

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

João Paulo Esper Spíndula

**Understanding Factors that Impact Individual Mobility Prediction with
Federated Learning**

Belo Horizonte
2024

João Paulo Esper Spíndula

**Understanding Factors that Impact Individual Mobility Prediction with
Federated Learning**

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Jussara Marques de Almeida Gonçalves
Co-Advisor: Aline Carneiro Viana

Belo Horizonte
2024

Spíndula, João Paulo Esper.

S757u Understanding factors that impact individual mobility prediction with federated learning [recurso eletrônico] / João Paulo Esper Spíndula. – 2024.

1 recurso online (64 f. il., color.) : pdf.

Orientadora: Jussara Marques de Almeida Gonçalves.
Coorientadora: Aline Carneiro Viana.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.

Referências: f. 58-64.

1. Computação – Teses. 2. Aprendizado do computador – Teses. 3. Algoritmos de predição – Teses. 4. Aprendizado Federado – Teses. 5. Mobilidade – Pessoas - Análise de Desempenho – Teses. I. Gonçalves, Jussara Marques de Almeida. II. Viana, Aline Carneiro. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*82(043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Understanding Factors that Impact Individual Mobility Prediction with
Federated Learning

JOÃO PAULO ESPER SPÍNDULA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Jussara Marques de Almeida Gonçalves

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

Aline Carneiro Viana

PROFA. ALINE CARNEIRO VIANA - Coorientadora
l'Ecole Polytechnique - INRIA

Pedro Olmo Stanciolli Vaz de Melo

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Luís Henrique Maciel Kosmowski Costa

PROF. LUÍS HENRIQUE MACIEL KOSMALSKI COSTA
Departamento de Eletrônica e Computação - UFRJ

Belo Horizonte, 8 de novembro de 2024.

Acknowledgments

This work is part of the Mob Sci-Dat Factory project (ANR-23-PEMO-0004) under the France 2030 program. It was supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), by *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and by the CAPES-STIC-AMSUD 22-STIC-07 LINT project.

Resumo

Os padrões de mobilidade são inerentemente vinculados à natureza humana (por exemplo: variabilidade individual, dinâmica temporal, fatores comportamentais, curiosidade, interação social), tornando a previsão de mobilidade um problema multifacetado e desafiador, que requer modelos sofisticados e dados abrangentes.

Os modelos de aprendizado de máquina se destacam na previsão da localização em que uma pessoa estará no próximo intervalo de tempo, mas geralmente levantam preocupações relacionadas à privacidade. Para abordar essas questões de privacidade, mantendo os benefícios dos modelos de aprendizado de máquina, o Aprendizado Federado oferece uma estrutura distribuída que permite o treinamento colaborativo de modelos de previsão de mobilidade humana, sem exigir o compartilhamento de dados de localização dos usuários, que são altamente confidenciais.

No entanto, no domínio do Aprendizado Federado para previsão de mobilidade *individual*, trabalhos anteriores carecem de uma compreensão completa dos diversos fatores que podem impactar o desempenho dos modelos de previsão baseados no Aprendizado Federado.

Nesta dissertação, fornecemos um estudo abrangente do impacto de vários aspectos relacionados ao comportamento humano, características dos dados, soluções algorítmicas de aprendizado de máquina e estruturação arquitetônica do Aprendizado Federado. Além disso, quantificamos o impacto desses fatores na eficácia (acurácia) e na eficiência (tempo de execução, consumo de memória e de energia) da previsão, revelando que ignorar esses fatores resulta em interpretações errôneas dos resultados, e acatá-los auxilia a obtenção de melhores resultados de eficácia e eficiência.

Palavras-chave: previsão de mobilidade de indivíduos; aprendizado federado; análise de desempenho; heterogeneidade do comportamento de mobilidade.

Abstract

Mobility patterns are inherently linked to human nature (e.g., individual variability, temporal dynamics, behavioral factors, curiosity, social interaction), making mobility prediction a multifaceted and challenging problem that requires sophisticated models and comprehensive data.

Machine learning models excel at predicting the location a person will be at the next time interval, but they often raise privacy concerns. To address these privacy issues while maintaining the benefits of machine learning models, Federated Learning offers a distributed framework that enables collaborative training of human mobility prediction models without requiring the sharing of highly sensitive location data.

However, in the domain of Federated Learning for *individual* mobility prediction, prior work lacks a thorough understanding of the many factors that may impact the performance of prediction models based on Federated Learning.

In this thesis, we provide a comprehensive study of the impact of various aspects related to human behavior, data characteristics, machine learning algorithmic solutions, and Federated Learning architectural structuring. Furthermore, we quantify the impact of such factors on effectiveness (accuracy) and efficiency (execution time, memory, and energy usage) of the prediction, revealing that, ignoring these factors lead to misleading result interpretation, and acknowledging them empowers both effectiveness and efficiency results.

Keywords: individual mobility prediction; federated learning; performance analysis; mobility behavior heterogeneity.

Contents

1	Introduction	8
1.1	Research scope	9
1.2	Thesis contributions	10
1.3	Thesis outline	12
2	Background	13
2.1	Individual mobility	13
2.2	Federated Learning	18
2.3	Summary	23
3	Key Factors Impacting FL4iMP	24
3.1	Human factors	25
3.2	Data factors	26
3.3	Algorithm factors	27
3.4	Architecture factors	28
3.5	Summary	29
4	Dataset: Overview and Analysis	30
4.1	Datasets	30
4.2	Data pre-processing	31
4.3	User mobility profiling	32
4.4	Data analysis	34
4.5	Summary	37
5	Evaluation	38
5.1	Experimental methodology	38
5.2	Impact of human and data factors	40
5.3	Impact of number of clients	45
5.4	Impact of the local algorithm	45
5.5	Impact of data distribution across clients	49
5.6	Summary	52
6	Conclusions and future directions	54
6.1	Conclusions	54
6.2	Future directions	55
	Bibliography	58

Chapter 1

Introduction

Human mobility reflects the inherent human *need to connect* with others (for family, work, or leisure) and is shaped by cultural norms and practices. Indeed, mobility is linked to *human nature* in various ways. For example, humans have a natural *curiosity* and desire to *explore*, however, they also *adapt to changing* environments and circumstances, leading to shifts in mobility patterns over time. Conversely, daily *routines and habitual behaviors* are a significant aspect of human nature, influencing the predictability of some mobility patterns, such as daily commutes. Naturally, there exists a vast literature in the human mobility domain, with solutions aiming at improving human life, e.g., for traffic forecasting (Boyce and Williams [10]), mobile network handovers (Yap et al. [65]), epidemic control (e.g., COVID-19) (Badr et al. [6]), urban planning (Wang et al. [61]), targeted advertising (Zhang et al. [67]) and recommendation systems (Ko et al. [33]).

In such context, many works were proposed to capture and model the complex nature of individual mobility, through the *investigation* of components related to routine or exploration moments in mobility (e.g., Amichi et al. [2, 3, 4], Teixeira et al. [58, 59, 60]). A complementary and related type of work on mobility is *predicting* mobility, where some applications do so at a population level (e.g., traffic forecasting) (Boyce and Williams [10]), while others focus on individual-level prediction (Feng et al. [22]). We here focus on the latter, the task of *individual mobility prediction*, i.e.: **given a history of locations visited by a person, the task is to predict the location she will be in the next time instant** (Belal et al. [8]).

Previous work on individual mobility prediction has explored a wide range of techniques, from traditional Markov-based (Gambs et al. [24], Mathew et al. [41], Amichi et al. [4]) to machine learning (ML) models, such as logistic regression (Cuttone et al. [16]) and, more recently, neural network models (Feng et al. [22], Yang et al. [64], Huang et al. [31]). The latter has often outperformed traditional approaches, especially in terms of accuracy (Feng et al. [22], Yang et al. [64], Huang et al. [31]).

Even though accuracy is of great importance, there is a growing concern about how (private) data (e.g., visited locations) is collected and processed, leading to government regulations worldwide to protect users against privacy violations, such as the General Data

Protection Regulation (GDPR)¹ in Europe and the Brazilian General Data Protection Law (LGPD)².

To meet those user concerns, new privacy-preserving approaches have been proposed to support the design of prediction models, such as *k-anonymity* (Sweeney [57]), *differential privacy* (Dwork [17]), and the *Federated Learning* (FL) framework (McMahan et al. [42]). The key privacy advantage of FL, when compared to the other approaches, lies in keeping private user data on local devices.

