Understanding individuals' proclivity for novelty-seeking in human mobility

Licia Amichi

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAX099

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Understanding individuals' proclivity for novelty-seeking in human mobility

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Paris, le 23 Novevmbre 2021, par

LICIA AMICHI

Composition du Jury :

Isabelle Guérin Lassous Professeur, Université Claude Bernard Lyon 1 et INRIA (DANTE)	Président
Martin Tomko Senior Lecturer, University of Melbourne	Rapporteur
Nathalie Mitton Directrice de recherche, INRIA (FUN)	Rapporteur
Isabelle Guérin Lassous Professeur, Université Claude Bernard Lyon 1 et INRIA (DANTE)	Examinateur
Luca Pappalardo Researcher, ISTI-CNR	Examinateur
Aline Carneiro Viana Directrice de recherche, INRIA (TRIBE)	Directeur de thèse
Mark Crovella Professor, Boston University	Invité
Antonio Alfredo Loureiro Professor, Universidade Federal de Minas Gerais (UFMG)	Invité

Abstract

Understanding and predicting how humans move within space and time is of fundamental importance for many scientific domains, such as epidemic propagation (e.g., the COVID-19 pandemics), mobile networks, urban planning (e.g., anticipatory temperature control of a home for energy consumption management), or ride-sharing. Yet, apprehending human mobility is intrinsically complex. On the one hand, human movements are constrained by physical presence in workplaces, gyms, or universities, in addition to the involvement in routine and social activities. On the other hand, the large variety of leisure places and the availability of modern means of transportation allow people to break their routinary patterns to discover new places. Understanding human mobility is a longstanding challenge that goes back to the 19^{th} century. Up to the 20^{th} century, scholars focused only on group and migration flows in view of the quality of the leveraged data (Census data or surveys). But recently, with the ubiquity of mobile devices, Internet connectivity, and positioning systems (e.g., GPS system), capturing individuals' daily whereabouts at very fine spatial and temporal scales has become possible. This offers the opportunity to observe and study human mobility at the individual level with an unprecedented level of detail. In particular, the increasing availability of such data has led to ongoing development in the field of mobility research. Nevertheless, the scientific literature on human mobility prediction is oblivious to individuals' tendencies for novelty-seeking, i.e., exploring and discovering new places. Conventional predictors relying on personal geographical data perform poorly when it comes to discoveries of new regions. The reason is explained by the prediction relying only on previously visited/seen (or known) locations. As a side effect, places that were never visited before (or explorations) by a user cause disturbance to known location's prediction. Neglecting novelty-seeking activities at first glance appears to be inconsequential on the ability to understand and predict individuals' trajectories. In this manuscript, we claim and show the opposite: exploration-like visits strongly impact mobility understanding and anticipation.

This thesis focuses on exploratory visits in human mobility. It first seeks to unveil the exploratory preferences of the population. Afterward, it evaluates the impacts of exploration-like visits on the theoretical and practical predictability extents. Finally, it proposes an exploration-aware mobility prediction framework that integrates the notion of exploration.

Throughout this manuscript, we reveal the existence of three distinct mobility profiles with regard to the exploration activity – *Scouters* (i.e., extreme explorers who are keener to discover new places), *Routiners* (i.e., extreme returners who limit their mobility to a few locations), and *Regulars* (i.e., without extreme behavior). Besides, we show that the novelty-seeking factor is an essential element to consider and should not be overlooked when designing predictors, particularly for specific categories exhibiting high exploration activities. Furthermore, by integrating the

notion of exploration in prediction, we demonstrate substantial improvements in prediction accuracy by dint of fruitfully forecasting coarse-grained zones used for exploration activities.

Résumé

La compréhension et la prédiction du déplacement des individus dans l'espace et le temps sont fondamentales dans de nombreux domaines comme la propagation des épidémies (notamment, la pandémie du COVID-19), les réseaux mobiles, l'urbanisme (par exemple, le contrôle anticipé de la température d'une maison pour la gestion de la consommation énergétique) ou encore le covoiturage. Cependant, appréhender la mobilité humaine est intrinsèquement complexe car d'une part, les mouvements des individus sont limités par l'obligation d'une présence physique sur les lieux de travail, les universités ou dans le cadre de la participation à des activités routinières et sociales. D'autre part, la grande variété de lieux de loisirs et la disponibilité de moyens de transport modernes permettent aux individus d'interrompre leurs routines pour découvrir de nouveaux lieux. La compréhension de la mobilité humaine constitue un enjeu de longue date, datant du 19^e siècle. Jusqu'au 20^e siècle, les chercheurs ne s'intéressaient qu'aux flux collectifs et migratoires, compte tenu de la qualité des données utilisées (données de recensement ou d'enquête). En revanche, depuis peu avec l'omniprésence des appareils mobiles (téléphones portables, bracelets fitbit), de la connectivité Internet et des systèmes de positionnement (par exemple, le système GPS), il est devenu possible de capturer les allers et venues quotidiennes des individus à des échelles spatiales et temporelles très précises. Cela offre l'opportunité d'observer et d'étudier la mobilité humaine au niveau individuel avec un niveau de détail sans précédent. En particulier, la disponibilité croissante de ces données a conduit à un développement continu dans le domaine de la recherche sur la mobilité spatiale et temporelle. Cependant, la littérature scientifique sur la prédiction de la mobilité humaine ne tient pas compte des tendances des individus à rechercher la nouveauté, c'est-à-dire à explorer et à découvrir de nouveaux lieux. Les prédicteurs conventionnels reposant sur des données géographiques personnelles fonctionnent mal lorsqu'il s'agit de découvrir de nouvelles régions. La raison est expliquée par la prédiction reposant uniquement sur des emplacements précédemment visités/vus (ou connus). Comme effet secondaire, des emplacements qui n'ont jamais été visités auparavant (ou des explorations) par un utilisateur perturbent la prédiction d'un emplacement connu. Négliger à première vue les activités de recherche de nouveauté apparaît sans conséquence sur la capacité à comprendre et à prévoir les trajectoires des individus. Dans ce manuscrit, nous affirmons et montrons le contraire : les visites de type exploration impactent fortement la compréhension et l'anticipation de la mobilité humaine.

Dans ce manuscrit, nous cherchons à dévoiler les préférences exploratoires de la population dans un premier temps. Ensuite, nous évaluons l'impact des visites exploratoires sur la prévision théorique et pratique. Enfin, nous proposons un prédicteur de mobilité qui intègre la notion d'exploration. À travers ce manuscrit, nous révélons l'existence de trois profils distincts de mobilité en ce qui concerne l'activité d'exploration :

1. *Scouters* : les explorateurs extrêmes qui sont plus enclins à découvrir de nouveaux endroits, passent plus de temps à explorer et couvrent généralement de longues distances.

2. *Routiners* : les routiniers extrêmes qui limitent leur mobilité à quelques endroits, passent la majorité de leur temps dans des lieux familiers et parcourent de courtes distances quotidiennement.

3. *Regulars* : les individus réguliers (c'est-à-dire sans comportement extrême).

En d'autres termes, nous montrons que le facteur de recherche de nouveauté est un élément essentiel à considérer et ne doit pas être négligé lors de la conception de prédicteurs, surtout pour des catégories spécifiques d'individus qui présentent d'importantes activités exploratoires. En intégrant la notion d'exploration dans la prédiction, des améliorations considérables de la précision sont obtenues grâce à la prévision d'exploration des zones étendues.

À ma mère, à ma grand-mère
et à toute ma famille.

Acknowledgments

It is a pleasure to thank the many people who made this thesis possible.

First and foremost, I would like to thank all members of the INRIA Saclay research center and the IPP Doctoral School for welcoming me during these last three years and for giving me the opportunity to conduct the research presented in this thesis.

My deepest gratitude goes out to the members of my jury: Assoc. Prof. Martin Tomko (Melbourne University) for his inspirational work in the area of mobility research and for reviewing my manuscript and providing pertinent comments. Thank you to DoR. Nathalie Mitton (INRIA FUN), for having read my manuscript and supplied relevant remarks. I also want to thank the jury president Prof. Isabelle Guérin Lassous (Université Claude Bernard Lyon 1 at INRIA- DANTE), and Ph.D. Luca Pappalardo (ISTI-CNR) as an examiner.

Thank you to my thesis director, DoR. Aline Viana Carneiro, for giving me the opportunity to do this Ph.D. I would like to also thank my collaborators Prof. Mark Crovella (Boston University) and Prof. Antonio Loureiro (UFMG) for their availability, ongoing support, interest in my Ph.D., and contributions.

I would like to express my deep gratitude to Assoc. Prof. Megumi Kaneko (NII) who made my first research experience a pleasure and drew my attention to research.

I also thank my friends for giving me the support and friendship I needed. Thanks to all of whom provided the encouragement, the support, and the inspiration needed to make this project possible.

Last but not least, thanks to my mother, to whom I dedicate this thesis. Thanks to my grandmother, my aunt, my sister Lynda, my brother Lamine and my cousin Melissa. I am truly thankful for their trust and love that have always carried me.

Contents

List of Figures

List of Tables

Acronyms

ALZ Active LeZi. 22, 23, 62, 64, 70, 71, 97

CCDF Complementary Cumulative Distribution Function. xi, 40

CDF Cumulative Distribution Function. 64, 71, 90, 115

CDR Call Detail Record. 9, 11, 12, 14, 21, 23, 24, 27–31, 40–42, 44, 45, 47, 95, 98, 113–115

CTRW Continuous-Time Random Walk. 3, 9, 16

EL Exploratory Location. xi, xiii, 36–38, 41–43, 67

EOD Exploration Origin-Destination. 81

EPR Exploration Preferential Return. 3, 4, 9, 16, 17

EVP Exceptionally Visited Places. xi, xiii, 38, 41–43

FSA Finite-State Automaton. 35

GMM Gaussian Mixture probabilistic Model. xi, 43, 44

GPS Global Positioning System. 9, 12, 14, 23–25, 27–31, 40, 41, 44, 45, 52, 95, 98, 113–115

GSM Global System for Mobile Communications. 24

HMM Hidden Markov Model. 3, 23, 24, 77

IEI Inter-Exploration Interval. xi, 53, 54, 80

IEIP Inter Exploration Interval Predictor. 77, 80, 84–92

JSP Joint Spatial Predictor. xii, xiii, 80, 82, 88–92

MC Markov Chain. xi, xii, 21–24, 62, 64–74, 81–91, 97

MCMC Markov Chain Monte Charlo. 22

MLE Maximum Likelihood Estimation. 22

MMM Mixed Markov chain Model. 24

MVP Mostly Visited Places. xi, 37, 38, 41, 42

OVP Occasionally Visited Places. xi, xiii, 37, 38, 41, 42

PCA Principal Component Analysis. 4, 26

PDF Probability Distribution Function. 14, 55, 56

POI Points Of Interest. 29, 62, 99, 113

PPM Prediction by Partial Matching. 22, 23, 62, 64, 70, 71, 97

PSP Personal Spatial Predictor. xii, xiii, 80, 82, 88–92

RL Routine Location. xi, 36–38, 41–43, 67

SPM Sampled Pattern Matching. 22, 23, 62, 64, 65, 70, 71, 97

STMP Successive Types of Movements Predictor. 77, 79, 80, 84–88, 90–92

TRIM TRip DIversity Measure. 26

Chapter 1

Introduction

Contents

Human mobility refers to the displacement and movement of human beings within space and time. It is used to refer to individual mobility as well as group flows [13]. Mobility is closely tied to the history and existence of human beings. Starting from the Cognitive Revolution about 70,000 years ago, Sapiens who already populated East Africa began to *explore* and overrun the rest of the Earth [50, 51]. Besides migration flows driven by environmental conditions and socio-economic factors, humans perform daily trips for different motives that evolved with urbanization and the availability of different means of locomotion [96, 13]. Nowadays, humans not only move to ensure their subsistence but also for leisure, such as visiting a museum, a shopping center, or going to a bar or a restaurant.

Understanding and accurately predicting human movements are key components for various domains and applications such as targeted advertising [94], epidemic prevention (e.g., the COVID-19 pandemics) [11, 22, 49], urban planning [90], or smooth resource and handover management for mobile networks [92, 66]. Nonetheless, comprehending how individuals move in space and time is a longstanding challenge that roots back to the 19^{th} century [21]. Up to the mid of the 20^{th} century, scholars focused only on group and migration flows in view of the leveraged coarse-grained data. However, recently, with the ubiquity of mobile devices, Internet connectivity, and positioning systems, capturing individuals' whereabouts at very *fine spatial* and *temporal* scales became possible. This offers the opportunity to observe and analyze human mobility behavior at an unprecedented level of detail. Notably, the increasing availability of such data has led to an ongoing development in the field of human mobility.

The last decades have witnessed an extensive examination of human mobility behavior. Advanced statistical analyses of mobility trajectories reveal that individuals' movements are characterized by (i) *deep-rooted regularity of visits dictated by routine* interrupted by (ii) *irregular sporadic visits to unknown or rarely visited*

places [19, 46, 85]. Indeed, on the one hand, human movements are constrained by physical presence in workplaces, gyms, or universities, in addition to the involvement in routine and social activities. On the other hand, the large variety of leisure places and the availability of modern means of transportation allow the people to break their routinary patterns to discover new places [33]. Given the apparent complexity of human movements, Song et al. [86] attempt to ascertain if the mobility patterns are potentially predictable or not. They estimate the predictability upper bound of human mobility trajectories and find a strikingly high potential predictability of 93%.

Building upon the above findings, many advanced mobility models and predictors are proposed [85, 19, 10, 41, 62, 43]. Yet, they all show limited performance [32]. Current mobility models systematically fail in reproducing individuals' movements and substantially deviate from empirical results [85, 73]. Existing predictors never reach the coveted 100% prediction accuracy and deviate from the theoretical upper bounds of predictability [32]. The reasons for these limitations are manifold: the lack of ground truth data, human beings' complex nature and behavior, and the difficulty in forecasting visits to new places, i.e., *explorations*. Whereas data quality and human behavior are not easy to tackle and can raise major privacy concerns [17, 91, 84], substantial advances can be made in mobility research by looking at the exploration phenomenon.

The goal of the thesis is to examine and understand the exploration problem that has rarely been tackled in the literature but represents a real issue and should be carefully addressed to propose realistic generative models and accurate predictors [85, 32]. It first investigates state-of-the-art models and proposes a mobility profiling that captures individuals' exploration tendencies. Next, it evaluates the impacts of exploration-like visits on prediction accuracy. Finally, it proposes an exploration-aware mobility predictor that, in addition to revisits to known places, attempts to forecast moments of novelty-seeking and gives a coarse-grained spatial intuition on where the individual is going to explore.

1.1 Novelty-seeking in Human Mobility

Intuitively, human movements can be split into two basic complementary types of movements: ***returns***, i.e., exploitation of familiar places and ***exploration***, i.e., discoveries of new places. Whereas the routine and regularity of patterns have been widely investigated in the literature, novelty-seeking (exploration-like visits) is an emerging topic attracting more and more the scientific communities' attention. In this section, we review how novelty-seeking is tackled in different aspects of human mobility, namely, mobility modeling, mobility prediction, and the quantification and evaluation of the exploratory behavior among the population.

1.1.1 Novelty-seeking and Mobility Modeling

Recent studies on human mobility report the existence of several statistical properties commonly appearing in real-world mobility traces. This provides an excellent opportunity for modeling human mobility behavior. The traces generated by mobility models are of great importance; they can be used to create realistic simulations to support many applications such as human-assisted mobile networks, epidemics, or urban planning.

Minimal models relying on Random walks' [1] properties were first used to mimic individuals' mobility patterns such as Lévy Flight or Continuous-Time Random Walk (CTRW). Nevertheless, they roughly reproduce individuals' movements and substantially deviate from empirical data.

Distinguishing explorations from returns in mobility models: By introducing the notions of *return* and *exploration* into mobility models, Song et al. [85] explain the discrepancies of traditional Random walk-based models. Specifically, they show that the failures of Random walk-based models in reproducing individuals' trajectories emanate from their disregard to what is novelty (exploration). In their proposed Exploration Preferential Return (EPR) model, Song et al. [85] model each type of movement separately. In the case of an exploration (i): the next location is randomly chosen. In the case of a return (ii): the next location is chosen among the set of known places with a probability proportional to the number of visits to the locations. *This categorization of movements allows generating more representative traces, emphasizing the necessity to carefully tackle individuals' inclinations for discoveries of new places when modeling human movements.*

1.1.2 Novelty-seeking and Mobility Prediction

Due to predictors' invaluable contributions, a plethora of mobility prediction methods and techniques are proposed and becoming more and more robust and accurate [10, 41, 44, 7].

Novelty-seeking a missing notion: Explorations are a major limiting factor for prediction tasks [32]. Indeed, forecasting discoveries of new places is ambitious and challenging to tackle as it is about predicting the unknown. Conventional *personal* predictors such as Markov-based models [41, 62, 44] or Hidden Markov Model (HMM) [63] *utterly rely on historic personal location data to predict future locations.* Moreover, they forecast an individual's next location on the assumption that it belongs to the set of her known places [93]. *This engenders erroneous forecasts at each occurrence of an exploration event, which is worsened by the fact that such events are numerous and largely present in the daily lives of individuals: on average, 70% of visits happen only once [32]. This representative rate highlights how impacting exploration-intended visits are for conventional personal predictors and puts in evidence the need for detecting such types of movements.*

[1] Mathematically a random walk is a path that consists of successive random steps [82].

1.1.3 Novelty-seeking Quantification and Similarity Identification

Pappalardo et al. [73] and Scherrer et al. [78] are the first to quantify the trade-off between exploitation and exploration in human mobility and split the population accordingly.

In [73], Pappalardo et al. seek to quantify repetitive/irregular visits' influence on overall mobility. They find that individuals' repetitiveness and irregularity of visits are *heterogeneous*. Whereas some individuals have a regularity limited to a few locations, others have an expanded regularity and tend to diversify their daily visits. Specifically, they report the existence of two distinct mobility profiles: *explorers* who are characterized by a dynamicity of visits and go to several places on a daily basis, and *returners* who are marked by a steadiness of visits limiting their mobility between a few places. Moreover, they point out the importance of the social dimension in explaining the propensity to explore within certain areas.

Similar to Pappalardo et al. [73], Scherrer et al. [78] propose a data-driven approach to identify sub-groups of individuals displaying similar mobility behavior. They derive several features commonly used in mobility research from the spatiotemporal footprints of the individuals, such as the number of stops, the number of visited locations, or stop duration. Then apply the unsupervised learning Principal Component Analysis (PCA) approach to the extracted features to identify the most significant clusters. Accordingly, they find the existence of two main categories: (i) *travelers*, who are active and have a more diffused mobility, and (ii) *locals*, who roam in small and compact areas and move slowly.

Impacts of return and exploration tendencies on mobility modeling: Pappalardo et al. [73] show that the EPR is unable to recover the mobility properties of the two classes of individuals: explorers and returners. Subsequently, the authors propose an enhancement of the exploration part of the EPR model by presuming that an individual is attracted by popular locations at the group level when she discovers new places. *The enhanced EPR model can reproduce the key features of the aggregated mobility patterns where the EPR model fails.*

Understanding, quantifying, and including novelty-seeking in mobility models and predictors have a short history. More and more studies point out the importance and impacts of exploration-like visits so far neglected in mobility modeling and prediction literature. Meanwhile, new research questions have arisen. Namely, *how can exploration visits be distinguished from return visits? How do the tendencies to explore fluctuate among the population? What are the impacts of explorations visits on prediction accuracy?* Or *can exploration-like visits be anticipated?*

1.2 Thesis Contributions

The contributions of the thesis aim to answer the questions raised in the previous section and mainly cover two different perspectives: understanding individuals' tendencies to explore and integrating the notions of exploration in personal mobility predictors. The contributions are organized in a progressive way.

Q_1 What is exploration?

- **Problem**: A clear definition of *what is exploration?* is missing from the literature. The naive state-of-the-art approach in considering the first occurrence of a location in the mobility trace as an exploration can cause bias in understanding and accurately quantifying such events.
- **Contribution**: This thesis proposes a new per-user approach based on the visitation frequency of the distinct locations. Specifically, only the first occurrences of rarely visited places are viewed as moments of novelty-seeking. This has the advantage of not considering the first appearances in the mobility trajectory of highly visited places such as home or workplace as moments of exploration (cf. Chapter 4, $\mathbf{P}_2$).

Q_2 Are the tendencies to explore homogeneous, or do they fluctuate among the population?

- **Problem**: Recent literature reports a heterogeneity in the exploratory habits among the population. Nonetheless, existing profiling approaches neglect the notion of exploration and are established by quantifying regularity or diversity of visits.
- **Contribution**: On the contrary, this thesis proposes a new mobility profiling that captures the spatiotemporal properties of both exploratory and return habits. The results reveal the existence of three main *visiting profiles* (cf. Chapter 4, $\mathbf{P}_3$):
  - **Scouters:** are keener to explore and have large sets of visited places. They limit their routinary mobility to a small set of locations and walk longer distances.
  - **Routiners:** rarely break their returning routine to discover new places, constantly visit their known locations, and have confined mobility.
  - **Regulars:** exhibit a medium behavior and have confined mobility. However, they tend to walk long distances when exploring.

Q_3 What are the impacts of exploration visits on prediction?

- **Problem**: Whereas the necessity to reflect exploration-like visits when designing mobility models is demonstrated in the literature, it is still neglected and overlooked in mobility prediction.

- **Contribution**: This thesis sheds light on this phenomenon and evaluates the impacts of individuals' exploration tendencies on the accuracy of prediction achieved by the two most widespread prediction tasks. Namely, (i) *the next-place prediction task* that aims to predict the next location of an individual and (ii) *the next-place prediction task* that aims to predict transitions between places. The results show that Scouters are the least predictable users due to their high exploration activities. Moreover, the results emphasize the role of novelty-seeking in making human mobility behavior less foreseeable and compel the need to thoroughly understand the exploration phenomenon, to allow the design of accurate predictors (cf. Chapter 5, $\mathbf{P}_2$).

$\mathbf{Q}_4$ Can exploration-like visits be anticipated?

- **Problem**: Conventional predictors utterly relying on individuals' geo-stamped location data are oblivious of users' inclinations to exploration, resulting in low forecasting accuracy, especially for highly exploratory users.
- **Contribution**: This thesis proposes an exploration-aware mobility prediction framework that solely relies on timestamped location data. The proposed framework splits the location prediction problem into two main steps: (i) *predicting the next type of movement* (exploration or return) (ii) *inferring the spatial location of the visit*. The results show that the proposed framework increases the prediction performance in general. Still, it can also be tuned according to the needs and requirements of the using applications or services. For instance, exploration forecasts decrease uncertainty in population mobility anticipation, which directly enhances resource allocation in network planning (as in the placement of Mobile Edge Computing (MEC) by telecom operators) and recommendation systems performance (cf. Chapter 6, $\mathbf{P}_1$).

The results listed above (and detailed in Chapters 4, 5, and 6) are summarized in the following publications:

$\mathbf{P}_1$ From motion purpose to perceptive spatial mobility prediction. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM SIGSPTIAL International Conference on Advances in Geographic Information Systems. 2021.

$\mathbf{P}_2$ Revealing an inherently limiting factor in human mobility prediction. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. IEEE Transactions on Emerging Topics in Computing (TETC). *Under review.*

$\mathbf{P}_3$ Understanding individuals' proclivity for *novelty seeking*. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM SIGSPTIAL International Conference on Advances in Geographic Information Systems. 2020.

P$_4$ Mobility profiling: Identifying scouters in the crowd. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM CONEXT International Conference on emerging Networking EXperiments and Technologies Student Workshop. 2019.

P$_5$ Explorateur ou Routinier: Quel est votre profile de mobilité?. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ALGOTEL Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications. 2020.

1.3 Thesis Outline

After the introduction, the thesis is divided into seven chapters.

Chapter 2 explains the prerequisite notions for the rest of the thesis.

Chapter 3 focuses on the leveraged datasets. It starts with a detailed description of the data sources. Then explains the pre-processing steps applied to the original mobility traces to uniformize the temporal and spatial granularity. Following, it presents the followed procedures to filter the effects of noising data and too much missing data.

Chapter 4 first attempts to answer the following question *what is exploration?* Following, it proposes a mobility profiling based on users' inclinations to explore. To sustain its profiling, it investigates the mobility traits of each profile. Next, it goes further and reports the profiles to the spatial and temporal exploitation for each of return and exploratory visits (**P**$_3$).

Chapter 5 evaluates the impacts of novelty-seeking on the two most widespread prediction formulations, namely the next-cell and the next-place prediction (**P**$_2$).

Chapter 6 proposes a new exploration-aware mobility predictor that utterly relies on the location data of the considered users. It first forecasts the next type of movement (an exploration or a return). Next, depending on the forecasted movement, it gives either a fine-grained intuition on the next location of the user in case of a return, else it provides a coarse-grained spatial unit in case of an exploration (**P**$_1$).

Chapter 7 provides overall conclusions and directions for future work.

Chapter 2

Background

Contents

The availability of large amounts of data capturing individuals' whereabouts enables the study and forecast of human mobility behavior. This chapter introduces concepts, related works, and necessary knowledge that will be used later in other chapters. It first outlines the main types of data sources available for mobility research. Second, it reminds the reader of core concepts, general metrics, and the fundamental statistical laws governing human motions. Next, it conducts a literature review on individual mobility models. Afterward, it outlines the different formulations of the mobility prediction task and presents the seminal predictability estimation method of mobility traces. Then, it describes some of the classical predictors adopted in the following chapters and provides a literature review on mobility prediction. Last but not least, it highlights the existing heterogeneity in human mobility behavior. Special attention is given to the exploration phenomenon and its impacts on thoroughly understanding, modeling, and predicting human mobility.

2.1 Data Sources

The surge of new forms of locomotion such as trains, planes, or cars enabled individuals to cover large distances on a daily basis. This engenders a major turn in mobilities and social science [31] and urges scientists to employ several instruments for data gathering ranging from traditional surveys to the most accurate satellite-based technologies to investigate and understand human motion. This section outlines the main types of data sources employed for mobility research and discusses their efficiency in capturing human mobility behavior.

2.1.1 Census Data

A census is an acquisition, recording, and enumeration of information about the population, typically household structures, workplace location, mode of transport, travel time, and housing data. Census data allow large-scale statistical investigations but are available at different temporal and spatial resolutions, i.e., coarse-grained data. For example, in commuting flows, the covered area can range from a home location to a whole city or state [13]. The data collection is usually held at least once every ten years by national statistics agencies such as the United States Census Bureau. The aggregated data flows are accessible to any interested entity, notably media, researchers, government, and businesses. For instance, the United States Census Bureau made available on request various population, household, economy, and business aggregated data flows [1]. Although the coarse-grained nature of census data, they play an invaluable role in studying human mobility and validating group-level flows [13]. For instance, Simini et al. [81] employ such data to propose a model that reproduces transport patterns and explains population flows.

In short, census data are large-scale data, publicly available, but coarse-grained, i.e., does not capture precise information on where and when people move [70].

2.1.2 Surveys

Surveys are traditional instruments for data gathering but essential tools for understanding human mobility behavior. Surveys' data acquisition campaigns allow the collection of detailed information such as the purpose of trips, social context, or answers to personality tests [88, 54]. However, the data is collected solely on specific users within small geographical areas, such as workers of a company or students of a university. Moreover, surveys are strongly interconnected to errors. For instance, many people have acquiescent personalities and are more likely to agree with statements regardless of the content [69]. Therefore, the statistics induced from surveys can be called into question and are often supported by other types of data. For example, Stopczynski et al. [88] utilize surveys in addition to Facebook, sensors, and WiFi data to relate human mobility to social interactions.

[1] https://www.census.gov/data/academy.html

Succinctly, surveys' data require a gathering campaign, provide short-term detailed information about the users and their spatiotemporal displacements, but generally have low numbers of respondents.

2.1.3 Dollar Bills

In 1998, currency tracking websites, capturing the natural geographic circulation of American banknotes began to gain popularity. The track of a banknote starts after an individual registers it in the database on a currency tracking website, i.e., the individual enters her local ZIP code, the serial number of the bill, and the series designation (the year appearing on the front of a banknote). When a registered banknote is re-entered, a *hit* occurs such that the time and the distance between the previous and current registrations are recorded in the database. The statistics generated by currency tracking websites are broadly used in analyzing human mobility [13]. More specifically, the movement of a banknote can be perceived as an individual's travel. Thus, the flux of money between a set of cities is proportional to the flux of individuals. For instance, Brockmann et al. [19] manipulate dollar bill tracks to investigate the characteristics of human mobility. The authors efficiently derive substantial laws explaining the distributions of the distances walked by the individuals and the elapsed time between consecutive movements.

In summary, the circulation of banknotes gives an aggregate view of human motion over large geographical areas and the data are publicly available. However, some hops can be missing, and thus it is only an approximation of the circulation flow.

2.1.4 Call Detail Record (CDR)

A CDR is a record enclosing various information about a phone activity and is saved in Mobile network operator databases. A record can be generated by different phone activities: a phone call, an SMS exchange, or a mobile data session. Typically, an instance contains the starting time of the activity (timestamp), identifiers of the initiator and the receiver, duration of the activity, type of activity, and the location of the cell tower to which the initiator's device is connected. CDR data sources leverage large numbers of users and allow per-user mobile activities analyses. Hence, it is possible to infer an individual's mobility trajectory and to offer relevant statistical outcomes about her mobility behavior. CDR data are extensively used to study human mobility [46, 86]. Namely, the paper of González et al. [46] employs CDR data to efficiently capture fundamental features characterizing human movements. Nevertheless, such datasets are not publicly available, and their acquisition requires a non-trivial process due to privacy issues [89]. The quality of the data depends on the frequency of phone activities, which implies that some trips and displacements can be overlooked. Besides, the recorded locations are the cell towers' and not the users', and given the large areas covered by cell towers (10 m to a few kilometers) [13], trips occurring within a cell area are not well captured.

In short, CDR data provide a per-user mobility capture, and depending on the frequency of phone activities, the mobility trace is more or less illustrative. Nonetheless, such datasets are not publicly available.

2.1.5 Global Positioning System (GPS)

The GPS is a satellite-based radio navigation system consisting of 31 satellites carrying clocks that constantly transmit signals. Each signal carries the position of the emitting satellite and its local time. GPS receivers, on the contrary, do not transmit any data (most of the time) [4]. When activated, the reception of three distinct signals allows a receiver to determine its geographical location. More precisely, the receiver computes the time difference between its local time and the emission time of the received signals to infer its localization [4]. Nowadays, GPS-equipped devices are omnipresent in our daily grinds, such as smartphones or fit bracelets. This allows accurate captures of individuals' whereabouts and hence thorough temporal and spatial investigations of human mobility behavior. For instance, using high spatial and temporal resolutions GPS dataset, Lin et al. [60] show that the predictability upper bound of individuals' trajectories increases with the increase in the temporal resolution.

In a nutshell, GPS traces represent an invaluable source for studying human mobility. They usually provide fine-grained temporal and spatial resolutions. However, such data sources contain sensitive information and are not publicly available. Although some projects try to collect GPS data, they only involve small numbers of volunteers within specific communities as in surveys, for instance, university students or research laboratory members.

Although multiple types of data sources are employed in mobility research, the current understanding of human motions is still partial, and data is one of the main limiting factors. Whereas census and CDR datasets enable large-scale but sparse investigations, surveys and GPS datasets provide detailed and comprehensive information but on a substantially smaller scale. The best scenario would be to have a large-scale GPS dataset. Nonetheless, such datasets contain very sensitive information and are rarely made available for researchers. Moreover, even though large-scale GPS datasets can be acquired after a challenging inquiry process, other dimensions than the temporal and the spatial are necessary for a thorough understanding of human movements, such as social aspects [28] or other contextual features [32].

2.2 Understanding the Mechanisms Governing Human Mobility

Is human mobility simply random? Or are there fundamental laws ruling our mobility? This is an essential question that calls the attention of several disciplines,

including anthropology, economics, geography, physics, computer science, and so on [19]. This section starts with a presentation of the principal metrics commonly used to capture human movements' dynamic and statistical properties.

2.2.1 General Metrics

To unveil human mobility and understand the intrinsic characteristics underlying people's movements, several quantitative features are used. In the following, we outline the most relevant and used ones.

- **Displacement** Δr: also referred to as trip distance, flight length, or jump size. It is a primary metric used for modeling human movements. Technically, the displacement Δr is the Euclidean distance covered by an individual in a given period of time δt [19]. It is given by,

$$\Delta r = |r(t + \delta t) - r(t)|, \tag{2.1}$$

  where $r(t)$ is the vector containing the geographical coordinates of the location where the individual was at time t. This metric allows the short-term (when δt takes small values) or long-term (when δt takes large values) capture of movement activities of individuals. It exhibits the dynamics of human movements in space and time.

- **Waiting Times** Δt: also known as pause time, it is defined as the elapsed time between two consecutive transitions, i.e., changes of locations [19]. It reflects the time spent by an individual visiting a location. This metric allows the identification of meaningful locations of the individuals, i.e., where they spend substantial parts of their time, such as workplaces or home locations.

- **Radius of Gyration** r_g: some individuals are inherently inclined to travel long distances on a daily basis, while others have most of their movements concentrated in a specific area and have confined mobility [46]. To quantify this characteristic, the radius of gyration r_g is used,

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_{cm})^2}, \tag{2.2}$$

  where N is the total number of records of the mobility trace of the considered individual, r_i is the vector containing the geographical coordinates of the i^{th} record, and $r_{cm} = \sum_{i=1}^{N} \frac{r_i}{N}$ is the center of mass of the individual. This metric captures the diffusion of an individual relative to her center of mass.

- **Number of Visited Locations** $S(t)$: it is defined as the number of distinct locations that an individual has discovered up to time t [85]. This metric allows the quantification of the diversity of human visits.

- **Mean Square Displacement** $MSD(t)$: it is used to characterize the diffusion of the walkers in space with time. It measures the deviation of the walker's position compared to an original position r_0 [85]. It is given by,

$$MSD(t) = \langle (r(t) - r_0)^2 \rangle \equiv \langle \Delta r(t)^2 \rangle. \tag{2.3}$$

  The MSD and radius of gyration are related but distinct metrics. While the former captures the *diffusion* in space relative to the center of mass, the second determines the *typical distance* walked compared to the center of mass [13].

- **Visitation Frequency** f_k: it is the probability for an individual to visit a location k [85]. This metric captures regularity as well as the diversity of visits. Regularity relates to highly frequented locations, whereas diversity is ascertained through the number of distinct visited places.

2.2.2 Scaling Properties of Human Mobility

Founded on the general metrics outlined above, several discoveries of human movements' mechanisms over various spatial and temporal scales are reported [19, 46, 85]. We present the major statistical features observed in real mobility traces in the following.

