}$, we use an *observation set* as $\Omega_u = \{t_1, \cdots, t_k\} \subseteq \mathcal{T}$ to represent the time slots in which the representative locations are observed, and thus have the observed part of the trajectory, represented as $L^u_{\Omega_u} = \{\mathbf{l}^u_i | i \in \Omega_u\}$. The idea of the representative location worths some explanation. Sometimes, more than one CDR or locations can be found in a time slot. In this case, we use the most frequent location as the representative location of the time slot. Besides, we employ the 2-dimensional Euclidean space to mark the locations; each location $\mathbf{l}^u_i$ is a 2-dimensional Cartesian coordinate.

Given the definition above, we can formalize the slotted CDR completion task, *i.e.*, to fill all the time slots in a slotted CDR-based trajectory with locations as precise as possible, into the following minimization problem:

$$\min \sum_{i \in \Omega^C_u} \frac{|\mathbf{l}^u_i - \hat{\mathbf{l}}^u_i|}{|\Omega^C_u|}, \quad s.t.\ L^u_{\Omega_u} \tag{4.5}$$

where $\Omega^C_u = \mathcal{T} - \Omega_u$ and $\hat{\mathbf{l}}^u_i$ is the estimation of the missing location in the i-th time slot.

Compared with the instant CDR completion task, the slotted CDR completion is a more changeable task: the former's objective is to expand instant locations to disconnected temporal cell boundaries, while the latter's is now to fill all the time gaps and recover a trajectory with 100% completion, which means to infer more mobility information from the same amount of data. Besides, our completeness analysis in Section 3.4 has shown that the incompleteness of a slotted CDR-based trajectory is typically in a high level, which means that we may have $|\Omega^C_u| \gg |\Omega_u|$.

Yet, the good news is that, it is still feasible to design solutions for this challenging task due to the fact that human mobility is highly redundant [16, 17, 118]. Particularly, the following characteristics of human mobility can significantly help us to infer those missing locations, and especially, the first two have already contributed to our proposals for the instant CDR completion task.

- **Inter-call stability**. A user tends to spend a lot of time at places where he makes calls and travels fast between consecutive communications [87]. Thus a user tends to be stable

at those locations captured by CDR, which supports the representativeness of the slotted CDR-based trajectory.

- **Nighttime sability**. A user is usually stable during nighttime because of the human nature. Compared with daytime hours, less CDR are expected to be captured during nighttime hours, but the nighttime stability can help identifying the user's temporal home boundary with a high degree of accuracy, as already shown in Section 4.2.3.

- **Redundancy**. Human mobility is highly redundant in various aspects [16, 17, 118]. In terms of the displacement, one returns frequently to a few highly visited locations [16]; in a CDR dataset, the majority of cell towers are barely marked, while a very small amount of them is frequently visited by the same subscriber [118]. Despite of the loss, these active locations are captured by CDR with a high probability as shown in Section 3.5.3. In terms of the temporal repetitiveness, the order of human visiting patterns contributes to the high degree of human mobility predictability [17]. Besides, Oliveira *et al.* [118] show that the trajectory of hourly cell tower locations has a strong spatiotemporal correlation. These findings indicate that the missing location may be recovered with a high potential degree from knowing the temporal order of visiting.

In summary, the characteristics above support the feasibility of addressing the slotted CDR completion task and serve as the foundation in the design of our approach. In the following, we first review the existing approaches for this task and then propose our novel approach.

4.3.2 Current Approaches to Slotted CDR Completion

We introduce three existing approaches for the slotted CDR completion task. The `static` approach serves as our baseline, as in the instant CDR completion analysis in Section 4.2.2. The `fit` and `matrix` approaches are practical proposals in the literature.

4.3.2.1 Basic Solution: `Static`

Recall the `static` approach introduced in Section 4.2.2. It assumes that a users stays in his last seen location until a new location is observed. Regarding the slotted CDR completion task, it means to fill all the empty time slots with the newest preceding locations. In addition, we let this approach to fill the missing locations at the very beginning of the trajectory with the first observed location, so as to have a complete trajectory.

4.3.2.2 Practical Solution: `Fit`

This technique is designed by Hoteit *et al.* [24] for estimating missing locations from CDR. They build the best fit of a trajectory by using the linear or cubic interpolation on the existing locations. In particular, to infer missing locations between two consecutive observed CDR, they use the linear interpolation on the two time-stamped geographical coordinates if the user has a daily radius of gyration less than 3 km (named a *sedentary* user in [24]), or the cubic polynomial interpolation if the user has a larger radius of gyration. This technique has an improved flexibility compared with the `static` approach.

4.3.2.3 Practical Solution: `Matrix`

This technique refers to the *matrix factorization* technique, which is a common practice in data completion scenarios. For applying this technique, measured data is organized into a matrix $X \in \mathbb{R}^{p \times q}$ for which only a partial set of entries are observed, *e.g.*, a sensory data matrix, in the shape of sensors and time slots, collected from a wireless sensor network [131], and a movie rating data, in the shape of viewers and movies, in the Netflix competition [132]. The problem of completing such a matrix is to learn an unknown parameter (*i.e.*, the matrix X) based on a relatively small number of its samples/entries. An assumption is made in order for such inference to be meaningful, *i.e.*, X has a *low-rank* approximation. For instance, in the d-rank approximation, we have $X \approx LR^T$ where $L \in \mathbb{R}^{p \times d}$ and $R \in \mathbb{R}^{q \times d}$.

To apply the `matrix` technique for our slotted CDR-based completion task, we convert the trajectory $L^u_{\mathcal{T}}$ to the matrix X^u as follows:

$$X^u = \begin{bmatrix} \mathbf{l}^u_1 & \mathbf{l}^u_2 & \cdots & \mathbf{l}^u_{n_d} \\ \mathbf{l}^u_{n_d+1} & \mathbf{l}^u_{n_d+2} & \cdots & \mathbf{l}^u_{2n_d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{l}^u_{kn_d+1} & \mathbf{l}^u_{kn_d+2} & \cdots & \mathbf{l}^u_{(k+1)n_d} \end{bmatrix} \tag{4.6}$$

In this matrix, the trajectory is divided into $k+1$ equivalent sub-trajectories that have n_d time slots, and each row of the matrix represents a sub-trajectory. In practice, to leverage the daily repetitive pattern of human footprints [16, 118, 133], we set n_d to the number of time slots in one day. Therefore, for the matrix X^u, we have $X^u \in \mathbb{R}^{p \times q}$, where $p = \frac{N}{n_d}$ and $q = 2n_d$.

The matrix X^u is also incomplete as the original trajectory $L^u_{\mathcal{T}}$; we still use the observation set Ω_u to represent the indices of the known values of the matrix X^u. To complete the slotted CDR-based trajectory $L^u_{\mathcal{T}}$ is equivalent to recover the incomplete matrix X^u. For that, we apply the `matrix` technique. We assume that the matrix X^u has a d-rank approximation $X^u \approx LR^T$. The `matrix` technique solves the following low-rank matrix factorization problem to estimate L and R:

$$(\hat{L}, \hat{R}) = \arg\min_{L,R} \sum_{(i,j) \in \Omega^u} |X^u_{ij} - (LR^T)_{ij}|^2 + \lambda(||L||^2_F + ||R||^2_F), \tag{4.7}$$

where $\hat{L} \in \mathbb{R}^{p \times d}$, $\hat{R} \in \mathbb{R}^{q \times d}$, $||\cdot||_F$ is the Frobenius Norm, *i.e.*, $||X||_F = \sqrt{\sum_i \sum_j |X_{ij}|^2}$, and λ is a penalty parameter to avoid overfitting. Then, we obtain an approximation $X^u = \hat{L}\hat{R}^T$ that contains the estimation of all the missing locations.

The problem above can be regarded as a combination of multiple standard linear least squares problems, and can be solved via ALS (alternating least squares) [134]. The basic idea of ALS is cyclically updating one matrix factor at a time while holding the other one constant. In particular, we initialize L and R by two random small matrices. For each iteration, we first treat R as a constant and solve each row of L by a convex least square optimization problem. Then we do reversely to solve R.

In the aforementioned techniques, the `static` and `fit` utilize the inter-call stability, but do not consider the redundancy of human mobility. The `matrix` technique leverages the redundancy, while is not specifically designed for human mobility inference.

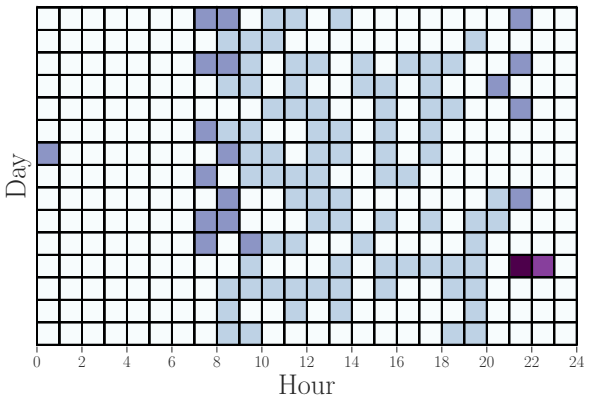


Figure 4.7: Actual incompleteness of a slotted CDR-based trajectory: each block represents the representative location of an one-hour time slot. The white color means *missing*; the others means cell tower locations observed.

4.3.3 Identifying Locations in Time Slots

Due to the limits in the existing proposals, we hereby propose our `CSCC` (Context-enhanced Slotted CDR Completion) approach, in order to fully utilize the human characteristics mentioned above and to complete a slotted CDR-based trajectory with a fairly good accuracy.

For the slotted CDR completion task, the most significant problem is *how to address the data loss appropriately?* Recall that the mobility data loss of CDR is in a high degree (*cf.* Section 3.4). Therefore, a slotted CDR completion approach should be first able to handle such loss. Further, the loss is not uniform. For instance, we portray an actual slotted CDR-based trajectory in the `CDR` dataset in Figure 4.7; the trajectory reveals the two typical features of all slotted CDR-based trajectories. The first is *heterogeneity*: locations during daytime hours are far more than during nighttime. The second is *redundancy*: the daily sub-trajectories in Figure 4.7 are highly repetitive. Both the features can contribute to the CDR completion to deal with the data loss. The heterogeneity provides more locations during daytime when human movement pattern is typically more complicated; the redundancy reduces the uncertainty of mobility information.

Based on the discussion above, we design the hierarchical `CSCC` approach, which receives an incomplete slotted CDR-based trajectory as input and generates the completed version as output via the three following steps.

1. **Nighttime data enhancement**. This step is to fill only nighttime time gaps in the trajectory with the home location identified. It is to deal with the heterogeneity and lighten the data loss during nighttime.

2. **Temporal improved tensor data completion**. This step is to fill the rest of time gaps from the context of observed locations. For that, we adopt the state-of-the-art *tensor factorization* technique, which utilizes the redundancy to recover the missing data.

3. **Cell tower estimation**. This is to provide more accurate geographical coordinates in the location resolution of cell towers. Using the context of all the cell towers observed, we propose a strategy to convert a free estimated coordinate to its most likely cell tower location.

In the following, we introduce each step of the approach in detail.

4.3.3.1 Nighttime Data Enhancement

As the first step of our approach, we adopt the `stop-by-spothome` strategy proposed in Section 4.2.3, which has been shown to be effective in identifying a user's nighttime $(10pm, 7am)$ locations. It can fill some nighttime slots with an accurate estimation, lighten the heterogeneity of the data loss, and ensure the performance of the tensor factorization technique in the next step. In order to have better accuracy, we add one constrain to the original proposal: we apply this step only if the user has a home location that contains $\geq 80\%$ of his nighttime CDR; otherwise, we skip the first step and begin directly with the next step. Particularly, this first step works as follows. Given a group of CDR, we apply `stop-by-spothome` to have the user's temporal home boundary along with the identified home location, construct the user's slotted CDR-based trajectory, and then fill the home location into the time slots that are contained by the identified temporal home boundary.

4.3.3.2 Temporal Improved Data Completion

So far, we have a slotted CDR-based trajectory of which the nighttime locations are partially completed by the first step. In this second step, we infer all the remaining locations. As in the `matrix` technique, we organize the trajectory in a structural form to utilize the redundancy of human mobility, while we use the *tensor* instead of the matrix to better capture the redundancy. For that, we employ the *tensor factorization* technique in this step for the location inference. It is an updated version of the `matrix` technique to recover redundant structural data [116, 135].

In the `matrix` technique, organizing a slotted CDR-based trajectory into a matrix makes it possible to utilize the daily repetitive pattern of the trajectory into its recovery. Similarly, we go further in this step. Since human movement also has a weekly repetitive pattern [16, 118], we divide the observing period into one-week sub-periods that contain one-day sub-trajectories, and then convert the trajectory $L^u_{\mathcal{T}}$ to a three dimensional tensor $\mathcal{X}^u$ as follows:

$$\mathcal{X}^u = [X^u_1, X^u_2, \cdots, X^u_{n_w}], \tag{4.8}$$

where n_w is the number of weeks in the observing period and X^u_i represent the location matrix in the i-th one-week sub-period. This tensor is a combination of the matrices that are converted from each one-week sub-periods and have the same form as in Equation (4.6). Therefore, we have $\mathcal{X} \in \mathbb{R}^{p \times q \times r}$ where $p = n_w$, $q = \frac{N}{n_w n_d}$, and $r = 2n_d$. For the known values in the tensor $\mathcal{X}^u$, we still use the set Ω_u to represent their indices in the tensor.

Then, we construct the optimization problem for the inference of missing values in the tensor $\mathcal{X}^u$. Similarly, for such inference to be meaningful, we assume that the tensor $\mathcal{X}^u \in \mathbb{R}^{p \times q \times r}$ has a CPD (canonical polyadic decomposition) [136] of three d-rank matrices $A \in \mathbb{R}^{p \times d}$, $B \in \mathbb{R}^{q \times d}$, and $C \in \mathbb{R}^{r \times d}$. In the CPD, each value $\mathcal{X}_{ijk}$ in the tensor is approximated as $\mathcal{X}_{ijk} = \sum_{f=1}^{d} A_{if} B_{jf} C_{kf}$; the tensor can be regarded as a combination of d 1-rank tensors. For simplicity, we employ the following concise expression of the CPD used by Kolda *et al.* [136], *i.e.*, $\mathcal{X}^u = [\![A, B, C]\!]$. With the CPD of the tensor $\mathcal{X}^u$, we can have the following tensor factorization minimization problem to estimate the decomposed matrices so as to infer the

missing values:

$$(\hat{A}, \hat{B}, \hat{C}) = \underset{A,B,C}{\arg\min} \sum_{(i,j,k)\in\Omega_u} (\mathcal{X}^u_{ijk} - [\![A, B, C]\!]_{ijk})^2 + \lambda(||A||_F^2 + ||B||_F^2 + ||C||_F^2) \tag{4.9}$$

where λ is a penalty parameter to avoid overfitting. Similarly, by solving this problem, we obtain an approximation of the tensor based on the CPD as $\hat{\mathcal{X}}^u = [\![\hat{A}, \hat{B}, \hat{C}]\!]$ and then have the inference of all the missing locations in the corresponding trajectory $L^u_{\mathcal{T}}$ from this approximation.

The CPD form in Equation (4.9) treats each dimension of the tensor equally. Yet, regarding the redundancy of human mobility, the daily repetitive pattern is usually stronger than the weekly one; also, the consecutive time slots may have the same cell tower location due to the inter-call stability. To leverage these features and have a more accurate inference than the `matrix` technique, we induct two more constrains into the optimization problem Equation (4.9) using the Toeplitz matrix. The first constrain is to emphasize the daily repetitive pattern. We construct a Toeplitz matrix $D \in \mathbb{R}^{q\times q}$ with central diagonal given by 1, and the first upper diagonal given by -1, and the others given by 0, as follows:

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{q\times q} \tag{4.10}$$

For each weekly matrix X^u_i in the tensor $\mathcal{X}^u$, $||DX^u_i||_F^2$ contains the differences between positions of the same hour in two consecutive days in the i-th sub-period. Importing all the consecutive differences into Equation (4.9) is equal to induct $||\mathcal{X}^u \times_q D||_F^2$, where $\times_q$ is the *tensor-matrix multiplication* [136] on the second dimension of the tensor $\mathcal{X}^u$, *i.e.*, $(\mathcal{X}^u \times_q D)_{imk} = \sum_{n=1}^{q} \mathcal{X}^u_{ink} D_{mn}$. Considering the CPD of the tensor, we have the following derivation:

$$\mathcal{X}^u \times_q D = [\![A, B, C]\!] \times_q D = [\![A, DB, C]\!]. \tag{4.11}$$

Therefore, the first constrain that we import into Equation (4.9) is $||[\![A, DB, C]\!]||_F^2$. The second constrain is to enhance the similarity between consecutive time slots. For that, we set another matrix $H \in \mathbb{R}^{r\times r}$ with central diagonal given by 1, the second upper diagonal given by -1, and the others given by 0, and construct the second constrain as $||\mathcal{X}^u \times_r H||_F^2 = ||[\![A, B, HC]\!]||_F^2$. Now we have the tensor factorization problem with the two constrains as follows:

$$\begin{aligned}(\hat{A}, \hat{B}, \hat{C}) = \underset{A,B,C}{\arg\min} & \sum_{(i,j,k)\in\Omega_u} (\mathcal{X}^u_{ijk} - [\![A, B, C]\!]_{ijk})^2 \\ & + \lambda(||A||_F^2 + ||B||_F^2 + ||C||_F^2) \\ & + \lambda_D||[\![A, DB, C]\!]||_F^2 + \lambda_H||[\![A, B, HC]\!]||_F^2\end{aligned} \tag{4.12}$$

where λ_H and λ_D are tradeoff parameters. We solve this problem to have an approximation of the tensor $\mathcal{X}^u$ instead of the one in Equation (4.9). The problem in Equation (4.12) is also a combination of multiple standard linear least squares problems as in the `matrix` technique

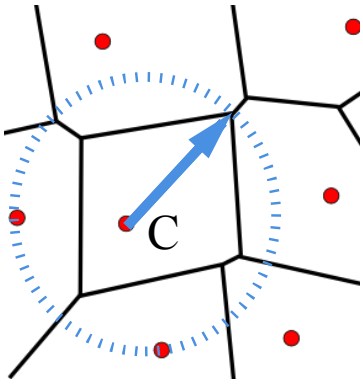


Figure 4.8: An example of computing the cell radius. For the cell C in the center of the figure, its radius is computed as the largest distance from the cell tower to the edge.

in Section 4.3.2.3. Thus, we also employ the ALS [134] to solve this optimization problem so as to have an approximation $\hat{\mathcal{X}}^u$.

Overall, the tensor factorization infers all the remaining missing locations. The last operation in this step is to extract from the tensor approximation $\hat{\mathcal{X}}^u$ the complete version of the slotted CDR-based trajectory $L^u_{\mathcal{T}}$. We use $\hat{L}^u_{\Omega^C_u}$ to represent the set of those inferred locations for the missing time slots, *i.e.*, $\hat{L}^u_{\Omega^c_u} = \{\hat{\mathbf{l}}^u_i | i \in \Omega^c_u\}$.

4.3.3.3 Cell Tower Estimation

After the second step, we already obtain a complete version of our target slotted CDR-based trajectory, while the locations inferred via the tensor factorization are geographical coordinates that do not bound to any cell towers. Considering that the spatial resolution of a CDR dataset is usually of cell towers, we design this third step to locate each inferred location in the set $\hat{L}^u_{\Omega^c_u}$ to its corresponding cell tower, in order to further enhance the accuracy.

The precondition of this step is that the deployment of cell towers is a prior knowledge. Usually, this condition is satisfied in a large-scale CDR dataset: locations that are captured by millions of CDR are highly likely to cover all the cell towers in the observing area. Therefore, we represent the cell tower deployment as a set C of all the cell tower locations, *i.e.*, $C = \{\mathbf{c}_j | 1 \leq j \leq N_{bs}\}$, where $\mathbf{c}_j$ is the j-th cell in the area; N_{bs} is the total number of cell towers in the observing area.

This step works as follows. Given an inferred location $\hat{\mathbf{l}}^u_i \in \hat{L}^u_{\Omega^c_u}$ from the tensor factorization step, the goal is to find a cell tower to replace $\hat{\mathbf{l}}^u_i$ as the new inferred location $\hat{\mathbf{l}}^u_{i,cell}$. Mathematically, we need a metric function $f(\mathbf{c}, \mathbf{l})$ to measure the correlation between the cell tower and the inferred location, and then solve the following problem for the cell tower estimation:

$$\hat{\mathbf{l}}^u_{i,cell} = \arg\max_{\mathbf{c}} f(\mathbf{c}, \hat{\mathbf{l}}^u_i), \quad s.t.\ \mathbf{c} \in C. \tag{4.13}$$

For the metric function $f(\mathbf{c}, \mathbf{l})$, we propose a function combining the distance and the possible handovers. We first estimate cell coverage by computing a Voronoi tessellation [137], and then compute for each cell tower the *cell radius* as the largest distance for the cell tower to its Voronoi polygon contour as illustrated in Figure 4.8. We define the metric function as

follows:

$$f(\mathbf{c}, \mathbf{l}) = \begin{cases} n_e^{\mathbf{c}} / n_e^{\text{total}} & \text{if } |\mathbf{c} - \mathbf{l}| \leq r_{\mathbf{c}} \\ 0 & \text{otherwise} \end{cases}, \tag{4.14}$$

where $r_{\mathbf{c}}$ represents the cell radius, $n_e^{\mathbf{c}}$ is the number of CDR observed in the cell tower $\mathbf{c}$, and n_e^{total} is the total number of CDR. Our metric function ensures that the cell towers are selected as candidates, if their distances to the inferred location are less than their own cell radii, and the candidate with the highest probability of appearance is chosen as the new inferred location.