To achieve that, in FL, a shared model is first *decentralized* trained on multiple devices, using only local data. That is, private data is only stored and analyzed locally on the corresponding device and used to train a local model. The local models are then uploaded to a central server. Then, a global model is generated by the server, that first aggregates the received local parameters, then chooses candidate devices and finally distributes the updated global model. The previous steps are repeated until convergence, and an optimal model is generated (Feng et al. [23]). Note that, in FL, only the locally learned models are shared with the remote server, instead of the private user data, significantly reducing the risk of data breaches. Furthermore, this decentralized ML technique has been shown to offer privacy preservation while still achieving competitive results (McMahan et al. [42]).

Indeed, given the increasing demand for user privacy protection, Federated Learning has been adopted in a plethora of domains, e.g.: Internet of Things (IoT) (Zhang et al. [68]), healthcare (Lee et al. [34]), Natural Language Processing (Hard et al. [26]) and cybersecurity applications (Lin and Huang [38], Hsu et al. [30]). Yet, applying FL to individual mobility prediction remains a significant challenge, as detailed in the next section.

1.1 Research scope

The literature on the use of the *Federated Learning (FL) framework for individual mobility prediction* (or FL4iMP for short) (Fan et al. [21], Feng et al. [23], Li et al. [35], Wang et al. [62], Ezequiel et al. [20], Proteasa et al. [47]), primarily focuses on prediction accuracy improvement, via the proposal of new FL4iMP models. However, some of these models (e.g., Fan et al. [21], Ezequiel et al. [20], Proteasa et al. [47]) lack the consideration of user heterogeneity, which is a key feature on individual mobility prediction, given how diverse the mobility of different individuals is (Amichi et al. [2, 3, 4]).

Indeed, more recently, some studies started addressing the heterogeneity of clients'

¹<https://gdpr-info.eu/>

²<https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/>

mobility patterns in the design of their FL models (Zhang et al. [69], Park et al. [46]). While these studies offer valuable insights, they mainly consider heterogeneity in terms of data sampling diversity or model parameter dissimilarity, i.e., not tackling heterogeneity from an individual mobility behavior point of view but rather from an indirect perspective.

Furthermore, no prior work, not even those that target user heterogeneity (Feng et al. [23], Zhang et al. [69], Park et al. [46]), report results separately for users with different mobility behaviors, so as to allow an assessment of how FL4iMP performance may vary across different behaviors. Instead, they focus on aggregated results, which can obscure performance differences that emerge for users with diverse mobility profiles.

Moreover, most studies also rely on Location Based Social Networks (LBSN) datasets which, as will be argued later (Chapter 3), tend to be very sparse and do not capture enough details of one’s mobility. As such, results are hardly generalizable. Also, the differences in evaluation setup across those studies make it hard to reach a common knowledge on FL4iMP performance.

Given the issues discussed previously, we notice that the FL performance is impacted by a plethora of factors (e.g., mobility behaviors and data), and the existing FL literature on individual mobility prediction lacks a *comprehensive understanding of the factors that most impact FL prediction results*, including aspects ranging from human behavior to solution design.

In this thesis, we fill the presented literature gaps by offering a much broader and insightful evaluation. A key research question driving our study is: *“How do various factors associated with the FL4iMP environment, from the human nature of its participants to architectural and algorithmic design decisions, impact its performance?”* Specifically, we aim to address: *“How does FL4iMP perform for people with very diverse mobility patterns?”*.

1.2 Thesis contributions

Our main contributions can be summarized as follows:

1. We identify different factors impacting FL4iMP performance.

On one hand, individual mobility is naturally heterogeneous, as it is driven by contextual aspects, personal characteristics, and motivations (Amichi et al. [2, 3, 4], Teixeira et al. [58, 59, 60], Gonzalez et al. [25], Smith et al. [54]). Thus, *human* factors related to individual mobility are key to our study. Additionally, for prediction purposes, human mobility is captured by the available data used to train (and test) the prediction models.

As such, different characteristics of the *data* may impact the prediction results. Diving deeper into the FL4iMP solutions, factors associated with the type of learning *algorithm* used to build the local and global models, as well as how the federated *architecture* is established, may also impact the FL execution and results. With that in mind, we assess how different factors associated with *human* behavior, *data* properties, learning *algorithms* as well as the FL *architecture* impact the FL4iMP results.

2. We evaluate each identified factor, on the context of FL4iMP, offering a much more comprehensive study of FL4iMP.

Our study relies on previously proposed features (Mucceli et al. [43], Teixeira et al. [59], Song et al. [55], Gonzalez et al. [25], Canzian and Musolesi [12]) and a profiling approach (Amichi et al. [2, 3, 4]) to identify and characterize different mobility patterns at the *individual* level.

Additionally, instead of relying on LBSN data, like previous work (Li et al. [35], Wang et al. [62], Ezequiel et al. [20], Proteasa et al. [47]), our analyses are performed on two cellular mobile phone datasets, namely, the *Shanghai* and *Shenzhen* datasets (Zhang et al. [66]). These cellular datasets are much less sparse when compared to LBSN datasets (as will be argued later in Chapter 3), and were collected in two different cities of heterogeneous properties in terms of spatial and temporal coverage, as well as population and area diversity (as will be discussed in Chapter 4).

Furthermore, we rely on *Flower*³, a state-of-the-art framework for FL (Beutel et al. [9]) to instantiate multiple FL4iMP solutions, by using two obtained literature FL4iMP models from Ezequiel et al. [20], and varying architectural parameters, within a unified (and thus comparable) setup. Our evaluation considers metrics related to both effectiveness (prediction accuracy) and efficiency (execution time, memory, and energy usage).

3. We provide insights into FL4iMP performance in various relevant scenarios.

In contrast to prior work, we here offer, within a *unified evaluation environment*, a much broader investigation of how different factors related to *human* behavior, *data*, FL *architecture* and *algorithms* impact FL4iMP performance. Moreover, unlike prior work, we offer insightful discussions on how FL4iMP performs for users with diverse mobility nature. To that end, besides using two datasets with very different properties, we cluster users into mobility profiles and employ various mobility-related features to characterize such groups as well as the complete user population in each dataset. Finally, we use the characterization results to guide our evaluation and discussions of FL4iMP performance.

The results of this thesis were published in two papers (Esper et al. [19, 18]).

³<https://flower.ai/>

1.3 Thesis outline

The rest of this thesis is organized as follows. In Chapter 2 we provide background foundations for individual mobility and Federated Learning, highlighting the literature strengths and shortcomings. Then, in Chapter 3 we categorize and discuss key factors impacting FL4iMP performance. Next, Chapter 4 describes the data, its processing, and methodology applied to isolate and understand the population heterogeneity of the datasets. Subsequently, Chapter 5 builds upon the previous chapters, evaluating the impact of key factors on FL4iMP, aided by data analysis, and offering accuracy, energy, computational resource, and time results. Finally, 6 concludes our work, summarizing our contributions, and provides directions for future work.

Chapter 2

Background

In this chapter, we go through a literature overview that provides background information on both individual mobility and Federated Learning. In Section 2.1, we first explain individual mobility modeling. Next, we discuss related work to this topic, detailing how individual mobility can be analyzed, in order to drive better solutions, improving context information. Finally, we present the task of individual mobility prediction, explaining its workflow and providing a timeline leading to the FL solutions for this task. In Section 2.2, we discuss the Federated Learning framework, and its applications outside the individual mobility domain. Next, we focus on the Federated Learning framework for individual mobility prediction (FL4iMP), by first explaining how its ecosystem works, followed by a literature review, pinpointing strengths and shortcomings of previous work.

2.1 Individual mobility

The literature on individual mobility covers an extensive range of studies, with applications in a variety of domains such as: epidemic dissemination (e.g., COVID-19) (Badr et al. [6]); traffic forecasting (Boyce and Williams [10]); urban planning (Wang et al. [61]); and mobile network handovers (Yap et al. [65]). In this Section, we first define how individual mobility has been *modeled* in the literature, as well as prior efforts to analyze mobility aiming at empowering solutions in this domain. Next, we define the individual mobility *prediction* task, followed by a literature overview of prior models proposed to tackle this task, which, in common, lack the privacy preserving mechanisms offered by Federated Learning.