- **Heavy-Tailed Displacements** and **Waiting-Times**: several studies examine the discrepancy of the displacement Δr employing various types of datasets. Brockmann et al. [19] find that the Probability Distribution Function (PDF) of jump size $P(\Delta r)$ for the trajectories of dollar bills follows a power-law [2]. Likewise, González et al. [46] and Song et al. [85] measure the PDF of displacements of CDR datasets. They report that $P(\Delta r)$ is better approximated by a truncated power-law distribution [3]. Further, using GPS data sets, Zhao et al. [95] show that human mobility can be modeled as a mixture of different transportation modes (walk/run, car/bus/taxi). Each of these movement patterns can be approximated by a lognormal distribution. Additionally, they reveal that the mixture of the decomposed lognormal jumps distributions associated with each modality is a power-law distribution.

  Examining the PDF of the waiting-time $P(\Delta t)$, Brockmann et al. [19] reveal that it follows a power-law. Withal, Song et al. [85] report that $P(\Delta t)$ follows a truncated power-law.

- **Ultra-Slow-Growth of the Radius of Gyration** and **Heterogeneously Bounded Mobility Areas**: using CDR records, González et al. [46] show that the radius of gyration r_g increases logarithmically in time. Besides, they

[2] A power law distribution has the form $Y = kX^{\alpha}$, where X and Y are variables of interest, α is the law's exponent, and k is a constant [12].

[3] A truncated power-law distribution is a power-law distribution multiplied by an exponential [25].

measure the conditional jump length distribution $P(\Delta r|r_g)$. They reveal that in addition to small travels, users with a large radius of gyrations perform few large jump sizes. In contrast, users exhibiting a small radius of gyration limit the majority of their displacement to nearby areas.

- **A Slow-Down Tendency for Location Discovery**: Song et al. [85] and Alessandretti et al. [6] show that the number of places that an individual has discovered up to time t grows as $S(t) \sim t^{\mu}$, where $\mu < 1$. This displays a sublinear growth in the total number of unique locations visited by an individual, i.e., individuals have a decreasing tendency to discover new places.

- **Ultra-Slow Spatial Diffusion**: Song et al. [85] assert that the MSD becomes saturated after a certain period. They show that the MSD grows even more slowly than the logarithmic diffusion law. This ultra-slow diffusion is rooted in individuals' need to return to their favorite locations.

- **Preferential Return**: González et al. [46] and Song et al. [85] report that the visitation patterns of humans are uneven. Individuals exhibit a high spatial regularity; they spend their time between a few locations and rarely visit non-routine locations. It was found that the probability f for a user to visit her k^{th} most favorite location follows Zipf's law [4], $f_k \sim k^{-\zeta}$, where $\zeta \approx 1.2 \pm 0.1$ [85]. Therefore, the visitation frequency distribution follows, $P(f) \sim f^{-(1+\frac{1}{\zeta})}$.

2.3 Individual-level Mobility Models

Individual mobility modeling is the mathematical formalization of individuals' trajectories. Various models are developed to incorporate some of the fundamental statistical features characterizing human mobility. This section describes some of the most commonly used individual mobility models and highlights what statistical properties they integrate and how they deviate from real traces.

2.3.1 Lévy Flight

Lévy Flight is a classical Random walk model, and a heavy-tailed jump lengths probability distribution characterizes it. In accordance with previous findings of Brockmann et al. [19], Lévy Flight can be used to model human displacements. However, it is marked with a super-diffusive MSD, which contradicts the *ultra-slow spatial diffusion* property (cf. Section 2.2.2). Additionally, in the Lévy Flight model, short and long trips have the same duration, which is unrealistic. In particular, the model does not verify the *slow-down tendency for location discovery* property (cf. Section 2.2.2). The probability of visiting any *new* location is expected to be asymptotically uniform everywhere [45]. Therefore, this model does not accurately

[4] Zipf's law is a relation between rank order and frequency of occurrence. This means when observations (e.g., locations) are ranked by their frequency, the frequency of a particular observation is inversely proportional to its rank [5].

mimic human mobility patterns and can only be a rough representation of human movements.

2.3.2 Continuous-Time Random Walk (CTRW)

CTRW is a Random walk with two independent random variables, the displacement Δr, and the waiting time Δt, both of them having *heavy-tailed* distributions. This suggests its relevance for modeling human movements. Besides, this modeling can engender either sub-diffusive or super-diffusive behavior depending on the parameters of displacement and waiting time distributions. However, using two real-world mobility traces, Song et al. [85] report that the CTRW model is in conflict with the main statistical laws governing individuals' mobility (cf. Section 2.2.2).

- **An almost steady tendency for location discovery**: while Song et al. [85] and Alessandretti et al. [6] confirm the sublinear growth of the total number of unique locations $S(t) \sim t^{\mu}$, with $\mu \leq 0.66 \pm 0.02$ for different types of data sources, the CTRW suggests that $\mu = 0.8 \pm 0.1$. This indicates that the CTRW model overestimates the inclination of individuals to discover new places.
- **A uniform visitation frequency:** the CTRW as a Random walk model integrates the property of uniformity of visit [85]. This contradicts the preferential return property. Individuals have a strong tendency to return to a few locations such as home or workplace while rarely visiting others such as a theater or a bar [46].
- **Unremitting diffusion:** opposing the ultra-slow diffusive human behavior, the scaling of the MSD in the CTRW suggests that an individual will asymptotically drift away from her original location. Although the impromptu mobility behavior of individuals, the ingrained homecoming regularity slows down diffusion excessively [46].

The CTRW model is a simple, tractable, and scalable model. Yet, using it to model individuals' mobilities is incomplete and induces considerable deviations from empirical data.

2.3.3 Exploration Preferential Return (EPR)

Song et al. [85] propose considering the notions of *Exploration* and *Preferential Return* when modeling human movements. For explorations, they suggest that the probability of visiting a new location decreases with time and consider the following exploration probability, $P_{new} = \rho S^{-\gamma}$, where S is the number of distinct visited places, and ρ and γ are parameters of the model determined by empirical data. The walking distance, as well as the waiting time, are in this case chosen from the distributions $P(\Delta r)$ and $P(\Delta t)$, respectively. The direction is randomly chosen. For returns, the authors consider the complementary probability of P_{new}, $P_{ret} = 1 - \rho S^{-\gamma}$. Moreover, they integrate *the preferential return* property. They

quantify the preferential attachment as follows, the probability Π_i for an individual to visit a known location i is proportional to her number of visits to i, i.e., $\Pi_i = f_i$.

The proposed EPR model is able to overcome the deviations of Random walk models in reproducing individual mobility. Specifically, it generates *heavy-tail trip distance and waiting time* distributions. Besides, it can reproduce *the ultra-slow-growth* property of the radius of gyration as well as *the heterogeneity of the total covered areas* feature (cf. Section 2.2.2).

Although the EPR model made significant improvements in mobility modeling, the generated traces still suffer from inconsistencies compared to real-world ones. Namely, the preferential attachment property as defined in the EPR model leads to two discrepancies [14].

- **Cumulative advantage:** the preferential attachment property stipulates that the earlier a location is discovered, the more it is visited. In other words, an early-visited place will seemingly be one of the most visited ones.

- **Preserved preference:** the EPR model assumes that individuals' spatiotemporal regularities and routines remain stable. Though individuals' preferences evolve and change over time, some places are integrated into the regularly visited places set while others are withdrawn.

Besides, the EPR's exploration modeling is elementary and not realistic. First, although the decreasing tendency of discoveries of new places reported in the literature, a clear definition of *what is exploration* is missing. This means that the decreasing trend to explore can be biased. Second, in the EPR, the waiting time and the walked distance are independently chosen. This implies that the walked distance can be irrealistic [73].

Pappalardo et al. [73] propose an extension for the EPR model to overcome the latter issue (i.e., preserved preference). The next location where an individual is going to explore is chosen depending on both its distance from the current position and its social relevance. In their empirical analysis, the authors show that the social context plays a significant role in discovering new places and that individuals are attracted by popular locations when exploring. The authors report substantial improvements and show that their generated traces are closer to real-world data and can reproduce the empirical heterogeneity of mobility behavior.

Further improvements in modeling returning habits are proposed in the literature, such as the recency model proposed by Barbosa et al. [14]. Likewise, modeling exploratory visits are gaining more and more attention, such as the recent work of Cornacchia et al. [27] that models exploratory visits by simultaneously considering spatial, temporal, and social dimensions.

In this thesis, we seek to better understand and characterize exploration-like visits that can serve in exploration mobility modelings.

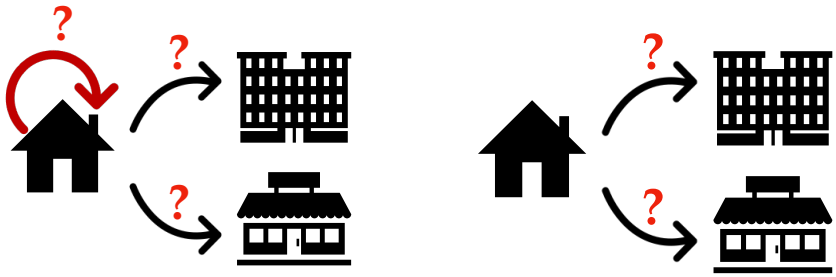
2.4 The Prediction of Human Mobility

Previous research has shown that human movements are far from being random [19], and the most startling is the significant regularity exhibited by individual trajectories. Individuals follow simple, reproducible patterns described mathematically on many spatiotemporal scales [46, 85]. This section outlines the most widespread prediction task formulations of human mobility. Next, it describes the most prevalent approach to measuring the theoretical predictability extents of human traces, adopted in the following chapters. Then, it reviews some of the prediction techniques developed to forecast future individuals' whereabouts. Finally, it lists the main factor impacting the predictability of human movement.

2.4.1 Mobility Prediction Tasks

There exist several ways to define the mobility prediction task depending on (i) the objectives of the forecasts and (ii) the contextual view, i.e., global/local view of the data.

According to the objectives of the forecast, human mobility prediction can be about (i) predicting the location of an individual within the next time-bin, i.e., *next-cell* prediction (cf. Figure 2.1a), or (ii) predicting the next location to which an individual will perform a transition, i.e., *next-place* prediction (cf. Figure 2.1b).



(a) Next-cell prediction.

(b) Next-place prediction.

Figure 2.1: Mobility Prediction Tasks.

- **Next-cell prediction:** Given an individual's sequence of timestamped travel events and considering a time window Δt, the next-cell prediction task attempts to answer the subsequent question, *where will the individual be at time* $t + \Delta t$*?* The triggering element is the time; after each period Δt, the system tries to forecast the future location of the individual (cf. Figure 2.1a). This prediction task is extensively adopted in the literature, with different time windows from a few minutes [60] to multiple hours [86, 19]. The accuracy of prediction achieved by this formulation can be very high, exceeding 90% [16, 32].

- **Next-place prediction:** This formulation is independent of the temporal dimension. It encompasses two main tasks: (i) predicting *when* an individual will make a transition (ii) predicting *where* the individual will go next (cf. Figure 2.1b). More precisely, the next-place prediction task aims at forecasting transitions between places [52]. Hence, the triggering element is the user's transition from her current location. The predictive performances achieved using this formulation are substantially lower than those achieved by the next-cell prediction with an accuracy of prediction of about 70%, mainly due to the elimination of self transitions [52, 32]. This prediction is one of the most challenging tasks [32, 93]. Admittedly, the stationary mobility behavior of the individuals is overlooked, i.e., the absence of consecutive records with the same location. Thus the system is more sensitive to the diversity of locations. Yet, predicting transitions between places can be more appealing for many domains.

According to the contextual view, the prediction approach can be (i) *personal*, i.e., solely based on the individual's whereabouts [86, 52, 62], or (ii) *joint*, i.e., by using both the individual's whereabouts and common patterns and information [43], [93], and [57].

- **Personal prediction:** In this prediction approach, the forecasts are solely based on self-information. Namely, considering the sequence of timestamped travel events of an individual, only her personal data are utilized to infer her future travel events. Personal-based models generally perform well in mobility prediction with dense and rich data sets [41, 62, 87, 44, 63]. Moreover, they are relatively privacy-preserving as they do not need a global view of the mobility traces, i.e., information of the surrounding individuals [38]. However, the predictive performances are highly impacted by the quality of the data. In particular, today's sparse and limited data records, such as in LSBN, are not always suitable with such high requirement approaches [38, 78].

- **Joint prediction:** In addition to the timestamped travel events of the considered individual, this prediction task requires a global view of the crowd and information about common mobility patterns to infer future travel events. Joint-based methods usually assume the existence of sub-groups sharing significant similarities in mobility behavior [10, 20, 7]. Following group identifications, the similarities observed within each group are used to support the forecasts of future travel events of individuals of these groups. This approach comes to overcome the difficulties encountered by personal-based methods with data quality [10, 38]. Promising performances are achieved with this approach. Nonetheless, these benefits come at the cost of centralized data training and hence raising privacy concerns [38].

2.4.2 Theoretical Predictability

Over the last decades, *entropy* has been extensively used to evaluate the randomness remaining in information flux [9]. *Entropy* is an information theory tool introduced by Shannon [80] that determines the minimum amount of storage and transmission required to give a complete description of the communicated data. It is defined as [80], $H(X) = -\sum_{x \in \mathcal{X}} p(x) \times \log_2(p(x))$, where X is a random variable taking values from the alphabet $\mathcal{X}$, $p(x) = Pr\{X = x\}$ is the probability that the outcome X equals x, $x \in \mathcal{X}$.

When the random variables form a *stationary process* [5], the entropy measure $H(X_1, X_2, \ldots, X_n)$ grows linearly with n at a rate $H(\mathcal{X})$. The rate $H(\mathcal{X})$ is noted as *entropy rate* and is interpreted as the best achievable data compression [29]. Given the broad use of stationary processes as mathematical models of various systems and phenomena [86, 18, 40], entropy rate arises as a natural and essential measure of uncertainty in a wide range of applications, notably in human mobility [86]. Entropy and its extensions are also used to measure *predictability*, that is, quantifying how much observations from the past can tell us about the future [15]. In practice, entropy-based predictability is proven to be one of the most effective tools to estimate predictability [36].

In the field of human mobility, Song et al. [86] establish the foundation to measure the predictability of human mobility trajectories. In what follows, we present the methodology proposed by Song et al. [86] that is extensively used in mobility research and *adopted in the coming chapters*.

Let $\mathcal{X} = \{X_i\}$ be a stationary stochastic process representing an individual's mobility trace of n records. Where X_i is a random variable describing the location of the individual at time i. The entropy rate $H(\mathcal{X})$, also denoted S can be written as [86],

$$S \equiv \lim_{n \to \infty} \frac{1}{n} S(X_1, X_2, \ldots, X_n). \tag{2.4}$$

By applying the chain rule to Eq. (2.4) and noting $S(n) \equiv S(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$, the entropy rate can be written as [29],

$$S \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} S(i). \tag{2.5}$$

Entropy rate is usually referred to as entropy. Subsequently, in the absence of confusion between the two concepts, as we only deal with mobility traces that are assumed to be stationary processes, we employ entropy to refer to entropy rate.

To evaluate the randomness present in mobility trajectories, Song et al. [86] assign three entropy measures to each individual having a mobility trace of size n with N distinct locations:

[5] A stationary process is a stochastic process that does not change its statistical properties with time [74].

- **Real entropy** S: It takes into account both the temporal order as well as the visitation frequency to the distinct locations (Eq (2.5)). The real entropy S is estimated by, $S^{\text{est}} = \left(\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\right)^{-1} \times \log_2(n)$ [55]. Λ_i is the length of the shortest subsequence that did not appear in the mobility sequence from index 1 to $i-1$, and which appears for the first time starting from index i.

- **Temporal-uncorrelated entropy** $S^{\mathbf{unc}}$: It captures the regularity of the individual in space, whereas the temporal dimension is overlooked, i.e., the order of visit is not taken into account. When the probability of the next location is independent of the current one, Eq. (2.5) is equivalent to $S^{\text{unc}} = -\sum_{i=1}^{N} p_i \times \log_2(p_i)$ [86], where p_i is the probability of visiting the location i.

- **Random entropy** $S^{\mathbf{rand}}$: This estimation stipulates that each location i in the mobility trace has the same probability of being visited, i.e., $p_i = \frac{1}{N}$. It is given by $S^{\text{rand}} = \log_2(N)$.

In reliance on the entropy measures, Song et al. [86] propose a method to estimate the upper bound predictability $\Pi^{\max}$ of human trajectories by solving the following equation, $S = H(\Pi^{\max}) + (1-\Pi^{\max})log_2(N-1)$ [86].

Leveraging a three-month-long CDR traces of 50,000 users, Song et al. [86] measure the three aforementioned versions of entropy (i.e., S real, S^{unc} temporal-uncorrelated, and S^{rand} random) and the corresponding upper bounds predictability ($\Pi^{\max}$, Π^{unc}, and Π^{rand}). They reveal a striking lack of variability in human mobility patterns. Notably, the distribution of the real entropy S shows a peak at around 0.8, indicating that if an individual chooses her next location randomly, it can be found on average among two locations $2^{0.8} \approx 1.74$. Moreover, $P(\Pi^{\max})$ has a peak at around 0.93, which means that only 7% of the time an individual's movements appear to be random. This has led to growing interests in understanding and predicting human mobility.

2.4.3 Practical Predictability

Building upon the above findings, many advanced *predicting algorithms* are designed attempting to approach the upper bound predictability, such as Markovian predictors [41, 62], Bayesian network models [43, 7], neural network algorithms [58], or advanced deep learning approaches [90]. In this section, we briefly present some of the practical prediction techniques that *we adopt in the following chapters*. Subsequently, we review the state-of-the-art on human mobility prediction.

2.4.3.1 Mobility Predictors

This thesis uses two categories of mobility predictors, namely, Markov-based methods and compression-based methods. Markov-based methods include Markov Chain

(MC) and compression-based methods include Prediction by Partial Matching (PPM), Sampled Pattern Matching (SPM), and Active LeZi (ALZ).

MC: It is a stochastic process that consists of a sequence of possible *states* and a *transition matrix* describing its states' changes. An MC satisfies the Markov property, i.e., predictions regarding future outcomes are solely based on the present state. The memory of an MC can be increased to include more past states k for its predictions and are referred to as $k-$order MC [41]. MCs are extensively used as statistical models of real-world processes. Namely, Liao et al. [59] and Song et al. [86] suggest using MCs for the modelization of human mobility traces. More precisely, a user's mobility trace can be viewed as a sequence of random variables $\{X_1, X_2, \ldots, X_n\}$, where a *state* corresponds to a location X_i, and a *transition* corresponds to the probability of moving from one place to another [86, 41]. The prediction of the current location of X_i is performed by building a transition matrix of probabilities and solving the following maximization problem: $x_i^* = \arg\max_x P(X_i = x | x_{i-1}, \ldots, x_{i-k})$, where $x_{i-1}, \ldots, x_{i-k}$ are the k previously visited locations [86]. Several methods are employed for the estimation of the probabilities, such as Markov Chain Monte Charlo (MCMC)-based estimators [30, 2, 59] or Maximum Likelihood Estimation (MLE) [87]; in this thesis, we employ the latter. MCs are simple to use, efficient, and have low computing costs [59, 41].

PPM: It is a widespread statistical data compression method. Considering a *sequence* of symbols, it utilizes variable size context to predict the next symbol or context [64]. A $k-$order PPM is a combination of MCs of orders ranging from k to 0 and an escape mechanism [64]. To predict the next symbol, the longest previously encountered context is used. If the current symbol is new in the context, an escape code is transmitted, and the size of the context is shortened by dropping one symbol. The escape mechanism and shortening of the context are repeated until reaching a match or a case with no match, i.e., $m = 0$ [26]. In human mobility prediction, the mobility trace of an individual can be viewed as a *sequence* of symbols, in which each symbol represents a location. In this thesis, we use the PPM scheme implementation proposed by Moffat et al. [64].

SPM: Jacquet et al. [53] propose a predictor based on pattern matching, referred to as SPM. Given a sequence of symbols $X_1, X_2, \ldots, X_n$, the algorithm starts by identifying the maximal suffix $X_i, X_{i+1}, \ldots, X_n$ that occurs earlier in the sequence, such that $X_i, \ldots, X_n = X_{i-j}, \ldots, X_{n-j}$ for a $j \in [1, n]$. The SPM scheme is similar to a $k-$order MC, but instead of using a context of length k, the suffix of interest is a fixed fraction α of the maximal suffix. This means that only an $\alpha \in (0, 1)$ fraction of the $X_i, X_{i+1}, \ldots, X_n$ is employed to predict the future context. In the field of mobility, individuals' traces can be viewed as sequences of symbols [87].

ALZ: It is a sequential scheme based on the LZ78 data compression algorithm [100] developed by Gopalratnam and Cook [47]. The ALZ algorithm can be viewed as a $k-$order MC that incorporates a sliding window approach. Specifically, to predict the next symbol of a sequence $X_1, X_2, \ldots, X_n$, ALZ maintains a window of the

immediately preceding symbols. The size of the window k is chosen to be equal to the length of the longest suffix seen in a classical LZ78 parsing. This means that, first, the sequence of contexts is parsed into $s(n)$ sub-contexts $Y_1, Y_2, \ldots, Y_{s(n)}$, where each Y_i is the extension of an Y_j sub-context with $1 < j < i$ [100]. Then, the length of the longest sub-context Y_i is taken as the size of the window. Following, within the window, ALZ gathers statistics on all possible contexts. This allows building a better approximation to the $k-$order MC [47]. In human mobility prediction, the ALZ manipulates mobility traces as sequences of symbols.

2.4.3.2 Literature on Mobility Prediction

Existing individual-level predictors seek to forecast the future location of a given user [41, 62, 10, 63]. They can be classified into two categories: *personal* or *joint* (cf. Section 2.4.1).

To enable the comparison and the evaluation of mobility predictors, a common metric is established, the accuracy of prediction, it is given by,

$$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}}. \tag{2.6}$$

Markov-based models are common models for personal predictors. Gambs et al. [41] propose an MC predictor that exploits the n previously visited locations to predict future visits. Using GPS mobility traces, they reveal that the next location can be predicted with a markedly high accuracy of 70% – 95%. Likewise, papers [62] and [87] use Markov-based models. The former uses a varying order MC model, where the order is high when the historical information is limited and becomes smaller when the historical trajectory is over 100 points. The achieved accuracy of prediction with a first-order MC model on a CDR dataset surpasses 90%. The latter exploits several methods 0-order MC, LeZi, PPM, and SPM to infer future locations based on past history. They show that *Markov predictors work as well or better than more complex compression-based predictors* and report an accuracy of prediction between 65% and 72% for Dartmouth's campus-wide Wi-Fi wireless network.

In contrast to the works mentioned above that tackle the next-cell prediction problem, Gidofalvi et al. [44] consider the next-place prediction (cf. Section 2.4.1). The authors propose an inhomogeneous continuous-time Markov model to predict when a user will leave her current location and where she will move next. The evaluation of the predictive performance with a GPS dataset reports that the accuracy of prediction is above 67%. Other more complex methods are employed for personal predictions. For instance, Mathew et al. [63] propose a hybrid approach that first clusters the locations according to their characteristics (temporal period in which they occurred), then trains HMM on each cluster. To predict the next visited place, the model starts by identifying the most likely cluster then infers the future location using the corresponding HMM. The authors measure a prediction accuracy of about 13.85% with GPS trajectories. A further example, Feng et al. [37] employ the Kalman filter to predict the future location of a vehicle.

Nonetheless, these conventional personal models fail to predict *visits to new locations* [93, 32]. Hence, several joint methods that train on aggregated data are more and more employed to enhance the predictive performance of individuals' mobility. Asahara et al. [10] propose a Mixed Markov chain Model (MMM) that first classifies the users into groups of similar users. Next, to predict the future location of a user, it trains an MC predictor for all users of the group to which she belongs. The authors compare the performance of the MMM with the MC and HMM models. They report an accuracy of prediction of 16.9% (45.6%) for MC, 4.2% (2.41%) for the HMM, and 74.1% (64%) for the MMM when predicting transitions between locations with simulation (real-world GPS) data. Calabrese et al. [20] introduce a predictor that combines an individual's past mobility choices with collective behavior to help (i) predict the likelihood the user changes location and (ii) infer the type of geographical areas that should be similar to the type that interest the collectivity at the given time. Using a CDR dataset, the authors report a good level of accuracy in terms of prediction error (60% of the errors are zero). Alhasoun et al. [7] propose a Dynamic Bayesian Network approach that couples friends "similar strangers" records to increase the accuracy in predicting the next location. They report an accuracy of prediction of 60.03% by relying on timestamped geographical data in addition to social contact contextual information.

While most of the previous works base their forecasts on timestamped geographical data, the recent availability of enriched geo-tagged datasets with various contextual information brings new opportunities to enhance the prediction task [32]. Nonetheless, such data sources are not publicly available. Moreover, with the emerging requirements of privacy protection, handling such data raises serious privacy concerns. Therefore, scholars strive to consider privacy issues straightforwardly in the modeling and system or bypass them by employing the least possible data [17, 91].

2.4.4 Factors Impacting Predictability and Prediction

Several studies attempt to dig out the significant factors that affect the ability to foresee human mobility and shed light on the origins of the limitations in predicting the next location:

Spatial and temporal resolutions: Jensen et al. [16] examine the upper bound predictability Π^{max} using various types of mobile sensor data. Namely, Global System for Mobile Communications (GSM), WLAN, Bluetooth, and acceleration of 48 days' records for 14 individuals. The distributions of the predictability of all datasets peak at values larger than 0.8, corroborating the high predictability extent of individuals' whereabouts reported by Song et al. [86]. Moreover, they evaluate the effects of varying the temporal resolution from a few hours to a few minutes and report that the highest predictability is achieved with a time window of 4 min to 5 min.

Likewise, using high-resolution GPS traces of 40 individuals, Lin et al. [60] investigate the effects of the spatial and the temporal resolutions on predictability

extent. They show that with a time window of one hour and spatial units of size equal to 1/20 of the average coverage area of cell towers ($3\,\mathrm{km}^2$), $\Pi^{\max}$ can be as high as 0.9. By varying the temporal window, they report surprisingly high scores for $\Pi^{\max}$, exceeding 0.98 with time windows of 20 min or less. Conversely, shrinking the size of spatial units leads to a decrease in performance.

Using cells of size of $563\,\mathrm{m} \times 563\,\mathrm{m}$, Smith et al. [83] compute the upper bound over six temporal quantizations and attest that the highest performance is achieved with the smallest time window, namely, 5 min. In contrast, the predictability $\Pi^{\max}$ drops considerably with the shrink of the size of the cells and approaches 75% for cells of size $\sim 35\,\mathrm{m} \times 35\,\mathrm{m}$.

Type of prediction: Using GPS traces, Ikanovic et al. [52] investigate the predictability extents with the next-place prediction task (cf. Section 2.4.1). They show that the predictability $\Pi^{\max}$ is significantly lower (peaking at around 0.71) than previously reported with the next-cell formulation.

Further, employing GPS traces, Cuttone et al. [32] evaluate the maximum achievable predictability with both next-cell prediction and next-place prediction formulations. They find that the maximum predictability for the next-cell prediction peaks at around 0.95, whereas it peaks at a substantially lower value, 0.68, for the next-place prediction.

Throughout their investigations, Ikanovic et al. [52] and Cuttone et al. [32] reveal that the low predictability achieved with the next-place prediction results from the removal of stationarity effects. Put differently, irregular visits, particularly *discoveries of new places*, lower the potential predictability.

Novelty-seeking: Recent studies show the importance of considering individuals' tendencies to explore and to discover new locations when predicting their mobility [32]. However, quantifying and evaluating the impacts of exploration-like visits on prediction are yet to be addressed and researched.

Human trajectories show a high degree of potential predictability. This inspires scientists to design and develop advanced predictors able to forecast individuals' future whereabouts. However, several factors can impact predictability, such as the quality of the data, the adopted prediction formulation, and, most importantly, individuals' preference for discoveries of new places. Whereas the formers have been widely investigated, the impacts of exploratory visits on prediction have rarely been addressed.

2.5 Heterogeneity of Routine, Diversity, and Exploratory Visits

Although the reported scaling properties ruling human movements, human mobility behavior is heterogeneous. For instance, through their analysis of regularity and

diversity of visits, Pappalardo et al. [73] reveal the existence of two distinct mobility profiles: $n-$*explorers* and $n-$*returners*, where n is the number of the most significant locations considered for the dichotomy. $n-$Explorers are characterized by a dynamicity of visits; they visit more than n distinct locations day-to-day. $n-$Returns are marked by a steadiness of visits; they limit their mobility between a few places (less than n on a daily basis).

Likewise, Naghizade et al. [68] propose a movement diversity measure TRip DIversity Measure (TRIM) able to distinguish between regular individuals and irregular ones. Specifically, given an origin and destination pair of locations, the proposed TRIM metric quantifies the regularity of an individual's movements by comparing the prefix tree of all the trips and the theoretical maximum trip diversity modeled using a theoretical full prefix tree. The proposed metric can capture the diversity of movements as well or better than the combination of several state-of-the-art metrics, such as the radius of gyration or the displacement.

Supporting the existing heterogeneity in human visits, particularly in terms of types of movements, Scherrer et al. [78] reveal the existence of two main categories of users (i) *travelers*, who are active and have a more diffused mobility, and (ii) *locals*, who roam in small and compact areas and move slowly. The authors propose a data-driven approach to identify sub-groups of individuals displaying similar mobility behavior. They derive several features commonly used in mobility research from the spatiotemporal footprints of the individuals, such as the number of stops, the number of visited locations, or stop duration. Then apply the unsupervised learning PCA approach on the extracted features to identify the most significant clusters.

Several mobility profiling approaches are proposed in the literature [10, 24]. Nevertheless, most methods are based on the regularity or diversity of visits, such as the dichotomy proposed by Pappalardo et al. [73] and the regularity quantification introduced by Naghizade et al. [68]. On the contrary, Scherrer et al. [78] propose a data-driven approach but do not provide insights on the exploration phenomenon or the related features.

2.6 Summary

This chapter provided the necessary background to understand individual mobility and highlighted the exploration phenomenon as being a main limiting factor in understanding, modeling, and predicting individuals' trajectories. First, it started with an overview of the data sources used in mobility research. Next, it outlined the general metrics commonly used to describe human movements as well as the statistical properties governing human mobility. Then, it presented state-of-the-art mobility models. Following, it gave an overview of the theoretical and practical predictability of human mobility traces. Finally, it emphasized the existing heterogeneity in routine, diversity, and exploratory visits.

In the next chapter, we present the leveraged data sources as well as the preprocessing procedure followed to prepare the data.

Chapter 3

Description of the used data

Contents

Our attempt to understand individuals' proclivity to explore relies on the availability of datasets. For an utter capture and study of novelty-seeking tendencies, it is highly recommended to leverage large-scale data sets or multiple data sources with detailed information on individuals' whereabouts. Such data sources ensure a reasonable assortment of human mobility trajectories and hence are eligible for statistical validations and generalization of the drawn conclusions. In this context, this thesis employs several fine-grained datasets.

This chapter starts with a description of the used data sources. Next, it presents the temporal sampling and spatial tesselation applied to the data to have uniform and comparable features. Then, it explains the procedure ensued to complete missing records. Finally, it outlines users' filterings.

3.1 Description of the Used Data

This thesis uses two categories of real-world mobility traces; three GPS datasets and one CDR dataset. The characteristics of the datasets are detailed in Table 3.1.

Table 3.1: Datasets description.

Dataset	Type	Users	Duration	Frequency	Records
Macaco [1]	GPS	132	34 months	5 min	900 k
Privamov [65]	GPS	100	15 months	few seconds	156 M
Geolife [97, 99, 98]	GPS	182	64 months	1 to 5 seconds	900 k
ChineseDB [23]	CDR	642k	14 days	1 h	150 M

3.1.1 GPS Data Sources:

Mobile devices equipped with at least proximity sensors allow tracking individuals' movements with the highest level of accuracy and temporal frequency [42]. In this work, three GPS datasets are used:

- ***Macaco Dataset*, From the European CHIST-ERA Macaco Project:** The dataset was collected by the EU CHIST-ERA MACACO project [1] through the Android crowdsensing mobile phone application MACACOApp. It collects the digital activities of 132 volunteers who gave informed consent from 6 different countries. The project follows the French Commission Nationale de l'Informatique et des Libertés privacy enforcement rules to warrant prticipants' privacy. It provides long-term and fine-grained digital activities logged with a fixed periodicity of 5 min. The data collection took place between May 2015 and April 2018 (34 months) and comprises more than 900k GPS tuples. Each GPS record (dev_id, timestamp, lat, lon) includes a user ID, a timestamp, and location information, i.e., GPS coordinates. For project-related privacy policies, this dataset is *not publicly available*.
- ***Privamov Dataset*, From the French Priva' Mov Project:** This *private data* was collected by the Priva'Moav project [65] through a crowdsensing application exploiting several sensors, namely, GPS. The crowdsensing campaign spans 15 months, from October 2014 to January 2016, with a frequency of sampling roughly equal to a few seconds. It contains around 156 million GPS records of 100 volunteers from the city of Lyon in France. Every tuple consists of four parts, an anonymized user ID, the date of collection, the latitude, and the longitude.
- ***Geolife Dataset*, Collected by Microsoft Research Asia:** Geolife is a *publicly available dataset* released by the Microsoft Research Asia project [97, 99, 98]. The dataset stores information about the mobility traces of 182 individuals broadly distributed in over 30 cities, mainly in China, the USA, and Europe. The project provides GPS tracks with varying frequency of sampling from 1 s to 5 s. The data collection spans more than 64 months time period, from April 2007 to August 2012, and contains more than 926k GPS tuples. In the dataset, a GPS trajectory is represented by a sequence of time-stamped location data, i.e., latitude and longitude.

3.1.2 CDR Data Sources:

Nowadays, mobile phones are ubiquitous technological devices in our day-to-day life. They offer a good proxy for capturing human footprints and hence the study of their mobility patterns [72]. We use one CDR dataset captured through phone calls and message exchanges.

- **ChineseDB Collected by Shanghai University:** This is the most significant dataset used in this thesis. It was collected by a major network operator

in China [23]. The data source gathers the mobile phone activities of 642k anonymized mobile phone subscribers from a central area in Shanghai. Per-user, the experiment merges the locations collected within each hour to ensure a frequency of one sample per hour. Each location of an hour represents the user's centroid within it, with the precision of 200 m according to the instruction of the data provider. It provides a higher spatial resolution compared to classical CDR data. The data collection covers a period of 14 days and comprises more than 150M tuples. Each tuple record contains the identifier of the user, the date of collection, and the location coordinates, i.e., the latitude and the longitude of the centroid of the locations where the user was for the last hour.