4.3.4 Performance Evaluation

We evaluate the performance of our `CSCC` approach by comparing with the other existing techniques introduced in Section 4.3.2, in terms of their accuracy, since they can all generate trajectories with 100% completion. For that, we need the ground-truth of slotted CDR-based trajectory. Thus, in the following, we first introduce the data preliminary of our ground-truth, then present the methodology of the evaluation, and finally show the evaluation results.

4.3.4.1 Data Preliminaries

For this evaluation, the ground-truth that we have employed in the evaluation of our instant CDR completion proposals are inappropriate: they either only cover daytime hours or have very small populations, while we need to verity our `CSCC` approach on a large population due to the fact that we will apply our approach on the large-scale `CDR` dataset in the next analysis. Thus, we construct two new ground-truth: one from the `Shanghai` dataset and the other from the `Flow` dataset. Both datasets are introduced in Section 3.2. In particular, we proceed as follows:

- From the `Shanghai` dataset, we select the period of 10 weekdays and extract approximately 28K users given the criteria that each user has at least 20 locations observed per day (*i.e.*, having overall completeness > 0.8). Recall that this dataset has a fixed temporal resolution, *i.e.*, one location per hour. Thus, we have 28K slotted CDR-based trajectories having $10 \times 24 = 240$ one-hour time slots.

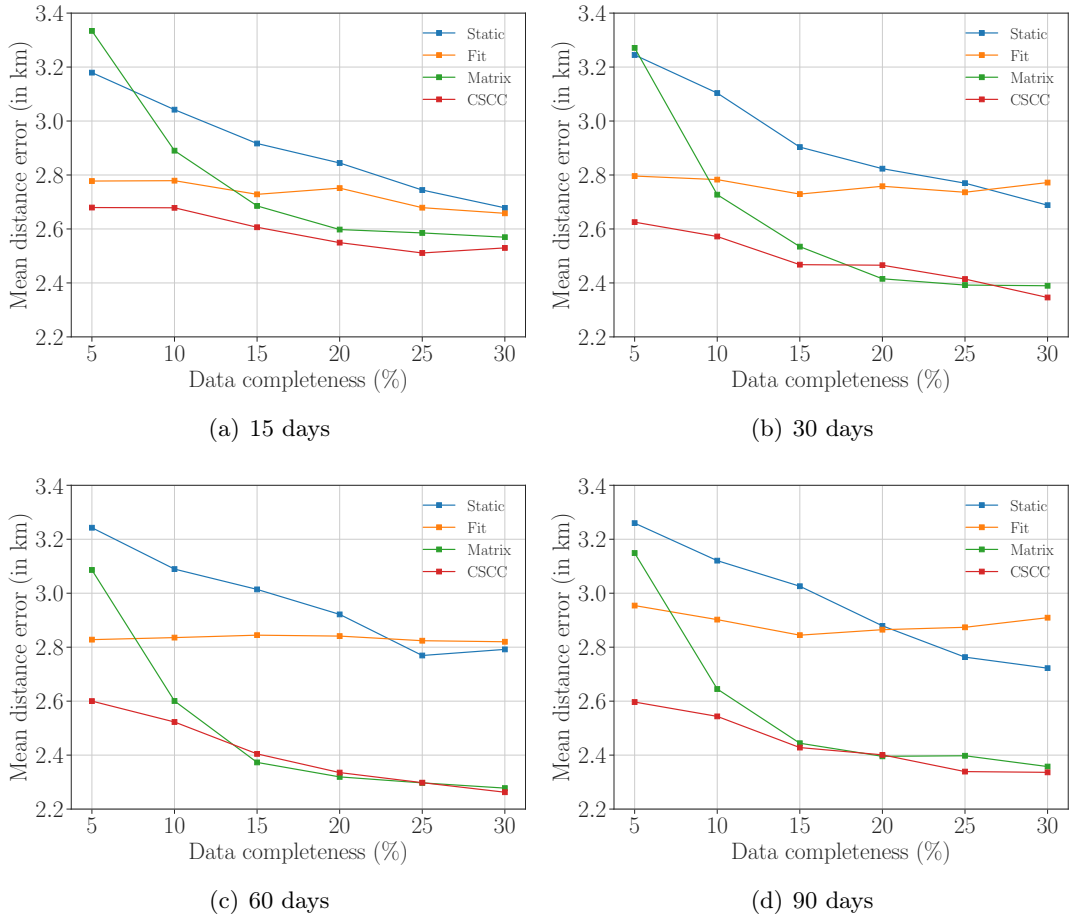
- From the `Flow` dataset, we select a consecutive period of 5 weekdays and $1,450$ users that have at least 18 locations observed per day. For these users, we also construct their slotted CDR-based trajectories having $5 \times 24 = 120$ one-hour time slots.

It is worth noting that we use weekdays and one-hour time slots as a common setting for both ground-truth in the evaluation, in order to have a fair tradeoff between the temporal resolution and the available population regarding the data completeness.

4.3.4.2 Methodology

To compare the slotted CDR completion techniques, we rely on the simulations driven by the two ground-truth datasets. In particular, the procedure of the simulation is designed as follows:

1. Duplicate each trajectory in the ground-truth datasets to "mimic" slotted CDR-based trajectories observed in $\{15, 30, 60, 90\}$ days.



(a) 15 days

(b) 30 days

(c) 60 days

(d) 90 days

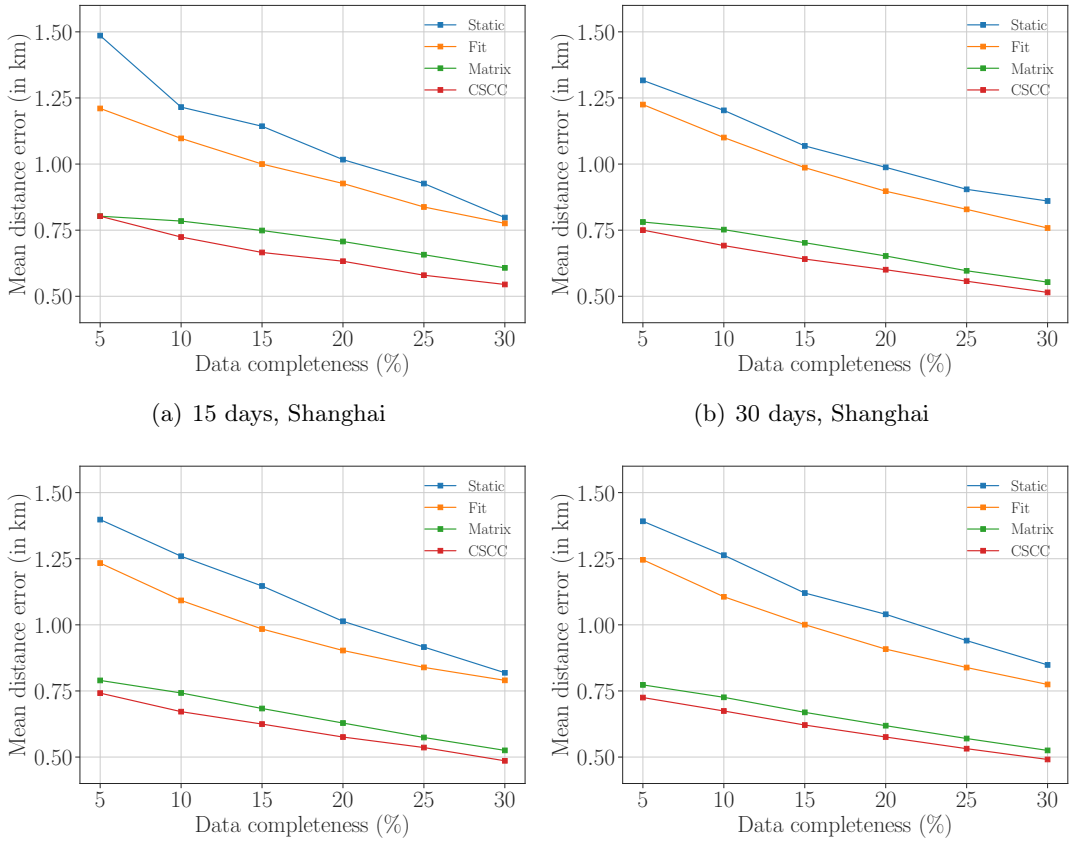
Figure 4.9: Mean distance error versus completeness on the slotted CDR-based trajectories from `Flow` dataset.

2. Generate the target incomplete slotted CDR-trajectories with the completeness percentage of $\{5, 10, 15, 20, 25, 30\}$.

3. Apply the baseline `static`, `fit`, and `matrix` techniques as well as our `CSCC` approach on the target CDR-based trajectories, and then obtain their inferred complete versions.

4. For each slotted CDR-based trajectories inferred from the techniques, compute the performance metrics introduced later on.

This procedure needs some explanation. First, on the generation of the target trajectories, we employ the `CDR` dataset to let those trajectories follow actual loss patterns in CDR. For instance, to make the target trajectories having 10% completeness, we select from the `CDR` dataset those having the same degree of completeness, compute the distribution of the hourly observation, and eliminating hourly locations from the ground-truth "mimic" slotted CDR-based trajectories according to the distribution until they have 10% completeness. Second, on the application of our `CSCC` technique, only the first two steps will be applied because we do not have the cell tower deployment in the observing area.



(a) 15 days, Shanghai

(b) 30 days, Shanghai

(c) 60 days, Shanghai

(d) 90 days, Shanghai

Figure 4.10: Mean distance error versus completeness on the slotted CDR-based trajectories from `Shanghai` dataset.

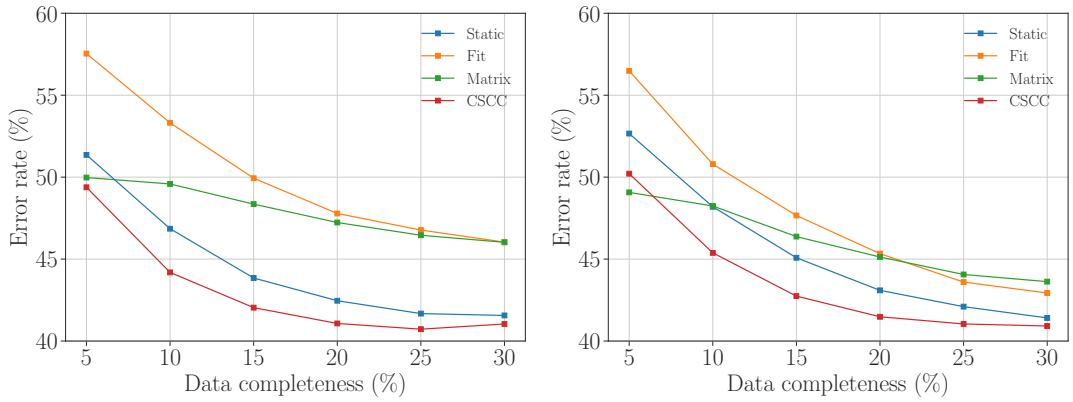
4.3.4.3 Accuracy Results

We evaluate the performance regarding two metrics: distance error and cell estimation accuracy. First, the distance error is computed as the average distance between the estimated and the actual location on all time slots having unknown locations. Mathematically, given a CDR-based trajectory L with a loss set Ω, the distance error of the trajectory is computed as follows:

$$\text{error}(\Omega, L) = \frac{1}{|\Omega|} \sum_{i \in \Omega} ||\mathbf{l}_i - \hat{\mathbf{l}}_i||_{\text{geo}} \tag{4.15}$$

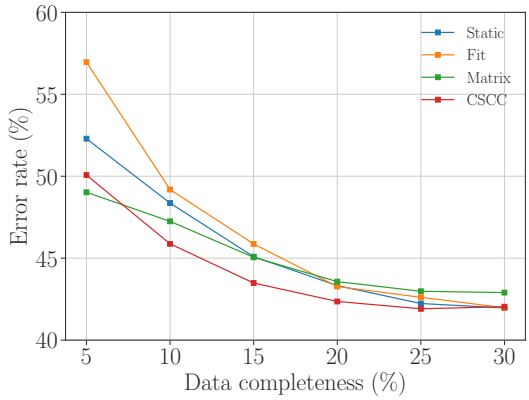
where $\mathbf{l}_i$ and $\hat{\mathbf{l}}_i$ represents the actual and estimated location at the i-th time slot respectively. Figure 4.9 and 4.10 show the mean distance error of all completed trajectories in each ground-truth dataset. We can clearly observe the following:

- The distance error of the `CSCC` approach is less than the ones of other comparison techniques. When the trajectory completeness $\geq 10\%$, the `CSCC` approach can almost have the distance error below 2.6 km on the trajectories of the `Flow` dataset and below 0.75 km on those of the `Shanghai` dataset. It is worth noting that the size of area that the cell tower covers around 2 km^2 in the former and the location represents $200m \times 200m$ in

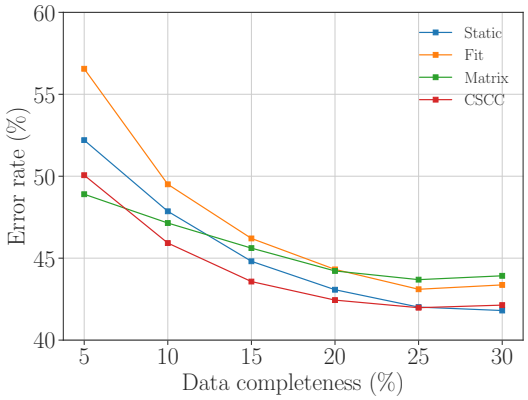


(a) 15 days, Mexico

(b) 30 days, Mexico



(c) 60 days, Mexico



(d) 90 days, Mexico

Figure 4.11: Mean cell estimation error versus completion on the slotted CDR-based trajectories from `Flow` dataset.

the latter (*cf.* Section 3.2). Such distance error is relatively good regarding the location resolution of the ground-truth datasets.

- The distance error decreases with the increasing of data completeness, and moreover, the differences among the techniques become smaller with the same increasing. This indicates that the increasing of mobility information contributes to all the techniques. Still, the distance errors of the `fit` and `static` approaches are higher than the ones of the rest, indicating that utilizing the redundancy of human mobility helps to the completion a lot, particularly when the completeness is low.

- The `matrix` approach has almost equivalent performance compared with the `CSCC` approach when the completeness $\geq 15\%$, which indicates that, when the completeness is high, the redundancy of human mobility makes major contribution in the data inference. Still, we conclude that the `CSCC` approach performs better and is more applicable considering the potential population in a CDR dataset.

Second, the cell estimation error is defined as the ratio of time slots having mistakes in estimating the cell tower, which mathematically defined as follows:

$$\text{cell-error}(\Omega, L, C) = 1 - \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbb{1}(\mathbf{l}_i = \mathbf{c}_i), \tag{4.16}$$

where $\mathbf{l}_i$ and $\mathbf{c}_i$ are the estimated and actual cell tower in the i-th time slot. Clearly, this metric is more strict than the distance error because it requires an absolute accurate estimation on each time slot. We plot the cell estimation error in Figure 4.11 only for the trajectories from the `Flow` dataset. The results show that, unless the completeness is extremely low (5%), the `CSCC` approach outperforms the other ones regarding the cell estimation error. We see that when the completeness $\geq 10\%$, more than 60% of the cell towers can be accurately recovered by the `CSCC` approach. Nevertheless, the simple `static` approach is enough when the completeness $\geq 25\%$.

Overall, these results support the advantage of our `CSCC` approach over the other competitors. Typically, for the slotted CDR-based trajectories having the temporal resolution of one hour and 10% of the completeness, which stand for 50% of the trajectories in the `CDR` dataset, our approach can accurately estimate 55% of the missing time slots and have fairly small distance errors which are in the same degree of the location resolution of the trajectories.

4.4 Summary

In this chapter, we propose novel approaches for the completion of both slotted and instant CDR-based trajectories, and evaluate them by comparing with other state-of-the-art techniques in terms of shortening time segments with unidentified locations and reducing spatial errors, leveraging the datasets presented in Chapter 3. Our data-driven simulations show that the proposed approaches outperformed previous proposals in the literature.

Chapter 5

Per-User Mobile Data Traffic Prediction

In this chapter, we mainly focus on the predictability of data traffic volumes consumed by any mobile network subscriber. The knowledge of the future consumption of mobile data traffic enables the understanding of human nature and enlightens the design of network management solutions, while the literature on this topic is fairly thin (*cf.* Section 2.1). Therefore, in this chapter, we fill this literature gap so as to achieve our primary goal. The content of this chapter and our contributions are summarized as follows:

1. Our study combines the theoretical evaluation and practical real-world prediction. For the former, we define the theoretical predictability and derive its confinement. For the latter, we state the practical predictability as the performance metric of prediction results attached to actual prediction techniques. Details are given in Section 5.1.

2. As the first preliminary step to our predictability analysis, we characterize individual mobile data traffic. We proceed from a novel viewpoint to understand the spatiotemporal correlation. Our results reveal that the three dimensions (*i.e.*, time, space, and data traffic) of a user are complementary and are all critical to achieving accurate prediction. Details are given in Section 5.2.

3. As the second preliminary step, we build time series of locations and data traffic volumes for each of our users. For that, we leverage the `CDR` and `Session` datasets to separately provide the information of mobility and mobile data consumption of the same users, and utilize our proposed CDR completion techniques to recover the incomplete mobility information. Our pioneer experience is introduced in Section 5.3.

4. We focus on the predictability of per-user mobile data traffic in isolation. In this scenario, our results show that practical algorithms that predict from the historical data volumes have high theoretical performance potential, *i.e.*, an expected average prediction accuracy of 81% over the users of study. However, real-world prediction can only achieve up to 65% by the legacy Markovian methods and up to 70% by the machine learning techniques. Details are given Section 5.4.

5. We then extend our study to the predictability with the mobility of each user jointly. We observe that, due to the strong spatiotemporal correlation, forecasting the data traffic and location jointly could achieve a better performance than doing separately. The theoretical analysis reveals that this improvement will be at maximum 10% on average, and our practical evaluation shows that the machine learning techniques can efficiently leverage the spatiotemporal correlation to improve the prediction accuracy in a degree of $1\% - 10\%$. Details are given in Section 5.5.

6. Next, we carry out our analysis from the opposite direction, *i.e.*, on the predictability of per-user locations instead of data volumes. We confirm the previous finding of the high degree of potential predictability of human mobility [17]. Moreover, it is inspiring to find that, for more than a half of our users, real-world prediction of their locations can benefit from the knowledge of their data traffic volumes as contextual information. Details are given in Section 5.6.

Finally, Section 5.7 summarizes this chapter.

5.1 Terminology and Definitions

Generally speaking, the predictability of a human behavior is the capacity of correctly forecasting it. We distinguish between two kinds of predictability. The *theoretical predictability* is the capacity that is only determined by the behavior's inherent uncertainty. It is not bound to any prediction method. The *practical predictability* is the maximum forecasting capacity that is achieved by a practical prediction method. It is measured by the prediction accuracy. We present their formal definitions in the following.

5.1.1 Theoretical Predictability

Definition 5.1: Theoretical predictability

Given a human behavior represented as finite discrete values, its theoretical predictability Π is defined as the maximum probability of correctly forecasting its current value from its historical outcomes.

The definition above aligns with the ones of Song *et al.* [17] (namely, predictability) and Feder *et al.* [65] (namely, minimal error probability). It is adopted by multiple human behavior studies (*cf.* Section 2.1.3). Our interest lies in the theoretical predictability of data traffic volumes generated by each mobile network subscriber (or equivalently, *user*), in isolation as well as jointly with the geographical locations where they consume associated mobile services. For that, we employ the information theory tools proposed in [17, 65, 67]. They aim at estimating the threshold of the theoretical predictability and utilize the *entropy* as an intermediary.

In information theory [64], entropy measures the degree of uncertainty or disorder of an information flow. Let a random variable X denote a discretized human behavior with a probability mass function $f_X(x) = \Pr(X = x)$. The entropy of this behavior is then formulated as $H(X) \equiv -\sum_{x \in X} f_X(x) \log f_X(x)$ [64]; we favor the binary logarithm so that the entropy's unit is *bit* in this thesis. Intuitively, entropy and predictability are negatively correlated variables: a random process with low (or high) uncertainty is highly (or little) predictable. Therefore, although the theoretical predictability of a human behavior is not directly observable due to the probabilistic nature, the intuition above can lead us to upper and lower bounds that confine the theoretical predictability as mathematical functions of entropy, presented next.