2.1.1 Individual mobility modeling and analysis

To solve such complex tasks (e.g., traffic forecasting), individual mobility is modeled via a time ordered sequence of *location records*, where each *location record* r represents a visit of a user to a certain location, and is composed by a tuple of: a user identifier (ID) u ; the timestamp of the visit t ; and a location ID l , i.e., $r = \langle u, t, l \rangle$, time-ordered. The combination of multiple location records generates a *mobility trajectory*, i.e., a set of ordered location records $\{r_1, r_2, \dots, r_n\}$, where a mobility trajectory s of a user u can also be defined as $s_u = \{\langle l_1, t_1 \rangle, \langle l_2, t_2 \rangle, \dots, \langle l_n, t_n \rangle\}$ (Feng et al. [23]).

The literature is quite rich in prior studies where the main goal is to analyze mobility patterns, by mining such patterns from mobility data (e.g., Gonzalez et al. [25], Amichi et al. [2, 3, 4]). Some of these studies aim at exploring such patterns to improve particular solutions (e.g., Brockmann et al. [11], Asgari et al. [5], Hess et al. [27]), whereas others propose metrics to characterize user mobility behavior with respect to various properties. In the latter direction, one example is the work by Teixeira et al. [58, 59, 60], where the authors proposed three metrics to model mobility behavioral patterns, namely, *Regularity*, *Stationarity* and *Diversity*:

Regularity aims to capture the preferences of a person to return to previously visited locations, quantifying per user mobility trajectory how prone the user is to return to previously visited locations. Formally, given a time-ordered sequence $X = (x_1, x_2, \dots, x_n)$ of locations visited by a person, the *Regularity* of the sequence is: $reg(X) = 1 - n_{unique}/n$, where n_{unique} is the number of distinct locations in X . Ranging from 0 to 1, the closer $reg(X)$ is to 1, the more prone the user who generated X is to return to previously visited locations.

Stationarity correlates to the number of observations for which the user stays continuously in the same location, also measuring per user mobility trajectory how inclined the user is to stay continuously in the same location. Given a time-ordered sequence $X = (x_1, x_2, \dots, x_i, x_{i+1}, \dots, x_n)$ of locations visited by a person, a *stationary transition* occurs at time i if $x_i = x_{i+1}$. Formally, for a time-ordered sequence $X = (x_1, x_2, \dots, x_n)$ of locations visited by a person, and assuming regular time intervals, the *Stationarity* of the sequence is: $sta(X) = sta_{trans}/(n - 1)$, where sta_{trans} is the number of *stationary transitions* in X . Ranging from 0 to 1, the closer $sta(X)$ is to 1, the more inclined the user who generated X is to stay continuously in the same location.

Diversity calculates how diverse a user's visitation pattern is, estimating the number of distinct sub-trajectories that can be found in the user's original mobility trajectory. Formally, given a time-ordered sequence $X = (x_1, x_2, \dots, x_n)$ of locations visited by a person, the *Diversity* of trajectories associated with X , $div(X)$, is given by the number of distinct trajectories in X . More specifically, if X is interpreted as a string, the *Diversity*

of trajectories is calculated by the ratio between the number of distinct sub-strings and the total number of sub-strings in X . Ranging from 0 to 1, the closer $div(X)$ is to 1, the more diverse and harder to predict the user who generated X is.

A different approach for *analyzing* mobility is to split users with similar mobility behavioral patterns into clusters, as done by Amichi et al. [2, 3, 4]. Their work proposes three categories of users: *Scouters*, *Regulars* and *Routiners*, given how prone they are to either explore new places or keep a routine-like behavior. Each group can be defined as follows:

- **Scouters:** defines the set of users more prone to explore new areas, making the *novelty* the most prevalent component in their mobility pattern.
- **Routiners:** defines the set of users who the *routine* as the most prevalent component in their daily mobility life. Although moments of exploration may also exist, users in this profile spend most of their time revisiting few previously visited locations with long periods of stationarity.
- **Regulars:** users in this profile have a mobility pattern well balanced between routine and novelty moments, with no stronger predominance of one of them.

To create these profiles, the authors first categorize each location appearing in a user mobility trajectory as either an exploration (i.e., a new or rarely visited location) or a return (i.e., a regularly visited location). One of their methods to do so is to compute the importance of each location as the frequency of its appearance in the user's mobility trajectory. A location is classified as a *return* (R) if it appears more than a certain level l of the average importance of all locations visited by the user; otherwise, it is classified as an *exploration* (U). Once each location is classified, two features are calculated for each user: *Intermittency* and *Degree of Return*.

Intermittency μ is the sum of the average number of successive explorations $\#U$ and the average number of successive returns $\#R$, $\mu = \#U + \#R$. This feature measures the frequency of shifts between explorations and returns. Higher values indicate fewer shifts, while lower values indicate frequent shifts, revealing whether (i) an individual visits several exploration-like (vs. return-like) locations in a row before revisiting a return-like (vs. exploration-like) one or (ii) is versatile, keep frequently switching between exploration- and return-like locations.

Degree of return α is the angle whose tangent is the ratio between the average number of successive returns $\#R$ over the average number of successive explorations $\#U$, $\alpha = \arctan(\frac{\#R}{\#U})$. This feature indicates how often the user revisits a return-like location after an exploration-like one. Higher values suggest that the average number of successive explorations is low, due to the user's preference for returning to regularly visited locations.

After both *Intermittency* and *Degree of return* are calculated for each user, either the *Gaussian Mixture probabilistic Model* (GMM) (Reynolds et al. [49]) or the *k-means* (Lloyd [39]) algorithm clusters each user according to the two metrics.

Some other techniques focus on *entropy* to model user mobility behaviors, such as the work by Song et al. [55], where authors utilize the full spatial-temporal order present in an individual's mobility pattern. To do so, it relies on the visitation frequency and on the order in which the locations were visited. Formally, considering a time-ordered sequence $X = (x_1, x_2, \dots, x_n)$, $S_i = -\sum_{X'_i \subset X_i} P(X'_i) \log_2[P(X'_i)]$, where $P(X'_i)$ is the probability of finding a particular time-ordered sub-sequence X'_i in the user's i trajectory X_i .

Finally, the work by Gonzalez et al. [25] measures mobility behavior from a different perspective, i.e., if a user has a confined mobility pattern, concentrated in a specific area, or if she tends to explore longer distances. Formally:

$$RG = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - r_{cm})^2}, \quad (2.1)$$

where N is the total number of records in the sequence, r_i the geographical coordinates of the i^{th} record, and $\sum_{i=1}^N (\frac{r_i}{N})$ the individual's center of mass. Ranging from 0 to ∞ and usually measured in kilometers, the bigger RG is, the less confined the user's mobility behavior is.

2.1.2 Individual mobility prediction

Solutions to the mobility prediction task often leverage mobility patterns for improving its results (Feng et al. [23], Zhang et al. [69], Park et al. [46]). The individual mobility prediction task is formally defined as follows: **given a past mobility trajectory of location records for user u , $s_u = \{r_1, r_2, \dots, r_n\}$, the mobility prediction task is to define the next location l_{n+1} visited by u in the next time interval t_{n+1}** (Feng et al. [23]). Early mobility prediction methods were based on Markov-based models (Gambs et al. [24], Mathew et al. [41], Amichi et al. [4]), in which a Markov chain is built from the user mobility trajectories, where a Markov chain is created with the transitions between visited locations, and prediction is based on each transition frequency.

Later, Machine Learning (ML) models, such as Logistic Regression models (Cuttone et al. [16]), started to include context information to improve prediction (e.g., the user's home location). However, more recently, the state-of-the-art relies on neural net-

works models (a type of ML model), that often outperform traditional approaches, especially in terms of accuracy (Feng et al. [22], Yang et al. [64], Huang et al. [31]). Neural network models, such as Recurrent Neural Networks (RNNs) (Rumelhart et al. [51]) work well with sequential data (the case of mobility trajectories), since they capture both spatial and temporal patterns in mobility prediction, where RNNs identify periodic trajectory patterns. However, RNNs notably suffer from *exploding* and *vanishing* gradients, which happen when the values of gradients either are too small (*vanishing*), preventing further learning, or too big (*exploding*) making the learning process unstable (Hochreiter et al. [29], Ezequiel et al. [20], Luca et al. [40]). To solve this problem, two variations, namely Long-Short-Term Memory (LSTMs) (Hochreiter and Schmidhuber [28]) and Gated Recurrent Units (GRUs) (Cho et al. [14]) were proposed with gates that mitigate the issue, by learning if the information is relevant keeping or not.