3.2 Spatial Tessellation and Temporal Sampling

We focus on the location data, i.e., latitude and longitude. Hence, we reconstruct the mobility trace $\mathcal{H}_u$ of each user u of the leveraged datasets by extracting the sequence of recorded locations along with the associated timestamps at fixed time periods δ, $\mathcal{H}_u = \langle (t_0, lon_0, lat_0), (t_1, lon_1, lat_1), \dots (t_n, lon_n, lat_n) \rangle$, where $t_i = t_0 + i \times \delta$.

Spatial Tessellation: Due to GPS and CDR range errors, locations are usually defined by spatial grid IDs or Points Of Interest (POI). Following customary practice [32], we superpose uniform grids of size c m $\times$ c m on the geographical maps (see Figure 3.1).

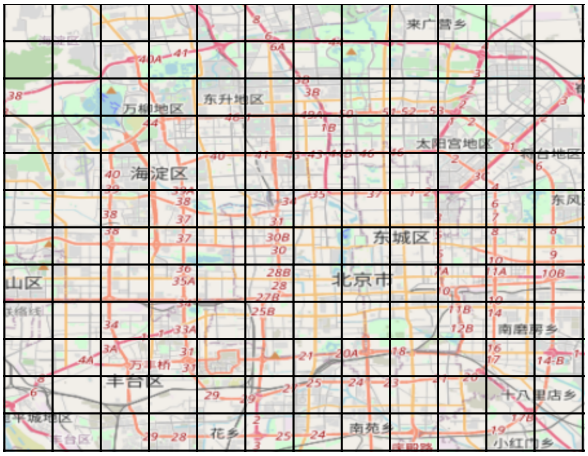


Figure 3.1: An overview on the partitioning of the city center of Beijing into squared cells of size 2 km$\times$ 2 km.

Next, we project the locations coordinates (lat_i, lon_i) to convert them into spatial grid IDs $(ID_i = \lfloor \frac{lon_i}{c} \rfloor, \lfloor \frac{lat_i}{c} \rfloor)$. Hence, the mobility trajectory of the individual u is converted into sequences of timestamped discrete symbols –a discrete mobility trajectory–, $\mathcal{T}_{u,c} = \langle (t_0, ID_0), (t_1, ID_1), \dots (t_n, ID_n) \rangle$.

Temporal Sampling: Recall that in each GPS data source, the locations data of the users are obtained at different temporal rates. To enable the comparison between outcomes of the different GPS datasets and ensure the validity and generalization of our conclusions, we re-sampled all the GPS data sources to have an *equal frequency of one sample every* δ_{GPS} *min.* For the CDR dataset, we use δ_{CDR} h, which values are specified in Section 7.3.

3.3 Data Completion

In CDR data, the sampling rate is inherently dependent on the phone activities (calls and messages) frequencies of an individual. Thereby, the mobility information provided by CDR is usually incomplete [76, 39]. Likewise, some records can be missing in the GPS datasets due to delayed measurements produced by the sleeping phases of mobile devices collecting the data [1]. Therefore, to enlarge the availability of human footprints and to have more *uniform and complete traces*, we comply with some steps proposed by Chen et al. [23] and complete with the most significant and visited locations as follows,

- **Workplace A:** Per individual u, we identify if existing the most frequent daily location ID_{wp_A} between 10 am and 11 am and name it *workplace A*.
- **Workplace B:** When possible, we locate the most visited location ID_{wp_B} between 2 pm and 5 pm and name it *workplace B*.
- **Home location:** If available, we determine the most prevalent place ID_h between 2 am and 6 am (night), which we refer to as *home location*.

Once *home* (ID_h), *workplace A* (ID_{wp_A}), and *workplace B* (ID_{wp_B}) locations are identified:

- If a record is missing at t_x between 10 am and 11 am and $ID_{wp_A} \neq \emptyset$, we complete the trace $\mathcal{T}_{u,c}$ with a new record (t_x, ID_{wp_A}).
- If a record is missing at $t_x \in [2\text{ pm}, 5\text{ pm}]$ and $ID_{wp_B} \neq \emptyset$, we add the tuple (t_x, ID_{wp_B}) to the mobility trajectory $\mathcal{T}_{u,c}$.
- If a record is missing at $t_x \in [2\text{ am}, 6\text{ am}]$ and $ID_h \neq \emptyset$, we add to the mobility trajectory $\mathcal{T}_{u,c}$ the record (t_x, ID_h).

3.4 Data Filtering

The limited granularity of the data as well as the low quantities of data generated per-individual lead to questionings about the reliability of the results and cause, in some cases, biased conclusions. Therefore, to capture a more thorough picture of the mobility behavior of the users, we filter our "idle" users. In this thesis, we consider two different filtering procedures, depending on the goals of each chapter.

- **Filtering F1**: We first define a *complete day* of data as a day in which a user has a record *at least each* δ^*_{GPS} min (δ^*_{CDR} h) for the GPS datasets (CDR dataset). This means for each day if the user does not have a record within the δ^*_{GPS} min (δ^*_{CDR} h) succeeding her last record, the day is incomplete. Following this, we filter out "idle" users and select only participants with at least x days of *consecutive* and *complete* days of data. This filtering is used when the quality of days of data is more priority than the number of users, such as users' characterization and features extraction that are sensitive to the contiguity and quality of days of data.

- **Filtering F2**: In this filtering, the condition for a complete day in **F1** is relaxed to a day in which a user has *on average* one record each δ^*_{GPS} min (δ^*_{CDR} h) for the GPS datasets (CDR dataset). Next, we filter out "idle" users and select only participants with at least x days of *complete* days of data with temporal gaps tolerance (the definition in **F2** of a complete day of data allows leveraging a larger number of users at the cost of having sparser traces). This filtering is used when the long-term data availability and the number of users are prioritized.

GPS data aggregation Agg_gps: due to the small number of individuals in GPS data sources and considering that these sources are of the same nature (i.e., with the same frequency of sampling and duration of analyses), we proceed as the following. We aggregate the filtered and manipulated GPS datasets and label this new dataset as Agg_gps. The aggregation consists of the simple concatenation of the datasets.

3.5 Summary

This chapter introduced the different real-world datasets of various size and sparsity levels leveraged in this thesis. Following, it described the data treatment procedure followed for the preparation of the data and users filtering. Further details on experimental settings are provided in Appendix 7.3.

In the next chapter, special attention is given to exploration-like visits. Specifically, it aims at understanding and quantifying individuals' tendencies to discover new places and capture the related mobility traits.

Chapter 4

Understanding Individuals' Proclivity for Novelty-seeking

Contents

Individual-level mobility research focuses on uncovering the mechanisms ruling individuals' movements [85]. Nonetheless, there exists a perplexity in understanding and predicting individuals' mobility patterns. Human beings' movements are a mixture of repetitive and regular transitions between known places and sporadic discoveries of new areas [46, 73, 79], both subject to a certain degree of uncertainty associated with free will and arbitrariness [6]. At each instant, an individual is confronted with an extensive list of choices with regard to *how* and *where* to spend her time and has two alternatives: she either returns to a place she visited in the past or explores a new location. Contrary to the extensive literature investigations on mobility regularity patterns, in this thesis, we focus on the discoveries of new places and endeavor to understand the exploration mechanisms.

This chapter is a centralized place to discuss the definition and the capture of individuals' tendencies to discover new places. It starts with a general definition,

identification, and characterization of individuals' proclivity for novelty-seeking and profiles the users accordingly. Next, it investigates the mobility traits of individuals of each identified mobility profile. Finally, it goes deeper into the investigation of the mobility behavior of each profile by reporting individuals of each group to the temporal and spatial exploitation.

4.1 Novelty-seeking Capture

In this section, we start with a general definition of novelty-seeking in human mobility. Next, we present the adopted two-dimensional modeling of individuals' visits. Following, we detail the state-of-the-art approach and our proposed method for the identification of novelty-seeking phases of the individuals. After this lightweight novelty-seeking definition and identification and visits' classification, we introduce two metrics with the potential to capture individuals' proclivity to explore. Finally, we use four real-world mobility traces to evaluate the proposed exploration identification and mobility profiling methods.

4.1.1 Definition of Novelty-seeking

Human movements can be divided into two primary types of movements: explorations and returns. An **exploration** can be defined as *a discovery of a new location* and a **return** as *a visit to a previously seen locality.* These definitions, however, comprise both **desired-transitions** and **forced-transitions**.

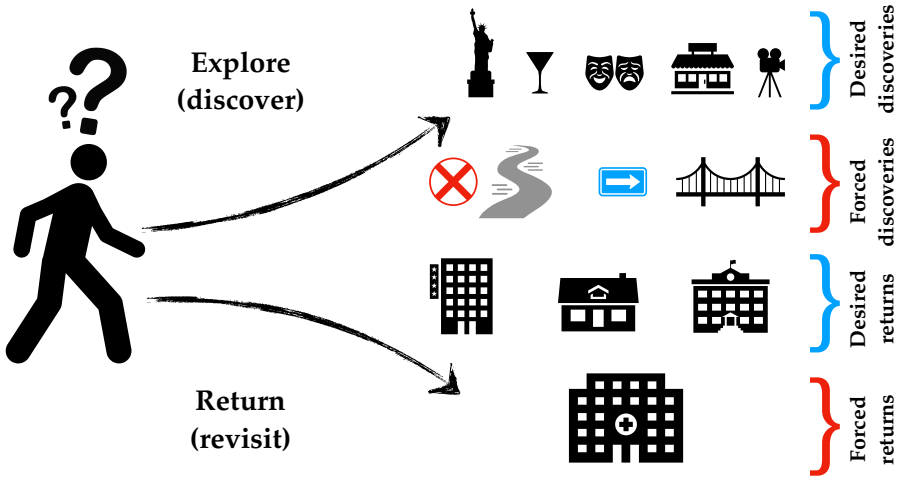


Figure 4.1: Human movements classification.

- **Desired-transitions:** are mainly aroused by the free movement will of individuals and are expected to be more regular in time and space.

- **Forced-transitions:** are inflicted by environmental circumstances, on which the individual has no control and are usually viewed as random events or noise while analyzing mobility traces.

By way of illustration of a forced-transition, the unexpected closing of a bakery. Consequently, regular clients of the bakery either go to another bakery they already know or search for a new one. Accordingly, explorations, as defined above, comprise both **desired-discoveries** and **forced-discoveries**, and returns comprise both **desired-returns** and **forced-returns**, as depicted in Figure 4.1.

To understand individuals' mobility, one should consider our last classification of transitions: **desired-discoveries**, **forced-discoveries**, **desired-returns**, and **forced-returns**. However, the differentiation between desired transitions and forced-transitions requires context-awareness along with an aggregated view of the population motion where the forced-transitions are more discernible. Due to the lack of contextual information in our datasets, hereafter, we consider the classification between **exploration** and **return** types of movements.

4.1.2 Formalization of the Generic Mechanisms Governing Individuals' Mobility

Let M be the Finite-State Automaton (FSA) describing an individual's movements, as shown in Figure 4.2, with two possible states: *exploring* (**E**) and *returning* (**R**). An individual u can either be in the exploring state (**E**) or the returning state (**R**). Two possible inputs can affect her state: *return* (T_R or S_R) by going back to historically known locations and *explore* by discovering new spots (T_E or S_E). In the exploring state **E**, discovering new areas (S_E) has no effect and keeps the individual in the state **E**. On the other hand, moving back to a known location (T_R), though recently explored, M shifts the state from **E** to **R**. In the returning **R** state visits to usual places (S_R) does not change the state; however, a discovery of a new spot (T_E), shifts the state back to the **E** state.

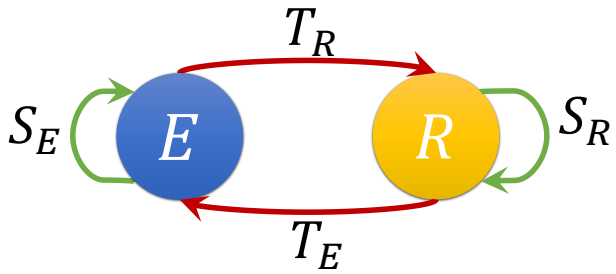


Figure 4.2: Finite-State Automaton M.

4.1.3 Novelty-seeking Identification

Strictly speaking, for an individual, an exploration is the discovery of a new geographical location, i.e., a place that is not included in her daily routine or where she was never seen before. But, *how can we distinguish novelty-seeking visits from routine visits?*

In what follows, we describe the state-of-the-art approach and our newly tailored per-user scheme to identify moments of novelty-seeking:

4.1.3.1 Baseline Identification

Existing works tackling the exploration problem have a naive approach to distinguishing between what is new and known. Given the mobility trace, $\mathcal{T}_{u,c} = \langle (t_0, l_0), (t_1, l_1), \ldots, (t_n, l_n) \rangle$ of the user u, the first occurrence of a location l_x in $\mathcal{T}_{u,c}$ is viewed as an exploration event (cf. Figure 4.3) [32, 6]. This implies that the first appearance of the *home* location or the *workplace* in the sequence is considered to be a moment of novelty-seeking. Yet, overvaluing the frequency of exploration

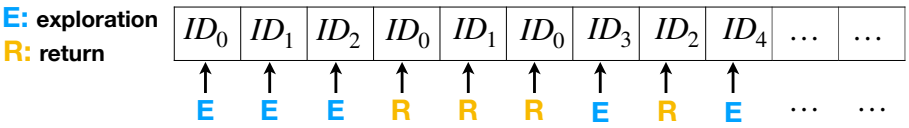


Figure 4.3: Novelty-seeking identification: Baseline approach.

events might twist the understanding of individuals' tendencies to explore. Hence, more accurate methods are essential for a more thorough identification of moments of novelty-seeking.

4.1.3.2 Proposed Identification

For more precise identification of moments of novelty-seeking, we first propose a per-user method to classify the visited locations into: (i) Routine Location (RL), i.e., locations used for return visits, and (ii) Exploratory Location (EL), i.e., locations visited when being in the exploring phase. Subsequently, only the first occurrence of EL locations are viewed as exploration events. We detail the two steps in the following,

1. **Location classification:** To avoid considering the first occurrences of highly visited locations such as *home* location or the *workplace* as moments of novelty-seeking, we base our location classification method on the visitation frequency metric (cf. Section 2.2.1).

   Proposed location classification:

   Let $F_u = \{l_1, l_2, \ldots, l_n\}$ be the set of the distinct locations visited by the user u. For each location $l_i \in F_u$, we assign a weight $f_u(l_i)$ outlining its visitation frequency in $\mathcal{T}_{u,c}$ (cf. Algorithm 1, Lines 4–5). It is given by,

Algorithm 1 Visitation-frequency-based threshold classification

1: **function Location_classification_1** ($\mathcal{T}_{u,c}$, *threshold*)
2: $f_u, T_{RL_u}, T_{EL_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ **Unique**($\mathcal{T}_{u,c}$) ▷ Extract the distinct visited locations
4: **for** j **in** F_u **do**
5: $f_u[j] \leftarrow$ **Frequency_of_appearance**($j, \mathcal{T}_{u,c}$), ▷ Eq. (4.1)
6: **end for**
7: **for** j **in** F_u **do** ▷ Classify the locations into RL and EL
8: **if** $f_u[j] \geq threshold$ **then**
9: T_{RL_u}. **Add**(j)
10: **else**
11: T_{EL_u}. **Add**(j)
12: **end if**
13: **end for**
14: **return** T_{RL_u}, T_{EL_u}
15: **end function**

$$f_u(l_i) = \frac{freq_u(l_i, \mathcal{T}_{u,c})}{\sum_{j=1}^{|F_u|} freq_u(l_j, \mathcal{T}_{u,c})}, \tag{4.1}$$

where $freq_u(l_i, \mathcal{T}_{u,c})$ is the number of occurrences of the location l_i in the mobility trace $\mathcal{T}_{u,c}$.

Next, we empirically determine a threshold "*threshold*", such that each location l_i holding a weight $f_u(l_i) \geq threshold$ is assigned to the RL set (cf. Algorithm 1, Lines 8–9). Otherwise, it is added to the EL set (cf. Algorithm 1, Lines 10–11).

The tuning of the parameter *threshold* is critical; high values for *threshold* can induce an overestimation of exploration events, while small values lead to a neglect of novelty-seeking moments. We quantify its impacts in Section 4.1.5.2.

Baseline location classification: To evaluate the performance of the proposed classification method, we compare it to the widespread location classification framework proposed by Papandrea et al. [71] described hereunder.

As in [71], we classify the visited locations according to their relevance (cf. Algorithm 2, Line 5). The Relevance $R_u(l_i)$ of a location l_i for a user u is given by,

$$R_u(l_i) = \frac{d_{visit}(l_i, u)}{d_{total}(u)}, \tag{4.2}$$

where $d_{visit}(l_i, u)$ is the number of days the individual u visited the location l_i, and $d_{total}(u)$ is the number of days the individual has been active.

Following, as in [71], we use the k-mean unsupervised approach with three components to classify the locations into: (1) Mostly Visited Places (MVP), i.e., locations most frequently visited by the user; (2) Occasionally Visited Places

Algorithm 2 Baseline location classification

1: **function Location_classification_2** ($\mathcal{T}_{u,c}$)
2: $T_{Relevance,u}, T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow \emptyset$
3: $F_u \leftarrow$ **Unique**($\mathcal{T}_{u,c}$) ▷ Extract the distinct visited locations
4: **for** j **in** F_u **do**
5: $T_{Relevance,u}[j] \leftarrow$ **Compute_relevance**(j) ▷ Eq. (4.2)
6: **end for**
7: $T_{MVP_u}, T_{OVP_u}, T_{EVP_u} \leftarrow k$-**means**($T_{Relevance_u}, 3$) ▷ Classify the locations into MVP, OVP and EVP
8: **return** $T_{MVP_u}, T_{OVP_u}, T_{EVP_u}$
9: **end function**

(OVP), i.e., locations of interest for the user, but visited just occasionally; (3) Exceptionally Visited Places (EVP), i.e., rarely visited locations (cf. Algorithm 2, Line 7).

2. **Exploration identification:** Initially, each user u has an empty set of known locations $\mathcal{L}_u(t_0) = \emptyset$ (cf. Algorithm 3, Line 2). After classifying the visited locations into RL and EL locations, we add the commonly visited locations, i.e., RL locations, to her set of known locations $\mathcal{L}_u$ (cf. Algorithm 3, Line 4). Following, when analyzing her mobility trace $\mathcal{T}_{u,c}$, if the current location l_i is not present in $\mathcal{L}_u$, the visit is considered to be an exploration and is then added to $\mathcal{L}_u$ (cf. Algorithm 3, Lines 6 – 8). Else it is considered to be a return (cf. Algorithm 3, Lines 9 – 10).

Algorithm 3 Novelty-seeking identification: Proposed approach

1: **function Novelty-seeking_identification** ($\mathcal{T}_{u,c}$)
2: $exploratory\_moments, return\_moments, \mathcal{L}_u \leftarrow \emptyset$
3: $T_{RL_u}, T_{EL_u} \leftarrow$ **Location_classification_1**($\mathcal{T}_{u,c}$)
4: $\mathcal{L}_u \leftarrow T_{RL_u}$ ▷ Add RL to the set of known places
5: **for** j **in** $\mathcal{T}_{u,c}$ **do**
6: **if** $j.id \notin \mathcal{L}_u$ **then** ▷ Identify moments of exploration
7: $exploratory\_moments$.**Add**($j.t$)
8: T_{EL_u}.**Remove**($j.id$)
9: **else** ▷ Identify moments of return
10: $return\_moments$.**Add**($j.t$)
11: **end if**
12: **end for**
13: **return** $exploratory\_moments, return\_moments$
14: **end function**

4.1.4 Mobility Profiling

After dissecting human visits into explorations and returns, for each user u, we first extract two sets:

- **Returning set ret_u:** is a set containing the sets of consecutive returns, $ret_u = \{r_0, r_1, \ldots, r_n\}$, where each $r_i = \{l_0, l_1, \ldots, l_x\}$ is a set containing the IDs of the cells where the user u performed successive returns.

- **Exploring set exp_u:** is a set of sets of consecutive explorations, $exp_u = \{e_0, e_1, \ldots, e_n\}$, where each $e_i = \{l_0, l_1, \ldots, l_x\}$ contains the ids of the cells where the user u performed successive explorations.

Next, we assign to each individual u two values: (1) $\#E = avg(|e_i|), e_i \in exp_u$, the average number of her successive explorations – the average number of consecutive self-transitions she made in state E, and (2) $\#R = avg(|r_i|), r_i \in ret_u$ the average number of successive returns – the self-transitions she made in state R.

To characterize how users balance the trade-off between revisits of familiar locations and new-places discoveries, we define the following metrics that utterly capture the exploration habits of an individual. The first metric captures the shifting habits between the exploration and the return modes. The second metric captures the susceptibility of users to remain in their routine rather than explore new places.

Definition 1 (*Intermittency* μ) *is the sum of the average number of successive explorations $\#E$ and the average number of successive returns $\#R$, $\mu = \#R + \#E$.*

The *intermittency* measure reveals whether an individual is versatile or prefers to remain steady concerning a category of location (i.e., return or exploration). Namely, it helps to recognize if a user is constantly fluctuating between visits to familiar places and discoveries of new spots, or once she starts a discovery, she does it repeatedly, before switching to revisits and vice versa.

Definition 2 (*Degree of return* α) *is the angle whose tangent is the ratio between the average number of successive returns $\boldsymbol{R}$ over the average number of successive explorations $\boldsymbol{E}$, $\alpha = \text{arctg}\left(\frac{\#R}{\#E}\right)$.*

The *degree of return* describes the exploration conducts of an individual compared to her returns. A high degree of returns suggests that: the average number of successive returns is higher than the average number of successive explorations $\#R > \#E$. Hence, the *degree of return* reveals what kind of explorer an individual is: whether she visits many new places on a row or goes back to a familiar location just after a few discoveries.

In what follows, we investigate whether the novelty-seeking habit is the same among the population or if it is a distinctive property. Namely, if there exist patterns followed by individuals while shifting between the exploration and return modes, or

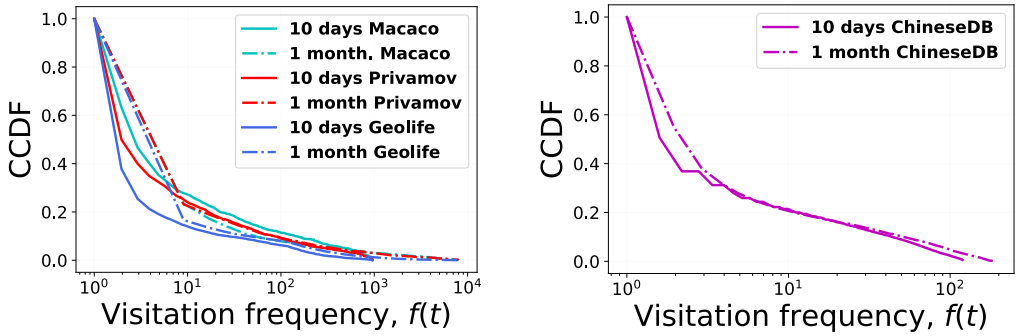
if there are several groups of users sharing the same habits but distinct from the others.

4.1.5 Experiments

4.1.5.1 Data Description

We recall the reader with a brief description of the characteristics of the data we use in the following experiments. We consider the filtering **F1** that selects only users that have at least 10 *complete* and *contiguous* days of data, and where a complete GPS day (CDR day) of data is a day in which an individual has a record at least each δ^*_{GPS}=15 min (δ^*_{CDR}=2 h). The number of users within each dataset are the following: 87 users for Macaco, 69 users for Privamov, 101 users for Geolife, and 3761 users for ChineseDB.

4.1.5.2 Novelty-seeking Identification



(a) GPS datasets.

(b) CDR dataset.

Figure 4.4: Complementary Cumulative Distribution Function (CCDF) of the visitation frequency.

Figure 4.4 reports the visitation frequency distributions obtained from all traces. For a general overview, we consider two different periods of data, 10 days and 1 month (this latter is obtained with the filtering **F2** cf. Section 3.4). The GPS Macaco, Privamov, and Geolife datasets, shown in Figure 4.4a, exhibit the same behavior, where a huge number of locations are visited only a few times. In contrast, a few places are frequently visited and have a very high visitation frequency score. Approximately 70% of the locations hold a visitation frequency score lower than 10, accounting for highly regular patterns of visits consisting of small sets of places. A less pronounced trend characterizes the visitation frequency distributions in the CDR Figure 4.4b. Here we measure a smaller value for the same proportion (70%), mainly due to the sparsity of the data. Although the datasets are different in nature,

these results are very similar. Thus, we use the same settings for the parameter *threshold* for all the datasets to classify the visited locations into EL or RL.

We consider two different settings for the parameter *threshold*,

1. **Relaxed exploration identification:** This setting aims at only recognizing highly visited locations, such as *home* location and *workplace*, as RL. Therefore, locations having a visitation frequency as high as $level = 90\%$ of the frequency of visit of the most visited location are assumed to be RL, i.e., $threshold = \max_{j=1}^{|F_u|}(f_u(l_j)) \times 0.9$.

2. **Refined exploration identification:** This setting aims for a more thorough identification of frequently visited locations. Hence, locations having a visitation frequency as high as $level\%$ of the average visitation frequency to the distinct locations are assumed to be RL, i.e., $threshold = \text{mean}_{j=1}^{|F_u|}(f_u(l_j)) \times level$.

Using the baseline (relevance-based) location classification approach (Algorithms 2), we categorize the visited places into EVP, OVP, and MVP. Next, using the refined exploration identification method, we measure the fraction of places within each category of places (i.e., EL and RL) with $level \in \{20, 80\}\%$, and evaluate their average visitation frequency, as shown in Figure 4.5.

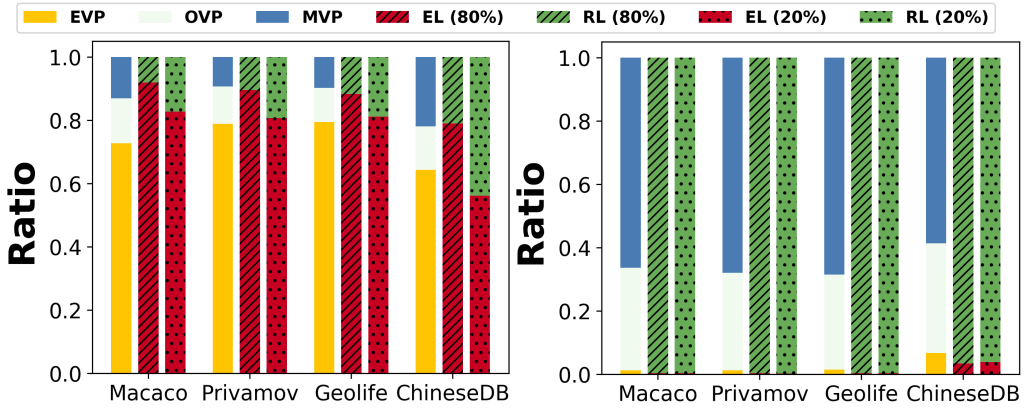


Figure 4.5: (a-left) Percentage of visited places. (b-right) Average visitation frequency. EVP, OVP, and MVP are categorized according to Algorithm 2. EL or RL are categorized according to Algorithm 1 for $level = 80\%$ and $level = 20\%$.

Figure 4.5 (a) reports the percentages of places classified within each category extracted from our datasets; EVP, OVP, and MVP by Algorithms 2, EL or RL by Algorithm 1.

First, we observe the high ratio of EVP jointly with OVP categorized by Algorithms 2. For all GPS datasets, more than 78% of the places, i.e., EVP $\cup$ OVP, are not integrated into the daily routines of the individuals. Note that the CDR dataset describes visits in a smaller temporal resolution (i.e., per hour), which naturally

impacts the precision in exploration inference of visits. Likewise, in all datasets, the proportion of locations used for EL surpasses 78% when *level* is set to 80% and is higher than 50% with *level* = 20%. Moreover, we can notice in the case where *level* = 80%, the proportion of places classified as EL by Algorithm 1 corresponds roughly to the percentage of places categorized as EVP ∪ OVP by Algorithm 2. In contrast, in Algorithm 1 with *level* = 20%, the fraction of places labeled EL is almost equal to the faction of locations classified as EVP by the baseline Algorithm 2.

Figure 4.5 (b) illustrates the proportion of the average frequency of visits towards each category of places. Firstly, we see a markedly high proportion of visits to locations used for RL, more than 90% of the visits are towards this category of places for *level* $\in \{20, 80\}$%. Whereas the same score is obtained by Algorithm 2 when taking MVP and OVP together. Additionally, the average frequency of visits held by EL for all datasets with *level* $\in \{20, 80\}$% is lower than the scores obtained by EVP. Indeed, in the relevance-based approach, the importance of a location is based on the number of days it was visited and not the amount of time she spent within it. This means, for a user u, if she weekly visits the municipal library for 4 hours, this latter will have the same relevance score as the bakery where she goes once a week for a few minutes only to buy a baguette.

In addition to the rate of places categorized in each group, we measure the percentage of intersection between EL places and EVP, then between EL and EVP ∪ OVP.

Table 4.1: Percentage of EL places present in EVP and in EVP ∪ OVP, with *level* $\in \{20, 80\}$.

	EL (80%) ∈ EVP	**EL (20%) ∈ EVP**	**EL (80%) ∈ EVP ∪ OVP**	**EL (20%) ∈ EVP ∪ OVP**
Macaco	60.1%	47.71%	78.33	68.38%
Privamov	50.75%	36.58%	76.92	65.38%
Geolife	41.19%	33.76%	67.82	59.18%
ChineseDB	88.78%	61.94%	98.27	84.23%

In Table 4.1, we report the percentage of overlap between the locations classified as EL with *level* $\in \{20, 80\}$% at first with EVP locations only then with EVP ∪ OVP. The fraction of places categorized as EL with *level* = 20% is closer to the fraction of places categorized as EVP when *level* is set to 80%. The overlap between EL and EVP is higher when *level* equals 80%. We can also observe that when measuring the degree of overlap of EL with EVP ∪ OVP, the obtained scores increase for both *level* = 20% and *level* = 80%, with a very high degree of overlap for CDR ChineseDB reaching 98.27% for *level* = 80%. Succinctly, the difference in our methodology quantifies the importance of a location in an individual's daily life. We notice the significant overlap between the classifications of our proposed method and the baseline approach ones'.

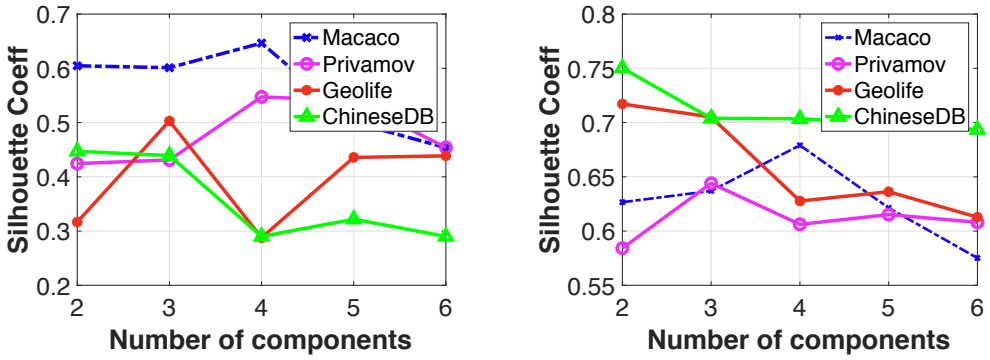
Thereby, setting *level* to 80% allows EL to capture exceptionally and occasionally visited places as the baseline classification approach.

In summary, *the proposed method, Algorithm 1, offers a satisfactory classification of the visited places.* First, it allows the detection of a higher number of places used for exploration visits (EL). On the other hand, the visitation frequencies to these locations are lower than the RL and EVP of Algorithm 2. Second, the performance of Algorithm 1 with $level = 80\%$ allows the identification of a higher number of places used for EL, and hence enables a more comprehensive detection of moments of exploration compared to the setting with $level = 20\%$. Indeed, the first occurrence of a location in the set of a user's EL locations is presumed to be a moment of exploration.

In the remainder of this chapter, we consider the relaxed exploration identification setting.

4.1.5.3 Mobility Profiling

After computing the intermittency μ and degree of return α for each individual, we use two clustering algorithms – the Gaussian Mixture probabilistic Model (GMM) and the k-means clustering method – to attest whether we can split the population into distinct cohesive and significant groups or not. To identify the best number of components of the clustering algorithms, and hence, the individuals' types, we use the silhouette score statistical test [77]. We run one hundred fits for five different sets of clusters (two to six). Then, we consider the mean value when choosing the best score.

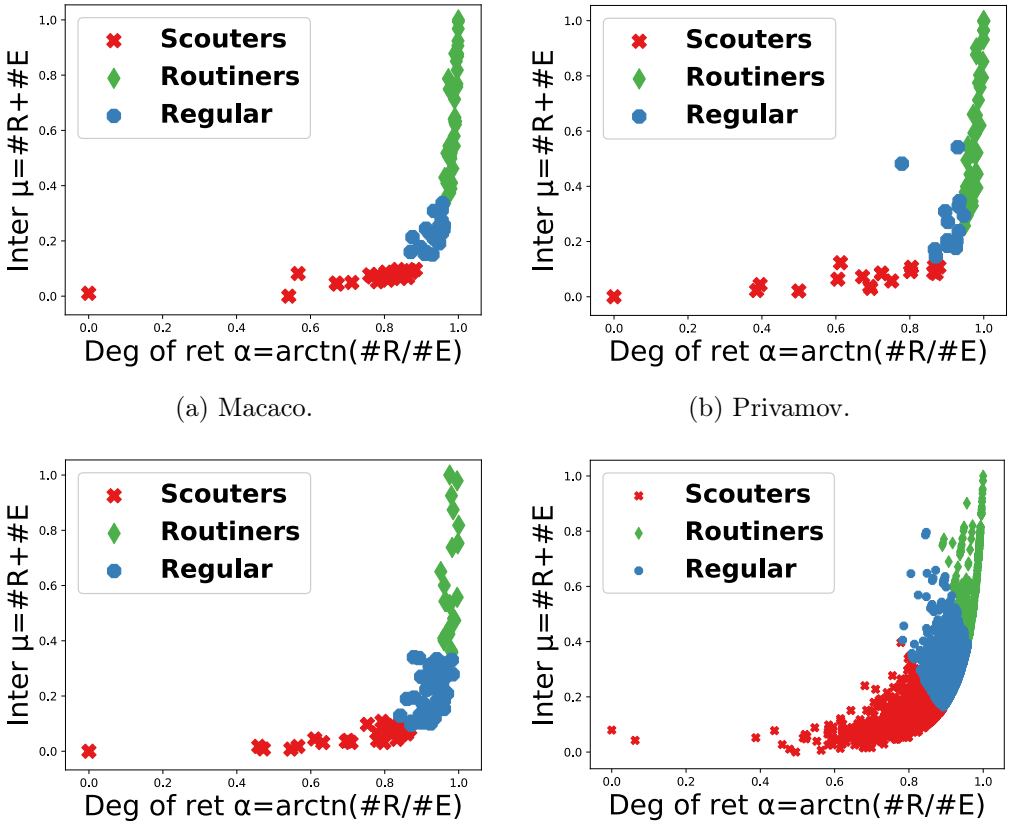


(a) GMM. (b) k-mean.