5.1.1.1 Upper Bound

Definition 5.2: Upper bound of theoretical predictability

Given a human behavior having N unique values and its entropy H, its theoretical predictability Π is always confined by the tight upper bound $\Pi^{\max}$, *i.e.*, mathematically,

$$\Pi \leq \Pi^{\max} \equiv \Phi^{-1}(H), \tag{5.1}$$

where Φ^{-1} is the inverse function of the function Φ defined as:

$$\Phi(x) \equiv x \log x + (1-x) \log \frac{(1-x)}{N-1}. \tag{5.2}$$

In the definition above, the derivation of Equation (5.1) and Equation (5.2) comes from rewriting the Fano's inequality [64], for which we refer the reader to [17, 65]. The function $y = \Phi(x)$ definitely has the inverse function $x = \Phi^{-1}(y)$ because it is a strictly monotonically increasing continuous function. Note that the upper bound $\Pi^{\max}$ is only determined by the entropy H and shows the highest prediction accuracy that any actual algorithms could probably achieve.

We focus on a user's time series of behaviors and predict such behaviors time slot by time slot in our analysis. Behaviors may have varying entropy for each time slot given their historical observations. Therefore, we employ the *entropy rate* to measure the average uncertainty of each time slot given the condition of being aware of the preceding behaviors. Mathematically, for a random process $\mathcal{X} = \{X_t\}$ denoting a time series of a discretized human behavior, its entropy rate $H(\mathcal{X})$ is defined as follows:

$$H(\mathcal{X}) \equiv \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} H(X_t | X_{t-1}, \cdots, X_1). \tag{5.3}$$

Note that we do not distinguish entropy and entropy rate in notation: both of them are represented by the symbol H because either of them measures the (average) uncertainty of behaviors in a single time slot.

Regarding the theoretical predictability, Song *et al.* [108] prove that the upper bound of Definition 5.2 still holds with the replacement of entropy by entropy rate in Equation (5.1). Accordingly, the upper bound limits the *overall* theoretical predictability over all the time slots. It means that, in the forecast of a time series, the accuracy of a predictor might exceed the upper bound during several sub-periods, but always falls behind the upper bound over the whole period. In order to align with real-world prediction, in the rest of this chapter, the symbol Π will always represent the overall theoretical predictability of a time series.

5.1.1.2 Lower Bound

When there are N possible values existing in a prediction, any prediction method will outperform the method that guesses one value randomly from the possible choices. Therefore, the theoretical predictability has a loose lower bound, *i.e.*, $\Pi \geq 1/N$. However, we favor a tighter lower bound derived from the common link between entropy rate and theoretical predictability. Both of them are determined by the nature of the probability mass function of a human

behavior. Once the behavior's entropy rate is fixed, the minimum value of its theoretical predictability with respect to the probability mass function is determined by the Kuhn-Tucker conditions [138, 139]. Thus, we have the following lower bound derived by Feder *et al.* [65][1].

Definition 5.3: Lower bound of theoretical predictability

Given a random process describing a human behavior that have N unique values and entropy rate H, its overall theoretical predictability has a lower bound $\Pi^{\min}$, *i.e.*, mathematically,

$$\Pi \geq \Pi^{\min} \equiv \Psi^{-1}(H) \geq \frac{1}{N}, \tag{5.4}$$

where the function Ψ^{-1} is the inverse function of the strictly monotonic function Ψ defined as follows:

$$\begin{aligned} \Psi(x) = (i+1)\log\left(\frac{i+1}{i}\right)\cdot(ix-i+1)+\log i, \\ \textbf{if } \exists\, i \in \{1,\cdots,N-1\},\quad \frac{i-1}{i} \leq x \leq \frac{i}{i+1}. \end{aligned} \tag{5.5}$$

5.1.1.3 Empirical Estimation

With the upper and lower bounds mentioned above, we can estimate a reasonable range of the overall theoretical predictability of a human behavior from the entropy rate. In our analysis, the user's locations or data traffic volumes are analyzed on per-user basis. Given a particular behavior of a user u as a random process $\mathcal{X}$, we collect a time series x_1^T of the studied behavior in T equivalent time slots, *i.e.*, $x_1^T \equiv \{x_1, \cdots, x_T\}$, and then compute the entropy rate $H_u(\mathcal{X})$ of the studied behavior. However, in any practical setting, the collected time series is finite and consequently, is unsuitable to apply Equation (5.3) on the entropy rate measurement directly. Instead, the entropy rate $H_u(\mathcal{X})$ is empirically estimated from the equation as follows:

$$H_u(\mathcal{X}) \approx H_u^{est}(x_1^T) \equiv \frac{T\log_2 T}{\sum_{i=1}^{T} L^i}, \tag{5.6}$$

where each L_i is the length of the shortest succeeding behavior subsequence that starts with the i-th time slot and never appears before. This estimator is built on the basis of the Lempel-Ziv encoding [140] and guarantees that $\lim_{T\to+\infty} E[|H_u^{est}(x_1^T) - H_u(\mathcal{X})|] = 0$.

Using the estimated entropy rate $H_u(\mathcal{X})$, we compute the upper and lower bounds (*i.e.*, $\Pi_u^{\max}(\mathcal{X})$, $\Pi_u^{\min}(\mathcal{X})$) according to their corresponding equations Equation (5.1) and Equation (5.4). Note that the theoretical predictability is probabilistic, *i.e.*, $0 \leq \Pi^{\min} \leq \Pi \leq \Pi^{\max} \leq 1$. Therefore, instead of using the corresponding inverse functions directly, we employ the numerical solver with the precision of two decimal places to compute them, *e.g.*, obtaining the upper bound $\Pi_u^{\max}(\mathcal{X})$ from the optimization problem as follows:

$$\Pi_u^{\max}(\mathcal{X}) = \arg\min_{0<\pi<1} |\Phi(\pi) - H_u(\mathcal{X})|. \tag{5.7}$$

Similarly, the lower bound can be computed by replacing the function Φ with the function Ψ in Equation (5.7).

[1]The lower bound in Definition 5.3 only stands for the predictors that follow the maximum-a-posterior rule

Consequently, the methodology presented above enables the estimation of the overall theoretical predictability of any discrete human behavior. Next, we turn to the definition of the practical predictability.

5.1.2 Practical Predictability

We rely on actual prediction methods (or in short, *predictors*) to forecast human behaviors. Although the predictability of a human behavior is determined by its uncertainty in substance, it is shown through the performance of predictors on the surface. Therefore, we define the practical predictability with respect to each predictor.

Definition 5.4: Practical predictability

Given a human behavior represented as finite discrete values, its practical predictability $\pi^{predictor}$ that corresponds to a real-world predictor is defined as the probability that this predictor can correctly forecast the behavior's current value.

In our setting, given a human behavior of a user u, we have its T observations as a finite time series $x_1^T \equiv \{x_1, \cdots, x_T\}$. Suppose that a predictor uses the first T_s values (*i.e.*, $\{x_1, \cdots, x_{T_s}\}$) to initialize itself and makes predictions of the remaining time slots. We estimate the practical predictability $\pi_u^{predictor}$ as the observed prediction accuracy using the $\{x_{Ts+1}, \cdots, x_T\}$ values as ground-truth, *i.e.*, mathematically,

$$\pi_u^{predictor} = \frac{1}{|T - T_s|} \sum_{t=T_s+1}^{T} \mathbb{1}(x_t = \hat{x}_t | x_1^{t-1}), \tag{5.8}$$

where x_t and $\hat{x}_t$ are the actual and predicted values in the t-th time slot.

Both theoretical and practical predictabilities are defined with respect to discrete human behaviors. In our predictability analysis, locations are readily discrete values because they represent cell towers, while data traffic volumes are almost continuous (*i.e.*, accurate up to kilobytes). Therefore, we discretize data traffic volumes into multiple quantizations in order to align with discrete locations, introduced later in Section 5.3.

We use the Markovian and machine learning techniques (*cf.* Section 2.1.3) as our predictors. It is worth noting that we exclude the ARIMA models because their original designs cannot accept discrete locations as input in the prediction, while we study the joint predictability in our analysis. The particular setting of each predictor will be presented on the corresponding analysis in the following sections.

So far, we have defined the theoretical and practical predictabilities and have introduced how to estimate them given a discrete human behavior. Before we step into the predictability analysis, we first present the characterization results.

5.2 Characterizing Individual Mobile Data Traffic

The characterization of per-user mobile data traffic has been addressed in the temporal and spatial dimensions, as introduced in Section 2.1.2. Generally, for mobile data traffic volumes of a user, we know that their temporal variation follows repetitive patterns that are different on weekdays and weekends [27, 34, 52], and most of them are generated in a few hours per

day [34, 52] and at a few locations [27]. These valuable findings are used in our predictability analysis. Nevertheless, the comprehensive joint understanding of how the three dimensions (space, time, and data traffic) are correlated is still missing in the picture [23]. Therefore, we characterize per-user mobile data traffic and observe the spatiotemporal variation in order to better perform our predictability analysis.

Our characterization requires the information of per-user mobile data traffic on both time and space dimensions. For that, we employ the `CDR` and `Session` datasets. We focus on 20 weekdays during a consecutive 4-week period and select 17,366 users and their days when they have at least 30 voice calls, to have enough location samples, and generate data traffic more than 1MB, to ensure "heavy" data usage. We convert all the time-stamped instant locations of each user to their temporal cell boundaries using the `stop-by-spothome` and `asym-adaptive` techniques (*cf.* Chapter 4), and then add locations into the data session records that belong to each temporal cell boundary. Finally, for each selected user, we have his data sessions consisting of locations, time-stamps, and data volumes, captured on a certain number of weekdays. Next, we begin with the characterization using these time-stamped and geo-referenced data sessions.

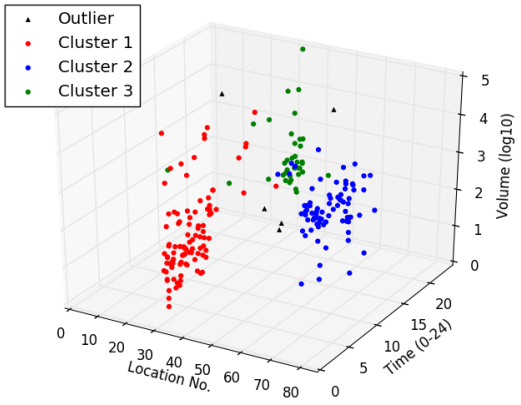
In order to have an intuitive spatiotemporal representation, we map each user's data sessions into a three-dimensional space of *location*, *time*, and *volume*. Each session becomes then a point $p(l, t, v)$ into this space. Note that we express l as the linear ordering of the corresponding bidimensional location, as returned by the Optics Algorithm [141]: a density-based cluster algorithm that places spatially close bidimentional locations as neighbors in the ordering l. This expression provides: (1) the number of each user's unique locations, *i.e.*, the maximum l; (2) the basic relation among locations, *i.e.*, two locations with close l values tend to be close in space. Then, time t is expressed by hours with decimals from 0 to 24, where the date is ignored. Finally, volume v is the magnitude of the data traffic volume, *i.e.*, $log_{10}(\cdot)$.

In this three-dimensional space, we can clearly observe how a user generates mobile data sessions. Examples are shown in Figure 5.1. For the user 1 in Figure 5.1(a), sessions are aggregated mainly on two major locations (*i.e.*with IDs 30 and 60), probably mapping to home and working place according to their time of visits. Besides, sessions containing large data traffic ($>$ 10MB) mostly occur at the location 30 during nighttime. In Figure 5.1(b) and 5.1(c), though the data traffic consumption of the two users are different, we also observe similar aggregating patterns of their sessions.

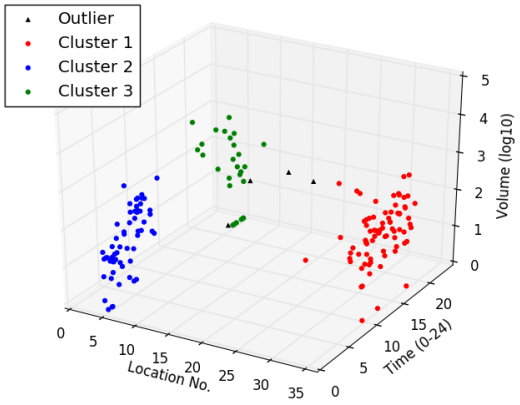
Overall, we find that the 3D space representation of a user's data sessions is typical for the vast majority of users. Hence, we investigate quantitatively the clustering of such points. For that, we use DBScan [142] to cluster each user's sessions in the three-dimensional space. The algorithm parameters MinPts (*i.e.*, the number of points required to form a cluster) and ϵ (*i.e.*, the cluster radius to consider) are set to MinPts $= 4$ and $\epsilon = 0.25$ after applying extensive tests. For the clustering, a weighted euclidean distance is measured between every two points $p_1(l_1, t_1, v_1)$ and $p_2(l_2, t_2, v_2)$, where the distance of each dimension is computed as follows:

- For the location dimension, $dist^{(location)}(p_1, p_2) = \omega_l |\mathbf{l_1} - \mathbf{l_2}|_{geo}$ in kilometers;
- For the time dimension, $dist^{(time)}(p_1, p_2) = \omega_t |t_1 - t_2|$ in hours;
- For the volume dimension, $dist^{(volume)}(p_1, p_2) = \omega_v |v_1 - v_2|$ $(|log_{10} \frac{Vol_1}{Vol_2}|)$.

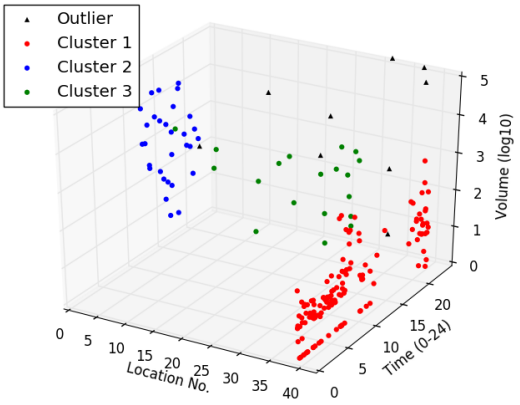
Each distance is normalized by the largest 1% of the distances on each dimension through the parameters ω_l, ω_t and ω_v, respectively. Examples of the clustering results returned by DBScan are shown in Figure 5.1(a-c) where colors are used to denote different clusters of session points.



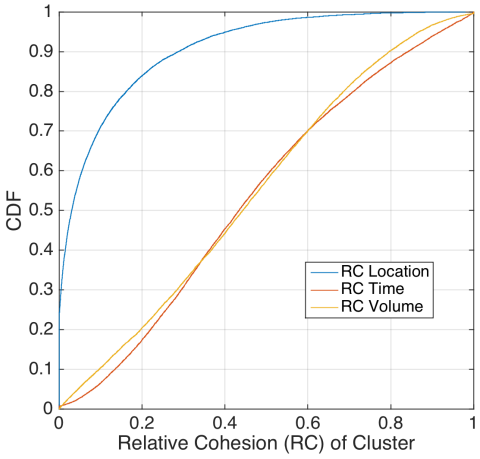
(a) Example: User 1



(b) Example: User 2



(c) Example: User 3



(d) Relative Cohesion

Figure 5.1: (a)(b)(c) Examples of mapping a user's sessions mapped the three-dimensional space. (b) CDF of each cluster's relative cohesion on the three dimensions. Figure best viewed in colors.

For each cluster, we use the *relative cohesion* (RC) to quantify the contribution of each dimension to a given cluster as $RC^{(*)} = \frac{\sum_{p \in C} dist^{(*)}(p,c)^2}{\sum_{p \in C} dist(p,c)^2}$, where C and c represent the cluster and its centroid $c = (l_{centroid}, t_{mean}, v_{mean})$, respectively. The RCs of the three dimensions satisfy $RC^{(loc)} + RC^{(time)} + RC^{(vol)} = 1$, where $0 < RC^{(*)} < 1$. Hence, if a cluster's RC in one dimension is significantly smaller than the other two dimensions, we can say that this dimension is contributing the most to the creation of the cluster.

Figure 5.1(d) shows the distributions of RCs along the three dimensions for all the selected users. The most striking behavior is the much lower RC in space than time or traffic volume. It means that *"where a user is"* drives the creation of the majority of clusters. In other words, the location of a mobile user has a high probability to trigger some routine service consumption activities. The future location of a user is usually critical in a solution aiming at predicting mobile service consumption activities. However, we also observe that some clusters are not mainly aggregated by locations. A non-negligible fraction of clusters have high RC in

the space dimension and low RCs in the time and data traffic dimensions. Accordingly, we conclude that the three dimensions are complementary, and, although with different weights, are all important for an accurate prediction of the behavior of mobile users.

5.3 Constructing Per-user Spatiotemporal Behavioral Data

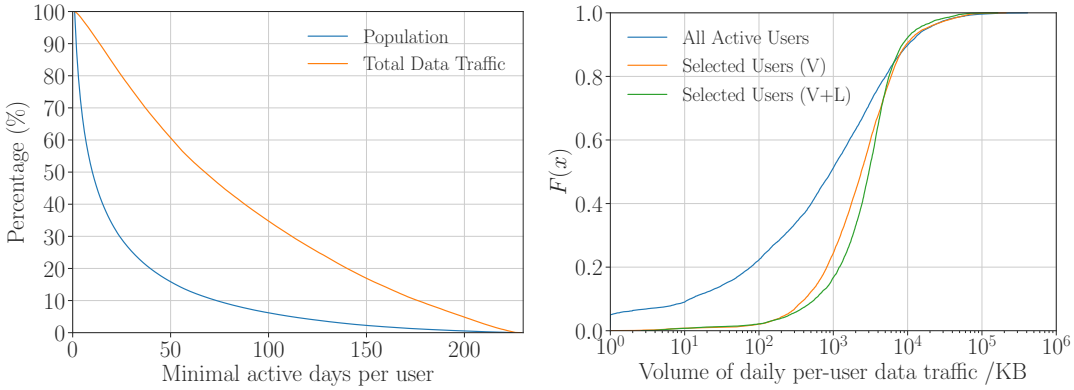
In this section, we present the data preliminary process, *i.e.*, how we select the target users and construct time series of locations and data traffic volumes for each selected user. As mentioned previously, we rely on two datasets (*i.e.*, `CDR` and `Session`) for the mobility and mobile data traffic information respectively. There is no readily choice of appropriate users or time series. We have to extract them via a data-driven analysis.

5.3.1 Active User Selection

Regarding the user selection, the basic rationale is as follows. First, we have to focus on the users who actively generate Interest data traffic through their mobile devices. Second, we need to reveal the full performance of utilized prediction methods under normal circumstances. For that, each user needs to have enough "regular" days in which he generates mobile data traffic. Therefore, we focus on weekdays and exclude holidays, vacation periods, and weekends (having insufficient days and different data traffic patterns). Third, we need to study the predictability along with mobility. Thus, each selected user must have enough location samples (*i.e.*, voice call CDR in the `CDR` dataset) to build his slotted CDR-based trajectory. In the following, we analyze the `CDR` and `Session` datasets and proceed the user selection.

We first discuss the criterion of the user selection with respect to mobile data traffic. During the observing period shared by the two datasets, there are 229 regular weekdays in total. We consider the daily activeness of each user in these weekdays. For that, an *active* day is defined as a day in which a user generates at least 1KB of mobile data traffic. We then portray in Figure 5.2(a) (blue line) the minimum number of active days versus the percentage of users. We observe that nearly 85% of the users are "inactive": they only generate data traffic in less than 50 days despite a 15-month observing period. For comparison, we plot in the same figure (orange line) the ratio of the total data traffic generated by these users. We see that these "inactive" users only take account for 40% of the total mobile data traffic, while approximately 5% of the users who have at least 150 active days generate almost 20% of the total mobile data traffic. The observation above confirm the existence of the so-called "heavy" users, as in [27, 52]. In our case, we need to focus on the heavy users and to have their time series of data traffic volumes as long as possible. Considering both the user activeness and the available number of users, we choose the users who have more than 150 active days, which provides us approximately 92K of the selected users.

We validate the criterion above in terms of the per-user data traffic consumption. Figure 5.2(b) portrays the CDF of the mean daily data volume of each user. We see that the distribution of our selected users (orange line) implies a more positive usage of mobile data traffic than that of all the users: 80% of the former users generate at least 1MB per day on average, while only 50% of the latter users do the same. Consequently, these selected users are "heavy" users which we need; they are highly active every weekday and generate a large amount of mobile data traffic.



(a) (b)

Figure 5.2: (a) Distributions of the number of minimum active days versus the number of users and the ratio of mobile data traffic. (b) CDF of the mean daily mobile data traffic of each user over all the users and the selected users.