Deepmove (Feng et al. [22]) and *Flashback* (Yang et al. [64]) are two examples of state-of-the-art RNN-based individual mobility prediction models. The first model utilizes a GRU network that is able to capture long-range dependencies within the trajectory, building upon contextual features like the regularity of mobility and spatial-temporal relationships. For example, *Deepmove* utilizes the periodical information of visits, i.e., which location was visited in the same hour of different days. *Flashback* also explores spatial-temporal contexts: it searches for historical hidden states with context similar to the current one, computing the weighted average of these historical hidden states to feed a RNN-based model. Both temporal and spatial contexts are explored, with the weights decreasing as the time between visits increases and the distance between past and current locations increases.

The aforementioned machine learning models (often referred to as *Centralized* or *Joint* models) stand out given their accuracy results, however, they also present concerning privacy issues. This type of modeling combine private user data (i.e., mobility trajectories) and collectively store them on a centralized server, leading to a powerful joint model that is able to leverage information from the whole population. However, this approach also leads to privacy and security risks, as a consequence of gathering and storing all raw user data in a centralized location. In our case, private information is the user trajectories, which contain the sequence of locations visited, and can be considered sensitive information. Fig. 2.1 depicts the workflow of *Centralized* solutions, where:

1. Private (raw) user data is uploaded to a centralized server, namely, the user trajectories with the sequence of locations visited;
2. An individual mobility prediction model (often a ML variation) is trained with the whole population personal data;
3. An optimal (e.g., converged) model is generated and distributed.

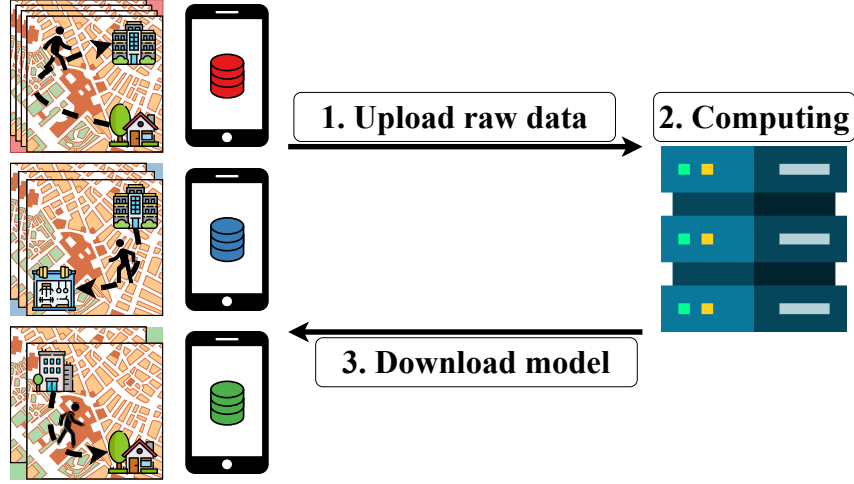


Figure 2.1: Overview of the *Centralized* workflow.

In this type of architecture, an adversary needs access to a single server to have access to the whole population private information. Furthermore, the information is stored as is, i.e., the original raw mobility trajectories are stored in the server, placing users in privacy and security risks. Federated Learning addresses the aforementioned issues by *keeping users' personal data locally stored*, as detailed next.

2.2 Federated Learning

In this section, we first discuss the Federated Learning framework, and its applications outside the individual mobility domain. Next, we explain the workflow of a FL framework for individual mobility prediction (FL4iMP), detailing how it increases user privacy when compared to traditional *Centralized* individual mobility prediction solutions. We then provide a comprehensive literature review, describing the strengths and shortcomings of previous work on the topic.

2.2.1 Overview of Federated Learning

Given the increased user concerns about how their private data is collected is processed, new privacy-preserving machine learning approaches have been proposed to support the design of ML-based prediction models. One such approach is the Federated

Learning (FL) framework (McMahan et al. [42]). FL offers privacy preservation while still achieving competitive ML results. To achieve that, a shared model is trained in a *decentralized* fashion on multiple devices, independently, i.e., private user data is only stored and modeled locally on the user’s device, creating *local models*. These local models (indeed, the model parameters) are then uploaded to a remote server that aggregates them to form a *global model*, typically using methods like averaging. The global model is then redistributed to devices. This process is repeated iteratively until the model converges. The key privacy advantage of FL lies in keeping private user data on local devices. Only the learned local models (i.e., model parameters) are shared with the remote server, significantly reducing the risk of data breaches (McMahan et al. [42]).

Indeed, a plethora of other workflows have been proposed for FL, to overcome the drawbacks of its decentralized nature, e.g.: *synchronous*, *asynchronous* and *semi-asynchronous* approaches (Xu et al. [63]), with different aggregation techniques, that utilize *model* or *gradient* aggregation (Qi et al. [48]). For the purposes of this thesis, we focus on a *synchronous* FL workflow with *model* aggregation.

Given the increasing demand for user privacy protection, many domains have adopted Federated Learning to address their tasks, e.g.: Internet of Things (IoT), healthcare, Natural Language Processing (NLP) and cybersecurity applications. For IoT applications, authors in Zhang et al. [68] considered FL solutions for: (i.) *Smart home applications*, e.g., speech recognition of Wake-Up-Word IoT devices; (ii.) *Smart city applications*, e.g., power consumption patterns learning; (iii.) *Autonomous driving applications*, e.g., limiting the vehicle communication with the Cloud via Federated Learning while still adapting to real-time changes; and (iv.) *Virtual reality applications*, e.g., wearable eye tracking data being trained locally and aggregated via Federated Learning. Healthcare applications with Federated Learning include the work by Lee et al. [34], in which, without sharing patient-level information, a Federated Learning model finds clinically similar patients in different hospitals. A state-of-the-art NLP application that adopts FL is Google’s keyboard (namely *Gboard*, proposed by Hard et al. [26]), that is already implemented and native to some Android devices. This next-word prediction model is still based on user typing history, however, it is embedded in the privacy preserving Federated Learning system. Finally, cybersecurity applications include the works by Lin and Huang [38] and Hsu et al. [30], in which authors used FL for malware classification and malware detection in Android devices, respectively.

For the purposes of this thesis, the domain in which Federated Learning is applied is for individual mobility prediction, where the framework learns spatial-temporal characteristics of users, to predict their whereabouts. This workflow will be detailed in the next section.

2.2.2 Federated Learning for Individual Mobility Prediction — the Ecosystem

As depicted by Fig. 2.2, a *Federated Learning (FL) framework for individual mobility prediction* — FL4iMP — solution is fed by traces of data describing the mobility trajectories of users. For the decentralized learning of the prediction model, user trajectories are stored in local devices and given as the main input to the FL4iMP solution. The goal of the FL4iMP model is then to predict the next location where the user will be at the next time instant.

The workflow of a FL4iMP solutions is as follows:

1. Each device trains a *local model* using only the local (*training*) data for a certain *number of epochs*. The *number of epochs* determines how many times each device will iterate over its entire local dataset during this local training step, ensuring the local model is thoroughly trained before being shared;
2. Instead of uploading their sensitive data, the devices only share the results of their locally trained models with the remote server, e.g., model parameters or weights;
3. The server aggregates multiple model weights/parameters, building a *global model*, which is the method of aggregating multiple local models. While different aggregation strategies have been proposed (Park et al. [46], Li et al. [36]), *FedAvg* (McMahan et al. [42]) is the most commonly used one, which *averages local model parameters to update the global model*;
4. The server distributes the aggregated model to local devices, with the previous steps being repeated for a certain number of *communication rounds*. The *communication rounds* refer to the number of times the server and the clients exchange model updates;
5. This iterative process continues until a *stop criterion* is met (e.g., convergence). The learned model can then be used to predict the next location visit of a (test) user.

Note that, as described, only the (local) model parameters are shared with the remote server, while the user's sensitive data (i.e., visited locations) are kept at the local devices. This fundamental property of FL makes it an attractive framework to address the increasing user concerns with privacy protection (McMahan et al. [42])¹. One could

¹Indeed, some works study how to reconstruct the user trajectory from the shared parameters (Feng et al. [23]), however, this analysis is not in the scope of this thesis.