Figure 4.6: **(a)** Silhouette score for the GMM. **(b)** Silhouette score for the k-means.

Figure 4.6 depicts the silhouette score obtained for the two clustering algorithms GMM Figure 4.6a and k-mean Figure 4.6b. Figure 4.6a shows that the optimal number of components for the GMM method varies from one dataset to another. However, a clustering with three elements appears to be more equitable, as all datasets have a score above 0.4. Likewise, the clustering with two components

is approximately just as effective. Figure 4.6b depicts that two, three, and four components are good candidates for the k-mean algorithm. Still, a clustering with three groups seems to be more balanced amid the datasets. Accordingly, we have two candidates for the best number of components. Nonetheless, we choose a clustering with three components as it maximizes the minimal score for both of the clustering algorithms and appears to be more meaningful for all of our data sources.



(a) Macaco.

(b) Privamov.

(c) Geolife.

(d) *ChineseDB* CDR.

Figure 4.7: Mobility Profiling.

We then apply the GMM and k-mean with three components on our data sources. We roughly obtain the same groups for both clustering algorithms. Thus, we only present the results obtained with the GMM algorithm. Figure 4.7 depicts the normalized intermittency of individuals against their normalized degree of return and displays the clusters resulting from applying the GMM algorithm on the GPS and CDR datasets. We can observe that our metrics can clearly capture the dissimilarity between the individuals in terms of human mobility dynamics. More importantly, the GMM identifies three distinct groups with identical *intermittency* and *degree of return* characteristics for all our data sources. We label the resulting groups as **Scouters** (red), **Routiners** (green), and **Regulars** (blue).

- Cluster 1: *Scouters or extreme explorers*, although holding varying degrees of return α, they are remarkably lower than others' scores. Moreover, they are notably intermittent, i.e., they are constantly shifting between the exploring and the returning states. These users are more prone to explore and discover new areas.

- Cluster 2: *Routiners or extreme-returners* have a surprisingly large degree of return. Besides, they tend to be steady in the different states of the automaton M, i.e., they rarely break their routine. Hence, we can deduce that these users rarely explore and prefer to stick among their common and known places.

- Cluster 3: *Regulars* adopt a medium behavior and have large degrees of return compared to the *Scouters*. Though, their intermittencies are distinctly smaller than those of *Routiners*. These users constantly alternate between explorations and revisits. Yet, their proclivity to explore is less critical than *Scouters*'.

The proposed approach captures two major mobility features that fully describe the exploration phenomenon, i.e., *intermittency between returns and explorations and the ratio of explorations compared to returners*, and allows a natural clustering of the individuals.

4.2 Spatiotemporal Characterization of Novelty-seeking

Here, we identify the specific mobility behavior traits of each profile: *Scouters*, *Routiners*, *Regulars*. Hence, we extract some of the fundamental features used to characterize human mobility from the spatiotemporal footprints of the individuals (cf. Table 4.2). The derived features are divided into three groups: *Relocation Activities*, *Temporal Activities*, and *Spatial Activities*.

- **Relocation Activities:** This category aims at quantifying and characterizing individuals' visits, transitions habits, and capturing uniqueness and repetitiveness of visits.

- **Temporal Activities:** This category relates to the behavior of individuals in time and captures the amount of time spent by individuals exploring, returning, and visiting distinct locations.

- **Spatial Activities:** The last category gives an intuition on the distances walked by individuals when performing each type of visit and the covered distances.

In what follows, due to the small number of users in each mobility profile for the GPS data sources, we use the Agg_gps dataset described in Chapter 3. In view of its different nature, we separately analyze the profiles resulting from the CDR dataset.

Table 4.2: Extracted features.

Category	N	Featture Name	Descrription
Relocation Activities	1: (a)	Number of successive **E**	The average number of successive explorations performed by the user
	2: (b)	Number of successive **R**	The average number of successive returns performed by the user
	3: (c)	Number of stops	The number of distinct areas visited by the user
	4: (d)	Ratio of unique places	The ratio of places visited only once
	5: (e)	Visitation frequency	The frequency of visits to each area known by the user
Temporal Activities	6: (a)	Total **E** time (min)	The total amount of time spent by the user when exploring new places
	7: (b)	Total **R** time	The total amount of time spent by the user when revisiting her known places
	8: (c)	Waiting time	The average amount of time spent by the user before making a transition to another place
	9: (d)	Duration of successive **E**	The average duration spent by the user when exploring new places
	10: (e)	Duration of successive **R**	The average duration spend by the user when revisiting her known places
Spatial Activities	11: (a)	Total **E** distance (km)	The total distance walked by the user when exploring new places
	12: (b)	Total **R** distance	The total distance walked by the user when revisiting her known places
	13: (c)	Radius of gyration r_g	The typical distance walked by the user compared to her center of mass (cf. Eq. (2.2))
	14: (d)	Ratio of distant **E**	The ratio of explorations located outside the circle of center r_0 and radius $R = r_g$ over the total number of visits
	15: (e)	Average displacement	The average distance the user walks when transiting from a place to another (cf. Eq. (2.1))

For the sake of comparing and displaying the variations of the different features among individuals of each mobility profile, we report the box-plot[1] of each feature for *Scouters*, *Routiners*, and *Regulars* as shown in Figs. 4.8, 4.9, 4.10, 4.11, 4.12,and 4.13.

4.2.1 Scouters' Mobility Traits

Scouters are energetic and dynamic when discovering new places. However, they become weary and flat while revisiting various areas they already know. Admittedly, when *Scouters* start exploring, they relish discovering many new places uninterruptedly compared to the rest of the population, as depicted in Figures 4.8a and 4.9a. On the contrary, after a few revisits of familiar spots, they are keen to break their returning routine and chase for new areas to expand their sets of known places, as shown in Figures 4.8b and 4.9b. Figures 4.8c and 4.9c depict that *Scouters* have remarkably large sets of known places. Indeed, this class of individuals performs many explorations, and by consequence, they get to know diverse places. *Scouters* have a surprisingly high ratio of places visited only once. Manifestly, they relish discovering new places. Yet, sometimes they do not revisit or include them in their routinary patterns, as can be perceived in Figure 4.8d and Figure 4.9d. Moreover, from Figure 4.8e and Figure 4.9e, we can observe that *Scouters* do not revisit the same places several times, except for some specific ones, which indicates that their routinary patterns consist of a small set of areas.

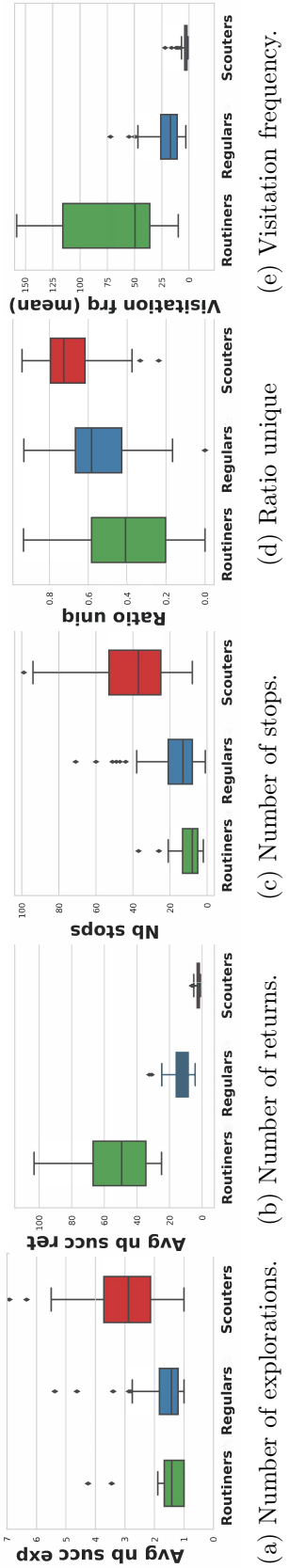
Figures 4.10a and 4.11a show that the total time amount of time spent by *Scouters* exploring is notably larger than the rest of the population, while their returning time is shorter, as depicted in Figures 4.10b and 4.11b.

Besides, *Scouters* wait a shorter amount of time before transiting from one place to another, as shown in Figures 4.10c and 4.11c. Furthermore, the average duration of successive explorations is higher for *Scouters*; on average, they spend more than 200 min $\approx$ 3 h exploring, as depicted in Figures 4.10d and 4.11d. Hence, individuals of this class not only relish discovering many places successively but also do it for longer periods. Conversely, their average returning time is shorter than the other profiles; approximately, they spend less than 1000 min $\approx$ 16 h returning, as depicted in Figures 4.10e and 4.11e.

Withal, *Scouters* are active and driven individuals. Figure 4.12a, Figure 4.13a, Figure 4.12b, and Figure 4.13b point out that they generally walk longer distances.

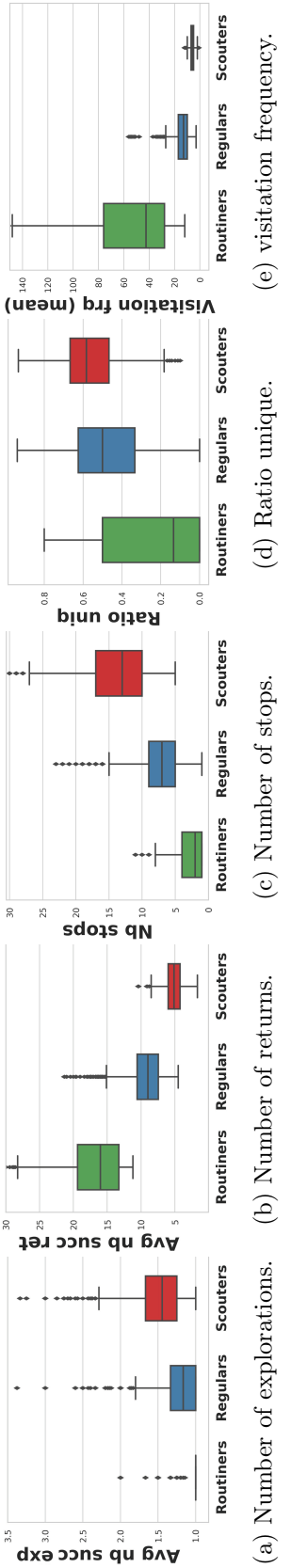
Particularly, they cover longer distances, as depicted by Figures 4.12e and 4.13e. Moreover, as shown in Figures 4.12c and 4.13c, unlike the other groups, *Scouters* are characterized by a larger radius of gyrations r_g, better seen in the CDR dataset. Namely, they cover larger areas on a daily basis.

[1] Some overlaps between the box-plots of the different groups can be noticed, yet this is essentially due to the limitation in the number of users. Though, the tendency is clearly discernible among the mobility profiles, especially in the CDR figures where we leverage a larger number of users.



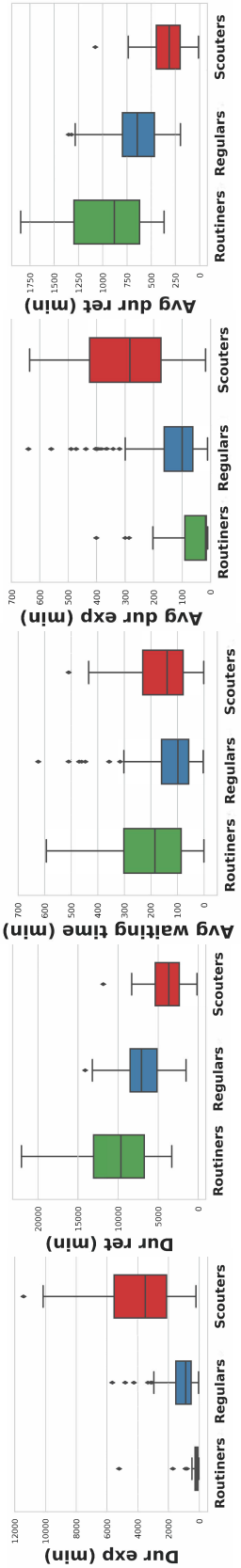
(a) Number of explorations. (b) Number of returns. (c) Number of stops. (d) Ratio unique (e) Visitation frequency.

Figure 4.8: Relocation Activities in Agg_gps dataset.



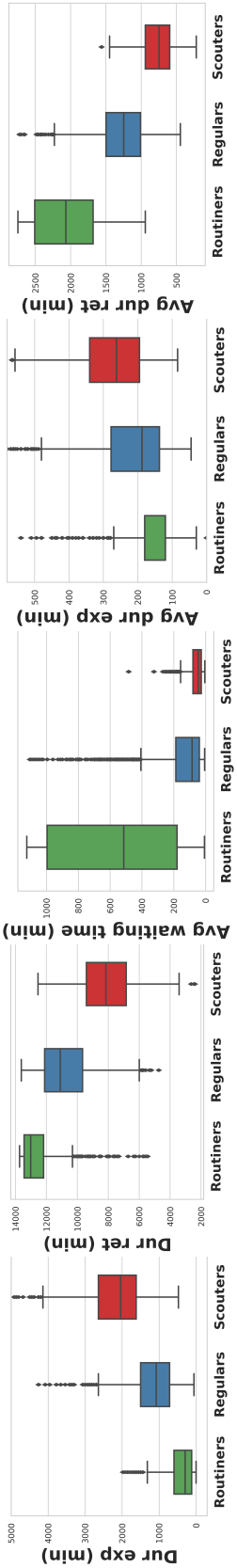
(a) Number of explorations. (b) Number of returns. (c) Number of stops. (d) Ratio unique. (e) visitation frequency.

Figure 4.9: Relocation Activities in ChineseDB dataset.



(a) Total exp time. (b) Total ret time. (c) Waiting time. (d) Avg exp duration. (e) Avg ret duration.

Figure 4.10: Temporal Activities in Agg_gps dataset.



(a) Total exp time. (b) Ttotal ret time. (c) Waiting time. (d) Avg exp duration. (e) Avg ret duration.

Figure 4.11: Temporal Activities in ChineseDB.

4.2.2 Routiners' Mobility Traits

Routiners are steady and rarely leave their zone of comfort. Unlike *Scouters*, they discover very few new places consecutively. Hence, once they explore, they either stay at the same place or go back to a familiar place, as shown in Figures 4.8a and 4.9a. Besides, they rarely interrupt their successive returns to discover new areas, and this can be observed in the very high value of successive returns in Figures 4.8b and 4.9b. Individuals of this profile have small sets of distinct visited places, meaning that they visit less and enjoy their routinary habits shifting between familiar locations as depicted by Figures 4.8c and 4.9c. They are also characterized by a small ratio of places visited only once and a large visitation frequency, as depicted by Figures 4.8d, 4.9d, 4.8e, and 4.9e. This indicates that *Routiners* frequently revisit many places they know.

Figures 4.10a, 4.11a, 4.10b, and 4.11b suggest that *Routiners* spend shorter amounts of time exploring. Additionally, they wait larger moments before making a transition to another place, as shown by Figures 4.10c and 4.11c. Likewise, Figures 4.10d, 4.11d, 4.10e, and 4.11e reveal that *Routiners* spend less than 300 min $\approx$ 5 h exploring. Accordingly, they usually prefer to return to their comfort zone before performing another discovery and spend large amounts of time returning before aspiring to discover new spots. Consequently, the total time allocated by these individuals for discoveries is smaller than the rest of the population, and on the contrary, they spend a large amount of time returning.

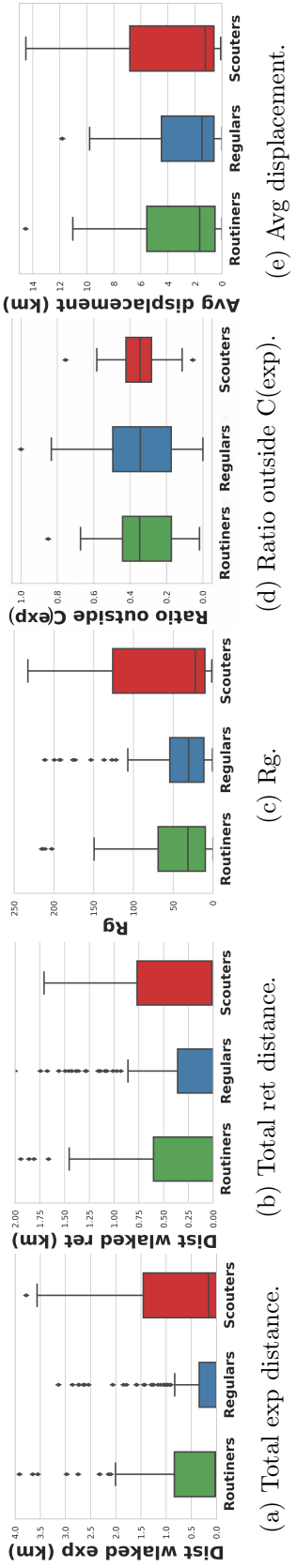
Routiners do not walk long distances in general, as depicted by Figure 4.12a, Figure 4.13a, Figure 4.12b, Figure 4.13b, Figure 4.12e, and Figure 4.13e, meaning that they go to close areas even when exploring. They are also characterized by a smaller radius of gyration r_g, as depicted in Figures 4.12c and 4.13c.

4.2.3 Regulars' Mobility Traits

From Figures 4.8a, 4.9a, 4.8b, and 4.9b, *Regulars* alternate between successive explorations and successive returns. In other words, they are constantly shifting between the exploring and the returning states. Besides, Figures 4.8c and 4.9c show that they have large sets of known places compared to *Routiners* but smaller than the *Scouters*'. From Figures 4.8d and 4.9d, we can observe the same thing concerning the ratio of places visited only once. Further, unlike, *Routiners* they do not equally visit their known locations but restrict their returns to a small set of places (cf. Figures 4.8e, and 4.9e).

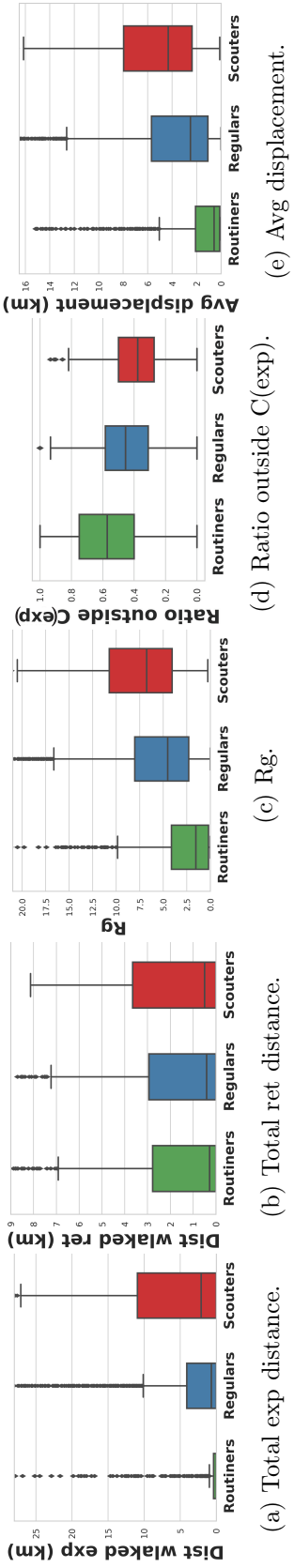
Regulars spend a larger amount of time exploring compared to the *Routiners* and a larger amount of time returning than *Scouters*, as shown in Figure 4.10a, Figure 4.11a, Figure 4.10b, and Figure 4.11b. The same is observed in terms of time spent on successive discoveries and revisits (cf. Figures 4.10d, 4.11d, 4.10e, and 4.11e). Besides, they usually wait a medium amount of time before performing transitions from one place to another (cf. Figures 4.10c and 4.11c).

Additionally, they walk larger distances when exploring compared to *Routiners*,



(a) Total exp distance. (b) Total ret distance. (c) Rg. (d) Ratio outside C(exp). (e) Avg displacement.

Figure 4.12: Spatial Activities in Agg_gps dataset.



(a) Total exp distance. (b) Total ret distance. (c) Rg. (d) Ratio outside C(exp). (e) Avg displacement.

Figure 4.13: Spatial Activities in ChineseDB.

as depicted in Figures 4.12a, 4.13a, 4.12b, 4.13b, 4.12c, 4.13c, 4.12e, and 4.13e.

Furthermore, we can also notice from Figures 4.13d and 4.12d without exclusion; all profiles have a high probability of going outside the circle of radius equal to their radius of gyrations $R = r_g$ when exploring.

4.3 Spatiotemporal Preferences

In this section, we verify if there exist temporal or spatial patterns followed by users of each profile when exploring. Admittedly, explorations are characterized by visits to new places that cannot be found in the past history of visited places of a user. However, such moments may present some patterns that can still be anticipated once the spatiotemporal features of a user's exploration behavior are well understood and modeled. This is motivated by the fact that such visits may have a temporal or a coarse-grained spatial regularity (e.g., users may like to visit different restaurants or bars but in the same neighborhood and usually on Saturday night) dictated by the user's motivations.

4.3.1 Temporal Patterns

We enrich our analysis with the *exploration of temporal semantics*, which refers to the interpretation of the occurring time of explorations, e.g., morning/evening, weekday/weekend. This dimension is essential for a thorough understanding of exploratory behaviors, as some discovery events occurring only in specific periods may remain hidden from global patterns. We define *temporal exploration regularity* as repeated explorations over time. For instance, a user exploring at very similar times each week is considered to have a highly regular exploratory temporal pattern at that moment of the week. Hereafter, we use a week-by-week comparison to determine the *temporal exploration regularity* of individuals. For this part, we only consider users with high temporal resolution (GPS) and who have at least four complete weeks of data according to the filtering **F1** (cf. Section 3.4). We are thus, left with 224 users.

To quantify the *temporal exploration regularity*, we adjust the ISI-Diversity [56] approach used in neural coding to our case of study and define:

Definition 3 (*Exploration timeline* T_u^w) *the exploration timeline is the ordered sequence of times the user u performed explorations during the week w.*

$$T_u^w = t_{u,1}^w, \ldots, t_{u,E_{w,u}}^w, \tag{4.3}$$

where $E_{w,u}$ is the total number of explorations made by u during w and t the offset in minutes from the origin "Monday 00:00" of the considered week.

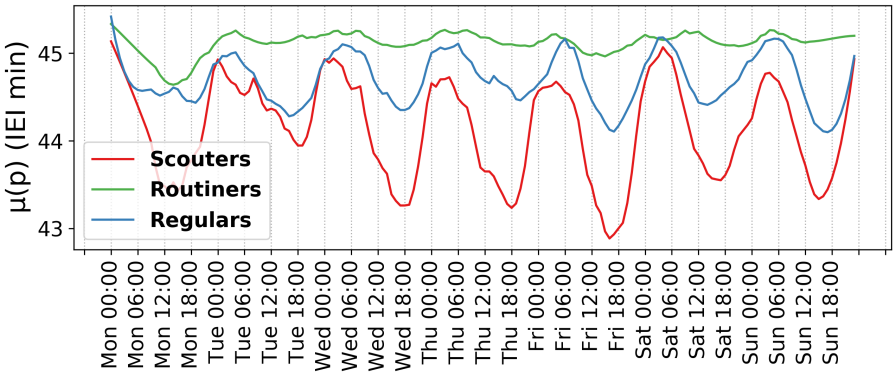


Figure 4.14: IEI instantaneous mean per periods of 1 h.

Definition 4 (*Inter-Exploration Interval (IEI)*) *it is the elapsed time between two consecutive explorations.*

We divide each week into periods of one hour. Each week comprises then 24×7 periods $P = [0, 1, \ldots, 24, \ldots, 72, \ldots, 168]$, and each period $p \in P$ has a starting time t^p_{min} and an ending time $t^p_{\text{max}} = t^p_{\text{min}} + 1$. Next, for each user u, we then measure the instantaneous inter-exploration interval function $I^w_u(t)$ that gives the IEI at time offset t of the week w, and is given by,

$$I^w_u(t) = \begin{cases} \min\left(\min(t^w_u | t^w_u > t), t^p_{\text{max}}\right) - \max\left(\max(t^w_u | t^w_u < t), t^p_{\text{min}}\right) & \text{, if } t \in [t^p_{min}, t < t^p_{max}], \\ 60 & \text{, else.} \end{cases} \tag{4.4}$$

If there are no exploration events within the period p, the instantaneous IEI will take the maximum possible value of 1 hour, i.e., $I^w_u(t) = 60$ min.

Next, for each individual, we compute the average instantaneous IEI per period:

$$I^w_u(p) = avg(I^w_u(t) | t^p_{\text{min}} \leq t < t^p_{\text{max}}) = \frac{1}{M} \sum_{t \in p} I^w_u(t), \tag{4.5}$$

where $M = |I^w_u(t)|$ and $t^p_{\text{min}} \leq t < t^p_{\text{max}}$. Last, we compute the instantaneous means per period p for each user u, given by, $\mu_u(p) = \frac{1}{W} \sum_{w=1}^{W} I^w_u(p)$, where W is the total number of weeks (exploration timelines).

Finally, we calculate the instantaneous mean per group as follows, $\mu(p) = \frac{1}{|U|} \sum_{u \in U} \mu_x(p)$, where U is the population of a mobility profile.

In Figure 4.14, we report the influence of the time of the week on the IEI instantaneous mean per period $\mu(p)$ for each mobility profile. We observe that individuals' exploration activities over the week contribute to their mobility profiles:

- *Scouters*' proclivity to explore is the highest for all week periods: They have a smaller IEI, which also means more explorations are performed. We can also notice that their exploration activities increase by the end of the week, reaching the maximum on Friday. Besides, *Scouters* tend to have a lower IEI from 4 pm to 8 pm during weekdays and hence explore more by the end of the day.

- *Routiners* have major discrepancies in exploration activities between Monday (cold start problem) and the other periods of the week. This reinforces our previous results on this group as being stationary and having a higher inclination to stay in their zones of comfort.

- *Regulars*' average instantaneous IEI means are nearly stable over the week during daytime with slightly higher exploration activity on Friday and Sunday.

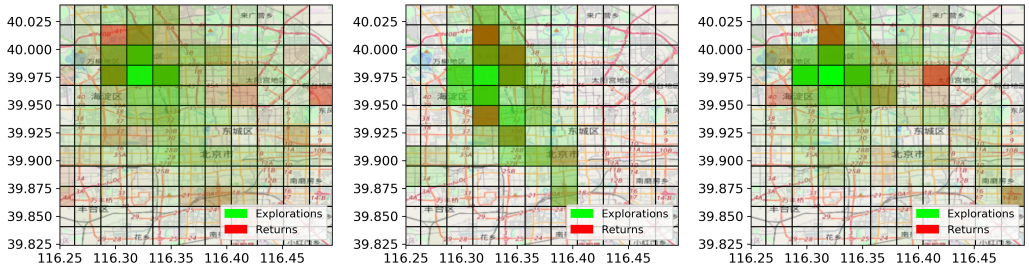
In summary, conversely to *Routiners* and *Regulars*, *Scouters* relish exploring all over the week, mainly in the afternoon and evenings. *Regulars*' proclivity to explore remains stable over the week with a slight increase for the weekends. A larger variation between Monday and the other days of the week can be noticed for *Routiners*.

4.3.2 Spatial Coverage

We here analyze and compare the spatial exploitation of *Scouters*, *Routiners*, and *Regulars*. Our main idea is to identify the geographical areas where individuals of each profile prefer to explore and how predictable they are in terms of types of visits in a coarse-grained resolution. In this regard, in addition to locations of type 3 (i.e., 300 m $\times$ 300 m), we consider zones of type 2 (i.e., 2 km $\times$ 2 km), and we refer to them as *Neighborhoods*. Following, for each individual, we compute the percentage of explorations and returns she performed within each Neighborhood. Because the datasets (i) are collected independently in different cities and (ii) each city has its own attraction areas and social gathering particularities, hereafter, we only present results for one city having the largest number of users, i.e., Beijing of the Geolife dataset.

In Figure 4.15, we make a zoom on the most visited areas in Beijing (the city center) and report the spatial coverage of each group among 132 Neighborhoods. The intensity of green/red corresponds to the percentage of explorations/returns in a given Neighborhood: The lighter shades of the colors indicate a low probability, while darker shades designate a high probability to explore/return. In the following, we list our main observations.

- *Scouters* have a high proclivity to explore in most Neighborhoods: Their explorations activities (i.e., green cells) are spread all around the city center. In particular, 83 of the 132 city-center Neighborhoods (i.e., more than 62%) are visited for explorations. Besides, their return activities (i.e., red cells) are



(a) Scouters. (b) Routiners. (c) Regulars.

Figure 4.15: Spatial use in Beijing (Downtown).

also dispersed: More than 67% of the Neighborhoods are used for returns (cf. Figure 4.15a)[2].

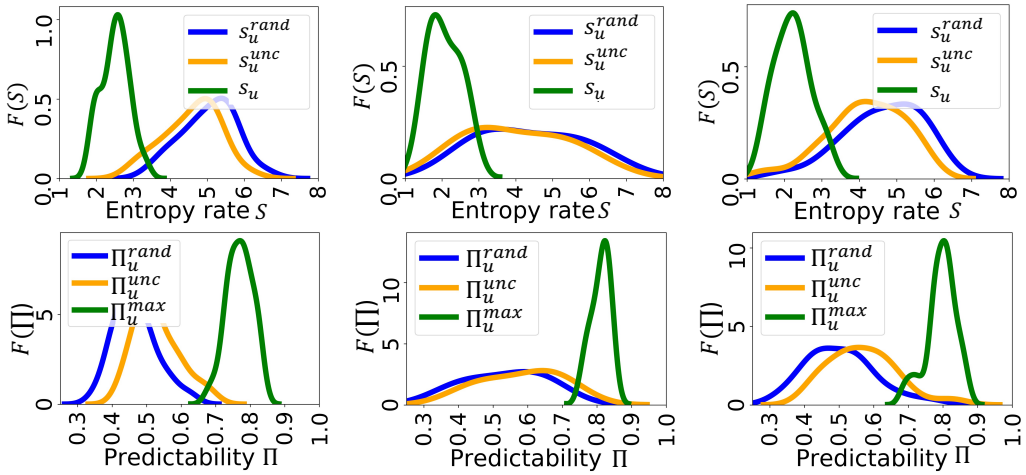
- *Routiners* relish exploring in specific areas and have compact spatial use when visiting. They use around 18% of the city center for explorations and also less than 19% for return activities, as shown in Figure 4.15b.
- *Regulars* favor visiting Neighborhoods within their vicinity when returning, but tend to go to more distant ones when exploring: 34% of the territory is used for both explorations and returns, as depicted in Figure 4.15c.

In what follows, we investigate the capacity of correctly forecasting exploring and returning activities with a coarse-grained spatial resolution. For each individual, we consider her sequence of visited Neighborhoods when exploring/returning as a stochastic process $\mathcal{X} = \{X_i\}$, where X_i is the i^{th} visited Neighborhoods.

Next, we assign the three entropy measures detailed in Section 2.4.2 to capture the degree of predictability of the sequences of visited Neighborhoods: (i) the random entropy S_u^{rand} that assumes that all Neighborhoods have the same probability of being visited; (ii) the temporal-uncorrelated entropy S_u^{unc}, which considers the visitation frequencies to the Neighborhoods but overlooks the temporal correlation; (iii) the actual entropy S_u that takes into account the visitation frequency of the Neighborhoods along with the order in which they were visited. Next, we evaluate the corresponding upper-bound theoretical predictability (cf. Section 2.4.2): (i) random predictability Π_u^{rand}; (ii) temporal-uncorrelated predictability Π_u^{unc}; (iii) real maximum predictability Π_u^{max}. Afterward, we compute the PDF of the three versions of the entropy and the corresponding predictability for the sequences of explorations (cf. Figure. 4.16) and the sequences of returns (cf. Figure. 4.17) for each mobility profile.

Figure 4.16 depicts the entropy rate distributions and the equivalent predictability distributions of individuals per profile *when exploring new places only*. We can

[2]Some Neighborhoods have light green shades; this implies that they were less visited compared to favorite ones and are not revisited as regularly visited places.



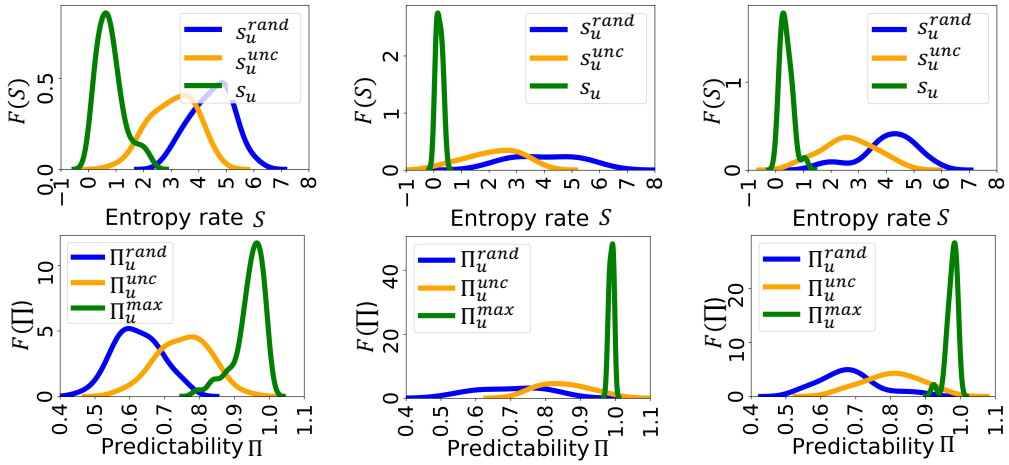
(a) Scouters. (b) Routiners. (c) Regulars.