Second, we have to select users in terms of the number of locations of each user. We need to build complete slotted CDR-based trajectories (*cf.* Section 4.1), in order to perform the joint predictability analysis. Due to our analysis on the completeness of the `CDR` dataset (*cf.* Section 3.4), there is no readily CDR-based trajectory having 100% of the completeness. Therefore, our slotted CDR completion technique proposed in Section 4.3 appears as a rescue, which is later used to reconstruct full complete trajectories. Our technique achieves its best average performance when the trajectories to be completed have at least 20% of the completeness (*cf.* Section 4.3.4). Besides, the completeness analysis shows that the number of available users having at least 20% of the completeness of is extremely small under high temporal resolutions. Therefore, we set the temporal resolution to one hour. In summary, our user selection criterion with respect to mobility information is that a user should have a slotted CDR-based trajectory having at least 20% of the completeness. Given this criteria, 7K users are selected from the overall 92K "heavy" users. These users, as shown in Figure 5.2(b), have the distribution of daily mobile data traffic volumes (green line) highly close to that of all the "heavy" users. It indicates that adding the mobility criteria into the user selection does not impact the activeness of mobile data traffic consumption heavily in the selected users.

5.3.2 Discretization of Volumes and Locations

As mentioned in Section 5.1, we need to construct time series of discrete data traffic volumes and locations for the select users, to meet with the definitions of the predictability.

Regarding mobile data traffic, for each select user u, we construct a time series represented by $v_1^T(u) = \{v_1^u, \cdots, v_T^u\}$, where v_i^u is his discrete mobile data traffic generated during the i-th time slot. We consider four different temporal resolutions for time slots: 15, 30, 45, or 60 minutes. In each case, we aggregate data traffic volumes from all the data sessions occurred in each time slot. For instance, when the resolution is 60 minutes, there are 24 time slots on each day and mobile data traffic is computed on an hourly basis: This will be our default setting, unless stated otherwise. For the discretization of data traffic volumes, we favor a representation that captures the data traffic magnitude over a uniform discretization. The rationale is that

Table 5.1: Users of Study

Group	Population	Time Series	Resolution	Days
User set $\mathcal{U}_1$	92K	$v_1^T(u)$	$\{15, 30, 45, 60\}$Min	≥ 150
User set $\mathcal{U}_2$	7K	$v_1^T(u)$, $l_1^T(u)$	60Min	

one is more interested in predicting whether a user will generate, *i.e.*, KiloBytes, MegaBytes or GigaBytes of traffic, rather than if a user's demand will be in the first (1 KB, 333 MB), second (334 MB, 666 MB) or third (667 MB, 1 GB) portions of one GB. Specifically, we employ the following five different quantizations of the data traffic volume spectrum, listed in order of increasing accuracy:

- **Q1**: four quantization levels, *i.e.*, *idle* (0 KB), *light* (1 KB, 1 MB), *heavy* (1 MB, 1 GB), and *extremely heavy* (1 GB, 10 GB).
- **Q2**: eight quantization levels, *i.e.*, 0, $(1, 10)$, $(10, 10^2)$, $\ldots$, $(10^6, 10^7)$, all values in KB. Once stated otherwise, **Q2** will be our default setting.
- **Q3**: twelve quantization levels, obtained by bisecting each level over 1 MB in Q2, *e.g.*, splitting (1,10) MB into (1,5.5) MB and (5.5,10) MB.
- **Q4**: sixteen quantization levels, obtained by trisecting each level over 1 MB in Q2.
- **Q5**: forty quantization levels, obtained by nine-secting each level over 1 MB in Q2.

Regarding mobility, each selected user u is collected as a time series of discrete locations, represented by $\ell_1^T(u) = \{\ell_1^u, \cdots, \ell_T^u\}$, where ℓ_i^u is the user's *representative location* of the i-th time slot. Such a time series can be converted from the user's corresponding slotted CDR-based trajectory (*cf.* Section 4.3.1) by replacing each location coordinate $\mathbf{l}_i^u$ by its corresponding cell tower identifier ℓ_i^u. Note that the distance between each two discrete locations is still measurable as we have their geographical coordinates. In particular, each day is split into 24 time slots as our default temporal resolution of data traffic volumes; each representative location is selected on an hourly basis. Even then, there is no readily full complete CDR-based trajectory that can be extracted from the `CDR` dataset. For that, we apply our proposed slotted CDR completion technique on the incomplete CDR-based trajectories of the selected users, and then convert them to time series of discrete locations.

In summary, we have two criteria of the user selection corresponding to mobile data traffic and mobility, respectively. Two groups of the "heavy" mobile data traffic users are then selected given the criteria, as shown in Table 5.1. The first user set $\mathcal{U}_1$ contains 92K users and their data session records extracted from the `session` dataset. For each user $u \in \mathcal{U}_1$, we have his time series of discrete mobile data traffic under four temporal resolutions and five data traffic volume quantizations. The user set $\mathcal{U}_1$ and its data will be used in the predictability analysis using temporal dynamics in Section 5.4. The second one $\mathcal{U}_2 \subset \mathcal{U}_1$ consists of 7K users and has their locations extracted from the `CDR` datasets in addition. For every user $u \in \mathcal{U}_2$, we have his time series of locations and quantized data traffic volumes in the temporal resolution of one hour. This data will appear in the rest of the predictability analysis in Section 5.5 and 5.6.

Table 5.2: Entropy Rate and Corresponding Theoretical Predictability

Entropy rate	Theoretical predictability
Actual entropy (rate) $H_u \approx H_u^{est}(x_1^T)$	$\Pi_u^{\max} \equiv \Phi^{-1}(H_u, N)$
	$\Pi_u^{\min} \equiv \Psi^{-1}(H_u, N)$
Random entropy (rate) $H_u^{\text{rand}} \equiv \log N$	$\Pi_u^{\text{rand}} \equiv \Phi^{-1}(H_u^{\text{rand}}, N)$
Temporal-uncorrelated entropy (rate) $H_u^{\text{unc}} \equiv -\sum_{x \in \text{Unique}(x_1^T)} \frac{N(x)}{T} \log(\frac{N(x)}{T})$	$\Pi_u^{\text{unc}} \equiv \Phi^{-1}(H_u^{\text{unc}}, N)$

5.4 Investigation through Temporal Dynamics

In this section, we study the predictability of mobile data traffic generated by individual users. For now, we focus on the forecasting scenario of data traffic volumes in isolation, and we will consider the joint predictability of mobile data traffic and mobility later on. In the following, we first study the theoretical predictability and then validate the performance of several predictors in real-world prediction.

5.4.1 Theoretical Predictability

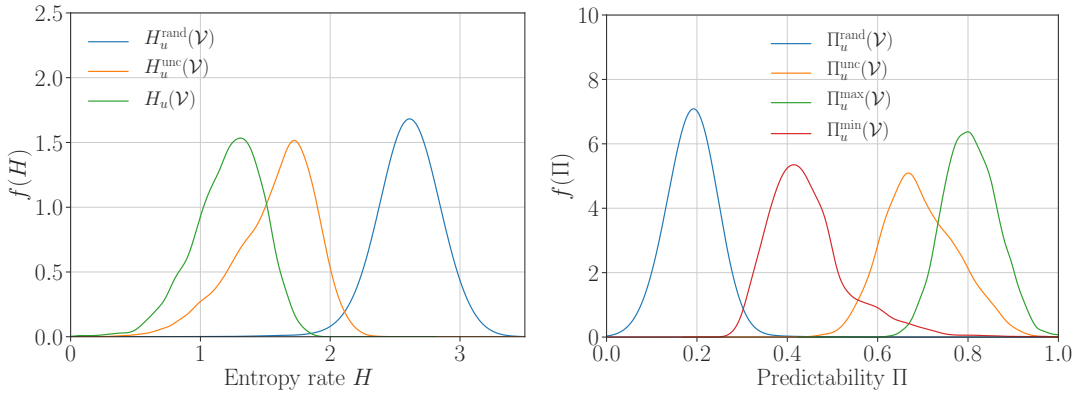
We evaluate the theoretical predictability $\Pi(\mathcal{V})$ of every "current" data traffic volume of a user predicted by his historical data traffic usage. We consider the mobile data traffic generation of each user as a stochastic process $\mathcal{V} = \{V_i\}$ and the discretized data traffic volume at the i-th time slot as a random variable V_i. For each user $u \in \mathcal{U}_1$, we have the observations of the stochastic process $\mathcal{V}$ during T consecutive time slots, represented by the time series $v_1^T(u)$. To have the confinement of the theoretical predictability $\Pi(\mathcal{V})$, we leverage $v_1^T(u)$ to derive three variants of the entropy rate which we will use to investigate the properties of the process $\mathcal{V}$. The variants are listed in Table 5.2 and are presented particularly as follows:

- The *actual entropy* (rate), denoted by $H_u(\mathcal{V})$, depends not only on the frequency of the appearance of each discretized data traffic volume but also on the order in which they appear, capturing the temporal order presented in a user's traffic usage pattern. The formal definition of $H_u(\mathcal{V})$ is given in Equation (5.3), while we employ the empirical estimator (*cf.* Section 5.1) to compute $H_u(\mathcal{V})$ using $v_1^T(u)$ in practice.

- The *random entropy* (rate) is computed by considering that each V_i is equally probable and time-independent in the process, defined as $H_u^{\text{rand}}(\mathcal{V}) \equiv \log N$, indicating the maximum degree of the uncertainty of the process $\mathcal{V}$ and the theoretical maximum value of the actual entropy $H_u(\mathcal{V})$.

- The *temporal-uncorrelated entropy* (rate) only considers the heterogeneity of the mobile data traffic of a user and characterizes the mobile data traffic demand $\mathcal{V}$ that has no temporal correlations, hence its name. With respect to the time series $v_1^T(u)$, it is computed as $H_u^{\text{unc}}(\mathcal{V}) \equiv -\sum_{v \in \text{Unique}(v_1^T(u))} \frac{N(v)}{T} \log(\frac{N(v)}{T})$, where $N(v)$ is the number of the appearance of the unique value v in the time series.

It is worth noting that the assertion $H_u \leq H_u^{\text{unc}} \leq H_u^{\text{rand}}$ stands for any time series.



(a) (b)

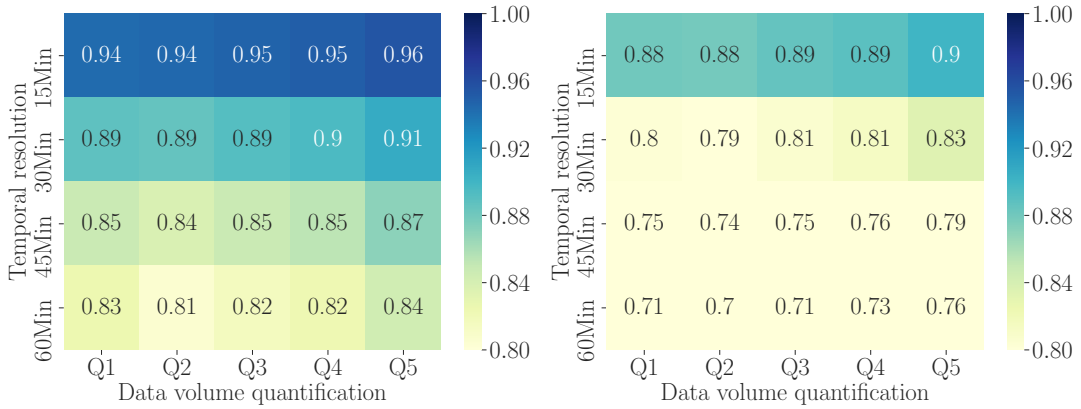
Figure 5.3: (a) Distributions of the random entropy $H_u^{\text{rand}}(V)$, the temporal-uncorrelated entropy $H_u^{\text{unc}}(V)$, and the actual entropy $\mathcal{H}_u(\mathbb{V})$ as observed in the individual traffic demand generated by the users $\mathcal{U}_1$. (b) Equivalent distributions of the upper bounds on the predictability $\Pi_u^{\text{rand}}(V)$, $\Pi_u^{\text{unc}}(V)$, and $\Pi_u^{\text{max}}(\mathbb{V})$.

As discussed in Section 5.1, we are able to evaluate any theoretical predictability by knowing its confinement as the upper and lower bounds computed from the entropy rate, as listed in Table 5.2. In our context, the upper bound Π_u^{max} is an estimation of the maximum achievable accuracy in the prediction of mobile data traffic volumes; the lower bound Π_u^{min} in reverse shows the lowest accuracy that can be expected if a predictor works properly [65]. Regarding the mobile data traffic, we compute three upper bounds, *i.e.*, $\Pi_u^{\text{rand}}(\mathcal{V})$, $\Pi_u^{\text{unc}}(\mathcal{V})$, and $\Pi_u^{\text{max}}(\mathcal{V})$, and one lower bound $\Pi_u^{\text{min}}(\mathcal{V})$. In addition, recall that multiple representations of $v_1^T(u)$ are possible, depending on the combination of time granularity and volume quantization. For each such combination, different results are obtained in terms of entropy rate and thus, predictability.

5.4.1.1 Baseline Results

Our baseline results are shown in Figure 5.3, and are obtained under 1-hour (time) and Q2 (data traffic volume) quantization level described in Section 5.3. Specifically, Figure 5.3(a) displays the PDF of the three versions of the entropy rate. Figure 5.3(b) portrays the PDF of the corresponding bounds of the theoretical predictability.

Let us start by considering the PDF of $H_u^{\text{rand}}(\mathcal{V})$ in Figure 5.3(a). Its range indicates that an equiprobable distribution of data traffic volumes during each time slot can be represented with up to three bits. This phenomenon is normal, as we consider eight traffic volume quantization levels as our default setting. When the temporal-uncorrelated entropy, i.e., $H_u^{\text{unc}}(\mathcal{V})$, is concerned, a sizable shift of probability occurs. The uncertainty decreases to $2^{H_u^{\text{unc}}(\mathcal{V})} = 2^{1.8} \approx 3$ bits. Under this model, each user tends to generate traffic that is described by just three quantization levels out of the eight available. For instance, at each time slot, a user may generate traffic by one order of MB or tens of MB, or stay idle; but typically she will not generate smaller or larger traffic volumes. The same holds for users who generate, e.g., order-of-KB or order-of-GB traffic. Ultimately, a reduced entropy rate implies higher regularity in the mobile data traffic demand. However, our main result is the significant shift presented by the PDF of $H_u(\mathcal{V})$, which is amassed around a value 1.24. When taking the temporal ordering of data



(a) $\overline{\Pi_u^{\max}}(\mathcal{V})$ (b) $\overline{\Pi_u^{\text{unc}}}(\mathcal{V})$

Figure 5.4: (a) Heatmap of the median predictability upper bound $\Pi_u^{\max}(\mathcal{V})$ for different quantizations of time and traffic volume. (b) Heatmap of the median predictability upper bound $\Pi_u^{\text{unc}}(\mathcal{V})$ for different quantizations of time and traffic volume.

sessions into account, one can reduce the uncertainty to just two quantization levels.

The probability distributions in Figure 5.3(b) confirm these findings and provide upper and lower numerical bounds to the predictability of the mobile data traffic demand generated by individual users. We observe that $\Pi_u^{\text{rand}}(\mathcal{V})$ peaks at 0.18, *i.e.*, it is very hard to guess the volume of traffic generated by such a stochastic model. The predictability grows for $\Pi_u^{\text{unc}}(\mathcal{V})$ which peak at 0.68. More importantly, $\Pi_u^{\max}(\mathcal{V})$ indicates that the demand of a user can be possibly predicted within 81% accuracy on average. It means that there is 19% of the time does the user generate a traffic volume in a manner that appears to be random, but in the remaining 81% of the time, we could predict his volume. Besides, the lower bound $\Pi_u^{\min}(\mathcal{V})$ peaks at 0.42, indicating that there are at least 42% of the data traffic volumes of a user appears in a regular pattern on average. These results prove that the traffic volume which users generate via their mobile devices is highly predictable.

The results in Figure 5.3(a) and Figure 5.3(b) refer to the case where the individual mobile data traffic is represented with a temporal resolution of 60 minutes and volume quantization Q2. In fact, data granularity has been shown to have a significant impact in the predictability of mobility [67]. We thus explore if the same is true in the case of predictability of the mobile data traffic demand.

The heatmaps in Figure 5.4(a) and in Figure 5.4(b) show the median of the upper bound on the mobile demand predictability, over four temporal and five traffic volume quantization levels. The two plots refer to $\Pi_u^{\max}(\mathcal{V})$ and $\Pi_u^{\text{unc}}(\mathcal{V})$, respectively.

In Figure 5.4(a), we observe that $\Pi_u^{\max}(\mathcal{V})$ is not significantly affected by the traffic volume quantization, i.e., our results appear to have general validity under different levels of accuracy in the representation of the mobile demand. In contrast, surprisingly, it grows with finer-grained temporal resolutions. The reason is that more idle time slots appear as the temporal resolution is increased; these idle slots tend to dominate the real-world distribution of mobile data traffic, reducing the entropy and improving the overall predictability but hiding the predictability of non-idle time slots. This is confirmed by Figure 5.4(b): $\Pi_u^{\text{unc}}(\mathcal{V})$ is slightly affected by variations in the time granularity. Instead, it is strongly dependent on the traffic volume quantization,

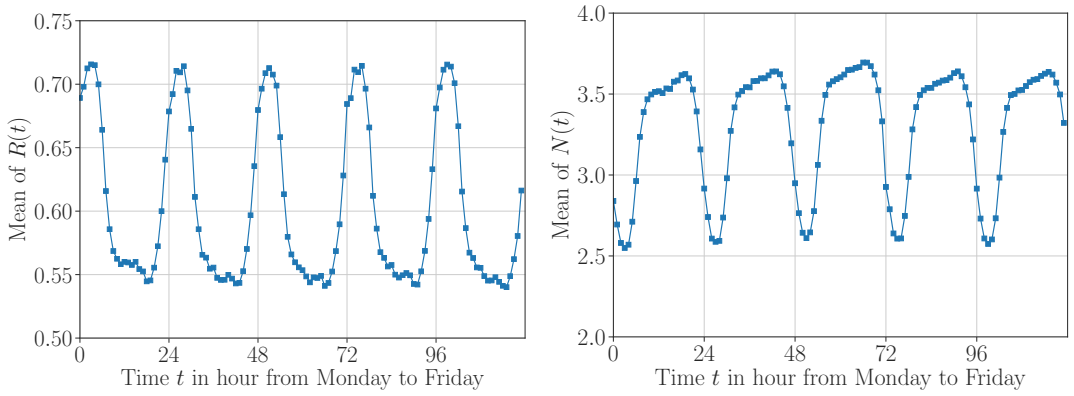


Figure 5.5: Temporal dynamics of individual mobile data traffic volume, during the average week. (a) Regularly $R(t)$. (b) Number of observed levels $N(t)$.

an artifact of the lack of temporal correlation in this model, which disappears in $\Pi_u^{\max}(\mathcal{V})$.

5.4.1.2 Temporal Variability

It is well-known that the demand generated by mobile phone users is time-dependent [34]. Thus, a relevant question is whether the predictability of mobile data traffic similarly undergoes temporal variations. Nevertheless, entropy does not allow a detailed view of such temporal variation.

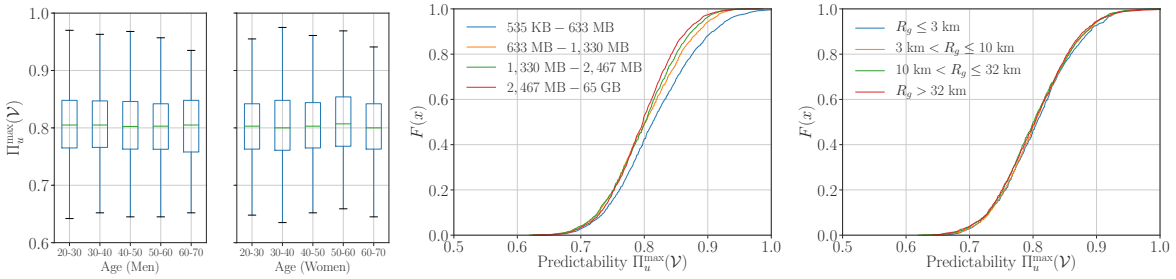
To that end, we compute, for each user and on an hourly basis, the regularity $R(t)$, *i.e.*, the probability that the user generates the most likely traffic volume observed during each hour. Similarly, we define $N(t)$ as the number of unique traffic volume levels observed at each hour. Regularity provides a lower bound on the predictability, as it ignores the temporal correlation of users' traffic demand patterns [17, 108], and is typically inversely correlated with $N(t)$. $R(t)$ and $N(t)$ can be seen as proxies of the predictability and entropy rate.

Figure 5.5 shows the evolution of $R(t)$ and $N(t)$ over the average week. We remark a clear circadian rhythm. At night, the mean of $R(t)$ rises to approximately 0.7, meaning that, on average, the user's demand matches the most likely traffic volume around 70% of the time. During the morning, working hours, and evening, $R(t)$ drops to 0.55. $N(t)$ shows opposite trends, as expected. As $R(t)$, $N(t)$ (as tied to the predictability) varies significantly over time as well. However, we do not observe significant variations from one day to another, which suggests that mobile data traffic volume predictability is not only imposed by the working schedule but is intrinsic to more generic human activities.