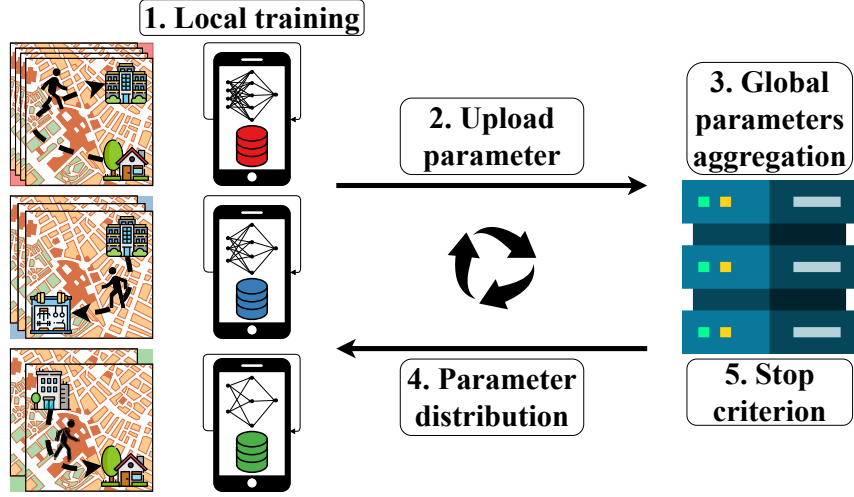


Figure 2.2: Overview of the FL4iMP workflow.

argue that privacy protection could be ensured by using only the user data to create a local model, without sharing the local parameters, however, users can benefit from one another by learning from the mobility patterns summarized in other local models.

Indeed, in an *ideal private scenario*, each user runs her own local model on her own local device, protecting her data from others. However, often due to hardware limitations, prior studies (Proteasa et al. [47], Ezequiel et al. [20], Li et al. [37]) on FL adopt a different architecture, using the abstraction of *data silos*, i.e., local entities that aggregate data from multiple users (e.g., users within the same department) (Stedman [56]). As later discussed in Section 3.4, this abstraction may greatly impact not only user privacy, but also the effectiveness and efficiency of the FL4iMP solution. Throughout this thesis, we use the term *client* to refer to such local entities, which could be either personal devices or data silos aggregating multiple users.

2.2.3 FL4iMP literature review

To our knowledge, one of the first FL4iMP solutions is *PMF* (Feng et al. [23]). One of its key features is enhanced user privacy protection, obtained by applying *differential privacy* (Dwork [17]) to parts of the original data. This approach mitigates attacks where the adversary can extract private user information from the shared model parameters. PMF also adds a personal user bias into each local model to address heterogeneous user behavior. Yet, according to the reported results, such personal user bias approach produced only marginal improvements (up to $\sim 3\%$) in the overall prediction accuracy.

Two recently proposed FL4iMP solutions that address heterogeneity in user mo-

bility more explicitly are *FR-HMP* and *FedGeo*. In *FR-HMP* (Zhang et al. [69]), clients with similar local model parameters are grouped into clusters, enabling “similar” clients to learn from one another rather than from the whole population. By doing so, the authors use the similarity between clients’ model parameters as an *indirect measure* of the similarity between the mobility patterns of the corresponding users. *FedGeo* (Park et al. [46]), in turn, introduces new components to the FL framework to mitigate the effects of client heterogeneity. One such component is a strategy to aggregate the currently learned global model into the selected clients’ local models based on the layer-wise similarity between those models. This approach aims to mitigate the disparity among client models arising from the heterogeneity of user trajectories. It also employs entropy-based sampling to select clients to participate in FL based on the diversity of their mobility patterns, aiming at building more generalizable global models.

Although both studies acknowledge the challenge brought by heterogeneous user mobility patterns, they address heterogeneity from a model gradient perspective (i.e., two clients are diverse if their uploaded model parameters are dissimilar) or a data sampling perspective (i.e., two clients are dissimilar if they report diverse distributions of visited locations). Thus, pinpointing heterogeneity is tailored to the local model’s capability (e.g., embedding, encoding, attention mechanisms) or to the clients’ data quality in capturing/describing diversity in mobility patterns. While such approaches report overall results under specific settings that are hard to generalize, they do not tackle heterogeneity from a individual behavior point of view but rather from an indirect data and model perspective. Moreover, *FR-HMP* and *FedGeo* have been evaluated in very restricted scenarios regarding user populations (only 200 and 10 users, respectively).

A few other studies focus on varying the algorithm used to learn the local models. In Proteasa et al. [47], the authors adopt Recurrent Neural Networks (RNN), Long-Short-Term Memory (LSTM), and also Gated Recurrent Unit (GRU) (Aggarwal [1]) as alternative learning algorithms. This work shed some light on the discussion of which neural network to use in FL4iMP, with results suggesting GRU as the best performer in most scenarios analyzed. In Ezequiel et al. [20], in turn, the authors evaluate *FL-Flashback* and *GRU-Spatial*. The former is a federated version of the aforementioned *Flashback*, which was initially proposed as a centralized mobility prediction approach (i.e., *without* FL) (Yang et al. [64]). The core of the (FL-) *Flashback* algorithm is the same as its centralized proposal. However, *GRU-Spatial*, keeps only the spatial perspective of the contexts, modeling temporal relations by the order of visits. In common, both studies give no attention to mobility heterogeneity, neither in the solution design nor the evaluation, and use the same strategy to aggregate local models, namely *FedAvg*.

As an alternative to RNN, LSTM, GRU based solutions, some works adopt *attention mechanisms* (Li et al. [35], Wang et al. [62], Fan et al. [21]), where the input information is scored (e.g., soft-max) according to the context (e.g., attention maps) and

the model focuses on the information with higher scores (Luca et al. [40]). In particular, *STSAN* (Li et al. [35]) integrates spatial-temporal information with a self-attention network as local model and employs a personalized FL approach mixing global and local models. In *STLPP* (Wang et al. [62]), an evolution of *STSAN*, clients cooperatively train their personalized models. However, the evaluation reported in both studies, restricted to Location Based Social Networks (LBSN) data, focuses primarily on prediction accuracy and communication rounds. It thus lacks an assessment of training times, which are quite relevant given that attention-based networks tend to have higher time complexity than other neural network approaches (Cunha et al. [15]). Finally, the work proposed in Fan et al. [21], adopts pre-training strategies, where the model is trained with small datasets to accelerate local training (Belal et al. [8]). Yet, their model required over 1,000 communication rounds to converge, which is considerably slower than the convergence reported in other studies.

2.3 Summary

In this chapter, we provided a literature overview and necessary background contexts and definitions for both individual mobility and Federated Learning. We first explained both individual mobility modeling and prediction, explaining how its correlation and privacy need of Federated Learning. Then, we also explained the FL framework, including applications outside the individual mobility domain.

Finally, we focused on FL4iMP, explaining its workflow, followed by comprehensive literature review, highlighting how prior work on FL4iMP focuses primarily on proposing new approaches, with evaluations that leave the FL4iMP research arena still very unsettled. Notably, no prior work, not even those that target user heterogeneity (Feng et al. [23], Zhang et al. [69], Park et al. [46]), report results separately for users with different mobility behaviors, so as to allow an assessment of how FL4iMP performance may vary across different behaviors. Most studies also rely on LBSN datasets (mainly from the *Foursquare* location-sharing service), which, as will be argued in Section 3.2, tend to be very sparse and do not capture enough details of one’s mobility. As such, results are hardly generalizable. Moreover, the differences in evaluation setup across those studies make it hard to reach a common knowledge on FL4iMP performance.

In contrast, we here offer, within a *unified evaluation environment*, a much broader investigation of how different factors related to human behavior, data, FL algorithms and architecture impact FL4iMP performance. In the next chapter, we categorize and discuss the identified key factors impacting FL4iMP performance.

Chapter 3

Key Factors Impacting FL4iMP

As depicted by 3.1, in this chapter we categorize and discuss key factors impacting FL4iMP performance in four categories:

Human factors (Section 3.1) present the impact that different mobility traits (exploratory or routine behaviors) have, where more exploratory users tend to be harder to predict. Furthermore, similar exploratory/routine behaviors tend to be different in each city, as each unique city urban planning (available transportation options, distribution of commercial and residential districts...) lead to more or less complex spatial-temporal explorations and returns (more/longer stops, longer distances, ...).