Figure 4.16: Entropy and predictability of profiles when considering only explorations.

observe the important shift of the real entropy S_u (green curve) in all groups compared with the random entropy S_u^{rand} (blue curve) and the temporal-uncorrelated entropy S_u^{unc} (yellow curve). The PDF of the random entropy $F(S^{\text{rand}})$ picks at 5.8 for the *Scouters* and around 5 for the *Routiners* and the *Regulars*. This indicates that the next *Neighbourhood* where a *Scouter* is going to explore can be found among $2^{5.8} = 56$ *Neighbourhoods* and among $2^5 = 32$ *Neighbourhoods* for the others if the individual chooses her next location to explore in a random way. Instead, $F(S)$ picks around 3 for the *Scouters*, 2 for *Routiners*, and 2.5 for *Regulars*. In other words, the actual uncertainty in terms of the number of *Neighbourhoods* is about $2^3 = 8$ for *Scouters*, $2^2 = 4$ for *Routiners*, and $2^{2.2} \approx 5$ for *Regulars*.

Additionally, the PDF of the maximum predictability $F(\Pi^{\text{max}})$ picks at $\Pi^{\text{max}} \approx 0.78$ for *Scouters* and at 0.8 for *Routiners* and *Regulars*. This means that *only* at least 22% (20%) of the time, a *Scouter* (a *Routiner* or a *Regular*) chooses her location in a manner that appears to be random. This suggests that, though the apparent randomness of individuals' explorations, a historical record of an individual's discoveries hides an unexpectedly high degree of potential predictability on a *coarse-grained spatial resolution scope*.

Figure 4.17 depicts the entropy rate distributions of the three versions of entropy and the equivalent distributions of the upper bounds on the predictability distributions for returns. We can note that the PDF of the maximum predictability $F(\Pi^{\text{max}})$ narrowly peaks around $\Pi^{\text{max}} \approx 0.98$ for *Routiners*, then comes *Regulars* with a pick at $\Pi^{\text{max}} \approx 0.96$ than *Scouters* with a pick at $\Pi^{\text{max}} \approx 0.94$. Accordingly, we corroborate our mobility profiling through the spatial exploitation analysis.



(a) Scouters. (b) Routiners. (c) Regulars.

Figure 4.17: Entropy and predictability of profiles when considering only returns.

4.4 Summary

In this chapter, we shed light on exploration-like visits. First, we detailed our exploration-identification methods. Second, we proposed a new mobility profiling method, with the potential to capture individuals' propensity to explore new areas, namely, *Scouters* (adventurous and prone to explore); (ii) *Routiners* (steady and routinary), and (iii) *Regulars* (with medium behavior). Third, we extracted the mobility traits of each group and strengthened the subsisting dissimilarity between them. Following, to sustain our profiling method, we reported the profiles to spatial and temporal use. We unveiled individuals' temporal patterns on a weekly basis and showed that *Scouters*' proclivity to explore is very significant throughout the week. Finally, we showed that explorations in a coarse-grained spatial scenario are far from being random.

In the next chapter, we claim and show that exploration-like visits strongly impact mobility understanding and anticipation. Based on the proposed mobility profiles, we evaluate the impacts of novelty-seeking, quality of the data, and the prediction task formulation on the theoretical and practical predictability extents.

Chapter 5

Impacts of Novelty-seeking on Predictability Extent

Contents

To what extent is human mobility predictable? A frequently tacked question in mobility research that led to different predictive studies, either to infer the theoretical upper bound of the predictability (i.e., theoretical predictability) [86, 60, 61] of the mobility traces or the prediction accuracy achieved by different predictors (i.e., practical predictability) [62, 43, 32]. The large fluctuations among predictability results and among the accuracy of prediction results bring a new question: *what are the origins behind these significant variations in the predictability measures?* Alternatively stated, *what are the essential factors that influence predictability?*

To answer this question, prior investigations demonstrate that the quality of the data considerably affects the predictability, namely the temporal and spatial resolutions [16, 60, 83, 32]. Indeed, human mobility is substantially more predictable when using finer-grained temporal resolution or when increasing the size of spatial units. Another impacting factor is the prediction formulation. The literature reports a range of task formulations of the mobility prediction, namely, the next-cell, the next-place, the next-activity, or still the next-cell combined with contextual data (cf. Section 2.4.1). When the stationary nature of human movements is contained in the prediction task, the achieved scores are higher.

Withal, a non-negligible impacting factor and focus of this chapter is the tendency of individuals to explore and discover new places. Admittedly, novelty-seeking events are highly present in our daily lives since we continuously hunt for new places and spots [32]. Moreover, the susceptibility to break the returning routine to explore and discover new places is heterogeneous among populations [73, 78].

This chapter aims to evaluate explorations' effects and the quality of the data on the two most widespread next location prediction *next-cell* and *next-place* tasks individually. It first starts with a description of the employed evaluation methodology. Next, it investigates the performance reached in the next-cell formulation and presents the impacts of each of (1) the spatial resolution, (2) the temporal resolution, and (3) and the proclivity to explore. Next, it examines the attainable accuracy of prediction in the next-place prediction task and shows the effects of the impacting factor, in this case, novelty-seeking.

5.1 Evaluation Methodology

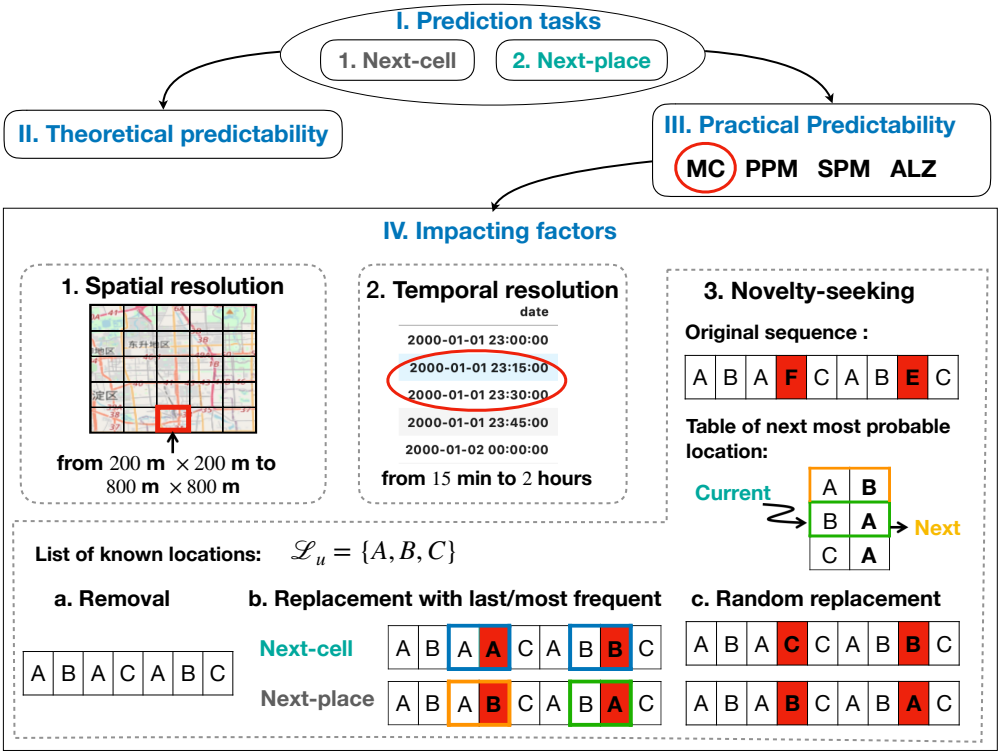


Figure 5.1: General overview of the applied methodology.

Hereafter, we describe the evaluation methodology to measure the impacts of data quality and individuals' tendency to explore on the two most widespread prediction tasks. Figure 5.1 gives a general overview of the applied methodology. Note

that the frequency of sampling for the Agg_gps is set to 15 min, i.e., $\delta_{Agg\_gps} = 15\,\text{min}$, and 1 h for the ChineseDB, i.e., $\delta_{ChineseDB} = 1h$. Besides, we use a squared tessellation with cells of size 200 m × 200 m, i.e., $c = 200\,\text{m}$ (cf. Table 7.1).

5.1.1 Prediction Tasks

There exist several ways to define the mobility prediction task depending on the quality of the available data and the objectives of the forecast (cf. Section 2.4.1). In what follows, we utilize the two most common prediction task formulations relying on personal location data only (cf. Figure 5.1–I):

- **Next-cell:** given a time window Δt, and let l_i be the current location of an individual. The next-cell prediction seeks to forecast the next location l_{i+1} of the individual at time $t + \Delta t$. This type of prediction can result in the current location as a future location for an individual, alternatively stated, the stationary nature of human trajectories is contained [83, 32].

- **Next-place:** this formulation is independent of the temporal dimension [52]. It seeks to answer the following question, *where will an individual go next?* The next-place prediction aims at forecasting changes in the spatial locality.

The three pillars of our evaluation methodology, i.e., Theoretical predictability, Practical predictability, and Impacting factors, as depicted in Figure 5.1, are detailed hereafter. The evaluations in Sections 5.2 and 5.3 follow such steps for the next-cell and next-place prediction tasks, respectively. An additional step is required for the next-place task (cf. Section 5.3) to remove stationary movements from the original sequences of visited locations.

5.1.2 Theoretical Predictability

For each prediction formulation, we start by measuring the theoretical predictability of the mobility behavior of each of the mobility profiles identified in Chapter 4, namely, the *Scouters*, *Regulars*, and *Routiners* (cf. Figure 5.1–II). This will provide insights into the capacity of correctly forecasting the mobility trajectories with an ideal and utter predictor. In this regard, we employ the state-of-the-art entropic-based approach proposed by Song et al. [86] (cf. Section 2.4.2) to estimate the upper bound of the theoretical predictability $\Pi^{\max}$.

For each user u of each mobility profile, given her discrete mobility trajectory $\mathcal{T}_{u,c}$, we consider the stochastic sequence $\mathcal{X} = \{X_i\}$, where X_i is the cell id of her i^{th} visited location. Then, we estimate the upper bound of the theoretical predictability $\Pi^{\max}$ of the X_1^n sequence of n records as in [86] (cf. Section 2.4.2).

5.1.3 Practical Predictability

Afterward, we estimate the practical predictability of each of the *Scouters*, *Regulars*, and *Routiners*. We compare the predictive performance of four state-of-the-art

predictors, namely, MC [30], PPM [64], SPM [53], and ALZ [47] (cf. Figure. 5.1–III, Section 2.4.3).

For the predictive performance comparison between the predictors, we measure the accuracy of the prediction achieved by each one (Eq. (2.6)). Given a stochastic sequence $X_1^n = \{X_1, \ldots, X_n\}$ of n observations capturing the trajectory of an individual u. For each predictor and each user u, we initialize (i.e., "warm-up") the considered predictor using the $n_s = \frac{2}{3} \times n$ first elements $X_1^{n_s}$ (i.e., 20 days for the Agg_gps and 10 days for the ChineseDB). Second, we use the predictor to forecast the next location X_{n_s+1}. After this forecast, we update the predictor by considering $n_s \leftarrow n_s+1$ first elements of the stochastic sequence X_1^n. We then repeat the second step while $n_s \neq n$. Finally, when $n_s = n$, we stop the iterations and compute the accuracy of prediction s_u (Eq. (2.6)) that can be written as,

$$s_u = \frac{1}{n - n_s} \sum_{t=n_s+1}^{n} \mathbb{1}(X_i = X_i^* | X_1^{i-1}), \tag{5.1}$$

where X_i is the actual location and X_i^* is the predicted value.

Experimental settings: For the MC(k) and PPM(k) predictors, we choose a $k \in [\![1, 2]\!]$. A k-th order MC predictor bases its forecast solely on the k previous observations. In contrast, a k-th order PPM model employs a combination of MC(j) models with $j \in [\![0, k]\!]$ [64]. For the SPM(α), we choose $\alpha \in \{0.1, 0.9\}$. α represents the fraction of the maximal suffix employed to predict the future location. Note that the *maximal suffix* is the immediately longest foregoing set of locations whose copy appeared in the previous location history.

5.1.4 Impacting Factors

Finally, we evaluate the impacts of each of the quality of the data and individuals' tendency to explore when relevant on the predictive performance achieved by each prediction task (cf. Figure 5.1–IV).

1. **Spatial variation procedure:** We investigate the effects of varying the spatial resolution on the accuracy of prediction s for individuals of each profile. We apply this variation with the next-cell formulation only, given that for the next-place prediction task, real POIs identification are favored [52, 32] (cf. Figure 5.1–IV.1).

2. **Temporal variation procedure:** In the case of next-cell prediction, we investigate the effects of varying temporal resolution on the accuracy of prediction s for each mobility profile. Provided that the next-place prediction task is independent of the temporal resolution, we do not investigate the impacts of the quality of the data factor on this formulation (cf. Figure 5.1–IV.2).

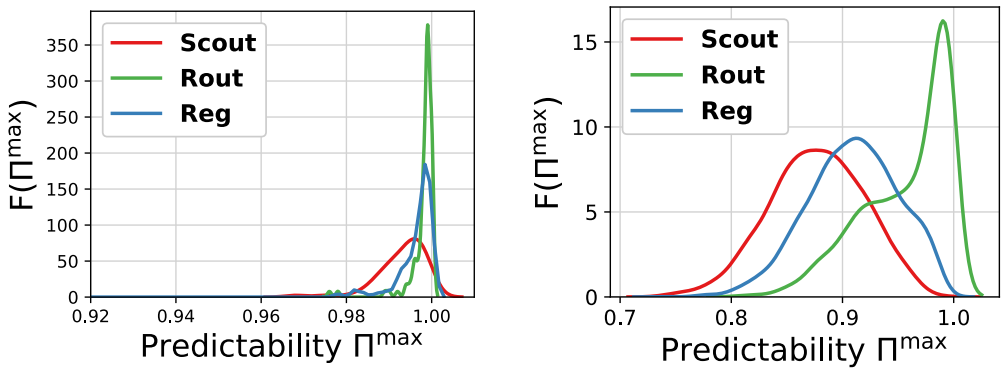
3. **Exploration-like visits isolation procedure:** We identify exploration-like visits using Algorithm 3 and remove them from the mobility trajectories or

replace them and observe how they affect the predictors' performances. These manipulations are performed for both prediction tasks but in different ways for the replacement procedures (cf. Figure 5.1–IV.3).

5.2 Next-cell

In this section, we tackle the next-cell prediction task. We first measure and analyze the theoretical and practical predictability of the mobility traces of individuals of each profile. Next, we investigate the effects of varying the spatial and temporal resolutions on the accuracy of prediction. Finally, we identify exploration-like visits and remove/replace them from/in the mobility trajectories, to probe the impacts of novelty-seeking on the predictive performance.

5.2.1 Theoretical Predictability



(a) Agg_gps.

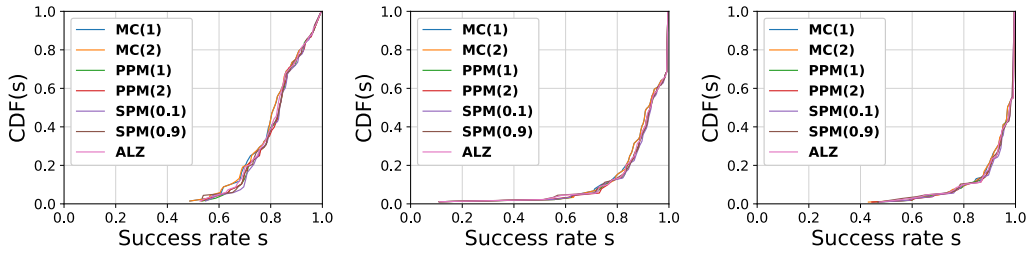
(b) ChineseDB.

Figure 5.2: Distributions of the upper bound of the theoretical predictability $\Pi^{\max}$ for individuals of each mobility profile.

Figure 5.2 portrays the distribution of the upper-bound predictability for each mobility profile for both the Agg_gps and the ChineseDB datasets. We can observe the high inherent predictability of the mobility trajectories of individuals of all profiles. Notably, individuals of the Agg_gps have a more eminent degree of potential predictability mainly due to the high frequency of sampling of the dataset $\delta_{Agg\_gps} = 15\,\text{min}$, while $\delta_{ChineseDB}$ is set to 1 h. Admittedly, a higher frequency of sampling allows a more complete capture of the stationarity and, consequently, increases the degree of predictability [32]. More importantly, from Figure 5.2b, we note that the predictability $\Pi^{\max}$ picks around 0.97 for the *Routiners*, 0.91 for the *Regulars*, and 0.87 for the *Scouters*. Taken together, these results indicate that *Routiners* are characterized by a very high degree of predictability while the *Scouters* are the least predictable individuals. Still, although presenting the lower predictability

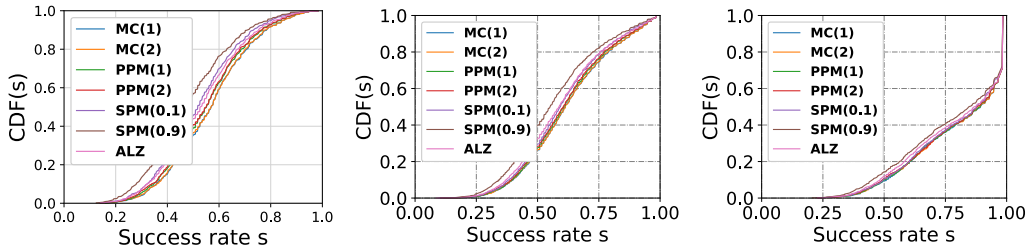
among the three mobility profiles, the *Scouters*' predictability is surprisingly high, mainly if considering the intuitive impossibility of predicting the uncertainties in *Scouters*' mobility. Indeed, as reported in Table 4.2, *Scouters* do have routines that consist of small sets of locations that they frequently visit.

5.2.2 Practical Predictability



(a) Scouters. (b) Regulars. (c) Routiners.

Figure 5.3: Distribution of the success rate score s_u of each predictor per mobility profile for the Agg_gps dataset.



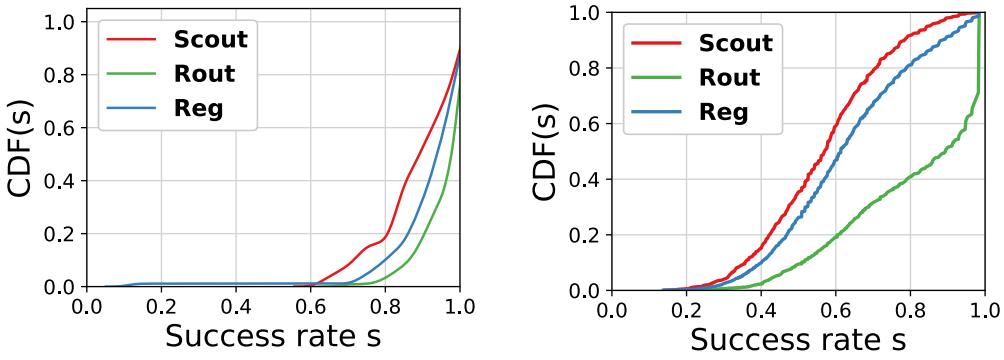
(a) Scouters. (b) Regulars. (c) Routiners.

Figure 5.4: Distribution of the success rate score s_u of each predictor per mobility profile for the ChineseDB dataset.

Estimating the predictability upper bound of individuals' trajectories reveals the high potential of predictability for all the profiles, with a lower score for Scouters (i.e., at most 0.87 in the ChineseDB dataset). Nevertheless, the prediction accuracy does not always reach the score provided by the theoretical measure [62] (see Section 2.4.3). Hereafter, we evaluate the accuracy of prediction achieved by each of MC, PPM, SPM, and ALZ.

Figures 5.3 and 5.4 plot the Cumulative Distribution Function (CDF) of the success score s of MC, PPM, SPM, and ALZ predictors concerning their possible parameters $k \in \{1, 2\}$ for MC and PPM and $\alpha \in \{0.1, 0.9\}$ for SPM. There is, however, little difference between the performance of the predictors. In the ChineseDB

dataset, where we leverage a large number of users, for both *Scouters* and *Regulars*, the best performances are achieved by the MC models. In contrast, the SPM achieves the lowest performances, particularly with $\alpha = 0.9$. For the *Routiners*, we observe that the performance of these predictors varies slightly with different settings. In general, the achieved performances by the distinct predictors are substantially comparable. Therefore, we only employ the MC(1) for our subsequent analyses.



(a) Agg_gps.

(b) ChineseDB.

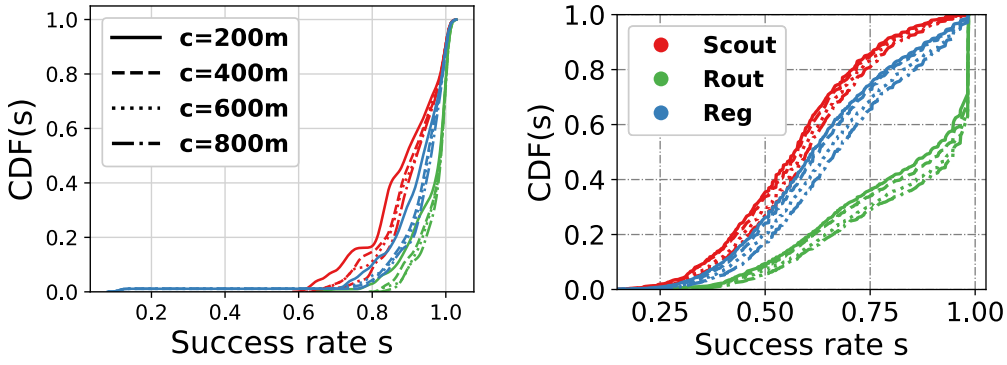
Figure 5.5: Distribution of the success rate score s_u of the MC(1) predictor per mobility profile.

For comparison simplification reasons, Figure 5.5 reports the distribution of the practical predictability of the MC(1) predictor for all of the *Scouters*, *Regulars*, and *Routiners*. We can notice that the best performances are obtained with *Routiners* and the lowest ones with the *Scouters*. We emphasize that *Scouters* are the hardest category of people to predict. However, they still present moments of regularity and, thus, with high accurate prediction results (i.e., 80% of *Scouters* have an accuracy of prediction s above 80%).

5.2.3 Impacting Factors

We now investigate the impacts of temporal frequency of sampling, spatial resolution, and exploration-like visits on the next-cell prediction formulation.

1. **Spatial resolution variation:** In Figures 5.6a and 5.6b, we investigate the correlation between the size of the geographical cells and the accuracy of prediction s per mobility profile. For this purpose, we vary the size of the squared tessellations $c \in \{200, 400, 600, 800\}$ meters. Intuitively and according to previous studies [62, 32, 67], the smaller the locations are, the less stationary behavior ascertained in the mobility trajectories of the individuals is. Hence, the less predictable they are.

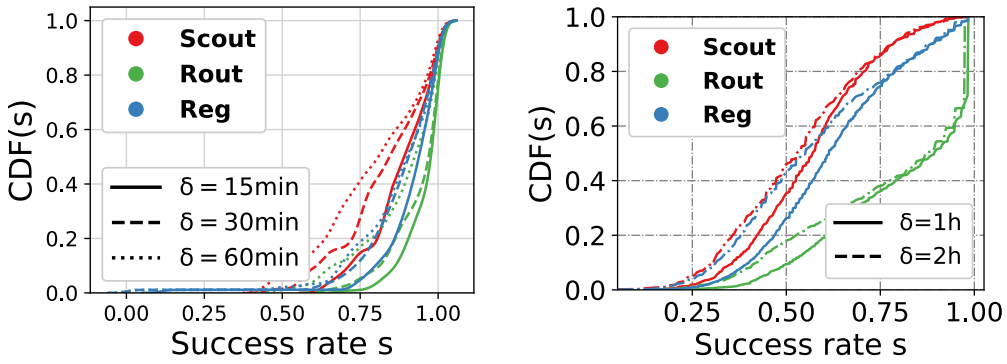


(a) Agg_gps. (b) ChineseDB.

Figure 5.6: Effect of spatial granularity on the success rate score s_u of the MC(1) predictor per mobility profile.

Not surprisingly and in agreement with previous studies, the prediction accuracy improves substantially with the increase in the size of the geographical cells. This is observed with individuals of all the profiles without any distinction.

2. **Temporal resolution variation:** We now examine how the frequency of sampling affects the ability to predict the mobility trajectories of each profile. We reset the spatial resolution to $c = 200\,\text{m}$, and vary the frequency of sampling $\delta_{Agg\_gps} \in \{15, 30, 60\}$ minutes for the Agg_gps dataset and $\delta_{ChineseDB} \in \{1, 2\}$ hours for the ChineseDB dataset.



(a) Agg_gps. (b) ChineseDB.

Figure 5.7: Effect of temporal granularity on the success rate score s_u of the MC(1) predictor per mobility profile.

Figures 5.7a and 5.7b show that the prediction accuracy decreases with the

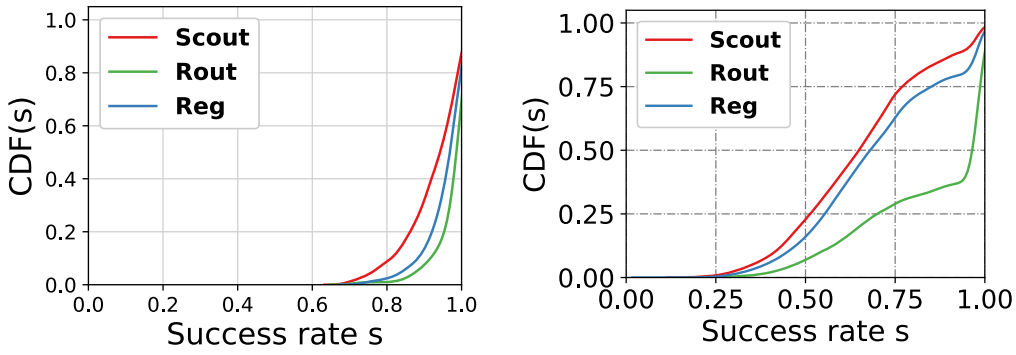
increase in the temporal resolution (when δ takes larger values). Indeed, the larger the sampling frequency, the harder the capture of the stationary behavior of individuals' mobility.

3. **Exploration-like visits isolation:** We want to scrutinize the impacts of novelty-seeking on the predictability of users' trajectories. In the following, we reset the spatial resolution to $c = 200\,\text{m}$ and the temporal resolution to $\delta_{Agg\_gps} = 15\,\text{min}$ and $\delta_{ChineseDB} = 1\,\text{h}$. For each user u, we use the proposed methodology presented in Algorithm 3 with the refined exploration identification (cf. Section 4.1.5.2) and $level = 80\%$ to classify her locations into EL and RL.

   To evaluate the impacts of novelty-seeking on the accuracy of prediction s achieved by MC(1), we adopt three approaches as detailed in Figure 5.1–IV.3:

   - ***1st proof-of-impact case***: We remove exploration-like records for all profiles and measure the accuracy of prediction s achieved by MC(1) with the new sequences (cf. Figure 5.1–IV.3.a). Clearly, this removal decreases the size of the trajectories and consequently can impact the accuracy of prediction. The corresponding results are depicted in Figure 5.8.
   - ***2nd proof-of-impact case***: As a first countermeasure to avoiding this size-related impact, we replace the exploration-like records with the last symbol met in the sequence (cf. Figure 5.1–IV.3.b). This action has the effect of adding a stationary period (equal to the size of each novelty-seeking period + 1). This approach is operated to assess whether the performance of the MC(1) predictor is only affected by the change in the length of the trajectories or if *the exploration-like visits play a role.* This substitution procedure favors the predictor once the stationary behavior is enhanced, as shown in Figure 5.9.
   - ***3rd proof-of-impact case***: As a second countermeasure to avoid both size-related impacts and stationarity increase, we identify exploration-like visits and substitute them with a random symbol found in the sequence (cf. Figure 5.1–IV.3.c). This procedure allows tackling both size-related effects and attenuating stationarity betterment impacts. Figure 5.10 shows the obtained results.

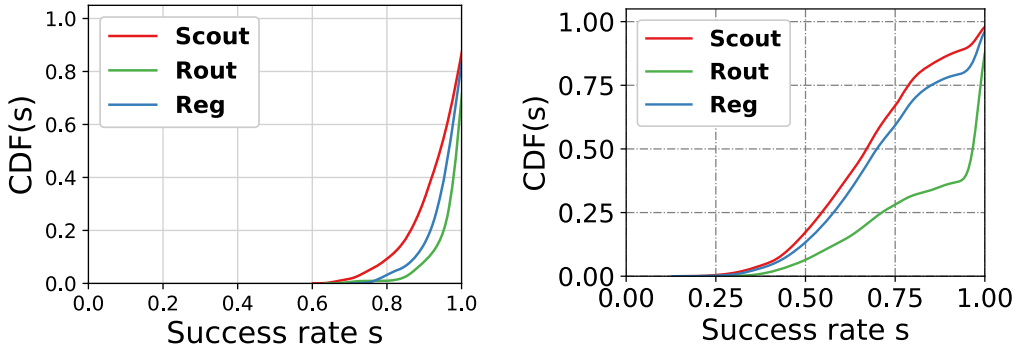
   The performance of MC(1) predictor indicates that: while the accuracy of prediction s is, on average, less than 60% (resp. 90%) for the least predictable class of users –i.e., *Scouters* – in the ChineseDB (resp. Agg_gps) dataset when considering exploration-like records (see Figure 5.5), Figure 5.8 shows that the predictor is considerably enhanced and can achieve an accuracy of prediction (on average) at least as high as 70% (resp. 95%) after removing exploration-like records. We have two hypotheses to explain this enhancement in the prediction accuracy: **H1:** the more irregular visits are omitted from the discrete mobility trajectory $\mathcal{T}$ of a user u, the more predictable she is. **H2:**



(a) Agg_gps. (b) ChineseDB.

Figure 5.8: Effect of novelty-seeking records removal on the success rate score s_u of the MC(1) predictor per mobility profile.

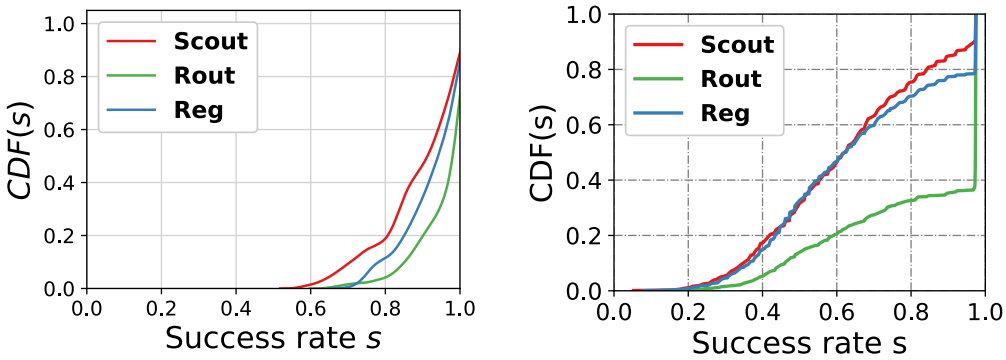
decreasing the lengths of a discrete mobility trajectory $\mathcal{T}$ allows the predictor to achieve better performance.



(a) Agg_gps. (b) ChineseDB.

Figure 5.9: Effect of novelty-seeking records replacement with stationarity stuffing on the success rate score s_u of the MC(1) predictor per mobility profile.

Replacing exploration-like visits allows us to assess one of the origins of the betterment in the predictive performance of the MC(1) predictor. Figures 5.9a and 5.9b show that when replacing exploration-like visits by adding stationarity, the accuracy of prediction is, in fact, further improved compared to the removal approach. Whereas the replacement of exploration-like visits with random locations does not necessarily improve the performance compared to the removal approach, it still achieves comparatively higher performances concerning the original trace (see Figure 5.10). Particularly, *Scouters* represent the most vulnerable category to the exploration phenomenon (their average



(a) Agg_gps.

(b) ChineseDB.

Figure 5.10: Effect of novelty-seeking records random replacement on the success rate score s_u of the MC(1) predictor per mobility profile.

prediction accuracy s is above 60%.). These findings allow us to corroborate the harmful effects that exploration-like visits have on the predictive performance of the classical MC predictor. Moreover, *Scouters* are more affected by these events, as shown in Figures 5.9 and 5.10. The isolation of these events engendered substantial improvements in the practical predictability of the *Scouters* compared to the other profiles.

Summarizing remarks: In a nutshell, in the next-cell prediction task, individuals of all profiles are impacted by both data quality and novelty-seeking. Increasing the temporal resolution of the data or enlarging the spatial cells' size allows achieving higher accuracies of prediction s. Moreover, high performances are usually achieved with this prediction task mainly due to stationarity effects, but moments of novelty-seeking do alter the predictive performance.

5.3 Next-place

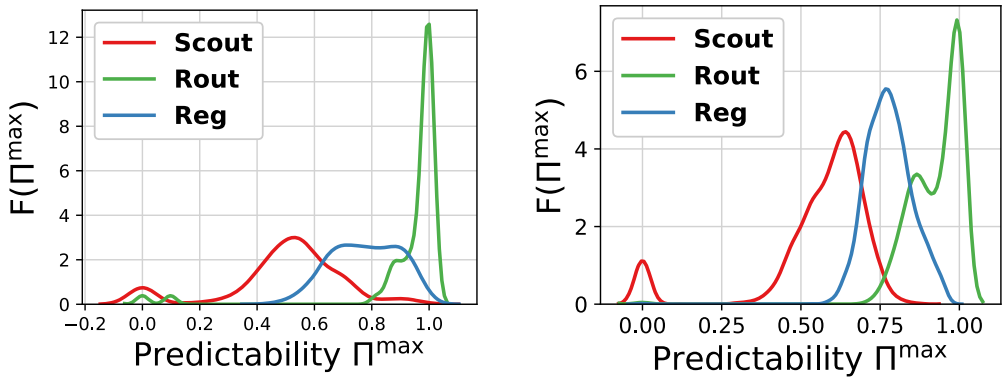
In this section, we tackle the next-place prediction formulation. We first reconstruct the discrete mobility trace $\mathcal{T}_{u,c}$ of each individual by removing stationarity records to fit the next-place prediction scenario. Next, we measure the theoretical $\Pi^{\max}$ and practical s predictability of the discrete mobility trace $\mathcal{T}_{u,c}$. After that, since this formulation of prediction is independent of the temporal resolution, we do not investigate the impacts of the data quality factor on this formulation of the prediction task. Finally, we measure the predictability of the three mobility profiles when removing and replacing exploration-like visits.