5.4.1.3 Variability across User Types

Our datasets let us explore several additional dimensions of the traffic volume predictability. These dimensions are related to the nature of the user. The results are shown in Figure 5.6 and discussed below.

- *Age and gender.* The age and gender of the mobile user are known to affect the way mobile services are consumed. However, Figure 5.6(a) shows that these do not affect in



(a) (b) (c)

Figure 5.6: (a) Boxplot of $\Pi_u^{\max}(\mathbb{V})$ categorized by user's age and gender. Each box denotes the median, $25^{th} - 75^{th}$ percentiles, and minimum and maximum values. (b) CDF of $\Pi_u^{\max}(\mathbb{V})$, when users are separated according to the total mobile data traffic volume of each user. Four groups are considered: 535 KB – 633 MB, 633 MB – 1,330 MB, 1,330 MB – 2,467 MB, and 2,467 MB – 65 GB. Each group contains 25% of the observed users. (c) CDF of $\Pi_u^{\max}(\mathbb{V})$, when users are separated according to their level of mobility. Four ranges of radius of gyration are considered, mapping to sedentary ($R_g \leq 3$km), urban (3 km $< R_g \leq 10$ km), peri-urban (10 km $< R_g \leq 32$ km), and long-range commuting ($R_g > 32$ km) profiles.

a remarkable manner the predictability of the traffic volume. Hence, age- and gender-induced behaviors remain similarly predictable when it comes to the traffic volume.

- *Overall mobile data traffic volume consumption.* We categorize the user set $\mathcal{U}_1$ into four groups, according to their data consumption during the observing period. Each group consists of 25% of the observed users. As shown in Figure 5.6(b), the predictability $\Pi_u^{\max}(\mathcal{V})$ tends to decrease as the data volume increases. Yet, the mean of $\Pi_u^{\max}(\mathcal{V})$ is 81.9% in the group of 535 KB – 633 MB and 79.81% in the group of 2,467 MB – 65 GB. The difference is thus small and can be imputed to the fact that a larger amount of data naturally entails more complex dynamics as well as may be characterized by less regularity. We conclude that the overall amount of generated traffic only marginally impacts the potential predictability of traffic volumes.

- *Mobility level.* Some correlations between mobility and mobile service usage were observed in the literature [27, 35, 40]. We study whether this occurs with mobile data traffic volume predictability as well. To that end, we compute, for each user, the radius of gyration [16] that measures the span of the user's movement, and classify our users into the following categories: sedentary, urban, peri-urban, and long-range commuters [24]. Figure 5.6(c) presents the CDF of $\Pi_u^{\max}(\mathcal{V})$ computed on each user category. Again, there exists a slight shift towards lower values in the $\Pi_u^{\max}(\mathcal{V})$ distribution, as the level of mobility grows. However, the variation is minimal, at approximately 0.4% between commuters and sedentary users. This implies that less mobile users may have slightly more regular data traffic patterns, yet the difference is marginal.

In conclusion, we find no significant correlations between dominant users' features and the predictability of the mobile data traffic volume they generate. In fact, all plots in Figure 5.6 indicate that the heterogeneity of $\Pi_u^{\max}(\mathcal{V})$ across all users is fairly low: the high predictability of traffic volume is a property shared throughout the whole user population.

5.4.2 Practical Predictability

So far, we have revealed the high theoretical predictability of per-user mobile data traffic volumes in isolation. In the following, we address whether or not this high theoretical predictability can be achieved. For that, we evaluate the practical predictability of several predictors in the real-world prediction of mobile data traffic volumes generated by the users of the set $\mathcal{U}_1$. It is worth noting that, for all the practical predictability analyses presented in this chapter, we employ our default setting on the time series, *i.e.*, the temporal resolution of one hour and the volume quantization Q2.

5.4.2.1 Methodology

We compute the practical predictability defined in Section 5.1.2. For each predictor and each user's time series, we initialize the predictor using data in a number of the time series' beginning days and evaluate the average prediction accuracy (defined in Equation (5.8)) on the rest of the time series as our estimation of the practical predictability. Recall that each user has a time series of discrete data traffic volumes collected in at least 150 days. To let each predictor "warm up" entirely and to exclude the accuracy impact brought by the lack of enough samples, we employ the first 100 days (*i.e.*, $T_s = 100 \times 24$ in Equation (5.8)) in the initialization of all the predictors. To ensure prediction substantial accuracy, we update each predictor periodically in the evaluation. In particular, given a user's time series of data traffic volumes $v_1^T(u)$, the practical predictability of a predictor is computed by the following procedure:

(1) Initialize the predictor using data from the first 100 days, *i.e.*, the partial time series $v_1^{T_s}(u)$ and then set $D = 100$ days.

(2) Use the predictor to make predictions of all time slots in the $(D+1)$-th day.

(3) Update the predictor using the data traffic volumes generated in the $(D+1)$-th day and set $D = D + 1$.

(4) Go back to (2) if D does not exceed the last day of $v_1^T(u)$. If it exceeds, stop the iteration and compute the practical predictability $\pi(\mathcal{V})$ defined according to Equation (5.8) .

In this section, we employ this procedure on the time series of data traffic volumes and the predictors introduced later on. Note that this procedure also holds for the remaining practical predictability analyses in this chapter, *i.e.*, to be applied on the time series of both data traffic volumes and locations.

5.4.2.2 Predictors

Building upon the procedure above, we evaluate the practical predictability $\pi(\mathcal{V})$ of several predictors. Since the theoretical predictability upper bound shows the highest expected performance of any predictor that leverages the regularity hidden in the temporal orders of a time series, we mainly choose our predictors that utilize such regularity, which are listed as follows.

- **Markovian predictors**. We utilize all the Markovian predictors presented in Section 2.1.3.3, *i.e.*, the `PPM`, `MC`, `SPM`, and `ALZ` predictors. The `PPM` and `MC` predictors make a prediction of the current data traffic volume from the preceding k previous data traffic volumes, the `SPM` predicts from the ratio α of the longest preceding data traffic volumes

that appears previously, and the ALZ decides the length of the preceding data traffic volumes via an automatic sliding window. Following the previous experience of the application of PPM, MC, and SPM on predicting locations [76] and aggregated data traffic volumes [33], we set their corresponding parameters as follows. For PPM(k) and MC(k), we choose $k \in [1, 5]$; for the SPM(α), we choose $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Regarding their implementations, the MC builds a transition matrix of probabilities by employing simple Markov chains and maximum likelihood probability estimation. For the PPM and SPM predictors, we favor their implementations in [78, Method C] and [79], respectively. The ALZ follows the algorithm presented in [80, Figure 3].

- **MLP (Multilayer Perceptron)**. This is the most classical machine learning technique that leverages artificial neural networks [83]. It has well-tested implementation and good flexibility, which can be even deployed on mobile devices equipped with mobile AI hardware. In particular, we employ a fully connected neural network that has three hidden layers, where each layer has 256 neurons and is activated by the ReLu function [83]. In the training phrase, the network is trained by the Adam optimizing method [143] with the initial learning rate 0.001. Regarding network input and output, we distinguish two predictors based on the MLP.

  - The MLP predictor has the same input and output format as in the PPM or MC predictor, *i.e.*, the preceding k data traffic volumes as input and the prediction of current data traffic volume as output. Compared with the Markovian predictors, it can accept a larger k and thus, we set $k \in [1, 8]$.
  - The MLP-CI predictor further employs the temporal contextual information as input along with the preceding data traffic volumes. Particularly, its input vector consists of k daily vectors that represent the data traffic consumption of a user in the previous days. Similarly, we set $k \in [1, 8]$. Each daily vector contains the discrete data traffic volume, the weekday via one-hot encoding, the time slot's hour, and the time difference with the target time slot in hours and days respectively. This predictor still generates a prediction of the current data traffic volume as output.

5.4.2.3 Prediction Results

Our results with respect to the practical predictability $\pi(\mathcal{V})$ of discrete data traffic volumes of each user $u \in \mathcal{U}_1$ are shown in Figure 5.7. For the per-user prediction accuracy of the PPM, MC, SPM, MLP, and MLP-CI predictors regarding their possible parameters, we plot the CDF of the prediction accuracy of each user categorized by the different settings of the same predictor in Figure 5.7(a-e). We observe that the performance of these predictors varies slightly with different settings. Particularly, the PPM and MC achieve their overall best performance when $k = 2$, so does the SPM when $\alpha = 0.25$. The reason is that the Markovian predictors have large probability space that increases quickly following a power law with the order k. Therefore, when $k > 2$, these predictors may suffer from lack of sufficient samples. Correspondingly, the MLP and MLP-CI achieve their best when $k = 4$, indicating the advantage of machine learning techniques clearly, *i.e.*, they can accept larger preceding data effectively in the prediction. Overall, we see that on each particular prediction, importing more historical data does not significantly enhance the prediction accuracy.

In our case, all the predictors are applied on a per-user basis, which means each user may have different setting of a predictor to have his own best prediction accuracy. For that, we employ a 3-fold cross validation process during the initialization of each predictor, to determine the best setting of each user. Here the practical predictability $\pi(\mathcal{V})$ of a certain predictor represents the best performance that it is achieved by each user on his own setting. By merging the results above, we portray in Figure 5.7(f) the CDF of the practical predictability $\pi(\mathcal{V})$ of each predictor in the prediction of discrete data traffic volumes, where we observe the following:

- Even the worst predictor, *i.e.*, `ALZ`, can still achieve the average prediction accuracy of 55%, which is approximately 10% below the best predictor and 26% below the theoretical predictability upper bound. This is consistent with the theoretical lower bound $\Pi_u^{\min}(\mathcal{V})$ (*i.e.*, the average accuracy of 42% in theory).

- The other Markovian predictors (`PPM`, `MC`, and `SPM`) have almost the same distributions of the prediction accuracy. This is reasonable because they are all designed upon the Markov chian. Particularly, they all have the mean prediction accuracy of 65%, which is still 16% below the upper bound .

- The `MLP` performs slightly better than the Markovian ones, achieving the average accuracy of 67%. In the distribution, the prediction accuracy per user varies more heavily than the latter. Combining the results, we conclude that among the predictors that only employ the regularity of the temporal orders of discrete data traffic volumes, it is hard to select one which has noticeable advantage over the others. In this context, although the `MLP` performs better, the simple `MC` is quite sufficient having a good trade-off between the computing complexity and achieved performance.

- The overall best performance comes from the `MLP-CI` predictor, which achieves the average prediction accuracy of 70%. Compared with the others, this predictor uses the temporal context as input, which provides more information in each time slot and capture the temporal regularity of mobile data traffic in a better manner.

Consequently, our results confirm that the high degree of the theoretical predictability is consistent with the practical predictability. Even a simple Markovian predictor can achieve a reasonably good performance in the prediction of per-user mobile data traffic, which is consistent with the observation on the aggregated mobile data traffic [33]. We also observe that leveraging machine learning and importing into prediction event time as contextual information can further enhance the prediction performance. So far, we have fully utilized information in time series of mobile data traffic volumes. To seek an opportunity to improve our prediction further, we will add into consideration spatial issues and study the joint predictability, presented in the next section.

5.5 Investigation through Spatiotemporal Dynamics

In this section, we push our analysis further to the study of the joint predictability of mobile data traffic volumes and visited locations on a per-user basis. We investigate how predictable is the combination of *how much* traffic is generated by a mobile phone user and *where* this happens. Our analysis provides a comprehensive understanding of whether it is possible to

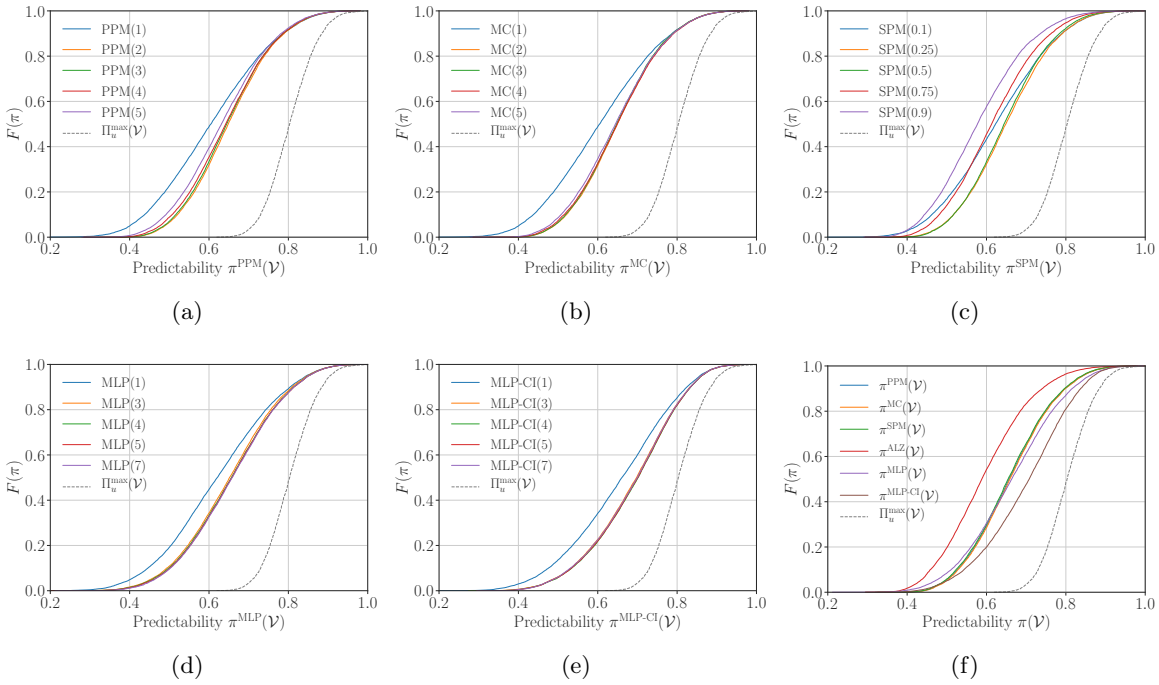


Figure 5.7: (a-e) Distributions of the prediction accuracy of each predictor with respect to its parameters. (f) Distribution of the practical predictability $\pi(\mathcal{V})$ of each predictor across the user set $\mathcal{U}_2$.

anticipate when, where, and how much mobile data traffic is generated by individual users. It is worth noting that, in this and the next sections, our analysis is based on the user set $\mathcal{U}_2$.

5.5.1 Theoretical Predictability

For each user $u \in \mathcal{U}_2$, we have the time series of data traffic volumes $v_1^T(u)$ and visited locations $\ell_1^T(u)$ during T one-hour time slots. Following the same methodology as in the previous section, we evaluate the theoretical predictability of forecasting data traffic volumes and locations at the same time. For that, we compute several necessary measures in the following.

The preliminary group of measures is related to the mobility of each user in isolation. In a similar way to we have done for mobile data traffic volumes, given a user's time series of locations $\ell_1^T(u)$, we calculate three entropy rate variants ($H_u^{\text{rand}}(\mathcal{L})$, $H_u^{\text{unc}}(\mathcal{L})$, and $H_u(\mathcal{L})$) and the corresponding bounds of the theoretical predictability. These measures are necessary for the comparison with the joint theoretical predictability that we focus in this section. We will explain in the next section the findings of these mobility-related measures from the viewpoint of pure mobility in the next section.

The major group of measures is computed by both the time series $v_1^T(u)$ and $\ell_1^T(u)$ of each user. Particularly, the data traffic process $\mathcal{V}$ and mobility process $\mathcal{L}$ are combined into a single joint process $\mathcal{M} = \{(V_i, L_i)\}$. Correspondingly, from a measurement data perspective, the time series are merged into $m_1^T(u) = \{(v_1^u, \ell_1^u), \cdots, (v_T^u, \ell_T^u)\}$. Then, the following variants of the entropy rate are calculated from $m_1^T(u)$, $\forall u \in \mathcal{U}_2$. *(i)* the *temporal-uncorrelated entropy* $H_u^{\text{unc}}(\mathcal{V}, \mathcal{L}) \equiv H_u^{\text{unc}}(\mathcal{M})$. It determines the heterogeneity deriving from simply considering the locations and data traffic volumes of each user together. *(ii)* the *joint actual entropy*

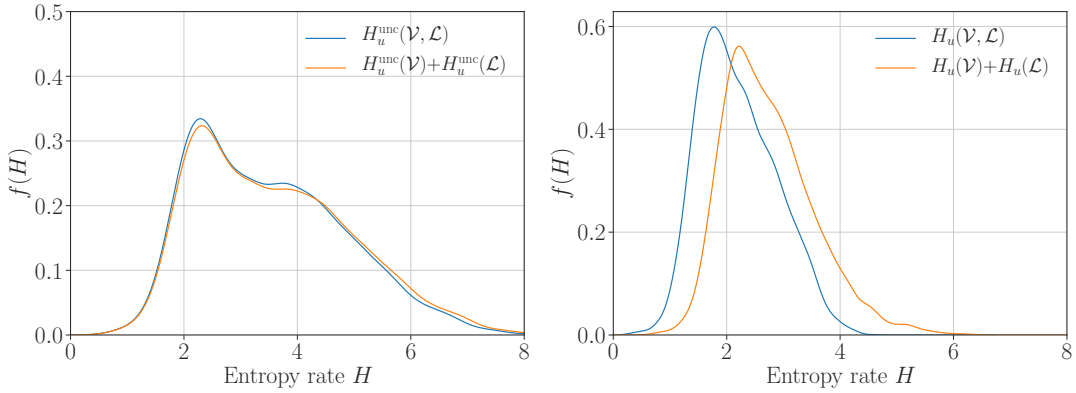


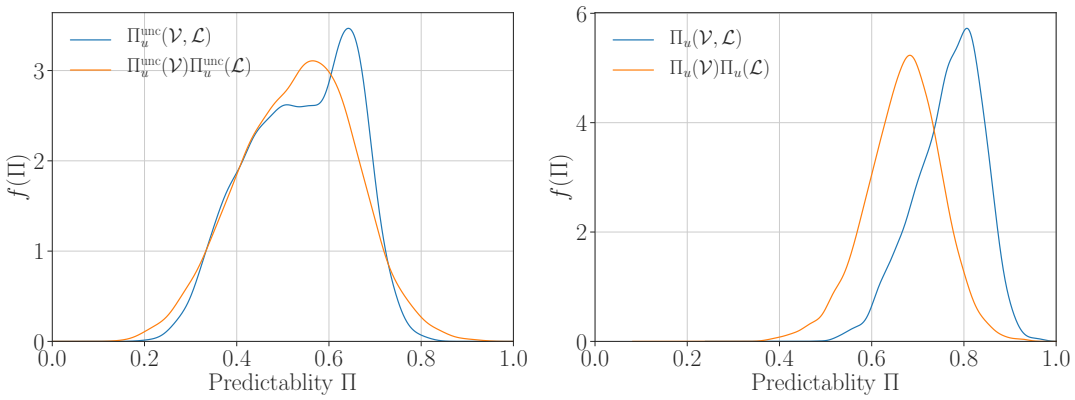
Figure 5.8: (a) Distributions of the different flavors of temporal-uncorrelated entropy variants: $H_u^{\text{unc}}(V,L)$, $H_u^{\text{unc}}(V)+H_u^{\text{unc}}(L)$ and $H_u^{\text{unc}}(V|L)$. (b) Distributions of the different flavors of entropy rates: $\mathcal{H}_u(\mathbb{V})$, $\mathcal{H}_u(\mathbb{V})+\mathcal{H}_u(\mathbb{V})$ and $\mathcal{H}_u(\mathbb{V}|\mathbb{L})$.

$H_u(\mathcal{V},\mathcal{L}) \equiv H_u(\mathcal{M})$. It is defined as the actual entropy rate of the joint stationary process $\mathcal{M}$ and expresses the combined uncertainty of a user's each current location and data traffic volume given his previous history of movements and mobile service usage. Correspondingly, we also compute the theoretical predictability upper bounds $\Pi_u^{\text{unc}}(\mathcal{V},\mathcal{L}) \equiv \Pi_u^{\text{unc}}(\mathcal{M})$ and $\Pi_u^{\text{max}}(\mathcal{V},\mathcal{L}) \equiv \Pi_u^{\text{max}}(\mathcal{M})$.

Our results are summarized in Figure 5.8 regarding the joint entropy rate variants and in Figure 5.9 regarding the theoretical predictability upper bounds. The figures portray the distributions for different ways of bringing together the two dimensions of data traffic volumes and locations. Each plot has two curves: *(i)* the joint entropy or associated predictability; *(ii)* the sum of entropy rates measured for data traffic volume and mobility separately, or associated predictability. The second curve represents the uncertainty (or theoretical predictability) in the case the stochastic processes driving mobility and traffic volume consumption are independent of each other. Figure 5.8(a) and 5.9(a) refer to temporal-uncorrelated versions, whereas Figure 5.8(b) and 5.9(b) concern our actual measures of interest.