Data factors (Section 3.2) highlight how different spatial-temporal resolutions can influence the results. For example, the exact same user trajectory can be captured in various ways given the frequency of the temporal resolution or the precision of the spatial resolution. These factors show how the Federated Learning algorithm can benefit from a more uniform, spatially accurate and frequent resolution.

Algorithm factors (Section 3.3) refer to the different FL models available for this task, both for *Global* and *Local* models, and **Architecture factors** (Section 3.4) represent the different ways in which a FL algorithm can be structured and/or parameterized.

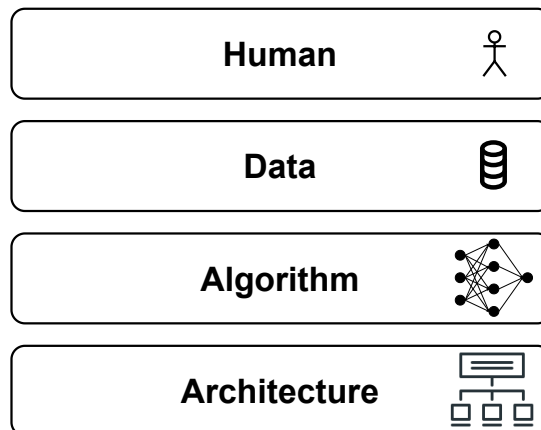


Figure 3.1: Categories of factors that impact individual mobility prediction with FL.

3.1 Human factors

User mobility behavior is a significant element influencing mobility prediction accuracy and thus, FL4iMP performance, being the source of data for learning the prediction model. As such, naturally, factors related to user mobility patterns are of primary importance.

The mobility trajectory of a person can be split into two main components (Teixeira et al. [59]). The *routine* comprises the locations the person often visits (e.g., home, work, sports clubs, school, etc). In turn, the *novelty* component consists of places she rarely or for the first time visits, related thus to moments of exploration. It is well known that individual mobility is highly heterogeneous with respect to these two components, even if restricted to a common (small) area (e.g., a given city). Some people naturally have a more exploratory nature, more often visiting places they have never been before. As such, their mobility patterns are characterized by a stronger novelty component. In turn, others are more driven by a daily circadian life routine, thus constraining their mobility to commonly known places that are regularly and repetitively visited (Gonzalez et al. [25], Mucceli et al. [43], Amichi et al. [2]).

From a prediction perspective, more exploratory users are naturally less predictable, being thus more challenging for mobility prediction. While this is true for any prediction model (even *Centralized* ones), in FL4iMP, additional challenges may emerge from greater user heterogeneity. As further discussed in Section 3.3, users with very heterogeneous mobility patterns may generate local models (e.g., model parameters) that are very different from one another. This asymmetry can significantly challenge the aggregation algorithm, which might struggle to generate a global model suitable enough for all dissimilar clients.

This thesis investigates how FL4iMP performs for users with heterogeneous mobility behavior. To that end, in Chapter 4, we employ a profiling approach proposed by Amichi et al. [2, 3, 4], to group users according to their mobility patterns, from exploratory to routine-like ones. We also characterize the observed mobility patterns with respect to several commonly used features.

3.2 Data factors

Spatial-temporal datasets are of enormous importance to studying human mobility, consisting of *time-stamped and geo-referenced samples* of users’ trajectories. Mobility datasets can be collected from a GPS-capable device or from a network, which gathers data when a mobile device interacts with it. For example, Telecom operator networks store Call Detailed Records (CDRs) containing information on generated events (e.g., calls, SMS) by mobile devices and the cell towers they were connected to and managed the event. In contrast, Location Based Social Networks (LBSNs), such as *Foursquare* and *Yelp*, record the point of interest (POI) where the user willingly chose to share her location (i.e., check-in). Other examples include eXtended Data Records (XDRs), which include greater detail in terms of captured parameters when compared to CDR data (Barrett et al. [7]).

To ensure unbiased and meaningful research on mobility prediction, datasets should capture: (1) peoples’ daily mobility behaviors, and (2) diversity of scenarios reflecting the natural heterogeneity that emerges across people, regions and time intervals.

Daily mobility behaviors: Datasets originated from different sources exhibit different characteristics, reflecting behaviors captured in the data. Such characteristics may impact how the user’s true mobility is represented and, thus may affect FL4iMP performance. We argue that, for the sake of learning a proper mobility prediction model, the dataset must capture the *routine* component of the user mobility as much as possible. The novelty component, comprising priorly unseen or rarely visited locations, though reflecting important moments in the users’ lives, has a low repetitive pattern, and few past history, which adds noise to the ML mobility learning.

GPS datasets offer a detailed view of users’ mobility, allowing a fine-grained investigation on mobility patterns. However, they are usually limited in the number of participants, time duration, and space coverage, raising concerns about the generality of the prediction results. In contrast, network-sourced (e.g., CDR and LBSN) datasets typically comprise larger populations, longer periods (e.g., months), and larger geographical areas (e.g., a city or region). Yet, their “interaction-demand” property raises concerns about how such actions relate to the routine component of a person’s mobility.

One could argue that users more often place check-ins in newly visited places, while check-ins at one’s home or workplace are rare (Silva et al. [53]). Thus, certain LBSN datasets – e.g., the ones related to check-ins in a location-sharing service (e.g., *Foursquare* – bring geographic information more often related to particular moments of leisure, with a specific or location (e.g., bar, restaurant). As such, LBSN datasets often describe mostly novelty-seeking moments in mobility, offering only a coarse a coarse view of user mobility, with low temporal and spatial granularity, which challenges its usage for

learning mobility patterns.

In CDR datasets, in turn, the granularity will be dictated by how often the user places a call or an SMS. Nevertheless, with mobile devices becoming proxies for human presence and activity, CDRs are acknowledged by the research community as a meaningful tool to study human mobility (Jiang et al. [32]), outlining many interesting properties regarding human mobility and activity patterns. Moreover, data completion strategies (Chen et al. [13]) have been commonly employed to reduce spatial-temporal sparsity in such datasets. Such observations makes CDRs a good candidate to investigate FL4iMP.

Diversity: Other important factors relate to the target population, spatial area, and the time period covered by the data. Naturally, more complex mobility patterns are expected over larger areas or regions with a greater diversity of locations. Similarly, the target area’s social, economic, and geographical characteristics (e.g., distribution of residential and commercial areas, availability of transportation modes, etc) may also directly influence how people move around it. Such characteristics will be reflected in data properties such as the total number of unique locations or the number of (unique) locations in each user trajectory. Naturally, in scenarios of more complex and diverse mobility patterns, more data is required so that the learning algorithms can build accurate and generalizable models.

In sum, mobility datasets can vary significantly concerning properties that may impact FL4iMP. In this work, we aim to investigate some of them. While prior FL4iMP studies relied mainly on LBSN data (Li et al. [35], Wang et al. [62], Ezequiel et al. [20], Proteasa et al. [47]), we use two CDR datasets with distinct spatial-temporal properties, as will be discussed in Chapter 4.

3.3 Algorithm factors

The algorithms employed to learn the local models (at the clients) and to aggregate them into a global model (at the remote server) are the heart of the FL4iMP solution and thus play a crucial role in its performance. Regarding the (local) learning algorithms, the choice of which neural network to adopt reflects how the local algorithm captures and models the spatial-temporal relationships present in the input data, potentially impacting FL4iMP performance. A more sophisticated model may be able to capture these relationships more accurately, improving its predictive power. Yet, it may also challenge the clients, which, if constrained in resources (e.g., memory, energy), may struggle to process complex models efficiently.

Regarding the aggregation algorithm, one key factor relates to how the multiple

local model weights are combined into a global model. The seminal *FedAvg* (McMahan et al. [42]) algorithm simply averages the received local model weights. However, prior work has shown that FL performance may degrade in face of clients handling non-IID (independent and identically distributed) data (Zhu et al. [70]). This degradation occurs because the divergences in local models due to non-IID data slow down the convergence of the global model and worsen the learning performance.