5.3.1 Discrete Mobility Trajectories Refurbishment

The next-place prediction formulation refers to the prediction of transitions between places. This formulation is more exposed to uncertainty as the stationarity behavior is omitted. Namely, the next-place prediction is about forecasting the next location where an individual will be, and that should be different from her current one. Thereby, given the discrete mobility trace $\mathcal{T}_{u,c} = \langle (l_0, t_0), (l_1, t_1), \ldots, (l_n, t_n) \rangle$ of a user u, we identify consecutive tuples that have the same location l and keep only the first tuple. Note that the frequency of sampling δ is not constant in this case, and the size of the mobility trajectories is smaller.

5.3.2 Theoretical Predictability

For each user u of each profile, as in Section 5.2, we estimate the upper bound of the theoretical predictability $\Pi^{\max}$ of the stochastic sequence $X_1^n = \{X_1, \ldots, X_n\}$ extracted from her refurbished discrete mobility trajectory $\mathcal{T}_{u,c}$ (cf. Figure 5.1–II).



(a) Agg_gps.

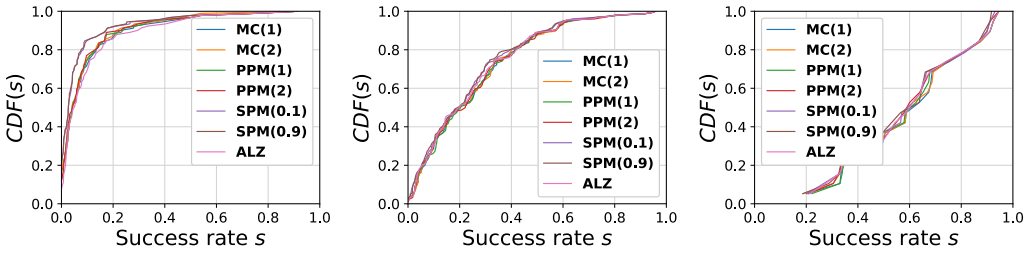
(b) ChineseDB.

Figure 5.11: Distributions of the upper bound of the theoretical predictability $\Pi^{\max}$ for individuals of each mobility profile.

The distributions of the upper bound of the theoretical predictability $\Pi^{\max}$ for individuals of each mobility profile are presented in Figure 5.11. Consistent with findings from previous studies [32], the predictability is markedly decreased for both Agg_gps and ChineseDB datasets. Additionally, Figure 5.11 reveals that *Scouters* are still the least predictable individuals, even in this formulation of human mobility prediction, while *Routiners* are the most predictable ones.

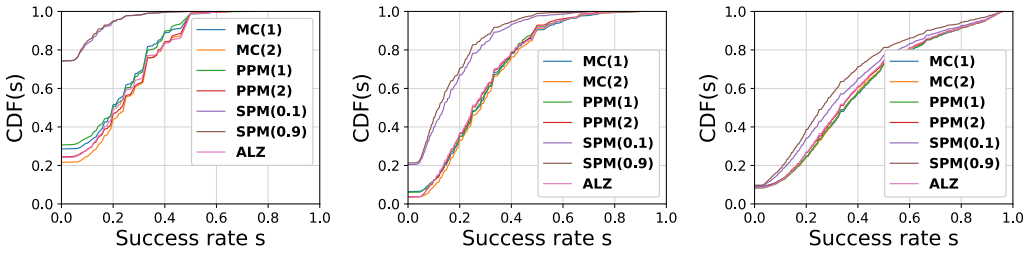
5.3.3 Practical Predictability

We evaluate the predictive performance achieved by the four predictors MC, PPM, SPM, and ALZ, with the next-place prediction task.



(a) Scouters. (b) Regulars. (c) Routiners.

Figure 5.12: Distribution of the success rate score s_u of each predictor per mobility profile for the Agg_gps dataset.



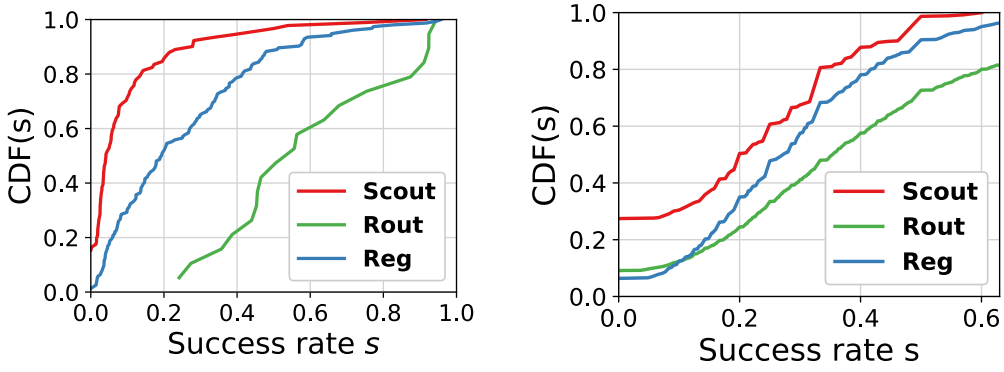
(a) Scouters. (b) Regulars. (c) Routiners.

Figure 5.13: Distribution of the success rate score s_u of each predictor per mobility profile for the ChineseDB dataset.

We apply the four predictors, MC, PPM, SPM, and ALZ, to the next-place prediction task (cf. Figure 5.1–III).

Figures 5.12 and 5.13 show the accuracy of prediction s achieved by each predictor with individuals of each profile. Clearly, the accuracy of prediction s is markedly lower than in the next-cell prediction task. In particular, the SPM performs poorly with the next-place prediction, especially with *Scouters*. The remaining predictors have comparable performances, with an average accuracy around 10%, 24%, and 60% (25%, 26%, 34%) for Agg_gps (ChineseDB) dataset for *Scouters*, *Regulars*, and *Routiners*, respectively. The achieved performances by the distinct predictors are substantially comparable. Therefore, to homogenize with the next-cell evaluation in what follows, we use MC(1).

For comparison simplification, Figures 5.14a and 5.14b display the accuracy of prediction of the MC(1) predictor in the next-place prediction scenario in CDF curves, one for each mobility profile: *Scouters*, *Regulars*, and *Routiners*. We can observe that the MC(1) predictor fares poorly, notably with the *Scouters*, where 85% of them have an accuracy of prediction below 20% in the Agg_gps dataset and below 40% for the ChineseDB. This conveys that the uncertainty in a typical



(a) Agg_gps.

(b) ChineseDB.

Figure 5.14: Distribution of the success rate score s_u of the MC(1) predictor per mobility profile.

individual's mobility trace is more significant than in the next-cell prediction.

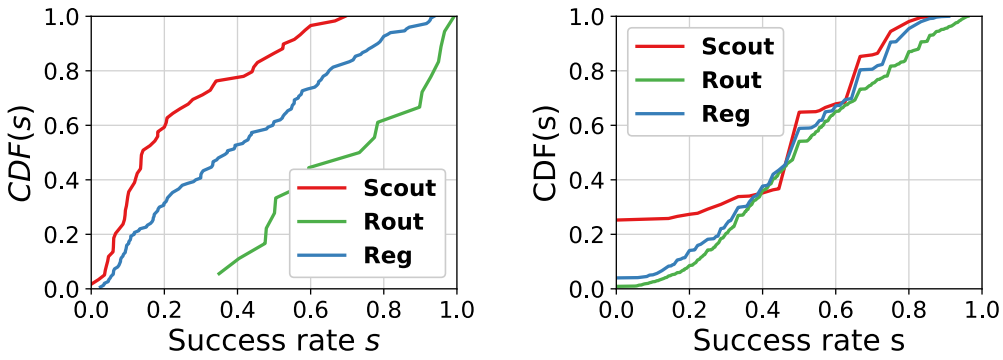
5.3.4 Impacting Factors

Recall that we only evaluate the impacts of exploration-like visits on the prediction accuracy in this prediction formulation. The next-place prediction task is independent of the temporal resolution and varying the spatial resolution is not adequate [52].

Exploration-like visits isolation: We now analyze the impacts of exploration events on the next-place prediction formulation. We start by identifying exploration-like visits per user using the visitation frequency-based methodology Algorithm 3 with $level = 80\%$. Next, we employ three methods to emphasize the impacts of novelty-seeking, also exemplified in Figure 5.1–IV.3:

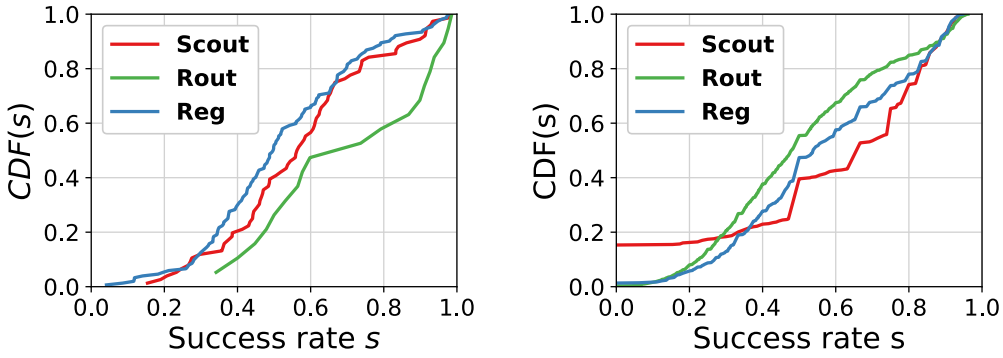
- ***1st proof-of-impact case***: As in the next-cell prediction analysis, we remove the exploration-like visits (cf. Figure 5.1–IV.3.a).
- ***2nd proof-of-impact case***: To avert size-related impacts, unlike in the previous prediction task, we do not replace exploration-like visits by adding stationary periods as it goes against the definition of the next-place formulation. Hence, given the last visited location "A" and the next visited one "C", if the current location "F" is assumed to be an exploration, we replace the exploration-like records "F" with the most frequent location that usually appears after "A" and different from "C" (which is "B" in Figure 5.1–IV.3.b). The results are depicted in Figure 5.16.
- ***3rd proof-of-impact case***: Slightly different from the 3rd proof of impacts of the previous prediction formulation, we replace exploration-like visits by a

random symbol met in the sequence that is different from the last and the next visited locations (cf. Figure 5.1–IV.3.c). Figure 5.10 shows the obtained results.



(a) Agg_gps. (b) ChineseDB.

Figure 5.15: Effect of novelty-seeking records removal on the success rate score s_u of the MC(1) predictor per mobility profile.



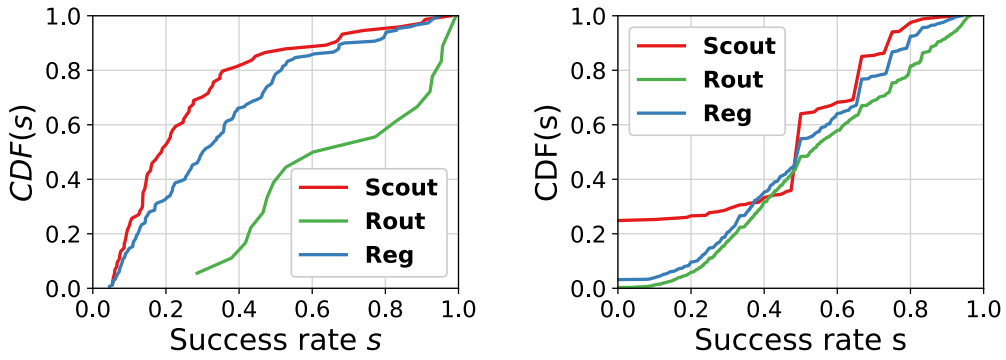
(a) Agg_gps. (b) ChineseDB.

Figure 5.16: Effect of novelty-seeking records replacement with stationarity stuffing on the success rate score s_u of the MC(1) predictor per mobility profile.

Figure 5.15 displays the accuracy for the MC(1) predictor while keeping only familiar visits in the mobility traces. The prediction accuracy is remarkably enhanced compared to the next-cell formulation case for all profiles. Notably, the average score is above 15% (above 50%) for the Agg_gps (ChineseDB) for *Scouters*.

The replacement of novelty-seeking places by the most probable known location further enhances the performance, particularly for *Scouters* (see Figure 5.16).

Further, we can discern the substantial harmful effects of exploration-like visits



(a) Agg_gps. (b) ChineseDB.

Figure 5.17: Effect of novelty-seeking records random replacement on the success rate score s_u of the MC(1) predictor per mobility profile.

on the predictability in the next-place prediction compared to the next-cell prediction. More importantly, the results show that *Scouters* are more impacted by the isolation of exploration-like records. The original median accuracy for *Scouters* is approximately less than 20% (see Figure 5.14), which is significantly lower than the performance of other profiles. Therefore, the removal or replacement of explorations events makes *Scouters* roughly as predictable as the other profiles.

Summarizing remarks: The next-place prediction task is a more challenging problem for individuals of all profiles. This formulation is more vulnerable to uncertainties as the stationarity behavior is overlooked. Therefore, the harmful effects of exploration activities are more discernible and have more impacts on the predictive performance. Essentially, understanding individuals' tendencies to explore and discover new places can benefit next-place-based predictors. Indeed, quantifying and anticipating individuals' inclinations for novelty-seeking can help predictors to enhance their performance by looking at further contextual data or collective mobility behavior.

5.4 Summary

In this chapter, we took a fresh look at the most significant factors affecting the predictability extent of individuals' mobility traces, namely, the prediction formulation, the spatial and temporal resolutions, and novelty-seeking. Utilizing the mobility profiling method developed in Chapter 4, we analyzed the effects of each factor on the predictability per profile. Following previous studies, we showed that regardless of the mobility profiles, the next-cell prediction achieves higher degrees of practical

and theoretical predictability compared to the next-place formulation. Mainly as a result of the high stationarity present in the next-cell prediction task. Besides, we asserted that increasing the size of the spatial cells leads to the increase of the stationarity and, hence, in the accuracy of prediction. Similarly, a fine-grained temporal resolution allows a higher capture of consecutive records with the same cell-id and, consequently, a growth in stationarity, implying higher prediction scores. More importantly, we shed light on the novelty-seeking phenomenon as a significant factor impacting predictability. Therefore, understanding the exploration phenomenon is fundamental to thoroughly predicting human movements.

In the next chapter, we propose a novel framework for adjusting prediction resolution when probable explorations are going to happen. Exploiting our previous findings, namely, the geographical occurrences of explorations are far from being random in a coarser-grained spatial resolution; instead of directly predicting a user's next location, we design a two-step predictive framework.

Chapter 6

Exploration-Aware Mobility Predictor

Contents

Conventional *personal* predictors such as Markov-based models [41, 62, 44] or HMM [63] utterly rely on historic personal location data to predict future locations (cf. Section 2.4.1). Moreover, they predict an individual's next location on the assumption that it belongs to the set of her known places [93]. This engenders erroneous forecasts at each occurrence of an exploration event, which is worsened by the fact that such events are numerous and largely present in the daily lives of individuals: on average, 70% of visits happen only once [32]. This representative rate highlights how impacting exploration-intended visits are for conventional personal predictors and puts in evidence the need for detecting such types of movements.

Advanced contextual information has recently been jointly used with mobility data to better tackle exploration visits in predictions [93, 63]. Examples are the semantics of the visited location, the activity performed within the location, the personality traits of the user [75], or her social circle [32, 93]. Such contexts request massive data collection and bring privacy concerns [17, 91]. Although possibility enhancing prediction, due to lack of contextual data and privacy concerns, we focus our investigations uniquely on users' mobility data, in this thesis.

This chapter proposes a newly tailored mobility prediction framework that tackles the exploration problem by leveraging the purpose of movements at prediction decisions, only using location data. Several works demonstrate human return visits' inherent temporal periodicity and spatial regularity [46]. Furthermore, Chapter 4 shows that, though the apparent randomness of exploration visits, their temporal

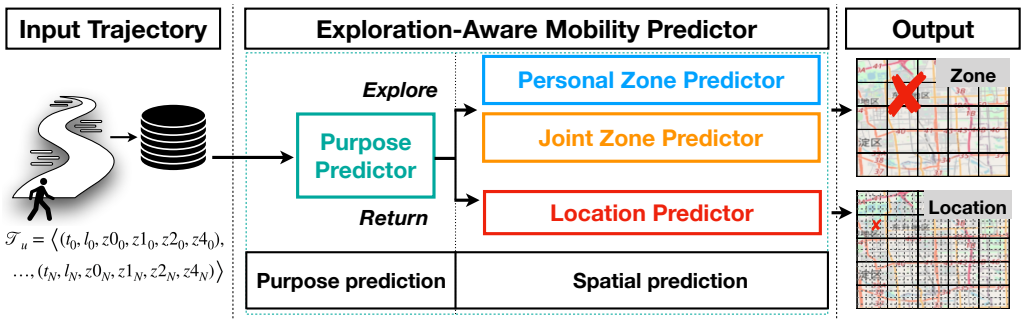


Figure 6.1: Exploration-Aware mobility Prediction Framework.

and geographical occurrences are far from random when considering coarser-grained spatial scopes [8]. Exploiting these properties – instead of conventionally predicting an individual's next location based on the history of known visited places –, we design a two-step prediction framework that encloses two modules: (i) the *purpose of movement predictor* and (ii) the *spatial predictor* (cf. Figure 6.1).

6.1 Purpose Prediction

Following recent practice [32, 85], we adopt the subsequent movement dichotomy presented in Chapter 4, i.e.,: (i) *explorations* or discoveries of new places e and (ii) visits of previously known locations termed *returns* r. This means the set of movements comprises two elements $\mathcal{M} = \{e, r\}$. The movement prediction task aims to answer the following question: *what will the individual do next? Explore or return?*

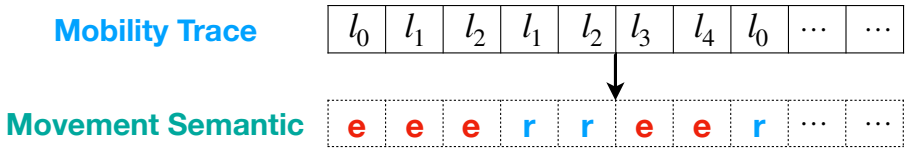


Figure 6.2: Adding movement semantic to a mobility trace.

In this chapter, as in [32], we consider the baseline identification of exploration events, i.e., the first occurrence of a location l_x in $\mathcal{T}_{u,c}$ is an *exploration* (e), else it is a *return* (r) (cf. Section 4.1.3.1). Thus, before browsing the mobility traces, each user u has an empty set of known locations $\mathcal{L}_u$. We then add movement semantic to each record $q_x \in \mathcal{T}_{u,c}$ in the mobility trace by associating the label r in case $l_x \in \mathcal{L}_u$. Otherwise, we associate the label e as depicted in Figure 6.2. When a location is first met, it is added to the set of known locations $\mathcal{L}_u$. Subsequently, we propose two approaches to forecast the next type of movement an individual will perform, described in the following sections.

6.1.1 Successive Types of Movements Predictor (STMP)

Algorithm 4 Successive Types of Movements Predictor (STMP)

1: **function** **STMP** $(\mathcal{T}_{u,c}, \mathcal{L}_u)$
2: **for** l **in** $\mathcal{T}_{u,c}$ **do** ▷ Successive same types of movement calculation
3: **if** $l \notin \mathcal{L}_u$ **then** ▷ Explorations
4: $nb\_exp_u \leftarrow nb\_exp_u + 1$
5: $\mathcal{L}_u$.**Add**(l)
6: $last \leftarrow$ **e**
7: **if** $nb\_ret_u > 0$ **then** ▷ Successive returns interruption
8: ret_u.**Add**(nb_ret_u)
9: $nb\_ret_u \leftarrow 0$
10: **end if**
11: **else** ▷ Returns
12: $nb\_ret_u \leftarrow nb\_ret_u + 1$
13: $last \leftarrow$ **r**
14: **if** $nb\_exp_u > 0$ **then** ▷ Successive explorations interruption
15: exp_u.**Add**(nb_exp_u)
16: $nb\_exp_u \leftarrow 0$
17: **end if**
18: **end if**
19: **end for**
20: $\mu_{exp}, \sigma_{exp} \leftarrow$**Stats**$(exp_u)$
21: $\mu_{ret}, \sigma_{ret} \leftarrow$**Stats**$(ret_u)$
22: **if** $last =$ **e** and $nb\_exp_u \in [\mu_{exp} \pm \sigma_{exp}]$ or $last =$ **r** and $nb\_ret_u \notin [\mu_{ret} \pm \sigma_{ret}]$ **then**
23: **return e** ▷ Predict an exploration
24: **else**
25: **return r** ▷ Predict a return
26: **end if**
27: **end function**

In the first proposed approach, we ignore the temporal dimension. In other words, only the order of occurrence of the types of visits is considered but not the elapsed time. To forecast the type of the $n + 1^{th}$ movement of the user u, we construct table exp_u that contains the number of successive explorations within the mobility trace $\mathcal{T}_{u,c}$ that comprises n records. When a user starts exploring a counter $nb\_exp \leftarrow 1$ is started. After each consecutive exploration, the counter is incremented $nb\_exp \leftarrow nb\_exp + 1$ until meeting a return or the end of the trace $\mathcal{T}_{u,c}$; the value of the counter is saved in table exp_u and reset to 0. Each time an exploration event occurs after a return, the process restarts again until reaching the end of the sequence (cf. Algorithm 4, lines 4–6). Likewise, we construct a table ret_u that contains the number of successive returns within the mobility trace $\mathcal{T}_{u,c}$ (cf. Algorithm 4, lines 12–13).

Following, we compute two values to characterize exploration visits, $\mu_{exp} = \text{mean}(exp_u)$ and $\sigma_{exp} = std(exp_u)$, which are the average and standard deviation (respectively) of successive explorations (cf. Algorithm 4, line 20). Similarly, we compute the average and standard deviation of successive returns (cf. Algorithm 4, line 21).

Final decision: According to the last type of movement, if the number of the successive same type of movement is included in the interval $[\mu_{type} \pm \sigma_{type}]$ with $type \in \{e, r\}$, STMP predicts the same movement as next. Otherwise, it predicts the opposite movement (cf. Algorithm 4, lines 22–26). For instance, if the last movement was an exploration e and the current number of successive exploration nb_exp_u is included in the interval $[\mu_{exp} \pm \sigma_{exp}]$, then an exploration e is predicted, else a return r is predicted.

6.1.2 Inter Exploration Interval Predictor (IEIP)

The temporal occurrence of exploration visits is the main and unique parameter considered in prediction decisions. To predict the $n + 1^{th}$ type of movement, we consider the trace $\mathcal{T}_{u,c}$ of size n. We focus on exploration events; as previously shown in Chapter 4, the temporal exploration activities appear to be regular. Therefore, we compute the IEI, i.e., the elapsed time between two consecutive explorations (cf. Algorithm 5, lines 2–5).

Final decision: If the elapsed time since the last exploration event is included in the interval $[\mu_{IEI} \pm \sigma_{IEI}]$, IEIP forecasts the next movement is an exploration. Else, it forecasts a return (cf. Algorithm 5, lines 9–13).

Algorithm 5 Inter Exploration Interval Predictor (IEIP)

```
1: function IEIP (T_{u,c}, L_u)
2: for (t, l) in T_{u,c} do                          ▷ IEI sequence computation
3:     if l ∉ L_u then
4:         tab_exp_interval.Add(t − last)
5:         last ← t
6:     end if
7: end for
8: μ_IEI, σ_IEI ←Stats(tab_exp_interval)
9: if t_{N+1} − last ∈ [μ_IEI ± σ_IEI] then
10:    return e                                      ▷ Predict an exploration
11: else
12:    return r                                      ▷ Predict a return
13: end if
14: end function
```

6.2 Spatial Prediction

Recall that return visits are highly foreseeable due to their high temporal periodicity and partial regularity [32]. Likewise, exploration visits are not completely random if we consider a coarse-grained spatial scope, as shown in Chapter 4. In what follows, we propose two spatial predictors leveraging the results of purpose predictors to improve exploration-like forecasts: (i) Personal Spatial Predictor (PSP), (ii) Joint Spatial Predictor (JSP). The description of such predictors is preceded by the introduction of two prediction methods used by PSP or JSP according to the type of

movement issued by the purpose predictors.

6.2.1 Prediction Methods

In this chapter, we consider spatial units of different sizes (cf. Appendix 7.3). Let $\mathcal{T}_u = \langle (t_0, l_0, z0_0, z1_0, z2_0, z4_0), \ldots, (t_n, l_n, z0_n, z1_n, z2_n, z4_n) \rangle$ be the mobility trace of the user u, with n records. We firstly leverage three distinct methods (cf. ME) for the next visits' prediction, accordingly adjusted to operate (i) on two spatial resolutions (i.e., location or zones) and (ii) with two mobility views (i.e., personal or joint):

- **(ME) MC Location-Predictor:** This predictor gets as input the mobility trace of an individual only. To predict the next location where the user will go, the *MC Location-Predictor* considers the stochastic sequence of the visited locations $X_0^n = l_0, l_1, \ldots, l_n$. Next, it trains a first-order MC predictor on the $n_s = \frac{2}{3} \times n$ first elements. Following, the *MC Location-Predictor* forecasts the next location l_{n+1} that the user will visit.

- **(ME) Personal Zone-Predictor:** We slightly modify the *MC Location-predictor* to predict the future coarse-grained zone, instead of locations, where the user will be. It considers the stochastic sequence of visited zones $X_0^n = zx_0, zx_1, \ldots, zx_n$. Then, similar to the *MC Location-Predictor*, it trains a first-order MC predictor using the first $n_s = \frac{2}{3} \times n$ elements of the coarse-grained sequence. Afterward, it forecasts the next visited zone zx_{n+1}.

- **(ME) Joint Zone-Predictor**: We go further here and design a benchmark predictor that leverages the collective exploratory mobility behavior of the population as input. Besides, prediction is made in terms of zones where a user will perform an exploration, in case she is more prone to discover a new place. First, it constructs an Exploration Origin-Destination (EOD) matrix $\mathbf{E}(t)$ at time t. The matrix gives an estimation of the probability of making a discovery in a zone j after visiting the location i. More precisely, the EOD matrix is of size $n \times m$, where n is the number of different "Origins" and m is the number of the different "Destinations". The "Origins" set contains *only locations* after which explorations happened. Hence, it is at most equal to the total number of locations visited by the users. The "Destinations" contains *all the distinct zones* where users explore. Each $\mathbf{e}_{i,j}$ gives the probability of exploring in the zone j after visiting the location i. For instance, in Figure 6.3, if the user in the location $L3$ is more prone to explore at time t, the *Joint Zone-Predictor* first constructs the EOD(t) matrix. Next, it identifies the most likely zone X where users usually explore after being in $L3$, i.e., the zone with the highest e_{L3X}. Finally, it suggests the zone $X = Z1$ as the next spatial unit where the user will explore, i.e., discover a new location.

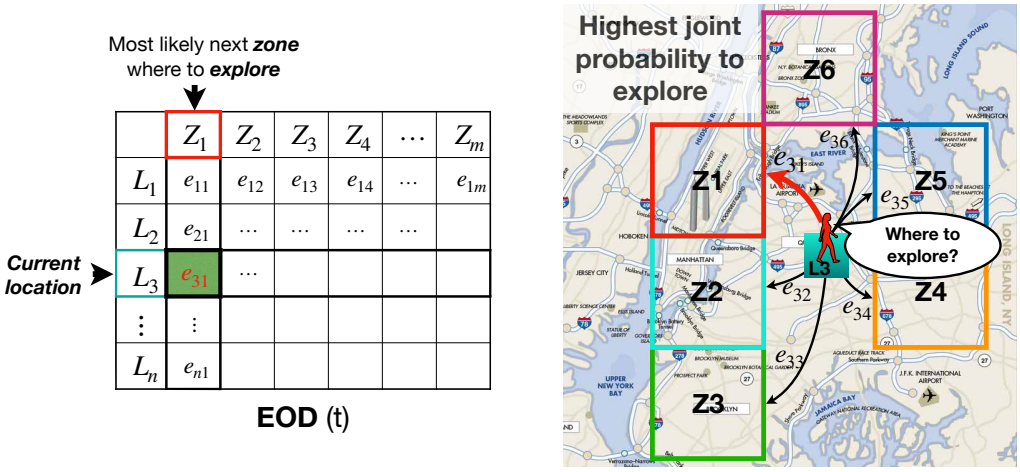


Figure 6.3: Joint Zone-Predictor.

6.2.2 Designed Spatial Predictors

Using the aforementioned prediction methods, we design two spatial predictors:

Personal Spatial Predictor (PSP): it takes the predicted type of movement as input. In case the forecasted movement is a return, it uses the *MC Location-Predictor* to forecast the next visited location. Hence it provides a fine-grained intuition on where the user will be next. On the other side, if the predicted movement is an exploration, the PSP employs the *Personal Zone-Predictor*. Accordingly, a coarse-grained spatial unit is returned.

Joint Spatial Predictor (JSP): similar to the PSP, the JSP takes the outcomes of a movement predictor as an entry. If the forecasted movement is a return, the *MC Location-Predictor* is used to infer the next location. Else, the *Joint Zone-Predictor* is used to infer the zone where an exploration might occur.

Note that the geographical accuracy of the proposed spatial predictors decays when the considered user is assumed to be exploring. Although this decay in performance, the inferred zones are of reasonable size that might lure and benefit many applications, such as recommendation systems or Multi-access Edge Computing infrastructures improvement.

6.3 Experiments

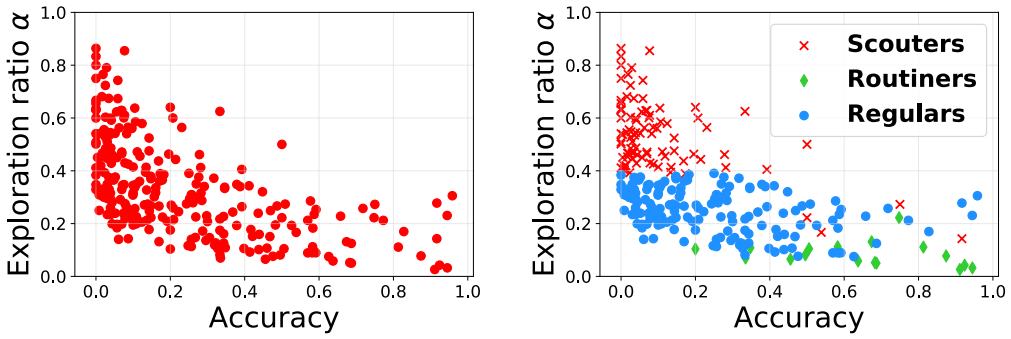
6.3.1 Identifying Hard- and easy-to-predict Users

Exploration activities are a key reason for the low accuracy of mobility prediction tasks (cf. Chapters 4 and 5). Individuals exhibiting a high tendency to explore are hence less likely to be foreseeable with predictors relying on past history. Therefore, to examine the efficiency of our proposed mobility prediction framework and investigate if it is more fitted to users who exhibit high exploration activities or is

beneficial to all users. We propose a simple mobility profiler that seeks to isolate hard-to-predict users because of their high exploratory activities from the rest of the population.

For each individual u with an n-long mobility trajectory, $\mathcal{T}_u$, we train a simple MC predictor on the first $n_s = \frac{2}{3}n$ records and predict the future location l_{n_s+1}. Then, we increment n_s and repeat until $n_s = n$. Afterward, we compute the accuracy of prediction that is a commonly used prediction evaluation metric (Eq. (2.6)), in this case, written as, $accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$. We also compute the exploration ratio for each user,

$$\alpha = \frac{\text{number of transitions of type exploration}}{\text{total number of transitions}}. \tag{6.1}$$



(a) All users.

(b) Per profile, as proposed in Chapter 4.

Figure 6.4: Accuracy of prediction vs. exploration ratio.

Figure 6.4a depicts the prediction accuracy achieved by a first-order MC predictor against the exploration ratio. In general, we can observe that the MC predictor performs poorly with users having a high exploration ratio, particularly those holding a ratio above 0.4.

In Figure 6.4b, we apply to the previous Figure 6.4a the profiling proposed in Chapter 4. *Scouters* are defined as users with a high tendency to explore, *Routiners* are individuals who rarely interrupt their returning routine to explore, and *Regulars* exhibit an intermediate behavior. We can observe that *Scouters* typically have an exploration ratio above 0.4 and hold the lowest predictive scores.

Accordingly, hereafter, we will first evaluate the performance of the proposed predictor on the whole population. Next, we will apply the proposed framework only to the hard-to-predict users and employ a simple MC with the rest of the population, as depicted in Figure 6.5.

We use the exploration ratio metric to determine if a user is hard-to-predict or not and variate the selection threshold in the set $Th = \{0.2, 0.4, 0.6\}$. So a user holding an exploration score above a given threshold is classified as hard-to-predict, else as an easy-to-predict individual. Hence, low Th results in a higher number of

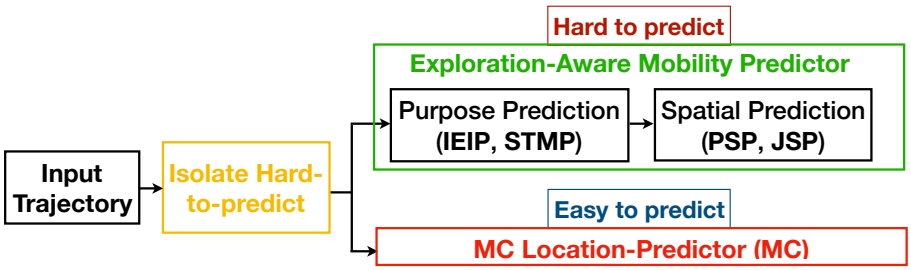


Figure 6.5: Global Prediction Framework.

hard-to-predict users to whom we apply a double prediction (movement and spatial). In contrast, high Th allows selecting users exhibiting high exploratory activities to be a candidate for a double prediction, while others are subject to a direct spatial prediction.

According to Figure 6.5, a ratio below 0.2 isolates *Routiners*, who get high accurate predictions with traditional MC due to their *easy-to-predict* mobility. The ratio range [0.2,0.4] isolates Regulars users. Finally, the values higher than 0.6 identify *Scouters* or *hard-to-predict* users (i.e., high probable users to get low traditional MC prediction accuracy) requiring the improvement of prediction methods. *Scouter* users trigger the exploration-enhanced MC predictor.

6.3.2 Purpose Prediction

Recall that our first goal is to infer an individual's next type of movement, given the movement history. In what follows, we evaluate the performance of each of the STMP and the IEIP. Given the imbalanced ratio between exploration and return transitions (cf. Appendix 7.3.1 Figure 7.3), to measure the performance of the proposed movement predictors in forecasting each type of movement, we employ three widely used information retrieval measures:

- **Precision** P: It is a measure of relevance. It shows the ability of a classifier not to label as positive a sample that is negative. It is defined as the number of true-positives T_p over the number of true-positives plus the number of false-positives F_p [3], $P = \frac{T_p}{T_p+F_p}$.
- **Recall** R: it measures a classifier's ability to find all positive samples. It is defined as the ratio between true-positive T_p samples over the number of true-positive T_p plus the number of false-negatives F_n [3], $R = \frac{T_p}{T_p+F_n}$.
- f_1**-score** $F1$: it is the harmonic mean of precision and recall [3], it is given by, $F1 = 2 \times \frac{P \times R}{P+R}$.