A first interesting remark is that $H_u^{\text{unc}}(\mathcal{V},\mathcal{L})$ and $H_u^{\text{unc}}(\mathcal{V})+H_u^{\text{unc}}(\mathcal{L})$ in Figure 5.8(a), and consequently $\Pi_u^{\text{unc}}(\mathcal{V},\mathcal{L})$ and $\Pi_u^{\text{unc}}(\mathcal{V})\cdot\Pi_u^{\text{unc}}(\mathcal{L})$ in Figure 5.9(a), are nearly indistinguishable. Instead, $H_u(\mathcal{V},\mathcal{L})$ and $H_u(\mathcal{V})+H_u(\mathcal{L})$ in Figure 5.8(b), and consequently $\Pi_u^{\text{max}}(\mathcal{V},\mathcal{L})$ and $\Pi_u^{\text{max}}(\mathcal{L})\cdot\Pi_u^{\text{max}}(\mathcal{V})$ in Figure 5.9(b), show significant differences. Hence, there exists some correlation between the mobility and traffic volume consumption processes, and such correlation mainly emerges when considering – and it is thus driven by – the temporal ordering of events. As observed in Figure 5.9(b), a joint prediction of the next consumed amount of traffic and of the future location where this occurs can yield a better accuracy than forecasting the two separately, when knowledge of the previous actions of the individual is taken into account. The shift between $\Pi_u^{\text{max}}(\mathcal{L})\cdot\Pi_u^{\text{max}}(\mathcal{V})$ and $\Pi_u^{\text{max}}(\mathcal{V},\mathcal{L})$ is of 10% on average.

More importantly, we note that the mean value of $\Pi_u^{\text{max}}(\mathcal{V},\mathcal{L})$ is at 0.82. Therefore, our main conclusion is that it is possible to anticipate how much mobile data traffic (as an order of magnitude) will be consumed by a given user and where this will occur in a very effective manner (*i.e.*, with an 82% accuracy on average), by knowing the past history of activities of



(a) (b)

Figure 5.9: (c) Distributions of the predictability upper bounds $\Pi_u^{\text{unc}}(V,L)$, $\Pi_u^{\text{unc}}(L)\cdot\Pi_u^{\text{unc}}(L)$ and $\Pi_u^{\text{unc}}(V|L)$ based on the corresponding temporal-uncorrelated entropies. (d) Distributions of the predictability upper bounds $\Pi_u^{\max}(\mathbb{V},\mathbb{L})$, $\Pi_u^{\max}(\mathbb{V})\cdot\Pi_u^{\max}(\mathbb{L})$ and $\Pi_u^{\max}(\mathbb{V}|\mathbb{L})$ based on the corresponding entropy rates.

the target individual.

5.5.2 Practical Predictability

Now we check whether the theoretical predictability is consistent with the practical performance. Following the methodology presented in Section 5.4.2.1, we evaluate the practical predictability of forecasting each "current" data traffic volume and visited location jointly.

5.5.2.1 Excluding Less Frequent Locations

For this analysis, we preprocess the time series of locations. For each user, we keep the most frequent fifteen locations in his time series and merge the rest into one "fake" location marked as "other." Thus, each time series $\ell_1^T(u)$ has at most 16 unique locations. This is to reduce the size of the probability space which increases with the order k and the number of unique locations in our predictors (*e.g.*, an MC(k) predictor needs to build $(N_\ell * N_v)^{k+1}$ probabilities in its transition matrix). We choose the threshold $k = 15$ due to our observation in Figure 5.10(a), where we see that the top 15 locations can occupy 95% of the time slots in the observing period. In this and next sections, we always employ this top-15 version of the time series $\ell_1^T(u)$ of our users instead of the original ones.

5.5.2.2 Prediction Results

Our evaluation on the joint practical predictability is still based on the predictors previously used and presented in Section 5.4.2.2, *i.e.*, the `PPM`, `MC`, `SPM`, `ALZ`, `MLP`, and `MLP-CI` predictors. Recall that we have for each user $u \in \mathcal{U}_2$ a mixed time series $m_1^T(u)$ consisting of $v_1^T(u)$ and $\ell_1^T(u)$. Based on these mixed time series, we proceed as follows.

First, we perform the procedure presented in Section 5.4.2.1: we initialize our predictors by the partial mixed time series $m_1^T(u)$ of the first 100 days and then predict the (volume,location) pairs in the rest of the mixed time series. In this case, the joint practical predictability

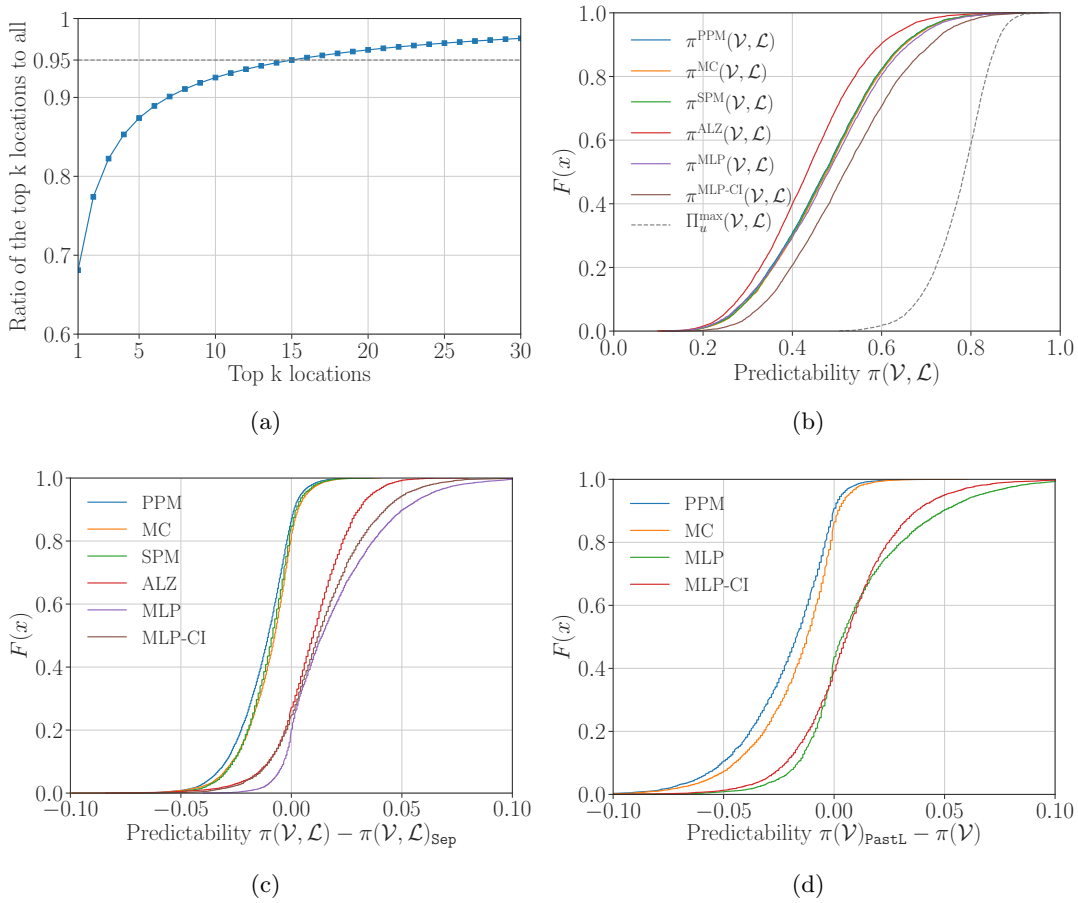


Figure 5.10: (a) Ratio of the top k locations to all the locations in the time series of locations owned by the user set $\mathcal{U}_2$. (b-d) Distributions of (b) the joint practical predictability $\pi(\mathcal{V}, \mathcal{L})$ across the user set $\mathcal{U}_2$, (c) the enhancement on the practical predictability $\pi(\mathcal{V}, \mathcal{L})$ by forecasting data traffic volumes and visited locations jointly compared with doing separately, and (d) the enhancement on the practical predictability $\pi(\mathcal{V})$ of each predictor by adding the information of the historical visited locations as a prior knowledge.

$\pi(\mathcal{V}, \mathcal{L})$ that we compute matches to the same joint forecasting scenario as the joint theoretical predictability $\Pi_u^{\max}(\mathcal{V}, \mathcal{L})$ above. Particularly, for the mixed time series, a success prediction of a time slot has to be correct in both the data traffic volume and visited location of that time slot. Our results with respect to the joint practical predictability $\pi(\mathcal{V}, \mathcal{L})$ are shown in Figure 5.10(b), which we observe the following.

- Among the predictors that only leverage the historical temporal orders of the mixed time series of each user, the `MLP` predictor performs the best with a quite little advantage over the others, while the `ALZ` does the worst. The other three Markovian predictors have almost the same performance.

- The `MLP-CI` predictor achieves the overall highest joint practical predictability; it has 50% of the average prediction accuracy to correctly forecast the data traffic volume and location at each time slot simultaneously. As in the forecast of data traffic volumes in

isolation, the improvement of the MLP-CI compared with MLP comes from the temporal contextual information.

- Still, there is a larger gap between the theoretical and practical predictabilities in the joint forecasting scenario than only predicting data traffic volumes. Even the performance of the best predictor is still far (*i.e.*, 30% on average) from the theoretical upper bound.

For comparison, we also evaluate the joint practical predictability *with the separate forecasting scenario*, represented by $\pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$. For that, we employ our predictors to predict the data traffic volumes and locations separately. Then we combine the predicted volume and location of each time slot into a prediction of the mixed time series. In this case, the maximum value of $\pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$ is limited by $\Pi_u^{\max}(\mathcal{V}) * \Pi_u^{\max}(\mathcal{L})$ in the theoretical predictability analysis. Compared with the joint forecasting scenario, we show the improvement of each predictor by computing $\pi(\mathcal{V},\mathcal{L}) - \pi(\mathcal{V},\mathcal{L})_{\texttt{Sep}}$ in Figure 5.10(c), from which we can clearly see the following.

- Regarding the Markovian predictors, only the ALZ can benefit from predicting the two behaviors jointly: almost 80% of the users have their performance improved up to 8%, while the problem is that this predictor performs the worst compared with the others. For the rest three Markovian predictors, only 20% of the users have better performance up to 3% of the improvement. A possible reason for this is that the number of samples is insufficient as the joint forecast scenario enlarges the probability space significantly.

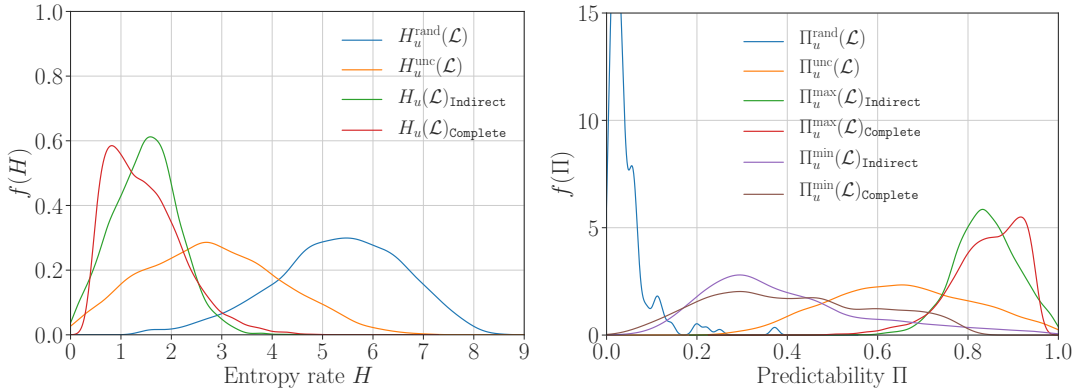
- The MLP and MLP-CI predictors are obviously enhanced by forecasting volumes and locations jointly: for 80% of the users, the improvement is at most 10% of the prediction accuracy. The practical predictability of these two predictors is consistent with the theoretical predictability estimated by the joint entropy rate.

The last thing we evaluate in this section, is the practical predictability of data traffic volumes by knowing the previous locations of the same user, which we mark as $\pi(\mathcal{V})_{\texttt{PastL}}$. We portray in Figure 5.10(d) the CDF of the enhancement (*i.e.*, $\pi(\mathcal{V})_{\texttt{PastL}} - \pi(\mathcal{V})$) of each predictor by adding this mobility information as an additional knowledge in the prediction. As in the joint forecasting scenario shown in Figure 5.10(c), the Markovian predictors cannot benefit from this additional information (only 10% of the users have improvements), while the MLP predictor can have upto 10% of the improvement for 60% of the users. Although the rest of the users have reduced performance, we guess that their performance may be also improved by adding more data as historical information in prediction. Still, we can say that the predictors based on machine learning can utilize the mobility information efficiently in the prediction of mobile data traffic.

Consequently, our results regarding the joint practical predictability measured above are consistent with those of the joint theoretical predictability. We see that forecasting jointly the data traffic volumes and locations performs better than doing so separately, due to the spatiotemporal correlation of mobile data traffic, as discussed in Section 5.2. The problem is, to have such benefit in real-world prediction, we need machine learning techniques to utilize the spatiotemporal correlation better, while legacy Markovian methods are insufficient.

5.6 Additional Investigation of Human Mobility

So far, we have carried out a detailed analysis of the predictability of per-user mobile data traffic. To perform the joint predictability analysis presented in the last section, we also derive



(a) (b)

Figure 5.11: (a) Distributions of the random entropy $H_u^{\text{rand}}(\mathcal{L})$, the temporal-uncorrelated entropy $H_u^{\text{unc}}(\mathcal{L})$, and the entropy rate $\mathcal{H}_u(\mathcal{L})_{\texttt{Complete}}$, $\mathcal{H}_u(\mathcal{L})_{\texttt{Indirect}}$ of each user u of the user set $\mathcal{U}_2$. (b) Distributions of the corresponding upper and lower bounds on the theoretical predictability $\Pi_u^{\text{rand}}(\mathcal{L})$, $\Pi_u^{\text{unc}}(\mathcal{L})$, $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Indirect}}$, $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Complete}}$, $\Pi_u^{\text{min}}(\mathcal{L})_{\texttt{Indirect}}$, and $\Pi_u^{\text{min}}(\mathcal{L})_{\texttt{Complete}}$ of each user $u \in \mathcal{U}_2$.

the theoretical predictability of mobility in isolation for each user $u \in \mathcal{U}_2$. In this section, we present the corresponding results in detail and extend our work to the practical predictability of human mobility.

Although there is a far larger body of the literature on the predictability of human mobility than the one of mobile data traffic [23], we still find two unique aspects to perform our analysis. First, we are going to validate the most popular finding of the theoretical predictability of per-user mobility performed by Song *et al.* [17, 108]. To the best of our knowledge, there is no validation using the CDR dataset at the same scale. Second, we study the prediction of locations of a user using his data traffic volumes as contextual information, inspired by the fact that knowing locations can contribute to predicting data traffic volumes.

5.6.1 Theoretical Predictability

We re-run the exact same study of Song *et al.* [17, 108] on the CDR owned by the user set $\mathcal{U}_2$ extracted from our `CDR` dataset. We compute for each user u the actual entropy $H_u(\mathcal{L})$ along with the other two entropy variants $H_u^{\text{unc}}(\mathcal{L})$ and $H_u^{\text{rand}}(\mathcal{L})$, as well as their corresponding theoretical predictability upper bounds, *i.e.*, $\Pi_u^{\text{max}}(\mathcal{L})$, $\Pi_u^{\text{unc}}(\mathcal{L})$, and $\Pi_u^{\text{rand}}(\mathcal{L})$. In addition, we also evaluate the lower bound of the theoretical predictability $\Pi_u^{\text{min}}(\mathcal{L})$. See Table 5.2 for all the variants.

The actual entropy $H_u(\mathcal{L})$ needs some explanation. It is impossible to obtain the actual value of $H_u(\mathcal{L})$ because some original locations are always missing in the CDR. Instead, we compute two alternatives of $H_u(\mathcal{L})$:

- $H_u(\mathcal{L})_{\texttt{Complete}}$ is computed from our completed time series of locations by the estimator defined in Equation (5.6);
- $H_u(\mathcal{L})_{\texttt{Indirect}}$ is calculated from the original incomplete time series by the indirect estimation algorithm proposed by Song *et al.* [17, 108]. This algorithm assumes that all the

missing locations belong to the same cell tower marked. On each iteration, it randomly replaces some locations by this cell tower and computes the estimated entropy rate. Finally, the algorithm builds a regression function between the ratio of missing locations and the estimated entropy rate, and estimates the actual entropy rate $H_u(\mathcal{L})_{\texttt{Indirect}}$ from the regression function by setting the missing ratio to zero.

Now we discuss the results corresponding to the theoretical predictability of mobility. We portray in Figure 5.11(a) the PDF of the entropy rate variants $H_u^{\text{rand}}(\mathcal{L})$, $H_u^{\text{unc}}(\mathcal{L})$, $H_u(\mathcal{L})_{\texttt{Indirect}}$ and $H_u(\mathcal{L})_{\texttt{Complete}}$. The variants corresponding to the results of Song *et al.* are $H_u^{\text{rand}}(\mathcal{L})$, $H_u^{\text{unc}}(\mathcal{L})$, $H_u(\mathcal{L})_{\texttt{Indirect}}$. Their distributions are bell-shaped and have statistical measures that are highly close to those found in [17, Fig.2-A] except for some minor shifts between their distributions and ours. Particularly, we can observe the following.

- $H_u^{\text{rand}}(\mathcal{L})$ has a mean of approximately 5 bit instead of 6 bit in [17]. It indicates that in our data, a user's entire movement space is composed of $2^{H_u^{\text{rand}}(\mathcal{L})} \approx 32$ cells on average, which is slightly smaller than that in [17].

- $H_u^{\text{unc}}(\mathcal{L})$ still has a significant variance as in [17], meaning that the geographical span of movements varies widely from person to person in our dataset. It has a peak at 2.8 bit instead of 3.5 bit as in [17]. This indicates that, users tend to favor some locations over others, and keeping this into consideration allows making more accurate forecasts on the next location they will visit: the uncertainty shrinks to just $2^{H_u^{\text{unc}}(\mathcal{L})} \approx 7$ locations on average.

- Most importantly, the peak of $H_u(\mathcal{L})_{\texttt{Indirect}}$ (*i.e.*, at 1 bit) is nearly equivalent to that in [17], which implies that the uncertainty in anticipating the next cell of a user is limited to $2^{H_u(\mathcal{L})} \approx 2$ options; this is consistent with what is reported in [17].

Consequently, our results are consistent with the findings of Song *et al.* [17], while our users have less visited locations and their mobility are slightly more certain in general.

The novel variant in Figure 5.11(a) is our version of the actual entropy $H_u(\mathcal{L})_{\texttt{Complete}}$. It has a peak at 0.8 bit which is less than $H_u(\mathcal{L})_{\texttt{Indirect}}$. This means that the next location of a user in our completed time series is more certain. We ascribe this result to the fact that we complete the original CDR data and our CDR completion approach enhances the probabilities of the appearance of some locations (*e.g.*, home) over the others.

With respect to the theoretical predictability of per-user mobility, we plot in Figure 5.11(b) the upper and lower bounds that are computed from the entropy variants above. Similarly, the distributions of $\Pi_u^{\text{rand}}(\mathcal{L})$, $\Pi_u^{\text{unc}}(\mathcal{L})$, and $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Indirect}}$ are consistent with their counterparts appeared in [17, Fig.2-B]. The slight differences between theirs and ours are caused by the shifts of the corresponding entropy variants, which thus we do not discuss in detail. The most important two variants are $\Pi_u^{\text{unc}}(\mathcal{L})$ and $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Indirect}}$. The former varies widely and peaks at 0.6, indicating that forecasting the next location from the historical probability of each location's appearance can only achieve up to the prediction accuracy of 60% on average. The latter, $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Indirect}}$, is the upper bound of the theoretical predictability of mobility. It peaks at 0.83, meaning that one can expect the prediction accuracy of 83% average, when utilizing the temporal orders of a user's historical visiting pattern into the prediction. Our estimation of this upper bound, *i.e.*, $\Pi_u^{\text{max}}(\mathcal{L})_{\texttt{Complete}}$, is higher, indicating that the forecast of each next location is slightly easier than the estimation applied on the original data. Besides,

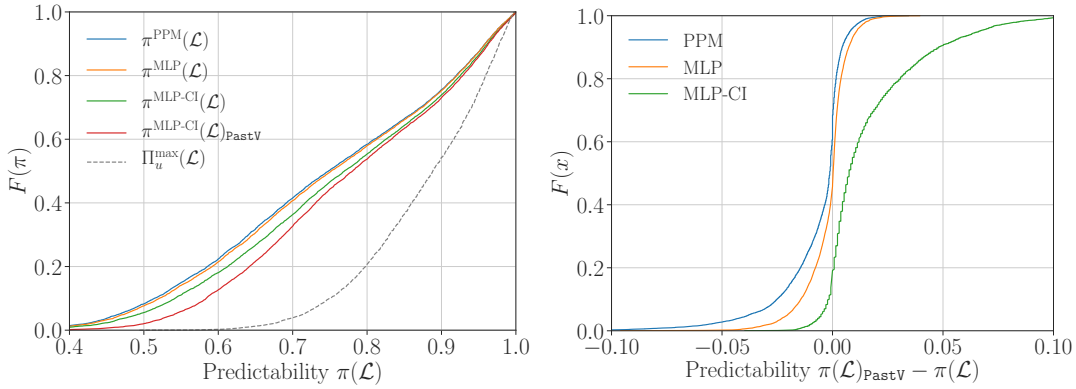


Figure 5.12: Distributions of (a) the practical predictability $\pi(\mathcal{L})$ of each user corresponding to the `PPM`, `MLP`, and `MLP-CI` predictors, and (b) the enhancement on the practical predictability $\pi(\mathcal{L})$ of each predictor by adding the information of the historical data traffic volumes as a prior knowledge.

we have two versions of the lower bound of the theoretical predictability, $\Pi_u^{\min}(\mathcal{L})_{\texttt{Indirect}}$ and $\Pi_u^{\min}(\mathcal{L})_{\texttt{Complete}}$. Their distributions are highly similar and both peak at around 0.3. This common peak indicates that any predictor that uses the MAP (maximum a posteriori) probability estimation on historical outcomes (such as `MC` or `PPM`) should achieve at least the prediction accuracy of 30% on average [65]. Compared with the lower limit estimated by $\Pi_u^{\text{rand}}(\mathcal{L})$ (peaking at 0.05) used in [17], this lower bound is more reasonable.