Alternative aggregation methods have been designed to reduce the instabilities due to non-IID data (Li et al. [36], Park et al. [46], Sattler et al. [52]) – i.e., the client drifts (Li et al. [37]) – by avoiding the aggregation of dissimilar local model parameters of clients. For instance, *FedProx* reduces the distances between local and global models by incorporating a Regularization term (i.e., μ) in the local objective function that restricts the local updates, bringing the averaged model closer to the global optima. However, prior related work have evaluated heterogeneous FL4iMP solutions under limited user populations (Li et al. [36], Park et al. [46]) and have not tackled heterogeneity from a human behavior perspective, but rather from the point of view of dissimilar model weights or data sampling distributions, i.e., model weights and data sampling are used as proxy metrics to infer heterogeneity in mobility patterns. By capturing human behavior representations (i.e., mobility profiles) before the model training, we are able to study the impact of IID vs. non-IID data in user mobility prediction according to distinct mobility profiles. Additionally, the profile characterization brings interpretability to non-IID data, helping us understand the FL4iMP results – an investigation lacking in prior work.

One aspect of particular interest is how fast the algorithm can generate a robust global model (i.e., converge). Considering that the aggregation algorithm generates the global model, and that aggregating similar model clients leads to faster global convergence, fewer communication rounds between server and local clients are needed, which can intuitively save resources.

In our study, we evaluate two local learning algorithms and two aggregation methods, considering scenarios of both IID and non-IID clients. Our results will be discussed in Chapter 5.

3.4 Architecture factors

One key FL architectural factor relates to how users are distributed among the clients. As argued in Section 2.2.2, from the perspective of *ideal private scenario*, each client would be a personal device. Yet, existing approaches use the *data silos* abstraction and thus consider multiple users per client. Such architectural design decision may greatly

impact not only user privacy, but also the effectiveness and efficiency of the FL4iMP solution. In fact, by leveraging data from multiple users into a local client, FL can train local models with superior generalization capabilities, potentially leading to more accurate predictions. This leads to a non-trivial trade-off between privacy and prediction accuracy.

In Chapter 5, we assume such *data silos* architecture and assess the impact of increasing the number of clients on FL4iMP. Furthermore, we also vary how the user population is distributed across the clients, building scenarios of IID and non-IID *mobility behaviors*, which indirectly leads to scenarios with diverse distributions of mobility patterns (IID and non-IID) across the clients.

3.5 Summary

In this chapter, based on how a typical FL4iMP solution works, as illustrated in Fig. 2.2, we identify key factors that may impact its performance. Fig. 3.1 presents how we grouped these factors into four categories, depending on whether they are related to: (1) the *human* mobility behavior; (2) properties of the *data* capturing such behavior and used as input to the FL4iMP solution; (3) the FL *architecture* itself; and (4) the learning *algorithms* employed to build both local and global models.

In the next chapter, we present the datasets used to evaluate (in Chapter 5) each factor category highlighted in this chapter.

Chapter 4

Dataset: Overview and Analysis

As argued in Section 3.2, our study relies on two CDR datasets with different spatial and temporal properties and diverse population characteristics. In this chapter, we first introduce the two datasets (Section 4.1) and how they were processed before being used (Section 4.2). These processed datasets exhibit a reasonable amount of data samples per user and consequently, across clients. This enables the investigation of heterogeneity stemming precisely from user mobility behaviors, rather than from data collection specificity (i.e., from the data sampling rate). Then, we investigate on an individual basis, the mobility behavior captured in both datasets. Our goal is not only to systematically assess whether the datasets exhibit heterogeneous properties (i.e., non-IID properties) but also to explain what aspects of users' mobility patterns make them dissimilar (Sections 4.3 and 4.4). To our knowledge, we are the first to *quantify* user heterogeneous mobility patterns before the FL training phase, i.e., directly from the raw data instead of from local or server model parameters (as in Zhang et al. [69], Park et al. [46], Li et al. [36]). We are also the first to recognize that understanding the trade-offs dictating FL4iMP performance requires looking beyond aggregated results, assessing how FL4iMP performs for users with diverse patterns separately. Next, we identify and characterize different mobility patterns in the two datasets. These results will be used to guide the evaluation of the FL4iMP factors in Chapter 5.

4.1 Datasets

The first dataset, referred to as **Shanghai dataset**, was gathered by a major Telecom operator in the metropolitan area of Shanghai, China. It provides location data for 58,502 users, recorded at a frequency of one location per hour over ten days. This dataset is fully anonymized, with no user identity or corresponding cell tower infrastructure included. Unlike classical CDRs, the recorded location is not the coordinates of the Base Station (BS) to which the user was connected. Instead, it represents the centroid of a cell

(in a grid representation) nearest to the BS the user was primarily connected to during each hour. In sum, the Shanghai dataset contains 10,396 distinct centroid coordinates and 9,135,780 records.

Our second dataset, referred to as **Shenzhen dataset**¹ (Zhang et al. [66]), was collected by several major operators in Shenzhen (China). It includes the location data of 414,271 users collected by 1,090 Base Stations, resulting in 38,218,717 records. This dataset is also fully anonymized, with no user identities provided. Additionally, while records are associated with their collection times, all specific date information was removed, making it impossible to determine if records generated by any two users were collected on the same day. The next section describes how we processed this issue.

4.2 Data pre-processing

We now discuss how both Shanghai and Shenzhen datasets were pre-processed for fair comparison and unbiased analysis. First, to handle the restricted temporal information in the Shenzhen dataset, we assume the starting day d_1 of records is the same for all users in the collection. Time-ordered timestamps belonging to the same user are then associated to day d_x until a timestamp equal or earlier than the previous one is found, which triggers the start of a new day d_{x+1} for that user.

Moreover, to mitigate the temporal gaps common to CDR datasets, we apply a data completion strategy to both datasets, as proposed in Chen et al. [13]. In such strategy, records are conveniently added to complete or fill the gaps, as described next. Completion is done only for short intervals when the devices were not used, which allows for increasing the temporal resolution of the dataset without biasing results analysis. Specifically, we first identify three types of places for each user: *home*, *workplace A*, and *workplace B*. We define as *home* the most frequent daily location in which the user was from 2 a.m. to 6 a.m. Similarly, *workplace A* and *workplace B* are defined as the recurrent daily locations from 10 a.m. to 11 a.m. and 2 p.m. to 4 p.m., respectively. Once each place is identified, if a record is missing during each hourly interval during the mentioned periods, a new record is added with the most frequent location associated with each place type. By doing so, our completion approach aims at a temporal resolution of one hour during the three completed periods.

Next, we employ a 200m $\times$ 200m square tessellation based on *OpenStreetMap* (OpenStreetMap contributors [44]), available in *Scikit-mobility* (Pappalardo et al. [45]), to map the location coordinates (latitude and longitude) into cells, we also replaced the original

¹<https://people.cs.rutgers.edu/dz220/data.html>

coordinates to the associated cell centroid coordinates. Ultimately, we conduct a data filtering that removes inactive users, who could lead to poor mobility behavior identification and distorted evaluation results, given their low number of records. To this end, we first keep per user, one record for each distinct location visited in each 1h interval; then, we filter out users with less than 10 days and 120 records, following the steps of Amichi et al. [2] and Teixeira et al. [59].

The Shanghai dataset comprises 58,471 users and 6,321 cells following the aforementioned data processing. Each user trajectory, on average, consists of 181 records (156 before the completion for the same users) spread over 10 days. In contrast, the processed Shenzhen dataset encompasses 42,266 users and 816 cells. On average, each user has 172 records (115 before completion) covering a period of 16.4 days. Thus, the two datasets exhibit very distinct properties concerning location diversity, temporal coverage, and user population. The Shanghai dataset has greater diversity in terms of unique locations and a larger user population. Also, mobility trajectories in Shanghai typically have more records. Yet, they tend cover shorter time periods, compared to those in Shenzhen. Finally, both processed datasets have the same spatial resolution (i.e., 200m $\times$ 200m cells) with the associated date and hour information. These similar spatial-temporal resolutions allow for a fair comparison and analysis between the two datasets, limiting the diversity of data sampling distribution to better focus on heterogeneity in mobility behaviors.