By considering returns as positive events and explorations as negative events, a true-positive result, T_p refers to the correct prediction of a return. A false-positive result F_p indicates that an exploration is predicted to be a return. A false-negative

F_n refers to a return predicted to be an exploration. A true-negative result T_n relates to the correct prediction of an exploration (see Table 6.1).

Table 6.1: Movement confusion matrix.

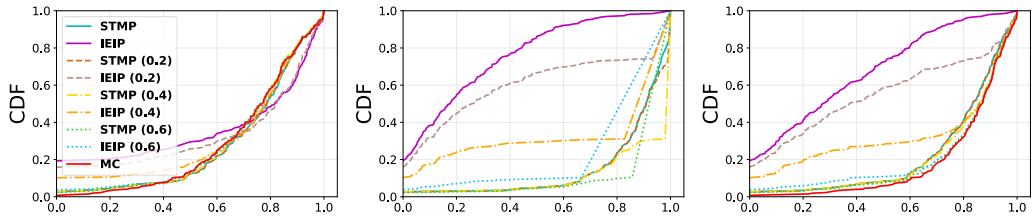
	Actual Return	**Actual Exploration**
Predicted Return	T_p	F_p
Predicted Exploration	F_n	T_n

We compare our proposed methods with the performance of the widely used state-of-the-art MC predictor. As in conventional personal predictors relying on location data, the MC predictor assumes that the next location a user will visit can be found in the set of known places [32, 93]. This means that the MC model is constantly forecasting returns. Thus, it holds the best scores in terms of predicting returns as a type of movement. Moreover, in view of the large proportion of returns compared to explorations (see Appendix 7.3.1 Figure 7.3), the MC allows evaluating how often the proposed algorithms are accurate in predicting the next type of movement. Alternatively stated, it helps to tune the movement predictors in favor of explorations/returns or to have global satisfying results.

Figures 6.6 and 6.7 represent the performance of the STMP and the IEIP in accurately predicting occurrences of return and exploration events, respectively. The proposed methods are first applied for the whole population, then only for hard-to-predict users, while an MC predictor is applied for the easy-to-predict users. For the hard-to-predict users' selection, as previously stated, we vary the exploration ratio's selection threshold α in the set $Th = \{0.2, 0.4, 0.6\}$, i.e., a lower threshold induces a higher labeling of hard-to-predict users.

Return visits: Figure 6.6 shows the precision, recall, and $f1$-score achieved by the STMP, the IEIP, and the conventional MC predictor. In Figure 6.6a, we can see that with more than 60% of the population, STMP and IEIP predictors achieve a precision above 70% in predicting returns. This indicates that more than 70% of the time, the forecasted returns are real revisits to known places. We also notice that the IEIP reaches the highest scores; the precision value is above 90% for more than 40% of the users. Furthermore, applying the proposed predictors only on users classified as hard-to-predict engenders slight changes in the precision scores when predicting returns.

Figure 6.6b depicts that the STMP succeeds at least 80% of the time in predicting returns for 80% of the population. On the contrary, the IEIP that focuses on exploration visits comes off badly in forecasting returns when applied to all users. Furthermore, we can observe that applying the proposed algorithms only on users exhibiting high exploratory activities allows improving the recall score, notably, the IEIP. Reducing the proportion of users to whom we apply the proposed algorithms leads to an increase in the probability of predicting returns. Hence, the obtained



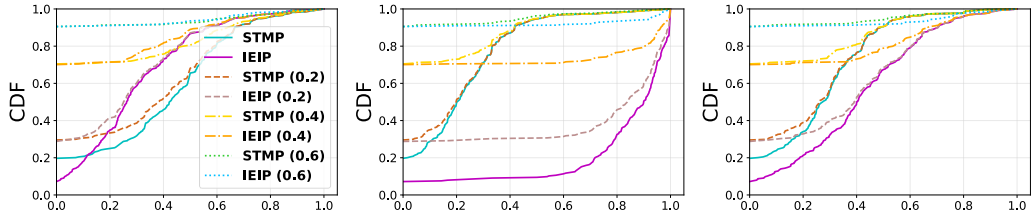
(a) Precision. (b) Recall. (c) F1-Score.

Figure 6.6: Performance comparison for returns forecasts.

recall scores are expected to improve with the decrease in the size of the selected hard-to-predict users. Curiously, the best average scores that we obtain with both of the STMP and IEIP movement predictors are when applying them to users having an exploration ratio above 40%, which approximately corresponds to the *Scouter* profile proposed in Chapter 4 [1].

As reported by the weighted average of precision and recall in Figure 6.6c, the STMP performs better than the IEIP in predicting returns. Namely, the average $f1$-score held by STMP exceeds 80%, and in general, its performance is very close to the MC's. On the contrary, the average $f1$-score reached by IEIP is less than 40% when applied to all users. Furthermore, Figure 6.6c reveals that increasing the exploration ratio for hard-to-predict user selection increases the achieved performance by both algorithms.

Exploration visits: Figure 6.7 presents the performance evaluation of the STMP and IEIP only, provided that a conventional predictor such as MC will always predict returns and fail at each discovery of a new location.



(a) Precision. (b) Recall (c) F1-Score.

Figure 6.7: Performance comparison for exploration type of movement forecasts.

Figure 6.7a shows that the precision achieved by STMP in predicting explorations for 60% of the population surpasses 35%. Whereas for the same proportion of the population, only 20% of the explorations inferred by IEIP are, in fact, dis-

[1]Note that in Figure 6.6b, the curve corresponding to the MC is not represented as the recall score is equal to 1 for all users.

coveries of new places. The partial application of the proposed STMP and IEIP to the population leads to a decrease in the attained precisions. Indeed, for users classified as easy-to-predict, the application of the MC to them induces null scores when predicting explorations.

Figure 6.7b depicts that for 60% of the population, at least 18% of the time explorations are correctly predicted by the STMP when applied for all users. Conversely, for the same percentage of users, more than 80% of moments of discovery are accurately foreseen by the IEIP. Similar to the measure of precision, the recall decreases with the decrease in the size of the selected population.

Figure 6.7c shows that the IEIP outperforms the STMP in predicting explorations, whether it is applied to all users or for the categories of users with a high proclivity to explore.

General visits: Next, we want to investigate how often the proposed movement predictors are accurate in predicting the next type of movement. We compute the accuracy of prediction achieved by the MC predictor and each of the proposed movement predictors. Here, the accuracy of prediction is the ratio between the number of correctly predicted types of movement and the total number of predictions.

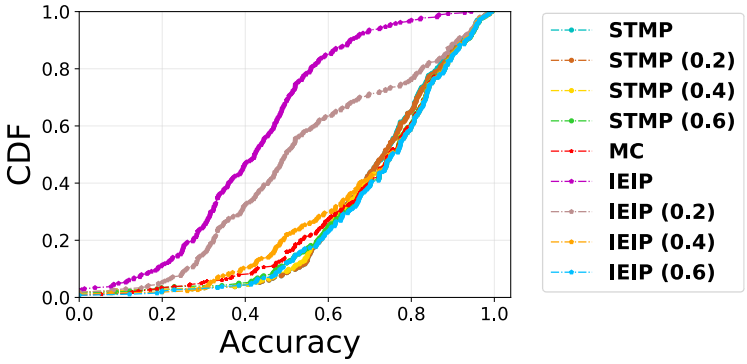


Figure 6.8: Accuracy of prediction of the type of movement.

Figure 6.8 shows the prediction accuracy in terms of types of movement achieved by STMP, IEIP, and MC. It is noted that apart from IEIP and IEIP (0.2), the other predictors perform almost equally well, but with a slight dominance of the IEIP (0.6). The accuracy of prediction is on average above 75% for the predictors except for the IEIP and IEIP (0.2) that have a score around 50%. Based on the aforementioned results, we can draw two main conclusions. First, the strategy adopted by conventional predictors, such as MC, assumes constant returns, which is a key lowering factor for the predictive performance. Figure 6.8 shows that, on average more than 20% of users' transitions are explorations. Second, the proposed

movement predictors are faced with a trade-off, gaining accuracy in predicting explorations at the cost of losing efficiency in forecasting returns. We point out that the proposed movement predictors are preliminary versions, that we aim to propose advanced versions in our future works. We claim here that the proposed movement predictors can be tuned to fit the requirements and needs of the using applications, unlike conventional models that focus on returns only.

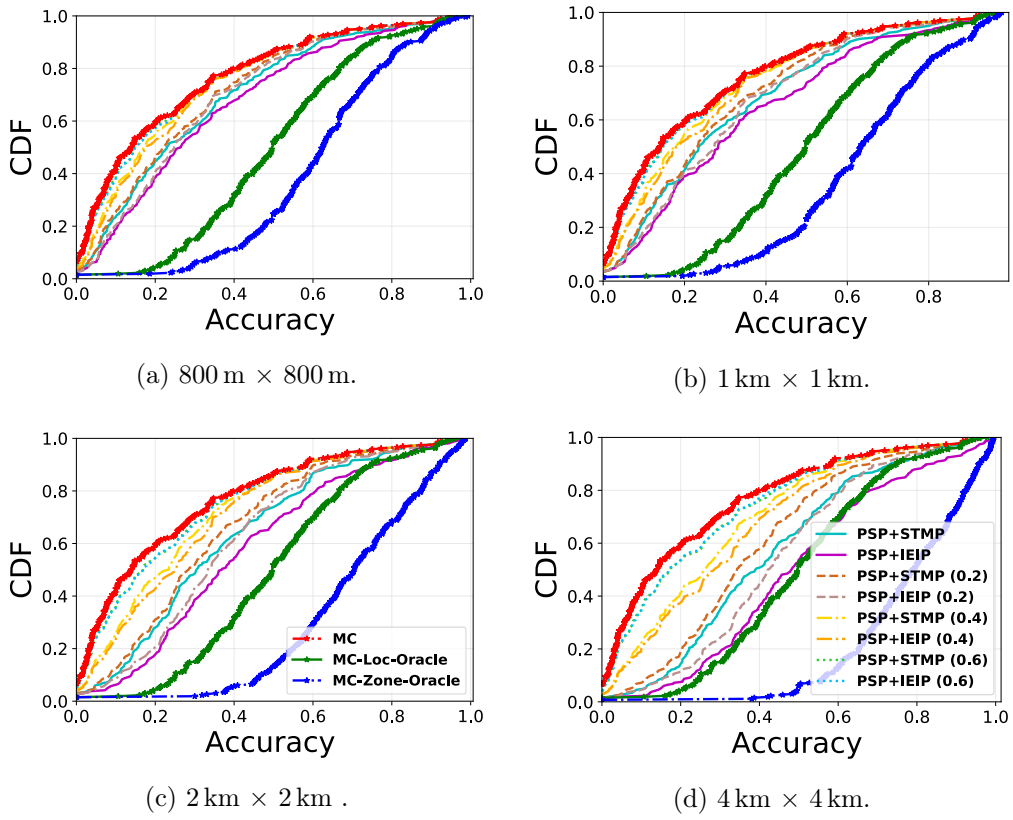
6.3.3 Spatial Prediction

We evaluate here the predictive power of the two proposed spatial predictors. First, the outcomes of the STMP and the IEIP predictors with their variations feed the spatial predictors. For the computation of the prediction accuracy, we consider a prediction to be correct if the inferred *location* or *zone* is correct. This is done through a comparison with three baseline predictors operating on different spatial scales: *MC Location-Predictor*, *MC Location-Oracle-Predictor*, and *MC Zone-Oracle-Predictor*.

- **MC Location-Predictor:** It is described in Section 6.2.
- **MC Location-Oracle-Predictor:** It performs as well as the *MC Location-Predictor* in predicting returns but reaches perfect scores in forecasting explorations. Given $\mathcal{T}_u = \langle(t_0, l_0, z0_0, z1_0, z2_0, z4_0), \ldots, t_x, l_x, z0_x, z1_x, z2_x, z4_x)\rangle$ the mobility trace of the user u, if next, the user makes a transition to a location l_{x+1} that is not present in $\mathcal{T}_u$, the *MC Location-Oracle-Predictor* will accurately predict l_{x+1}. This predictor holds the best feasible scores that an MC predictor endowed with a perfect movement predictor and spatial exploration forecaster can achieve.
- **MC Zone-Oracle-Predictor:** It follows the same strategy as the *MC Location -Oracle-Predictor*. Yet, it operates on a coarse-grained spatial resolution. Instead of predicting the next visited location, it predicts large zones. Given the trace $\mathcal{T}_u = \langle(t_0, l_0, z0_0, z1_0, z2_0, z4_0), \ldots, (t_x, l_x, z0_x, z1_x, z2_x, z4_x)\rangle$, this predictor forecasts the next zone zc_{x+1} of size c km $\times$ c km where the user is going to move. Besides, it is always accurate in predicting the right zone in case the user is exploring a new place. It holds the best achievable performance, and we take it as a reference to evaluate the efficiency of the proposed framework.

In what follows, we compare the performance of the PSP and JSP spatial predictors with the three baselines. As input, the spatial predictors receive the results from the STMP and the IEIP purpose schemes with their different settings. As previously indicated, locations are squared cells of size 200 m $\times$ 200 m. We consider 4 distinct sizes for the zones: 800 m $\times$ 800 m, 1 km $\times$ 1 km, 2 km $\times$ 2 km, and 4 km $\times$ 4 km.

PSP: Figure 6.9 reports the prediction accuracy of the PSP predictor with the different inputs and zone sizes. First, we can see that the prediction accuracy of



(a) 800 m × 800 m.

(b) 1 km × 1 km.

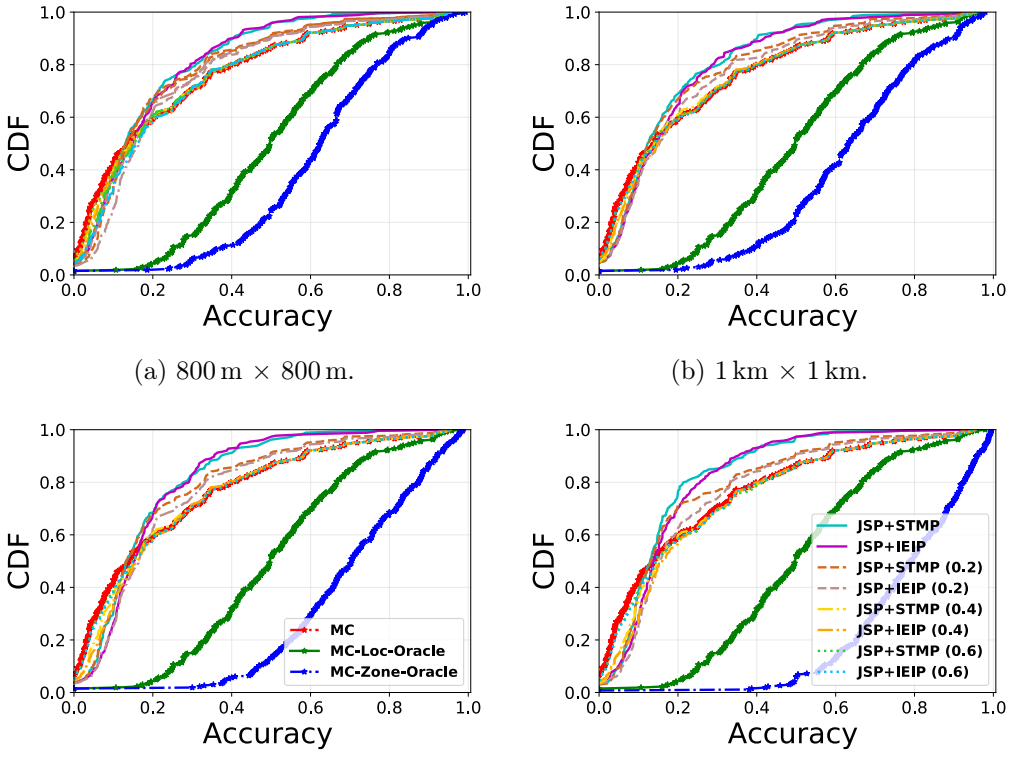
(c) 2 km × 2 km .

(d) 4 km × 4 km.

Figure 6.9: Accuracy of the PSP (common labels).

the PSP with the distinct inputs outperforms the *MC Location-Predictor*'s scores. Second, applying the proposed framework to larger proportions of users allows increasing the overall accuracy of prediction. We have two hypotheses with regard to the last observation: (a) applying the proposed framework is relevant to the whole population (b) by increasing the number of considered users, the number of false-negative forecasts (i.e., returns predicted to be explorations) increases and given the coarse-grained spatial prediction, in this case, the predictive performances are enhanced [32]. Third, with the expansion of the size of the zones, the accuracy of prediction of the PSP combined with the different movement predictors grows to approach the *MC Location-Oracle-Predictor* performance. Notably, the accuracy of prediction is substantially improved when considering zones of size 2 km × 2 km or zones of size 4 km × 4 km. We can also see that the PSP fed with the IEIP algorithm applied to the whole population slightly surpasses the score obtained by the *MC Location-Oracle-Predictor*. This is mainly due to false-negative forecasts.

JSP: Figure 6.10 depicts the prediction accuracy of the JSP and its different settings and the baseline spatial predictors. Unexpectedly, the accuracy of prediction is at most equal to the *MC Location-Predictor*. Besides, expanding the zone size helps in



(a) 800 m × 800 m.

(b) 1 km × 1 km.

(c) 2 km × 2 km.

(d) 4 km × 4 km.

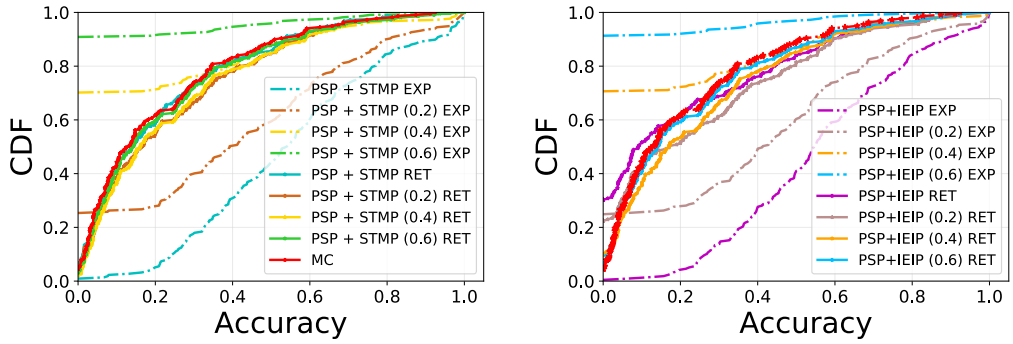
Figure 6.10: Accuracy of the JSP (common labels).

improving the minimal achieved scores only, but not the overall performance.

Spatial prediction for each type of movement: Next, to understand how the spatial predictors perform in predicting the locality of each type of movement. We report in Figure 6.11 and Figure 6.12 the CDF of the accuracy of the spatial prediction for each type of movement by the PSP and JSP, respectively. We depict the results with zones of size 4 km × 4 km. For other spatial resolutions of the zones, we provide the descriptive Tables 6.2 and 6.3.

Figure 6.11 shows the accuracy of predicting returns and explorations by the PSP predictor. First, we can see that the performance achieved by the combination of the PSP with the STMP in predicting returns is very close to the MC's, with slight improvements when applying the proposed framework partially to users exhibiting a high exploration activity. Specifically, when applying it to users having an exploration ratio α above 40% (see Figure 6.11a). When the PSP takes the IEIP's outcomes as input, the improvement in the predicting returns is more noticeable. Namely, when using the proposed framework with hard-to-predict users (see Figure 6.11b).

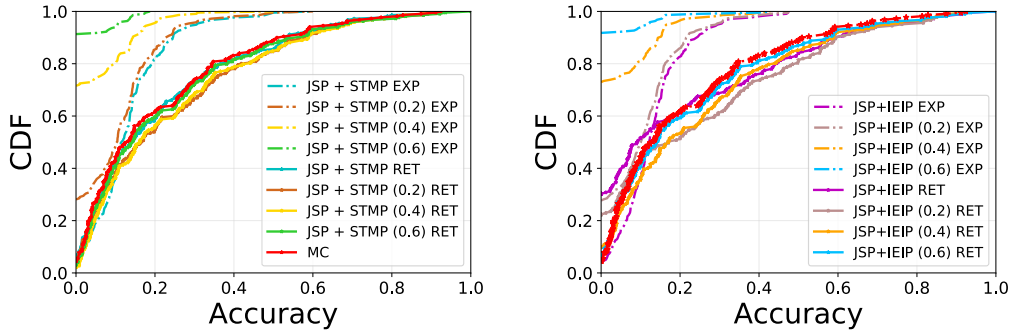
On the contrary, whereas the prediction accuracy of the *MC Location-Predictor*



(a) STMP. (b) IEIP.

Figure 6.11: Accuracy of the PSP for each type of movement.

in forecasting explorations is equal to zero, the PSP achieves appealing performance. Notably, it is more beneficial when using it with all users. Recall that for users classified as easy-to-predict we apply the *MC Location-Predictor* for the forecasts.



(a) STMP. (b) IEIP.

Figure 6.12: Accuracy of the JSP for each type of movement.

In Figure 6.12, we depict the accuracy of predicting the locality of each type of movement by the JSP. First, for returns forecasts, the JSP and PSP are alike, in view of the fact that they rely on the same location predictor and take the same inputs. Second, the difference in performance between the JSP and PSP emanate from exploration forecasts. Compared to the PSP, the JSP works poorly in predicting the locality of explorations. This suggests that the overall weak predictive power of the JSP presented in Figure 6.10 follows from the low potential of the Joint Zone-Predictor in forecasting the locality of explorations. *Unlike return patterns that can be common to many individuals that are strangers to each other, exploration patterns are more personal. Furthermore, we measured the Jaccard similarity for the top 5 most visited zones when exploring between users within the same city (Beijing). We*

report very low scores; the similarity is at most equal to 0.22. This implies that when it is about forecasting explorations, it is better to rely on the individual's mobility behavior than looking at the global patterns. Note that similarities in exploration spatial patterns might be observed among users within the same social circle and who are not perfect strangers to one another as discussed in [73].

In Table 6.2 and Table 6.3, we report the average accuracies of prediction achieved by the PSP and JSP in predicting the spatial occurrence of explorations while varying the spatial resolution.

PSP	**Explorations**			
Spatial Units	800 m × 800 m	1 km× 1 km	2 km× 2 km	4 km× 4 km
STMP	0.39	0.41	0.50	0.63
STMP (0.2)	0.26	0.28	0.35	0.44
STMP (0.4)	0.10	0.11	0.14	0.18
STMP (0.6)	0.03	0.03	0.05	0.06
IEIP	0.34	0.36	0.45	0.58
IEIP (0.2)	0.22	0.24	0.32	0.4
IEIP (0.4)	0.09	0.09	0.12	0.16
IEIP (0.6)	0.02	0.02	0.04	0.05

Table 6.2: Ratio of correctly predicted explorations and returns for the PSP.

JSP	**Explorations**			
Spatial Units	800 m × 800 m	1 km× 1 km	2 km× 2 km	4 km× 4 km
STMP	0.17	0.17	0.16	0.14
STMP (0.2)	0.11	0.11	0.11	0.10
STMP (0.4)	0.03	0.05	0.05	0.04
STMP (0.6)	0.01	0.01	0.02	0.04
IEIP	0.15	0.16	0.15	0.14
IEIP (0.2)	0.11	0.11	0.11	0.10
IEIP (0.4)	0.04	0.04	0.05	0.04
IEIP (0.6)	0.01	0.01	0.02	0.02

Table 6.3: Ratio of correctly predicted explorations and returns for the JSP.

From Table 6.2, we can see that the accuracy of predicting the spatial units where explorations might occur is relatively high for the PSP. Additionally, it increases with the increase of zones size. On the contrary, Table 6.3 shows that the JSP

performs poorly in predicting the locality of exploration events, and the changes in performance are not noticeable with the change in the spatial resolution.

6.4 Summary

In this chapter, we designed a novel 2-step adaptive prediction framework composed of (i) a *purpose prediction* (i.e., exploration or return) that feeds (ii) a *spatial prediction*. Contrary to existing methods, the proposed framework does not naively rely on mobility return's regularity, but adjusts its forecasts when explorations are more probable to happen. We first designed two purpose prediction algorithms that base their forecasts on coarse- or fine-grained regularities observed in exploratory and return visits. We then developed two spatial prediction tasks that take the outcomes of the purpose predictors as input and operate on different spatial scales. The two spatial predictors are similar in predicting returns, but employ different strategies in case the input is an exploration. The first relies its forecasts on the personal data of the considered user, whereas the second one exploits common exploratory tendencies among the population to infer the next coarse-grained zones. The proposed framework achieves interesting results both in inferring the next type of movement as well as in forecasting the spatial occurrence of the visits. Moreover, they confirm that exploration visits are not wholly random if the spatial resolution is increased. Besides, we find that explorations are more personal contrary to returns where common patterns are shared between users. Unlike conventional methods that are always wrong when explorations occur, the proposed framework does predict new visits for spatial units of excellent precision when networks plannings are concerned; and reasonable precision for more personalized applications, as for recommendation services.

In the next chapter, we present a summary of our results, discuss the limitations of our techniques, and provide general directions for research on the topic of human mobility prediction.

Chapter 7

Conclusion

Contents

This thesis has focused on exploratory visits in human mobility. The study has targeted the understanding and the quantification of exploration-like visits and their integration in mobility prediction.

7.1 Conclusions

The purposes of the research reported in this thesis were threefold:

Purpose 1: The first was to *understand and capture individuals' inclinations for discoveries of new places*. This led us to develop metrics able to characterize individuals' preferences for each type of visit: exploration and return. The quantification and evaluation of individuals' inclinations with regard to each type of visit led to the identification of three main mobility profiles (cf. Section 4.1). Namely, *Scouters* (adventurous and prone to explore), (ii) *Routiners* (steady and routinary), and (iii) *Regulars* (with medium behavior). The experiments were conducted using four real-world trajectories data (3 GPS and one CDR) described in Section 3.1.

The proposed profiling was further investigated and strengthened through the computation of multi-dimensional metrics (cf. Table 4.2) that asserted the subsisting dissimilarity between the mobility behavior of individuals of each group [8] (cf. Section 4.2):

- **Scouters:** are dynamic, relish discovering many new places successively, and are keen to break their returning routine to explore. They are marked by large sets of known places, resulting from their constant quest for new locations. Besides, they do not revisit the same locations several times except for specific ones such as home location or workplace. Moreover, *Scouters* spend large amounts of time exploring; on average, when they start exploring, they do it for more than 3 h. Additionally, they tend to walk long distances in general, either for return visits or exploratory visits.

- **Routiners:** are steady and rarely break their routine to discover new places. Once they start exploring new areas, they either stay in the discovered place or return to a familiar location. Unlike individuals of other profiles, *Routiners* have small sets of known places that they constantly revisit. Additionally, they spend short amounts of time exploring. Besides, they wait long periods before making transitions to other areas. Moreover, they have confined mobility; they walk small distances either when exploring or when returning.

- **Regulars:** alternate between exploration visits and return visits. They know more places compared to *Routiners*, but less compared to *Scouters*. The same is observed in terms of the duration of visits. *Regulars* explore less than *Scouters*, but spend long periods chasing new places compared to *Routiners*. Unlike when moving back and forth between familiar places, *Regulars* cover longer distances when they are in the exploring phase.

Adding temporal semantics to exploration activities showed that *Scouters*' proclivity to explore is significant throughout the week, with an increased tendency to discover new places from 4 pm to 8 pm. On the contrary, the *Routiners* have a trifling exploration activity during weekdays that increases on weekends, namely, on Sundays (cf. Section 4.3.1).

The analysis of the spatial exploitation of individuals having traces in the city center of Beijing revealed that *Scouters* have a spread activity all over the city; they use more than 62% of the neighborhoods for exploration visits and 67% for returns. In contrast, *Routiners* have confined mobility and use around 18% of the neighborhoods for both types of visits. *Regulars* favor visiting places within their vicinity when returning, but tend to go to more distant ones when exploring and use 34% of the territory for both types of visits. The computation of the entropy and predictability measures when increasing the spatial resolution (spatial units of size $2\,\text{km} \times 2\,\text{km}$) showed that explorations visits in a coarse-grained spatial scenario are far from being random (cf. Section 4.3.2).

Purpose 2: While the impacts of the prediction formulation and the quality of the data on prediction extents have been widely investigated in the literature, the limiting factor that arises from the intrinsic nature of human mobility –exploration tendencies – was rarely addressed. The second purpose of the dissertation was to *evaluate the impacts of exploration-like visits on the ability to foresee individuals' trajectories*. Using the two most widespread prediction formulations, namely, the next-cell prediction and the next-place prediction (cf. Section 2.4.1), we emphasized the role of novelty-seeking in making human mobility less foreseeable (cf. Section 2.4.1).

Specifically, the computation of the predictability revealed that regardless of the prediction task, individuals exhibiting a high tendency to explore, namely the *Scouters*, are the least predictable. For the next-place prediction formulation, the upper bound predictability peaked at values lower than 60% for the *Scouters*. In

contrast, it peaked at values above 90% for *Routiners*. Similarly, the prediction accuracy achieved by practical predictors, namely, MC, PPM, SPM, and ALZ, reached the lowest scores with *Scouters*. On average, 75% of an MC predictor forecasts are incorrect for the *Scouters* vs. less than 40% for the *Routiners* when considering the next-place prediction formulation.

Additionally, we showed that the next-place prediction formulation is more vulnerable to exploratory visits; the median of the prediction accuracy for *Scouters* is below 20% for the next-place prediction vs. 90% for the next-cell prediction.

Moreover, the removal and substitution of exploration-like visits from individuals' mobility traces led to substantial enhancements in the prediction accuracy, notably with *Scouters* – the prediction accuracy scores achieved with *Scouters* were enhanced to approach the performance achieved with other profiles.

These findings compel the need to thoroughly understand the exploration tendencies of the users and integrate the notion of exploration when designing mobility predictors, particularly when considering the next-place prediction formulation.

Purpose 3: The third purpose was to *tackle conventional predictors' limitation of being oblivious to users' explorations*. The recent separation of exploration visits from return visits in mobility modelings led to substantial improvements in the generated traces; the generated trajectories are closer to real-world mobility trajectories. Nevertheless, the notion of exploration is still missing from mobility predictors, and as shown in Chapter 5, exploration-like visits do decay the ability to foresee human mobility. To this end and relying on the previous findings: (i) individuals' tendencies to explore vary through the population, (ii) in a coarse-grained spatial resolution exploration-like visits are far from being random, (iii) the next-place prediction is highly affected by exploratory visits, we introduced a newly tailored exploration-aware mobility prediction framework that tackles the exploration problem by leveraging the purpose of the movements at prediction decisions, only using location data.

Specifically, the proposed framework splits the mobility prediction task into two steps: (i) first predicting the next type of movement: an exploration or a return, (ii) depending on the forecasted type of movement, it either predicts a fine-grained spatial unit (cell) in case of a return, or it predicts a coarse-grained spatial unit (zone) in case of an exploration. The experimental results demonstrated that the proposed framework achieves interesting results both in inferring the next type of movement as well as in forecasting the spatial occurrence of the visits. Contrary to conventional methods that focus on return visits, the proposed approach allows the using services and applications to tune to predictors according to their needs and requirements. Moreover, we found that exploration visits are more personal contrary to returns where common patterns are shared between users.

7.2 Limitations

Throughout this manuscript, we discussed and addressed the exploration phenomenon in human mobility. However, the current understanding of human mobility and, in particular, individuals' inclinations for discoveries of new places is still partial. One major limiting factor is data availability. Indeed, to accurately quantify and capture exploration activities, fine-grained, long-term, and large-scale geo-stamped data are necessary. Moreover, to determine circumstances and factors influencing users to explore, additional contextual data are needed, such as the semantics of the visited places, social circle information, personal calendar entries, or personality traits.

7.2.1 Privacy Issues in Data Acquisition

Mobility-related data contain potentially sensitive professional and personal information; they are among the most sensitive data currently being collected [34]. For instance, by solely relying on location data, a person's home location, attendance to a particular religious or political building, or presence in a motel or abortion clinic can be inferred [34]. Moreover, pseudonymization and standard de-identification are not sufficient to prevent users from being re-identified [35, 48]. Recent studies revealed that using only four location records (i.e., place and time), it is possible to re-identify 95% of the time the individuals [35].

Accordingly, mobility-related data are rarely made available for researchers and, more particularly, contextual information. In this manuscript, we only leveraged anonymized fine-grained geo-stamped data of individuals who provided informed consent. Hence, other dimensions such as the role played by social ties or personality traits in triggering the exploratory phases of the users were not investigated.

7.2.2 Data Sparcity

We leveraged two types of datasets: GPS and CDR. Although their suitability for mobility research, their sparsity was a significant limiting factor leading to a considerable bulk of information being dropped. As discussed in Chapter 3, the capture of the mobility trajectories using CDR data is highly dependent on the phone activity of the user. Likewise, GPS datasets are also filled with gaps as a consequence of the sleeping phases of the collecting devices. Moreover, the users can also willingly stop sharing their location for privacy reasons or battery savings.

Sparse data limit the understanding of the exploration tendencies. Yet, filtering users with insufficient data can engender biases in the induced conclusions. For instance, in CDR datasets selecting only individuals with a high phone activity might not be representative of the population.

7.3 Extensions

As indicated in Section 1.1 of the introduction, understanding individuals' inclinations to explore has a short history. After this study, we believe that there are still

unresolved issues. In this section, we will discuss the potential research perspectives in the context of our analytical insights and technical contributions.

Long-term evaluation of the tendencies to explore: We have not investigated whether individuals switch their mobility profile and, if so, under which circumstances they are more likely to hop from one group to another? Suppose an individual is identified as an explorer *(*Scouter) at a given time. Does it mean that she will always be an explorer? Or will she become a returner (*Routiner*) after a given period or at a certain season of the year? Such studies require a long-term and large-scale availability of mobility traces of individuals, preferably within the same region, and a global overview of the population mobility behavior.