In summary, our results validate the nature about human mobility originally presented in [17], *i.e.*, a non-trivial level of predictability, encoded in temporal orders of a user's visiting patterns, is across the population. On the other hand, these results lead to a slightly higher upper bound over a majority of the observed users and a notable higher upper bound than the original expectation in [17].

5.6.2 Practical Predictability

We evaluate the practical predictability of visited locations of each user via the same predictors (*cf.* Section 5.4.2.2) and methodology (*cf.* Section 5.4.2.1) that we employ in the analysis of data traffic volumes. Differently, we replace the data traffic volumes by the locations in the input and output of each predictor. Accordingly, the practical predictability is marked as $\pi(\mathcal{L})$. After that, each predictor is given a new setting: we add the previous data traffic volumes into its input. For the Markovian predictors, it means to build their transition matrices from both locations and volumes and extract the probabilities of the current location. We refer the practical predictability of the predictors with this new setting as $\pi(\mathcal{L})_{\texttt{PastV}}$. Recall that we still use the time series of the top 15 locations and "other" to ensure a fairly small size of the probability space for the Markovian predictors (*cf.* Section 5.5.2.1).

We plot our results concerning the practical predictability of mobility in Figure 5.12, which includes the distributions of the practical predictability and its enhancement by knowing the past data traffic volumes of each predictor. It is worth noting that, although we proceed the corresponding analysis, we do not illustrate the results of the `ALZ`, `MC`, and `SPM` predictors

in Figure 5.12. For the ALZ predictor, it does receive a huge relative improvement from the historical data traffic volumes, but even its improved accuracy is still far worse than the others. For the MC and SPM predictors, their performance is highly close to that of the PPM predictor. Therefore, we only show the latter in Figure 5.12 as the representative of all the Markovian predictors.

Figure 5.12(a) shows the CDF of the practical predictability of the PPM, MLP, and MLP-CI predictors as well as the improved-by-volume version of the last predictor. Although the theoretical predictability of per-user mobility shows an 85% of the maximum expected accuracy on average, the predictors that only leveraging the temporal orders of historical visiting patterns have their limit on the practical predictability. The Markovian predictors can only achieve 73% of the average practical predictability. Even the MLP predictor, which shows a fairly good performance in the prediction of data traffic volumes, only performs slightly better (1-2%) than the Markovian ones. Approximately 76% of the average practical predictability is achieved by the MLP-CI predictor which further leverages the time as the context information. The best performance is achieved by the MLP-CI predictor with the knowledge of previous data traffic volumes, which has 79% of the practical predictability.

We then focus on the practical predictability of each user in detail. Figure 5.12(b) plots the CDF of the enhancement on the practical predictability of each user brought by the use of historical data traffic volumes in the prediction. We note that for the PPM and MLP predictors, only less than 50% of the users have such enhancement up to 5%, while the practical predictability of the results even decreases at most 10% surprisingly. According to the nature of the Markovian predictors, we believe that the main reason is that considering volumes into the prediction significantly increases the size of the probability space, while the number of our samples is not enough. Since we have used time series of 100 days in each prediction, adding more locations and volumes will not be a practical solution in this case. For the MLP-CI predictors, 80% of the users can receive their enhancement from the previous data traffic consumption information by up to 10%. It indicates that, the context information, such as time and data traffic consumption, do have the capability of achieving a better prediction of a user's locations, while only the machine learning techniques can absorb and utilize such information efficiently.

5.7 Summary

In this chapter, we analyze the predictability of per-user mobile data traffic, in isolation and jointly with mobility, using our datasets introduced in Chapter 3 and our CDR completion approaches presented in Chapter 4. Our data-driven analysis exploits the theoretical and practical performance in the prediction. In short, we conclude that there is a high degree of the predictability of individual mobile data traffic and it can be further enhanced by the knowledge of users' mobility. The main reason of the high predictability is the existence of the spatiotemporal correlation in each user's mobile data traffic dynamics. In real-world prediction, the legacy Markovian approach could achieve a fairly good prediction accuracy, while the key to further enhance the performance is the use of context information (*e.g.*, time of events or locations). For that, the novel machine learning techniques are quite useful. Nevertheless, there is still a gap between the real-world prediction accuracy and the theoretical accuracy upper bound.

Chapter 6

Conclusion and Outlook

With this thesis, we provide the technical foundation on two research perspectives (*i.e.*, the utilization of operator-collected datasets and the prediction of per-user mobile data traffic) by addressing the central research questions of each of them. In this chapter, we first summarize the insights and contributions with our answer to these research questions. After that, we outline the limitations and future research directions based on our results. Finally, this thesis finishes with the last concluding remarks.

6.1 Summary of the Thesis

First, we focused on the utilization of operator-collected datasets. It was our secondary objective but had to be addressed in advance in order to support the achievement of our primary objective (*cf.* Chapter 1). Regarding this topic, we addressed three central research questions.

Whether and to what extent does the use of CDR data affect human mobility studies? We answered this question by applying several data-driven studies on our datasets (*cf.* Chapter 3). We studied the biases of the essential human mobility characterizing features with respect to the spatial resolution and the temporal heterogeneity and sparsity of typical CDR datasets. Our results revealed that: *(i)* the limited spatial resolution caused the distance shift around $200-500$ meters from the users' CDR-logged locations to their actual positions in metropolitan areas; *(ii)* the short-term CDR collection had the limited capacity of capturing human mobility information while the long-term CDR gathering could yield enough details to mark and rank overall significant locations and to model the population distribution of the human movement span.

What degree of (in)completeness can one expect in CDR-based trajectories inferred from real-world large-scale mobile phone datasets? We evaluated the (in)completeness by mining the nationwide large-scale `CDR` dataset that consists of CDR of the most common type. As presented in Chapter 3, we highlighted that common CDR hardly captured fully complete trajectories, however, under the hourly temporal resolution, the majority of the CDR-based trajectories could provide the completion around $10\% - 80\%$. These valuable resources might be rejected and filtered out by the legacy data preliminary techniques. To utilize these non-negligible trajectories, we developed the CDR completion techniques and addressed the next research question. Besides, we found several long-tail population models that could fit the distribution of the (in)completeness. They opened opportunities to characterize the trajectory completeness in a CDR dataset from only the length of the observing period and the target temporal resolution without carrying out a time-consuming computing task to mine all the records.

How efficient can the CDR completion fill spatiotemporal gaps in CDR-based trajectories? Building upon the state-of-the-art studies and our insights presented in Chapter 3, we formalized the CDR-based trajectory and proposed several completion approaches to infer human mobility information and to reconstruct CDR-based trajectories in Chapter 4. The data-driven simulations showed that our proposed approaches could achieve a far increased completion and retain essential spatial accuracy compared with the typical proposals in the literature. The approaches offered useful tools for the utilization of CDR and more importantly, served as the foundation to deal with our primary objective.

Second, based on the techniques that we developed for the operator-collected dataset utilization, we addressed the prediction of per-user mobile data traffic, which was our primary objective (*cf.* Chapter 1). We aimed at answering two central research questions.

How predictable is the individual consumption of mobile data traffic? To answer this question, we evaluated the theoretical and practical predictability of individual mobile data traffic. First, we provided a first study of the theoretical predictability and derived a promising upper bound (*i.e.*, 81% on average) to the performance of practical algorithms for the prediction of individual consumption of mobile data traffic. Second, we evaluated the state-of-the-art prediction techniques and revealed that the average performance of the Markovian prediction techniques could achieve up to 16% below the theoretical predictability while the machine learning techniques could reach 11% close to the upper bound. The reason is that the former requires the Markov property which actual variations does not always have, while the latter utilize artificial neural networks that have more flexibility on the recognition of patterns and regularities. Still, there was a performance gap between the theoretical and practical predictability.

Whether can and how does the information of individual mobility contribute to the prediction of mobile data traffic? We also answered this question through the predictability by applying an extensive data-driven analysis. Our results revealed that forecasting per-user mobility and mobile data traffic jointly can have 10% of the enhancement on the theoretical predictability thanks to the strong correlation, while in real-world prediction, the Markovian prediction techniques could hardly achieved improved accuracy while the machine learning techniques had a $1\% - 5\%$ degree of accuracy improvement varying across the users. Surprisingly, our analysis showed that knowing mobile data traffic of a user significantly helped the prediction of his whereabouts for 80% of the users, leading to an improvement up to 10% regarding accuracy. Consequently, although the spatiotemporal information could further improve the design of prediction model, the gain was not dramatic with respect to a technique that only relied on temporal information. Our prediction analysis concluded the feasibility of anticipating how much mobile data traffic would be consumed by a given subscriber and where this would occur in a very effective manner.

6.2 Limitation and Outlook

In this section, we discuss the limits and potential research perspectives in the context of our analytical insights and technical contributions on the two topics proposed in this thesis.

6.2.1 Operator-collected Dataset Utilization

The data gathering by mobile network operators remains a very hot enabler of multiple research disciplines such as human behavior understanding, mobile communication, and machine learning. Future research directions are listed in the following.

Human Mobility Reconstruction Using Mixed Data Building upon the long-term CDR collection of voice calls, we propose in this thesis the approaches to reconstruct human mobility, while they are designed to learn mobility information from a single dataset. A natural progression is to study the mobility reconstruction using multiple datasets. Although new datasets are being released to research communities, which have higher spatial and temporal resolutions than standard CDR datasets, their observing periods are still short due to the primary concern. Therefore, one can reasonably expect having coarse-grained long-term mobility information of users and finer-grained short-term mobility information of the same users only in some partial observing periods. It would be interesting to assess the long-term mobility reconstruction problem using such mixed information.

Deep Utilization of Contextual information Our CDR completion approaches leverage the repetitive human mobility patterns and are powered by the context, including the visiting importance of locations and visiting regularity of cell towers. More research is needed to utilize the available contextual information in a deeper level. Two examples are listed in the following.

- **Environmental Context.** The POIs (places of interest) of a cell tower location, rather than its appearance or importance, are more essential factors to forge a user's mobility: a user's repetitive pattern of visited locations comes from his regular interests or habits to visit some POIs. Thus, future mobility reconstruction research might characterize these POIs and develop techniques that first build an upper-level profile about the POIs of a missing location and then infer that location using both the profile and repetitive pattern.

- **Population Context.** The reconstruction of a user's trajectories may benefit from knowing similar trajectories of other users. Jeong *et al.* [77] make an effort on this aspect and propose a novel technique to recover the visiting sequence of a user using the location displacements of his similar users. However, a new challenge arises when we deal with slotted CDR-based trajectories which are strict periodical location sequences with time slots. It is shown that raw slotted CDR-based trajectories are not only incomplete but also highly unique: users may share some locations while they do not appear at these locations in the same time slots [144]. Future trails on the mobility reconstruction with the population context should assess both the uniqueness among trajectories and the required temporal resolution, which remains a challenging task.

6.2.2 Per-user Mobile Data Traffic Prediction

With our quantitative study on the prediction of per-user mobile data traffic, we make an essential step towards the understanding of to what degree the prediction accuracy can be expected, which prediction techniques are effective, and how their parameters affect the prediction quality. Here we outline future research perspectives that can begin with our results.

Data Content Estimation Due to the limits of our datasets, our prediction analysis addresses time-slotted data volumes. Nevertheless, in each time slot, the data volume is the sum of data traffic of multiple content types. We may update existing prediction techniques, provided that a deeper understanding of mobile data traffic, such as future data content types and volumes, is available. We believe that it will be an important topic in future research.

Population Prediction Models An issue that is not addressed in this thesis is to measure and utilize the similarity of the data traffic usage among different users. Our prediction techniques are built upon individual prediction models while population prediction models can use the data of multiple users and may help to solve the cold-start problem and provide better prediction accuracy. Therefore, it remains an open research topic to understand how to derive a population prediction model with user collaboration and to identify users whose mobile data traffic can be predicted together.

Service Integration of Mobility and Mobile Data Traffic Predictions The limitation of our prediction analysis is that we only focus on the pure prediction of locations or data traffic volumes while we do not consider to integrate prediction methods into practical services. For instance, the dynamic data offloading requires a deep understanding of data traffic, content, and user mobility in mobile networks. It is important to know how much data traffic will appear and which content types can be safely offloaded to alternative communication channels. In this case, the prediction of both aggregated and individual mobile data traffic may be leveraged. Therefore, future prediction research would combine with real-world application scenarios.

Real-world Deployment of Prediction Models on Mobile Devices Deploying the prediction techniques in a real-world application scenario is a natural consequent step. Our results regarding the predictability can be applied on mobile devices: an appropriate prediction method can support those applications that need future knowledge of mobility and mobile data demand. The challenge on the implementation of prediction models comes from the limited computing power and energy capacity of mobile devices, which need to be investigated in future research.

6.3 Concluding Remarks

Our contributions underline the efficient utilization of operator-collected datasets and the prediction capacity of per-user mobile data traffic. Our CDR completion techniques that leverage the human mobility pattern and the environmental context can reconstruct CDR-based trajectories with fairly good temporal completion and spatial accuracy and thus, along with our pioneer experience, support the utilization of human footprints in operator-collected datasets effectively. Our insights and contributions in the prediction analysis enhance the understanding of mobile data traffic and support both mobile network operators and customers. Our results with respect to the theoretical and practical predictabilities provide opportunities for mobile applications or mobile network operators to build their services on the top of them in a more effective manner. Therefore, we conclude that our insights and technical contributions will support both the mobility-related analyses and the designs and implementations that rely on mobile data traffic prediction.

Bibliography

[1] G. Goggin, *Cell Phone Culture*. Routledge, Sept. 2006. (Cited on page 1.)

[2] C. V. N. Index, "Global mobile data traffic forecast update, 2016-2021." https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2013. (Cited on page 1.)

[3] W. Su, S.-J. Lee, and M. Gerla, "Mobility prediction in wireless networks," in *MILCOM 2000 Proceedings. 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No.00CH37155)*, vol. 1, pp. 491–495, IEEE, IEEE, 2000. (Cited on page 1.)

[4] P. N. Pathirana, A. V. Savkin, and S. Jha, "Mobility modelling and trajectory prediction for cellular networks with mobile base stations," in *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing - MobiHoc '03*, pp. 213–221, ACM, ACM, 2003. (Cited on page 1.)

[5] W.-S. Soh and H. S. Kim, "QoS provisioning in cellular networks based on mobility prediction techniques," *IEEE Communications Magazine*, vol. 41, pp. 86–92, Jan. 2003. (Cited on pages 1 and 2.)

[6] P. Baier, F. Durr, and K. Rothermel, "TOMP: Opportunistic traffic offloading using movement predictions," in *37th Annual IEEE Conference on Local Computer Networks*, pp. 50–58, IEEE, IEEE, Oct. 2012. (Cited on page 1.)

[7] V. A. Siris and D. Kalyvas, "Enhancing mobile data offloading with mobility prediction and prefetching," in *Proceedings of the seventh ACM international workshop on Mobility in the evolving internet architecture*, pp. 17–22, ACM, 2012. (Cited on pages 1 and 2.)

[8] H. Petander, "Energy-aware network selection using traffic estimation," in *Proceedings of the 1st ACM workshop on Mobile internet through cellular networks - MICNET '09*, pp. 55–60, ACM, ACM, 2009. (Cited on pages 1 and 2.)

[9] X. Zhuo, W. Gao, G. Cao, and S. Hua, "An incentive framework for cellular traffic offloading," *IEEE transactions on mobile computing*, vol. 13, pp. 541–555, Mar. 2014. (Cited on page 1.)

[10] L. Wang, D. Zhang, and H. Xiong, "effSense," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication - UbiComp '13 Adjunct*, pp. 1075–1086, ACM, ACM, 2013. (Cited on page 1.)

[11] R. Zhu, B. Liu, D. Niu, Z. Li, and H. V. Zhao, "Network latency estimation for personal devices: A matrix completion approach," *IEEE/ACM transactions on networking (TON)*, vol. 25, pp. 724–737, Apr. 2017. (Cited on pages 1, 2 and 10.)

[12] Z. Li, J. Bi, and S. Chen, "Traffic prediction-based fast rerouting algorithm for wireless multimedia sensor networks," *International Journal of Distributed Sensor Networks*, vol. 9, p. 176293, Jan. 2013. (Cited on pages 1 and 2.)

[13] Y. Xu, J. Winter, and W.-C. Lee, "Prediction-based strategies for energy saving in object tracking sensor networks," in *IEEE International Conference on Mobile Data Management, 2004. Proceedings. 2004*, pp. 346–357, IEEE, IEEE, 2004. (Cited on page 1.)

[14] D. G. Taylor and M. Levin, "Predicting mobile app usage for purchasing and information-sharing," *Intl J of Retail & Distrib Mgt*, vol. 42, pp. 759–774, Aug. 2014. (Cited on page 1.)

[15] C. Zhang, X. Ding, G. Chen, K. Huang, X. Ma, and B. Yan, "Nihao: A predictive smartphone application launcher," in *International Conference on Mobile Computing, Applications, and Services*, pp. 294–313, Springer, 2012. (Cited on page 1.)

[16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, June 2008. (Cited on pages 1, 2, 3, 16, 17, 18, 35, 39, 42, 57, 58, 59, 61 and 83.)

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, Feb. 2010. (Cited on pages 1, 2, 3, 4, 11, 16, 17, 18, 19, 39, 42, 57, 58, 70, 71, 82, 92, 93 and 94.)

[18] M. W. Horner and M. E. O'Kelly, "Embedding economies of scale concepts for hub network design," *Journal of Transport Geography*, vol. 9, pp. 255–265, Dec. 2001. (Cited on page 1.)

[19] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas," *Scientific Reports*, vol. 2, p. 1001, Dec. 2012. (Cited on page 1.)

[20] L. Pappalardo, D. Pedreschi, Z. Smoreda, and F. Giannotti, "Using big data to study the link between human mobility and socio-economic development," in *2015 IEEE International Conference on Big Data (Big Data)*, pp. 871–878, IEEE, IEEE, Oct. 2015. (Cited on page 1.)

[21] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, "A theory of power-law distributions in financial market fluctuations," *Nature*, vol. 423, pp. 267–270, May 2003. (Cited on page 1.)

[22] F. Rebecchi, M. Dias de Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti, "Data offloading techniques in cellular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 580–603, 2015. (Cited on page 1.)

[23] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2016. (Cited on pages 1, 2, 3, 7, 8, 13, 14, 16, 17, 21, 41, 42, 74 and 92.)

[24] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, May 2014. (Cited on pages 2, 3, 4, 15, 16, 18, 19, 39, 58 and 83.)

[25] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proceedings of the 18th international conference on World wide web - WWW '09*, pp. 791–800, ACM, ACM, 2009. (Cited on pages 2 and 27.)

[26] P. Baumann, W. Kleiminger, and S. Santini, "How long are you staying?," in *Proceedings of the 19th annual international conference on Mobile computing & networking - MobiCom '13*, pp. 231–234, ACM, ACM, 2013. (Cited on pages 2 and 17.)

[27] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *2011 Proceedings IEEE INFOCOM*, pp. 882–890, Apr. 2011. (Cited on pages 2, 7, 8, 9, 10, 14, 24, 35, 42, 73, 74, 76 and 83.)

[28] S. Ostring and H. Sirisena, "The influence of long-range dependence on traffic prediction," in *ICC 2001. IEEE International Conference on Communications. Conference Record (Cat. No.01CH37240)*, vol. 4, pp. 1000–1005, IEEE, IEEE, 2001. (Cited on page 2.)

[29] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM transactions on networking (TON)*, vol. 5, pp. 835–846, Dec. 1997. (Cited on page 2.)

[30] A. Sang and S.-q. Li, "A predictability analysis of network traffic," *Computer Networks*, vol. 39, pp. 329–345, July 2002. (Cited on page 2.)

[31] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of cellular networks traffic," in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, pp. 973–978, IEEE, Oct. 2012. (Cited on pages 2, 8, 11 and 17.)