Adding spatial semantics: We have not ventured to investigate the characteristics of the places where individuals go when they are in the exploring phase. The reason is that we leveraged mobility traces of users from different countries and continents with no spatial semantics or important concentration of users within the same regions. Extracting POI and the types of the visited places is a step forward that requires further data collection and examination. Particularly, POIs are massive in contemporary cities and vary in time; many new infrastructures are built while old ones are removed or replaced [96]. The linkage between exploration-like visits and spatial semantics for the population as a whole or per mobility profile can bring new insights on the ability to understand and infer the potential locality of future novelty-seeking events. For instance, if an individual is more likely to go to a restaurant when exploring, this information can assist mobility predictors in being more accurate in inferring the regions where the individual might explore.

Exploration similarity: Of additional interest is the problem of including the similarity of trajectories before profiling the users or predicting the spatial locality of explorations. Indeed, individuals exhibiting spatial similarities of visits are likely to have common interests and hence comparable mobility behaviors. Whereas the similarity in repetitive and regularities of visits have been widely studied for different types of data sources, the exploratory *spatial* similarity is yet to be investigated.

Social information context: Leveraging social circle information in addition to geolocation data can bring further understanding of what triggers individuals' propensity to explore. Friends or family members are more likely to discover the same new places compared with strangers. In the paper [73], the authors showed that exploration events are highly correlated with the social circle. In Section 6.3.3, we attempted to find similarities in exploratory visits on the population level. Nevertheless, unlike returns, explorations are proper to the individuals. Hence, the social dimension appears to be of invaluable contribution to studying the novelty in visits.

Bibliography

[1] EU CHIST-ERA Mobile context-Adaptive CAching for COntent-centric networking (MACACO) project. https://macaco.inria.fr/. (Cited on pages 27, 28 and 30.)

[2] *Markov Chain Monte Carlo*, pages 330–343. Springer New York, New York, NY, 1999. doi:10.1007/0-387-22724-5_24. (Cited on page 22.)

[3] Goutte Cyril ; and Gaussier Eric. A probabilistic interpretation of precision recall and f-score with implication for evaluation. In Losada David E; and Fernández-Luna; Juan M., editors, *Advances in Information Retrieval*, pages 345–359, Berlin Heidelberg, 2005. Springer Berlin Heidelberg. doi:10.1007/978-3-540-31865-1_25. (Cited on page 84.)

[4] J. S. Abel and J. W. Chaffee. Existence and uniqueness of gps solutions. *IEEE Transactions on Aerospace and Electronic Systems*, 27(6):952–956, 1991. doi:10.1109/7.104271. (Cited on page 12.)

[5] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS computational biology*, 12(12):e1005110, 2016. (Cited on page 15.)

[6] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491, 2018. doi:10.1038/s41562-018-0364-x. (Cited on pages 15, 16, 33 and 36.)

[7] Fahad Alhasoun. City scale next place prediction from sparse data through similar strangers. 2017. (Cited on pages 3, 19, 21 and 24.)

[8] Licia Amichi, Aline Carneiro Viana, Mark Crovella, and Antonio A. F. Loureiro. Understanding individuals' proclivity for novelty seeking. In Chang-Tien Lu, Fusheng Wang, Goce Trajcevski, Yan Huang, Shawn D. Newsam, and Li Xiong, editors, *SIGSPATIAL '20: 28th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 3-6, 2020*, pages 314–324. ACM, 2020. doi:10.1145/3397536.3422248. (Cited on pages 78 and 95.)

[9] Michael O'Neill Anthony Brabazon. *Natural Computing in Computational Finance*. Studies in Computational Intelligence. Springer, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-77477-8. (Cited on page 20.)

[10] Akinori Asahara, Kishiko Maruyama, Akiko Sato, and Kouichi Seto. Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances*

in Geographic Information Systems, GIS '11, pages 25–33, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/2093973.2093979. (Cited on pages 2, 3, 19, 23, 24 and 26.)

[11] Hamada S. Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M. Squire, and Lauren M. Gardner. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11), nov 2020. doi:10.1016/S1473-3099(20)30553-3. (Cited on page 1.)

[12] Yaneer Bar-Yam. Concepts: power law. *New England Complex Systems Institute*, 2016. (Cited on page 14.)

[13] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018. Human mobility: Models and applications. URL: https://www.sciencedirect.com/science/article/pii/S037015731830022X, doi:https://doi.org/10.1016/j.physrep.2018.01.001. (Cited on pages 1, 10, 11 and 14.)

[14] Hugo Barbosa, Fernando B. de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. The effect of recency to human mobility. *EPJ Data Science*, 4(1):21, 2015. doi:10.1140/epjds/s13688-015-0059-8. (Cited on page 17.)

[15] William Bialek and Naftali Tishby. Predictive Information. *arXiv e-prints*, pages cond–mat/9902341, February 1999. arXiv:cond-mat/9902341. (Cited on page 20.)

[16] Bjørn Sand Jensen, Jakob Eg Larsen, K. Jensen, J. Larsen, and Lars Kai Hansen. Estimating human predictability from mobile sensor data. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 196–201, 2010. doi:10.1109/MLSP.2010.5588997. (Cited on pages 18, 24 and 59.)

[17] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3133956.3133982. (Cited on pages 2, 24 and 77.)

[18] Giulio Bottegal, Alessandro Chiuso, and Paul M.J. Van den Hof. On dynamic network modeling of stationary multivariate processes. *IFAC-PapersOnLine*, 51(15):850–855, 2018. 18th IFAC Symposium on System Identification SYSID 2018. URL: https://www.sciencedirect.com/science/article/

pii/S2405896318317798, doi:https://doi.org/10.1016/j.ifacol.2018.09.118. (Cited on page 20.)

[19] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006. doi:10.1038/nature04292. (Cited on pages 2, 11, 13, 14, 15 and 18.)

[20] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Human mobility prediction based on individual and collective geographical preferences. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 312–317, 2010. doi:10.1109/ITSC.2010.5625119. (Cited on pages 19 and 24.)

[21] H.C. Carey. *Principles of Social Science*, volume 3. J. B. Lippincott & Company, 1867. URL: https://books.google.fr/books?id=-yAWAAAAYAAJ. (Cited on page 1.)

[22] Chang, Serina and Pierson, Emma and Koh, Pang Wei and Gerardin, Jaline and Redbird, Beth and Grusky, David and Leskovec, Jure. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021. doi:10.1038/s41586-020-2923-3. (Cited on page 1.)

[23] G. Chen, A. Carneiro Viana, M. Fiore, and C. Sarraute. Complete Trajectory Reconstruction from Sparse Mobile Phone Data. *EPJ Data Science*, October 2019. (Cited on pages 27, 29 and 30.)

[24] Xihui Chen, Jun Pang, and Ran Xue. Constructing and comparing user mobility profiles. *ACM Trans. Web*, 8(4), November 2014. doi:10.1145/2637483. (Cited on page 26.)

[25] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. (Cited on page 14.)

[26] J. Cleary and I. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32(4):396–402, 1984. doi:10.1109/TCOM.1984.1096090. (Cited on page 22.)

[27] Giuliano Cornacchia and Luca Pappalardo. Sts-epr: Modelling individual mobility considering the spatial, temporal, and social dimensions together. *Procedia Computer Science*, 184:258–265, 2021. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops. URL: https://www.sciencedirect.com/science/article/pii/S1877050921006645, doi:https://doi.org/10.1016/j.procs.2021.03.035. (Cited on page 17.)

[28] Giuliano Cornacchia, Giulio Rossetti, and Luca Pappalardo. Modelling Human Mobility considering Spatial,Temporal and Social Dimensions. *arXiv e-prints*, page arXiv:2007.02371, July 2020. arXiv:2007.02371. (Cited on page 12.)

[29] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. (Cited on page 20.)

[30] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476956, arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476956, doi:10.1080/01621459.1996.10476956. (Cited on pages 22 and 62.)

[31] Tim Cresswell. *On the Move: Mobility in the Modern Western World*. Routledge, 2006. (Cited on page 10.)

[32] Andrea Cuttone, Sune Lehmann, and Marta C. González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7(1):2, 2018. doi:10.1140/epjds/s13688-017-0129-1. (Cited on pages 2, 3, 12, 18, 19, 24, 25, 29, 36, 59, 60, 61, 62, 63, 65, 70, 77, 78, 80, 85, 89 and 114.)

[33] Giannotti F. ; Pappalardo L.; Pedreschi D.; and Wang D. *A Complexity Science Perspective on Human Mobility*, pages 297–314. Cambridge University Press, 2013. doi:10.1017/CBO9781139128926.016. (Cited on page 2.)

[34] Yves-Alexandre De Montjoye, Sébastien Gambs, Vincent Blondel, Geoffrey Canright, Nicolas De Cordes, Sébastien Deletaille, Kenth Engø-Monsen, Manuel Garcia-Herranz, Jake Kendall, Cameron Kerry, et al. On the privacy-conscientious use of mobile phone data. *Scientific data*, 5(1):1–6, 2018. (Cited on page 98.)

[35] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013. (Cited on page 98.)

[36] Timothy DelSole. Predictability and information theory. part i: Measures of predictability. *Journal of the Atmospheric Sciences*, 61(20):2425 – 2440, 01 Oct. 2004. URL: https://journals.ametsoc.org/view/journals/atsc/61/20/1520-0469_2004_061_2425_paitpi_2.0.co_2.xml, doi:10.1175/1520-0469(2004)061<2425:PAITPI>2.0.CO;2. (Cited on page 20.)

[37] Huifang Feng, Chunfeng Liu, Yantai Shu, and Oliver W. Yang. Location prediction of vehicles in vanets using a kalman filter. *Wirel. Pers. Commun.*, 80(2):543–559, Jan. 2015. doi:10.1007/s11277-014-2025-3. (Cited on page 23.)

[38] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. Pmf: A privacy-preserving human mobility prediction framework via federated learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1), Mar. 2020. doi:10.1145/3381006. (Cited on page 19.)

[39] 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*, pages 469–477, 2012. doi:10.1109/INFCOM.2012.6195786. (Cited on page 30.)

[40] Jonas Fransson. *Non-equilibrium nano-physics: a many-body approach*, volume 809. Springer, 2010. (Cited on page 20.)

[41] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. MPM '12, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2181196.2181199. (Cited on pages 2, 3, 19, 21, 22, 23 and 77.)

[42] R. K. Ganti, F. Ye, and H. Lei. Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine*, 49(11):32–39, 2011. doi:10.1109/MCOM.2011.6069707. (Cited on page 28.)

[43] Huiji Gao, Jiliang Tang, and Huan Liu. Gscorr: Modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1582–1586, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2396761.2398477. (Cited on pages 2, 19, 21 and 59.)

[44] Győző Gidófalvi and Fang Dong. When and where next: Individual mobility prediction. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, MobiGIS '12, pages 57–64, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2442810.2442821. (Cited on pages 3, 19, 23 and 77.)

[45] Joseph E. Gillis and George H. Weiss. Expected Number of Distinct Sites Visited by a Random Walk with an Infinite Variance. *Journal of Mathematical Physics*, 11(4):1307–1312, April 1970. doi:10.1063/1.1665260. (Cited on page 15.)

[46] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782,

2008. doi:10.1038/nature06958. (Cited on pages 2, 11, 13, 14, 15, 16, 18, 33 and 77.)

[47] Karthik Gopalratnam and Diane J. Cook. Active lezi: an incremental parsing algorithm for sequential prediction. *Int. J. Artif. Intell. Tools*, 13(4):917–930, 2004. doi:10.1142/S0218213004001892. (Cited on pages 22, 23 and 62.)

[48] Marco Gramaglia and Marco Fiore. On the anonymizability of mobile traffic datasets. *arXiv preprint arXiv:1501.00100*, 2014. (Cited on page 98.)

[49] Grantz, Kyra H. and Meredith, Hannah R. and Cummings, Derek A. T. and Metcalf, C. Jessica E. and Grenfell, Bryan T. and Giles, John R. and Mehta, Shruti and Solomon, Sunil and Labrique, Alain and Kishore, Nishant and Buckee, Caroline O. and Wesolowski, Amy. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, 11(1):4961, 2020. (Cited on page 1.)

[50] Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014. (Cited on page 1.)

[51] John F. Hoffecker. The spread of modern humans in europe. *Proceedings of the National Academy of Sciences*, 106(38):16040–16045, 2009. URL: https://www.pnas.org/content/106/38/16040, arXiv:https://www.pnas.org/content/106/38/16040.full.pdf, doi:10.1073/pnas.0903446106. (Cited on page 1.)

[52] Edin Lind Ikanovic and Anders Mollgaard. An alternative approach to the limits of predictability in human mobility. *EPJ Data Science*, 6(1):12, 2017. doi:10.1140/epjds/s13688-017-0107-7. (Cited on pages 19, 25, 61, 62 and 72.)

[53] P. Jacquet, W. Szpankowski, and I. Apostol. A universal pattern matching predictor for mixing sources. In *Proceedings IEEE International Symposium on Information Theory,*, pages 150–, 2002. doi:10.1109/ISIT.2002.1023422. (Cited on pages 22 and 62.)

[54] K. Jaffrès-Runser, G. Jakllari, Tao Peng, and V. Nitu. Crowdsensing mobile content and context data: Lessons learned in the wild. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 311–315, 2017. doi:10.1109/PERCOMW.2017.7917579. (Cited on page 10.)

[55] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998. doi:10.1109/18.669425. (Cited on page 21.)

[56] Thomas Kreuz, Daniel Chicharro, Ralph G. Andrzejak, Julie S. Haas, and Henry D.I. Abarbanel. Measuring multiple spike train synchrony. *Journal of Neuroscience Methods*, 183(2):287–299, 2009. URL: https://www.sciencedirect.com/science/article/pii/S0165027009003616, doi:https://doi.org/10.1016/j.jneumeth.2009.06.039. (Cited on page 52.)

[57] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, pages 243–260, Berlin, Heidelberg, 2006. Springer-Verlag. doi:10.1007/11853565_15. (Cited on page 19.)

[58] Vaibhav Kulkarni, Arielle Moro, and Benoît Garbinato. A mobility prediction system leveraging realtime location data streams: Poster. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom '16, pages 430–432, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2973750.2985263. (Cited on page 21.)

[59] Lin Liao, Dieter Fox, and Henry Kautz. Location-based activity recognition using relational markov networks. IJCAI'05, pages 773–778, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc. (Cited on page 22.)

[60] Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 381–390, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2370216.2370274. (Cited on pages 12, 18, 24 and 59.)

[61] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012. URL: https://www.pnas.org/content/109/29/11576, arXiv:https://www.pnas.org/content/109/29/11576.full.pdf, doi:10.1073/pnas.1203882109. (Cited on page 59.)

[62] Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific Reports*, 3(1):2923, 2013. doi:10.1038/srep02923. (Cited on pages 2, 3, 19, 21, 23, 59, 64, 65 and 77.)

[63] Wesley Mathew, Ruben Raposo, and Bruno Martins. Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 911–918, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2370216.2370421. (Cited on pages 3, 19, 23 and 77.)

[64] A. Moffat. Implementing the ppm data compression scheme. *IEEE Transactions on Communications*, 38(11):1917–1921, 1990. doi:10.1109/26.61469. (Cited on pages 22 and 62.)

[65] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, et al. Priva'mov: Analysing human mobility through multi-sensor datasets. In *NetMob 2017*, 2017. (Cited on pages 27 and 28.)

[66] Apollinaire Nadembega, Abdelhakim Hafid, and Tarik Taleb. Mobility-prediction-aware bandwidth reservation scheme for mobile networks. *IEEE Transactions on Vehicular Technology*, 64(6):2561–2576, 2015. doi:10.1109/TVT.2014.2345255. (Cited on page 1.)

[67] Elham Naghizade, Jeffrey Chan, Yongli Ren, and Martin Tomko. Contextual location imputation for confined wifi trajectories. page 14, Melbourne, Australia, 2018. Springer. (Cited on page 65.)

[68] Elham Naghizade, Jeffrey Chan, and Martin Tomko. From small sets of gps trajectories to detailed movement profiles: quantifying personalized trip-dependent movement diversity. *International Journal of Geographical Information Science*, 34(10):2004–2029, 2020. (Cited on page 26.)

[69] Jovancic Nemanja. 4 types of bias in research and how to make your surveys bias-free, Mar. 1999. URL: https://www.leadquizzes.com/blog/types-of-bias-in-research/. (Cited on page 10.)

[70] J. R. Palmer, T. Espenshade, F. Bartumeus, Chang Y. Chung, N. E. Ozgencil, and K. Li. New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50:1105–1128, 2012. (Cited on page 10.)

[71] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. On the properties of human mobility. *Computer Communications*, 87:19–36, 2016. URL: https://www.sciencedirect.com/science/article/pii/S0140366416300901, doi:https://doi.org/10.1016/j.comcom.2016.03.022. (Cited on page 37.)

[72] 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)*, pages 871–878, 2015. doi:10.1109/BigData.2015.7363835. (Cited on page 28.)

[73] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1):8166, 2015. doi:10.1038/ncomms9166. (Cited on pages 2, 4, 17, 26, 33, 60, 92 and 99.)

[74] Peyton Z Peebles and Bertram Emil Shi. *Probability, random variables, and random signal principles*, volume 3. McGraw-Hill New York, NY, USA:, 2001. (Cited on page 20.)

[75] Chan R. The cambridge analytica whistleblower explains how the firm used facebook data to sway elections, May 2020. URL: https://www.businessinsider.fr/. (Cited on page 77.)

[76] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *SIGMOBILE Mob. Comput. Commun. Rev.*, 16(3):33–44, December 2012. doi:10.1145/2412096.2412101. (Cited on page 30.)

[77] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. (Cited on page 43.)

[78] Luca Scherrer, Martin Tomko, Peter Ranacher, and Robert Weibel. Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science*, 7(1):19, 2018. doi:10.1140/epjds/s13688-018-0147-7. (Cited on pages 4, 19, 26 and 60.)

[79] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013. doi:10.1098/rsif.2013.0246. (Cited on page 33.)

[80] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, Jan. 2001. doi:10.1145/584091.584093. (Cited on page 20.)

[81] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012. doi:10.1038/nature10856. (Cited on page 10.)

[82] J. G. Skellam. Random dispersal in theoretical populations. *Biometrika*, 38(1/2):196–218, 1951. URL: http://www.jstor.org/stable/2332328. (Cited on page 3.)

[83] 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)*, pages 88–94, 2014. doi:10.1109/PerCom.2014.6813948. (Cited on pages 25, 59 and 61.)

[84] Gürkan Solmaz and Damla Turgut. A survey of human mobility models. *IEEE Access*, 7:125711–125731, 2019. doi:10.1109/ACCESS.2019.2939203. (Cited on page 2.)

[85] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010. doi:10.1038/nphys1760. (Cited on pages 2, 3, 13, 14, 15, 16, 18, 33 and 78.)

[86] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010. URL: https://science.sciencemag.org/content/327/5968/1018, arXiv:https://science.sciencemag.org/content/327/5968/1018.full.pdf, doi:10.1126/science.1177170. (Cited on pages 2, 11, 18, 19, 20, 21, 22, 24, 59 and 61.)

[87] Libo Song, D. Kotz, Ravi Jain, and Xiaoning He. Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649, 2006. doi:10.1109/TMC.2006.185. (Cited on pages 19, 22 and 23.)

[88] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *P L o S One*, 9(4), 2014. doi:10.1371/journal.pone.0095978. (Cited on page 10.)

[89] Moritz von Mörner. Application of call detail records - chances and obstacles. *Transportation Research Procedia*, 25:2233–2241, 2017. World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016. URL: https://www.sciencedirect.com/science/article/pii/S2352146517307366, doi:https://doi.org/10.1016/j.trpro.2017.05.429. (Cited on page 11.)

[90] Jinzhong Wang, Xiangjie Kong, Feng Xia, and Lijun Sun. Urban human mobility: Data-driven modeling and prediction. *SIGKDD Explor. Newsl.*, 21(1):1–19, May 2019. doi:10.1145/3331651.3331653. (Cited on pages 1 and 21.)

[91] Andy Yuan Xue, Rui Zhang, Yu Zheng, Xing Xie, Jin Huang, and Zhenghua Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *Proceedings of The 29th IEEE International Conference on Data Engineering*, April 2013. URL: https://www.microsoft.com. (Cited on pages 2, 24 and 77.)

[92] Khong-Lim Yap, Yung-Wey Chong, and Weixia Liu. Enhanced handover mechanism using mobility prediction in wireless networks. *PLOS ONE*, 15(1):1–31, 01 2020. doi:10.1371/journal.pone.0227982. (Cited on page 1.)

[93] C. Yu, Y. Liu, D. Yao, L. T. Yang, H. Jin, H. Chen, and Q. Ding. Modeling user activity patterns for next-place prediction. *IEEE Systems Journal*, 11(2):1060–1071, 2017. doi:10.1109/JSYST.2015.2445919. (Cited on pages 3, 19, 24, 77 and 85.)

[94] Dongxiang Zhang, Long Guo, Liqiang Nie, Jie Shao, Sai Wu, and Heng Tao Shen. Targeted advertising in public transportation systems with quantitative evaluation. *ACM Trans. Inf. Syst.*, 35(3), January 2017. doi:10.1145/3003725. (Cited on page 1.)

[95] Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. Explaining the power-law distribution of human mobility through transportationmodality decomposition. *Scientific Reports*, 5(1):9136, 2015. doi:10.1038/srep09136. (Cited on page 14.)

[96] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3), September 2014. doi:10.1145/2629592. (Cited on pages 1 and 99.)

[97] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321, 2008. (Cited on pages 27 and 28.)

[98] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data(base) Engineering Bulletin*, June 2010. URL: https://www.microsoft.com/en-us/research/publication. (Cited on pages 27 and 28.)

[99] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1526709.1526816. (Cited on pages 27 and 28.)

[100] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977. doi:10.1109/TIT.1977.1055714. (Cited on pages 22 and 23.)

Appendix A

Experimental Settings

Depending on the goals of each part of the thesis, we use different settings for the spatial and temporal resolution, as well as the followed filtering procedure for users selection.

Spatial Tesselation:

- In Chapter 4, we consider cells with c=300 m that we refer to as $ID = l3$. There are two reasons to consider cells of size 300 m × 300 m. First, in this thesis, we focus on the discoveries of new places on a daily basis, for instance, going to a new restaurant or a new shop. Considering the imprecision and uncertainty of GPS and CDR systems, we claim a cell of such size roughly corresponds to daily regions of interest and can still capture discovery moments. Hence, in Chapter 4, the mobility trace of each user u has the shape $\mathcal{T}_{u,300} = \langle (t_0, l3_0), (t_1, l3_1), \ldots, (t_N, l3_n) \rangle$.
- In Chapter 5, we study the impacts of the spatial resolution on the predictive performance and consider spatial units of different sizes. We start with smaller spatial units compared to Chapter 4, but reasonable and meaningful and also roughly corresponding to daily POI. Then we consider larger spatial units with a uniform increase in size $c = c+$ 200 m. The considered spatial units in Chapter 5 are the following:
  - Cells of size 200 m × 200 m that we refer to as *locations* of type 2, $l2$.
  - Cells of size 400 m × 400 m that we refer to as *locations* of type 4, $l4$.
  - Cells of size 600 m × 600 m that we refer to as *locations* of type 6, $l6$.
  - Cells of size 800 m × 800 m that we refer to as *zone* of type 0, $z0$.

  The records of the mobility traces $\mathcal{T}_{u,200}$ are extended with the IDs of the different spatial units, i.e., the mobility trace of each user u is extended to $\mathcal{T}_u = \langle (t_0, l2_0, l4_0, l6_0, z0_0), (t_1, l2_1, l4_1, l6_1, z0_1), \ldots, (t_n, l2_n, l4_n, l6_n, z0_n) \rangle$.
- In Chapters 6, we propose to tackle the prediction of the spatial occurrence of exploration events by increasing the spatial granularity. In addition to cells of size 200 m × 200 m, we consider larger spatial units that we refer to as zones. The considered units are the following:
  - Cells of size 200 m × 200 m that we refer to as *locations* of type 2, $l2$.
  - Cells of size 800 m × 800 m that we refer to as *zones* of type 0, $z0$.

– Cells of size 1 km × 1 km that we refer to as *zones* of type 1, $z1$.

– Cells of size 2 km × 2 km that we refer to as *zones* of type 2, $z2$.

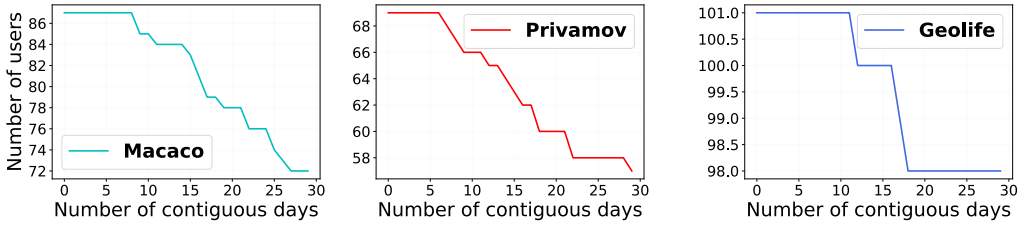
– Cells of size 4 km × 4 km that we refer to as *zones* of type 4, $z4$.

The records of the mobility traces $\mathcal{T}_{u,200}$ are extended with the IDs of the different spatial units. Hence, the mobility trace of each user u has the shape $\mathcal{T}_u = \langle (t_0, l2_0, z0_0, z1_0, z2_0, z4_0), \ldots, (t_n, l2_n, z0_n, z1_n, z2_n, z4_n) \rangle$.

Temporal Sampling: For the temporal resolution, we set δ_{GPS} to 5 min, which is the highest common achievable frequency of sampling for the GPS datasets and we set δ_{CDR} to 1 h, which is the highest achievable frequency of sampling for the GPS dataset. We choose the highest possible frequencies of sampling (smallest possible values) as they allow a more thorough capture and understanding of individuals' mobility behavior [32].

Data filtering: For the definition of a complete day of data, we set δ^*_{GPS} to 15 min and δ^*_{CDR} to 2 h. In what follows, we detail the filtering procedures used within each chapter.

- In Chapter 4, we use the filtering **F1**. This means each selected user must have x *consecutive* and *complete* days of data, where a complete day of data for a GPS dataset (CDR dataset) is a day with at least one record each 15 min (2 h). To set x the number of days of data required for users selection, we report in Figure 7.1 the number of leveraged users with the variation of the parameter x from 1 day to 1 month for the GPS datasets. For the CDR dataset, given the large number of users compared to GPS datasets, by default, we consider x=14 days, which is the maximal possible value.



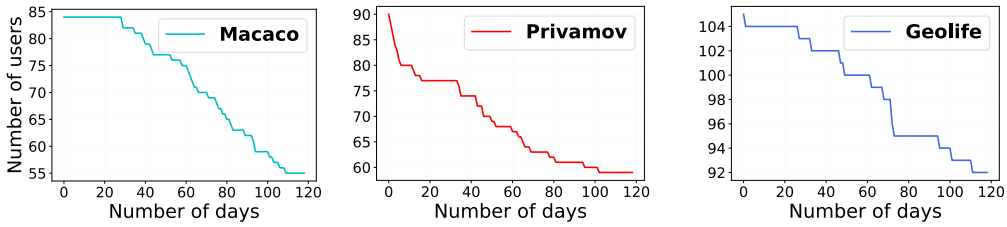
(a) Macaco. (b) Privamov. (c) Geolife.

Figure 7.1: Number of users vs. number of contigious complete days of data.

Although the fine-grained nature of the GPS data, Figure 7.1 reveals the presence of many temporal gaps. Therefore, to ensure a relatively high number of users per-data set while keeping a reasonable number of contiguous data, we set x to 10.

- In Chapters 5 and 6, we use the filtering **F2**. This means we select only users

having x days, with each having at least 96 records in the GPS trace (12 records in the CDR trace).



(a) Macaco. (b) Privamov. (c) Geolife.

Figure 7.2: Number of users vs. number of complete days of data.

In Figure 7.2, we report the number of the leveraged users with the variation of x from 1 day to 120 days, i.e., 4 months. There is a trade-off between the number of leveraged users and the number of complete days of data. To keep a reasonable number of users while having a long enough study period, we select only individuals with 1 month of complete days of data.

Table 7.1 summarizes the characteristics of the data used within each chapter after data completion and users filtering.

7.3.1 Dichotomy of Movements

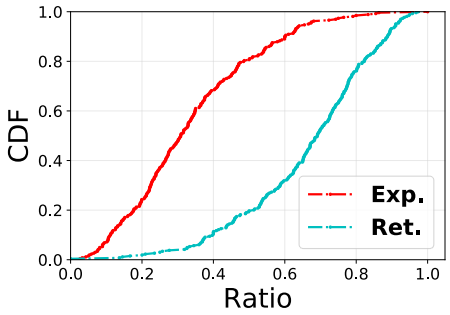


Figure 7.3: Ratio of types of transition.

Figure 7.3 reports the CDF of the proportions of transitions of types exploration (Exp) and returns (Ret). We can see that the majority of the population has a low exploration activity; more than 60% of the population has an exploration ratio lower than 40%. Inversely, all users depict a high returning activity; 80% of the users have a degree of return above 50%. Indeed, routine patterns are embedded in today's societies.

Table 7.1: Summary about the four datasets used within each chapter after data completion and users selection.

Dataset	Filtering	Spatail units & Duration	Macaco	Privamov	Geolife	Agg_gps	ChineseDB
Chapter 4	**F1**	**Spatial Units:** • 300 m×300 m • 2 km^2 in Sec. 4.3.2 **Duration:** • 10 days • 1 month in Sec. 4.3.1	87	69	101	257 (224 in Sec. 4.3.1)	3761
Chapter 5	**F2**	**Spatial Units:** • 200 m×200 m • 400 m×400 m • 600 m×600 m • 800 m×800 m **Duration:** 1 month	84	77	105	266	4860
Chapter 6	**F2**	**Spatial Units:** • 200 m×200 m • 800 m×800 m • 1 km×1 km • 2 km×2 km • 4 km×4 km **Duration:** 1 month	84	77	105	266	/

Appendix B

List of Publications produced during the Ph.D. Studies

- From motion purpose to perceptive spatial mobility prediction. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM SIGSPTIAL International Conference on Advances in Geographic Information Systems. 2021.
- Revealing an inherently limiting factor in human mobility prediction. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. IEEE Transactions on Emerging Topics in Computing (TETC). *Under review.*
- Understanding individuals' proclivity for *novelty seeking.* Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM SIGSPTIAL International Conference on Advances in Geographic Information Systems. 2020.
- Mobility profiling: Identifying scouters in the crowd. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ACM CONEXT International Conference on emerging Networking EXperiments and Technologies Student Workshop. 2019.
- Explorateur ou Routinier: Quel est votre profile de mobilité?. Licia Amichi, Aline C. Viana, Mark Crovella, Antonio Loureiro. ALGOTEL Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications. 2020.

Outside the scope of this thesis, we also obtained other results:

- Joint allocation strategies of power and spreading factors with imperfect orthogonality in LoRa networks. Licia Amichi, Megumi Kaneko, Ellen Hidemi Fukuda, Nancy El Rachkidy, and Alexandre Guitton. IEEE Transactions on Communications 68, 2020.
- Spreading factor allocation strategy for LoRa networks under imperfect orthogonality. Licia Amichi, Megumi Kaneko, Nancy El Rachkidy, and Alexandre Guitton. IEEE International Conference on Communications (ICC), 2019.

ECOLE
DOCTORALE

Titre : Comprendre la tendance exploratoire de la mobilité humaine

Mots clés : Mobilité individuelle, Exploration, Profilage, Prédiction

Résumé :
La compréhension et la prédiction du déplacement des individus dans l'espace et le temps sont fondamentales dans de nombreux domaines comme la propagation des épidémies (notamment, la pandémie du COVID-19), les réseaux mobiles, l'urbanisme ou encore le covoiturage. Cependant, appréhender la mobilité humaine est intrinsèquement complexe car d'une part, les mouvements des individus sont limités par l'obligation d'une présence physique sur les lieux de travail, les universités ou dans le cadre de la participation à des activités routinières et sociales. D'autre part, la grande variété de lieux de loisirs et la disponibilité de moyens de transport modernes permettent aux individus d'interrompre leurs routines pour découvrir de nouveaux lieux. Depuis peu avec l'omniprésence des appareils mobiles, de la connectivité Internet et des systèmes de positionnement, il est devenu possible de capturer les allers et venues quotidiennes des individus à des échelles spatiales et temporelles très précises. Cela offre l'opportunité d'observer et d'étudier la mobilité humaine au niveau individuel avec un niveau de détail sans précédent. Cependant, la littérature scientifique sur la prédiction de la mobilité humaine ne tient pas compte des tendances des individus à rechercher la nouveauté, c'est-à-dire à explorer et à découvrir de nouveaux lieux. Les prédicteurs conventionnels reposant sur des données géographiques personnelles fonctionnent mal lorsqu'il s'agit de découvrir de nouvelles régions. La raison est expliquée par la prédiction reposant uniquement sur des emplacements précédemment visités/vus (ou connus). Comme effet secondaire, des emplacements qui n'ont jamais été visités auparavant (ou des explorations) par un utilisateur perturbent la prédiction d'un emplacement connu. Négliger à première vue les activités de recherche de nouveauté apparaît sans conséquence sur la capacité à comprendre et à prévoir les trajectoires des individus. Dans ce manuscrit, nous affirmons et montrons le contraire : les visites de type exploration impactent fortement la compréhension et l'anticipation de la mobilité humaine.

Title : Understanding individuals' proclivity for novelty-seeking in human mobility

Keywords : Individual Mobility, Exploration, Profiling, Prediction

Abstract :
Understanding and predicting how humans move within space and time is of fundamental importance for many scientific domains, such as epidemic propagation (e.g., the COVID-19 pandemics), mobile networks, urban planning, or ride-sharing. Yet, apprehending human mobility is intrinsically complex. On the one hand, human movements are constrained by physical presence in workplaces, gyms, or universities, in addition to the involvement in routine and social activities. On the other hand, the large variety of leisure places and the availability of modern means of transportation allow people to break their routinary patterns to discover new places. But recently, with the ubiquity of mobile devices, Internet connectivity, and positioning systems, capturing individuals' daily whereabouts at very fine spatial and temporal scales has become possible. Nevertheless, the scientific literature on human mobility prediction is oblivious to individuals' tendencies for novelty-seeking, i.e., exploring and discovering new places. Conventional predictors relying on personal geographical data perform poorly when it comes to discoveries of new regions. The reason is explained by the prediction relying only on previously visited/seen (or known) locations. As a side effect, places that were never visited before (or explorations) by a user cause disturbance to known location's prediction. Neglecting novelty-seeking activities at first glance appears to be inconsequential on the ability to understand and predict individuals' trajectories. In this manuscript, we claim and show the opposite: exploration-like visits strongly impact mobility understanding and anticipation.

Institut Polytechnique de Paris
91120 Palaiseau, France