[32] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," *IEEE Communications Magazine*, vol. 52, pp. 234–240, June 2014. (Cited on pages 2 and 8.)

[33] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems - SIGMETRICS '11*, pp. 305–316, ACM, ACM, 2011. (Cited on pages 2, 7, 8, 9, 13, 85 and 86.)

[34] E. Mucelli Rezende Oliveira, A. Carneiro Viana, K. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Computer Networks*, vol. 112, pp. 176–193, Jan. 2017. (Cited on pages 2, 9, 10, 73, 74 and 82.)

[35] H.-H. Jo, M. Karsai, J. Karikoski, and K. Kaski, "Spatiotemporal correlations of handset-based service usages," *EPJ Data Science*, vol. 1, pp. 1–18, Nov. 2012. (Cited on pages 2, 4, 9, 10, 19, 42, 44 and 83.)

[36] A. Nika, A. Ismail, B. Y. Zhao, S. Gaito, G. P. Rossi, and H. Zheng, "Understanding and predicting data hotspots in cellular networks," *Mobile Network Application*, vol. 21, pp. 402–413, Oct. 2015. (Cited on pages 2 and 9.)

[37] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, "Big data driven mobile traffic understanding and forecasting: A time series approach," *IEEE transactions on services computing*, vol. 9, pp. 796–805, Sept. 2016. (Cited on pages 2, 7, 8 and 9.)

[38] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM transactions on networking (TON)*, vol. 25, pp. 1147–1161, Apr. 2017. (Cited on pages 2, 7, 8 and 9.)

[39] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *INFOCOM, 2012 Proceedings IEEE*, pp. 1341–1349, IEEE, 2012. (Cited on pages 2 and 8.)

[40] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: Connecting people, locations and interests in a mobile 3G network," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference - IMC '09*, IMC '09, (New York, NY, USA), pp. 267–279, ACM, 2009. (Cited on pages 2, 8, 14, 16 and 83.)

[41] C. L. Williamson, E. Halepovic, H. Sun, and Y. Wu, "Characterization of CDMA2000 Cellular Data Network Traffic.," *LCN*, pp. Z000–719, 2005. (Cited on page 2.)

[42] Y. Li, J. Yang, and N. Ansari, "Cellular smartphone traffic and user behavior analysis," in *2014 IEEE International Conference on Communications (ICC)*, pp. 1326–1331, IEEE, June 2014. (Cited on pages 2, 9 and 10.)

[43] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, pp. 1–6, VDE, 2014. (Cited on pages 2, 10 and 11.)

[44] N. Bui and J. Widmer, "Modelling Throughput Prediction Errors as Gaussian Random Walks," in *The 1st KuVS Workshop on Anticipatory Networks*, Sept. 2014. (Cited on pages 2 and 10.)

[45] B. Liu, D. Niu, Z. Li, and H. V. Zhao, "Network latency prediction for personal devices: Distance-feature decomposition from 3D sampling," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 307–315, IEEE, Apr. 2015. (Cited on pages 2 and 10.)

[46] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, p. 33, Dec. 2012. (Cited on pages 3, 13, 14, 16, 17, 18, 24, 36 and 39.)

[47] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proceedings of the 20th annual international conference on Mobile computing and networking - MobiCom '14*, (New York, USA), ACM, 2014. (Cited on pages 3 and 17.)

[48] E.-H. Chung and A. Shalaby, "A trip reconstruction tool for GPS-based personal travel surveys," *Transportation Planning and Technology*, vol. 28, pp. 381–401, Oct. 2005. (Cited on pages 4 and 19.)

[49] "EU CHIST-ERA Mobile context-Adaptive CAching for COntent-centric networking (MACACO) project." https://macaco.inria.fr/. (Cited on pages 4 and 27.)

[50] G. L. Ulmer, *Internet invention: From literacy to electracy*. Longman New York, 2003. (Cited on page 7.)

[51] S. Hoteit, S. Secci, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, and G. Pujolle, "Content consumption cartography of the Paris urban region using cellular probe data," in *Proceedings of the First Workshop on Urban Networking*, UrbaNe '12, (New York, NY, USA), pp. 43–48, ACM, 2012. (Cited on pages 7 and 8.)

[52] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 230–235, IEEE, IEEE, Mar. 2015. (Cited on pages 8, 9, 10, 14, 73, 74 and 76.)

[53] A. Fumo, M. Fiore, and R. Stanica, "Joint spatial and temporal classification of mobile traffic demands," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pp. 1–9, IEEE, 2017. (Cited on page 8.)

[54] Y. Zang, F. Ni, Z. Feng, S. Cui, and Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks.," *ChinaSIP*, pp. 458–462, 2015. (Cited on page 9.)

[55] Z. Yi, X. Dong, X. Zhang, and W. W. 0007, "Spatial traffic prediction for wireless cellular system based on base stations social network.," *SysCon*, 2016. (Cited on page 9.)

[56] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3g network using hourglass co-clustering," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, pp. 341–352, ACM, 2010. (Cited on page 9.)

[57] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, pp. 13–18, ACM, 2012. (Cited on page 9.)

[58] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 3899–3912, June 2017. (Cited on page 9.)

[59] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, pp. 180–186, ACM, 2017. (Cited on page 9.)

[60] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting whatsapp through large-scale measurements in mobile networks," in *ACM SIGCOMM Computer Communication Review*, vol. 44, pp. 133–134, ACM, 2014. (Cited on page 9.)

[61] Q. Deng, Z. Li, Q. Wu, C. Xu, and G. Xie, "An empirical study of the wechat mobile instant messaging service," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 390–395, May 2017. (Cited on page 9.)

[62] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck, "Over the top video: the gorilla in cellular networks," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 127–136, ACM, 2011. (Cited on page 9.)

[63] Z. Li, X. Wang, N. Huang, M. A. Kaafar, Z. Li, J. Zhou, G. Xie, and P. Steenkiste, "An empirical analysis of a large-scale mobile cloud storage service," in *Proceedings of the 2016 Internet Measurement Conference*, pp. 287–301, ACM, 2016. (Cited on page 9.)

[64] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991. (Cited on pages 11, 70 and 71.)

[65] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, 1994. (Cited on pages 11, 70, 71, 72, 80 and 94.)

[66] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, p. 2923, Oct. 2013. (Cited on pages 11 and 17.)

[67] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 88–94, IEEE, Mar. 2014. (Cited on pages 11, 27, 70 and 81.)

[68] J. Wang, Y. Mao, J. Li, Z. Xiong, and W.-X. Wang, "Predictability of road traffic and congestion in urban areas," *PLoS ONE*, vol. 10, p. e0121825, Apr. 2015. (Cited on page 11.)

[69] G. Ding, J. Wang, Q. Wu, Y.-d. Yao, R. Li, H. Zhang, and Y. Zou, "On the limits of predictability in real-world radio spectrum state dynamics: From entropy theory to 5G spectrum sharing," *IEEE Communications Magazine*, vol. 53, pp. 178–183, July 2015. (Cited on page 11.)

[70] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*. Springer New York, 2000. (Cited on page 11.)

[71] K. Papagiannaki, N. Taft, Z.-L. Zhang, and C. Diot, "Long-term forecasting of internet backbone traffic: Observations and initial models," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 2, pp. 1178–1188, IEEE, 2003. (Cited on page 11.)

[72] N. Sadek and A. Khotanzad, "Multi-scale high-speed network traffic prediction using k-factor gegenbauer arma model," in *Communications, 2004 IEEE International Conference on*, vol. 4, pp. 2148–2152, IEEE, 2004. (Cited on page 11.)

[73] Y. Qiao, J. A. Skicewicz, and P. A. Dinda, "An Empirical Study of the Multiscale Predictability of Network Traffic.," *HPDC*, 2004. (Cited on page 11.)

[74] B. Zhou, D. He, Z. Sun, and W. H. Ng, "Network traffic modeling and prediction with arima/garch," in *Proc. of HET-NETs Conference*, pp. 1–10, 2005. (Cited on page 11.)

[75] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer US, 1971. (Cited on pages 11, 12 and 13.)

[76] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *IEEE transactions on mobile computing*, vol. 5, pp. 1633–1649, Dec. 2006. (Cited on pages 11, 13 and 85.)

[77] J. Jeong, M. Leconte, and A. Proutiere, "Cluster-aided mobility predictions," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, Apr. 2016. (Cited on pages 11 and 99.)

[78] A. Moffat, "Implementing the PPM data compression scheme," *IEEE Transactions on communications*, vol. 38, no. 11, pp. 1917–1921, 1990. (Cited on pages 12 and 85.)

[79] P. Jacquet, W. Szpankowski, and I. Apostol, "An universal predictor based on pattern matching, preliminary results," *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities*, pp. 75–85, 2000. (Cited on pages 12 and 85.)

[80] K. Gopalratnam and D. J. Cook, "Active lezi: An incremental parsing algorithm for sequential prediction," *International Journal on Artificial Intelligence Tools*, vol. 13, pp. 917–929, Dec. 2004. (Cited on pages 12 and 85.)

[81] "Ensemble methods – scikit-learn." `http://scikit-learn.org/stable/modules/ensemble.html`. (Cited on pages 12 and 53.)

[82] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Annals of statistics*, vol. 29, pp. 1189–1232, Oct. 2001. (Cited on pages 12 and 53.)

[83] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015. (Cited on pages 13 and 85.)

[84] G. Rutka, "Neural network models for internet traffic prediction," *Elektronika ir Elektrotechnika*, vol. 68, no. 4, pp. 55–58, 2006. (Cited on page 13.)

[85] "Specifications - 3GPP." `http://www.3gpp.org/specifications`, 2018. (Cited on pages 13, 14 and 15.)

[86] C. Iovan, A.-M. Olteanu-Raimond, T. Couronné, and Z. Smoreda, "Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies," in *Geographic Information Science at the Heart of Europe*, pp. 247–265, Springer, 2013. (Cited on pages 13 and 18.)

[87] M. Ficek and L. Kencl, "Inter-call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model," in *2012 Proceedings IEEE INFOCOM*, pp. 469–477, IEEE, Mar. 2012. (Cited on pages 13, 18, 44, 49 and 57.)

[88] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: Beyond power-law and lognormal distributions," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 596–604, ACM, 2008. (Cited on page 13.)

[89] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *INFOCOM, 2014 Proceedings IEEE*, pp. 1806–1814, IEEE, 2014. (Cited on page 13.)

[90] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla, "How long will she call me? distribution, social theory and duration prediction," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 16–31, Springer, 2013. (Cited on pages 13 and 18.)

[91] P. O. V. De Melo, L. Akoglu, C. Faloutsos, and A. A. Loureiro, "Surprising patterns for the call duration distribution of mobile phone users," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 354–369, Springer, 2010. (Cited on page 13.)

[92] T. Cheng, J. Wang, J. Haworth, B. Heydecker, and A. Chow, "A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling," *Geographical Analysis*, vol. 46, no. 1, pp. 75–97, 2014. (Cited on page 14.)

[93] J. Ma, H. Li, F. Yuan, and T. Bauer, "Deriving operational origin-destination matrices from large scale mobile phone data," *International Journal of Transportation Science and Technology*, vol. 2, no. 3, pp. 183–204, 2013. (Cited on page 14.)

[94] Q. Xu, A. Gerber, Z. M. Mao, and J. Pang, "AccuLoc: Practical localization of performance measurements in 3G networks," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pp. 183–196, ACM, 2011. (Cited on page 14.)

[95] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 329–344, ACM, 2011. (Cited on page 14.)

[96] Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin, "Understanding the bias of call detail records in human mobility research," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1738–1762, 2016. (Cited on pages 14, 16 and 18.)

[97] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, "A review of urban computing for mobile phone traces: Current methods, challenges and opportunities," in *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, p. 2, ACM, 2013. (Cited on pages 14 and 15.)

[98] W. Wu, Y. Wang, J. B. Gomes, D. T. Anh, S. Antonatos, M. Xue, P. Yang, G. E. Yap, X. Li, S. Krishnaswamy, *et al.*, "Oscillation resolution for mobile phone cellular tower data to enable mobility modelling," in *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, vol. 1, pp. 321–328, IEEE, 2014. (Cited on page 14.)

[99] R. S. Campos, "Evolution of positioning techniques in cellular networks, from 2G to 4G," *Wireless Communications and Mobile Computing*, vol. 2017, 2017. (Cited on pages 14, 15 and 16.)

[100] J. Schlaich, T. Otterstätter, M. Friedrich, *et al.*, "Generating trajectories from mobile phone data," in *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010. (Cited on page 15.)

[101] "OpenCID." http://wiki.opencellid.org/wiki/FAQ. (Cited on page 15.)

[102] "France Open Data." https://www.data.gouv.fr/fr/datasets/. (Cited on pages 15 and 34.)

[103] "Google map geolocation API." https://developers.google.com/maps/documentation/geolocation/intro. (Cited on page 15.)

[104] "Unwired Labs Location API." http://unwiredlabs.com/. (Cited on page 15.)

[105] "OpenSignal." http://opensignal.com/. (Cited on page 15.)

[106] "Mozilla Location Service." https://location.services.mozilla.com/. (Cited on page 15.)

[107] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Ranges of human mobility in Los Angeles and new York," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 88–93, IEEE, IEEE, Mar. 2011. (Cited on pages 15, 16 and 17.)

[108] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Supplementary material." http://www.sciencemag.org/cgi/content/full/327/5968/1018/DC1, Jan. 2005. (Cited on pages 16, 17, 18, 71, 82 and 92.)

[109] R. Ahas, S. Silm, E. Saluveer, and O. Järv, "Modelling home and work locations of populations using passive mobile positioning data," *Location based services and TeleCartography II*, pp. 301–315, 2009. (Cited on page 17.)

[110] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Lecture Notes in Computer Science*, pp. 133–151, Springer Berlin Heidelberg, 2011. (Cited on pages 17, 18, 51 and 52.)

[111] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking - MobiCom '07*, (New York, USA), pp. 123–134, ACM, Sept. 2007. (Cited on page 17.)

[112] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010. (Cited on page 18.)

[113] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, Mar. 2013. (Cited on pages 18 and 42.)

[114] N. B. Ponieman, A. Salles, and C. Sarraute, "Human mobility and predictability enriched by social phenomena information," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, (New York, NY, USA), pp. 1331–1336, ACM, 2013. (Cited on page 18.)

[115] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, pp. 96–100, Feb. 2012. (Cited on page 18.)

[116] D. Zhang, J. Zhao, F. Zhang, and T. He, "coMobile: Real-time human mobility modeling at urban scale using multi-view learning," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, (New York, NY, USA), pp. 40:1–40:10, ACM, 2015. (Cited on pages 19, 42 and 61.)

[117] R. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, pp. 32–39, Nov. 2011. (Cited on page 22.)

[118] E. Mucelli Rezende Oliveira, A. Carneiro Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin, "On the regularity of human mobility," *Pervasive and Mobile Computing*, vol. 33, pp. 73–90, Dec. 2016. (Cited on pages 24, 57, 58, 59 and 61.)

[119] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Pers Ubiquit Comput*, vol. 10, pp. 255–268, Nov. 2005. (Cited on page 27.)

[120] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, Feb. 2010. (Cited on pages 29 and 35.)

[121] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, pp. 596–604, ACM, ACM, 2008. (Cited on pages 31 and 32.)

[122] M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi, "Optimal spatial resolution for the analysis of human mobility," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 248–252, IEEE, Aug. 2012. (Cited on page 34.)

[123] S. Hoteit, S. Secci, S. Sobolevsky, G. Pujolle, and C. Ratti, "Estimating real human trajectories through mobile phone data," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 2, pp. 148–153, IEEE, IEEE, June 2013. (Cited on page 35.)

[124] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PLoS ONE*, vol. 7, p. e39253, June 2012. (Cited on page 38.)

[125] C. Iovan, A.-M. Olteanu-Raimond, T. Couronné, and Z. Smoreda, "Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies," in *Geographic Information Science at the Heart of Europe*, pp. 247–265, Springer International Publishing, 2013. (Cited on page 42.)

[126] L. M. Silveira, J. M. de Almeida, H. T. Marques-Neto, C. Sarraute, and A. Ziviani, "MobHet: Predicting human mobility using heterogeneous data sources," *Computer Communications*, vol. 95, pp. 54–68, Dec. 2016. (Cited on page 42.)

[127] S. Lu, Z. Fang, X. Zhang, S.-L. Shaw, L. Yin, Z. Zhao, and X. Yang, "Understanding the representativeness of mobile phone location data in characterizing human mobility indicators," *IJGI*, vol. 6, p. 7, Jan. 2017. (Cited on page 42.)

[128] G. Khodabandelou, V. Gauthier, M. El-Yacoubi, and M. Fiore, "Population estimation from mobile network traffic metadata," in *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, IEEE, June 2016. (Cited on page 43.)

[129] J. A. Hartigan, "Clustering," *Annual review of biophysics and bioengineering*, vol. 2, pp. 81–102, June 1973. (Cited on page 51.)

[130] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer New York, 2001. (Cited on pages 53 and 54.)

[131] L. Kong, M. Xia, X.-Y. Liu, G. Chen, Y. Gu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in wireless sensor networks," *IEEE Transactions Parallel Distributed Systems*, vol. 25, pp. 2818–2828, Nov. 2014. (Cited on page 59.)

[132] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Matrix factorization and neighbor based algorithms for the netflix prize problem," in *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08*, pp. 267–274, ACM, ACM, 2008. (Cited on page 59.)

[133] C. M. Schneider, V. Belik, T. Couronne, Z. Smoreda, and M. C. Gonzalez, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, pp. 20130246–20130246, May 2013. (Cited on page 59.)

[134] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," *Lecture Notes in Computer Science*, vol. 5034, pp. 337–348, 2008. (Cited on pages 59 and 63.)

[135] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation," in *Proceedings of the fourth ACM conference on Recommender systems - RecSys '10*, pp. 79–86, ACM, ACM, 2010. (Cited on page 61.)

[136] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, pp. 455–500, Aug. 2009. (Cited on pages 61 and 62.)

[137] J. Portela and M. Alencar, "Cellular network as a multiplicatively weighted voronoi diagram," in *CCNC 2006. 2006 3rd IEEE Consumer Communications and Networking Conference, 2006.*, vol. 2, pp. 913–917, IEEE, IEEE, 2006. (Cited on page 63.)

[138] H. Kuhn and A. Tucker, "Proceedings of 2nd berkeley symposium," 1951. (Cited on page 72.)

[139] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. (Cited on page 72.)

[140] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Transactions on Information Theory*, vol. 44, pp. 1319–1327, May 1998. (Cited on page 72.)

[141] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics," in *Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD '99*, vol. 28, pp. 49–60, ACM, ACM, 1999. (Cited on page 74.)

[142] P.-N. Tan, M. Steinbach, V. Kumar, *et al.*, *Introduction to data mining*, vol. 1. Pearson Addison Wesley Boston, 2006. (Cited on page 74.)

[143] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 85.)

[144] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, Mar. 2013. (Cited on page 99.)

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : L'étude des habitudes humaines : de la reconstruction de la mobilité à la prédiction du trafic mobile

Mots clés : mobilité humaine, trafic de données mobiles, prédiction, enrichissement des données

Résumé : Caractériser et prédire la mobilité humaine et le trafic de données mobiles est important dans les réseaux cellulaires, par exemple, pour l'optimisation du réseau, la gestion du réseau et la conception de l'application. La mobilité et le trafic de données mobiles des individus sont deux aspects significatifs dans ce contexte. Dans cette thèse, les jeux de données collectés à partir d'appareils mobiles sont exploités pour étudier les problèmes de recherche clés concernant les deux aspects. En termes de mesure de la mobilité individuelle, nous étudions l'incomplétude des données de mobilité et étudions les biais correspondants causés par l'utilisation de données de mobilité incomplètes. En termes d'enrichissement des données de mobilité individuelles, nous proposons des solutions innovantes pour reconstruire des trajectoires individuelles, puis prouvons leur efficacité en utilisant les simulations de données réelles. Enfin, en tant que contribution la plus significative de cette thèse, l'analyse détaillée est réalisée pour montrer la prévisibilité théorique et pratique de la consommation de trafic de données mobiles des individus.

Title : Human habits investigation: from mobility reconstruction to mobile traffic prediction

Keywords : human mobility, mobile data traffic, prediction, data enrichment

Abstract : Characterizing and predicting human mobility and mobile data traffic serves as an important component in cellular networks, and plays a key role in network optimization, network management, and application design. The mobility and mobile data traffic of individuals are two significant aspects in this context. In this thesis, the datasets collected from mobile devices are mined to study key research problems regarding the two aspects. In terms of the individual mobility measurement, we study the mobility data incompleteness and study the corresponding biases caused by the use of such incomplete mobility data. In terms of the individual mobility data enrichment, we propose novel solutions to reconstruct individual trajectories and then prove their effectiveness via data-driven simulations. Finally, as the most significant contribution of this thesis, the detailed analysis is performed to show the theoretical and practical predictability of the mobile data traffic consumption of individuals.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France