4.3 User mobility profiling

To capture the human factors discussed in Section 3.1, we perform a profiling of the users in each dataset. To this end, we apply the mobility modeling and profiling strategy described in Amichi et al. [2, 4], yielding a powerful user characterization according to both routine and novelty components in their daily trajectories. The strategy consists in grouping users in the following three classes, whose temporal and spatial patterns will be analyzed in the next section:

- **Routiners:** defines the set of users that have the *routine* as the most prevalent component in their daily mobility life. Although moments of exploration may also exist, users in this profile spend most of their time revisiting few previously visited locations with long periods of stationarity.
- **Scouters:** defines the set of users more prone to explore new areas, making the *novelty* the most prevalent component in their mobility pattern. Contrarily to the

Routiner profile, users in this profile very often break their routine while seeking for novelty, resulting in mobility patterns that are more diverse and less Regular.

- **Regulars:** users in this profile have a mobility pattern well balanced between routine and novelty moments, with no stronger predominance of one of them.

To create these profiles, we first employ the classification method named *Visitation-Frequency-Based Identification*, as described in Amichi et al. [4], to categorize each location appearing in a user trajectory as either an exploration (i.e., a new or rarely visited location) or a return (i.e., a regularly visited location). The method first computes the importance of each location as the frequency of its appearance in the user’s mobility trajectory. A location is classified as a *return* if it appears more than a level $l\%$ of the average importance of all locations visited by the user; otherwise, it is classified as an *exploration*. In this work, we experimented with various values of l and set $l = 80\%$, allowing for a more precise exploration location distinction.

Once each location is classified, we calculate two features for each user: *Intermittency* and *Degree of Return*. *Intermittency* measures the frequency of shifts between explorations and returns. Higher values indicate fewer shifts, while lower values indicate frequent shifts, revealing whether (i) an individual visits several exploration-like (vs. return-like) locations in a row before revisiting a return-like (vs. exploration-like) one or (ii) is versatile, keep frequently switching between exploration- and return-like locations. The *Degree of Return* feature indicates how often the user revisits a return-like location after an exploration-like one. Higher values suggest that the average number of successive explorations is low, due to the user’s preference for returning to regularly visited locations.

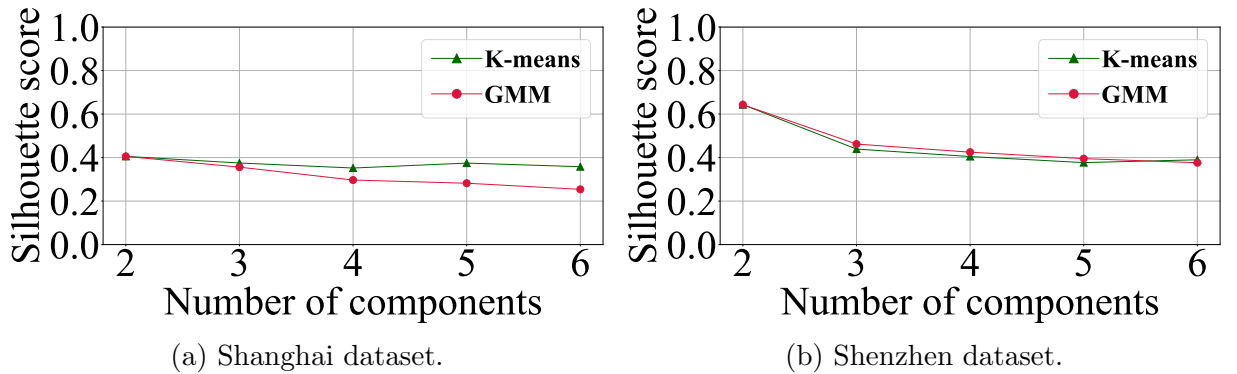


Figure 4.1: *Silhouette* scores per dataset.

Computed values of *Intermittency* and *Degree of Return* are then scaled from 0 to 1 via a min-max normalization, i.e., $x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$. Finally, we cluster the users using the two normalized features. As in Amichi et al. [2, 4], we experimented with different clustering algorithms (*Gaussian Mixture probabilistic Model* (GMM) and *k-means*) and variable number of clusters, using the *Silhouette* score (Rousseeuw [50]) as metric of

clustering quality. As depicted by Fig. 4.1, the best results were obtained with the *k-means* algorithm, which split the users into the three aforementioned groups/profiles. Fig. 4.2 shows clustering results: each dot represents a user, colored according to the uncovered profile. In a nutshell, we observe the relationship between the profile definitions and the features (axes): *Scouters* and *Routiners* are on opposite sides and *Regulars* fall between them.

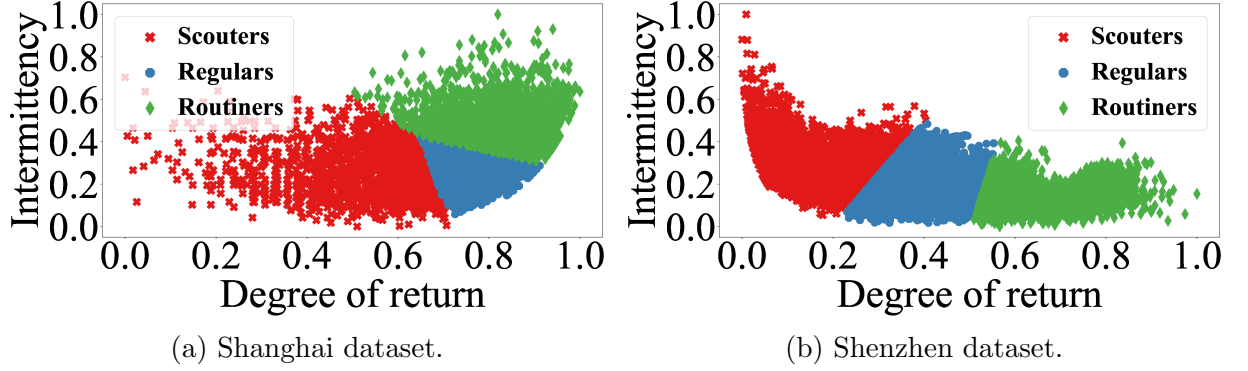


Figure 4.2: Profiling of users' mobility behaviors per dataset.

4.4 Data analysis

As presented in Section 4.2, the two datasets are very diverse with respect to user population, unique locations, and time periods. We now explore these differences by analyzing the user trajectories in each dataset using various mobility-related features. Our characterization considers all users and users in each identified profile, aiming to highlight their main differences and provide insights to help us understand the FL results presented in Chapter 5.

Our analysis relies on the following five features: (1) **Regularity** (Teixeira et al. [58, 59]) captures if an individual is prone to return to previously visited locations; (2) **Stationarity** (Teixeira et al. [58, 59]) quantifies if an individual stays continuously in the same location; (3) **Diversity** (Teixeira et al. [59]) measures how diverse an individual's visitation pattern is; (4) **Entropy** (Song et al. [55]) provides a complementary view to Diversity, and builds upon the visitation frequency, the order in which the locations were visited, and the time spent at each location; and (5) **Radius of gyration** (Gonzalez et al. [25]) measures whether a user has a confined mobility pattern (smaller values) concentrated in a specific area, or if she tends to travel longer distances (bigger values).

Regularity, *Stationarity* and *Diversity* are measured in the 0-1 scale: the closer the value is to 1, the more likely the individual is to exhibit the described behavior. Higher

values of *Entropy* suggest harder to predict users' mobility. Typically, trajectories with high entropy tend to have greater Diversity, lower Regularity, and Stationarity.

Mobility profile	Number of users	Number of records	Number of unique locations visited
Scouters	13,728 (24%)	179.23 ± 0.07	19.21 ± 0.14
Regulars	28,679 (49%)	181.36 ± 0.05	11.38 ± 0.07
Routiners	16,064 (27%)	182.84 ± 0.07	8.72 ± 0.09
All users	58,471 (100%)	181.27 ± 0.04	12.49 ± 0.06

(a) Shanghai dataset.

Mobility profile	Number of users	Number of records	Number of unique locations visited
Scouters	20,447 (48%)	153.95 ± 0.45	8.89 ± 0.10
Regulars	11,716 (28%)	159.39 ± 1.12	8.17 ± 0.14
Routiners	10,103 (24%)	223.33 ± 1.23	5.56 ± 0.09
All users	42,266 (100%)	172.04 ± 0.55	7.90 ± 0.07

(b) Shenzhen dataset.

Table 4.1: Average (with 95% C.I.) number of users, records, and unique locations visited per user on each profile for the two datasets.

Table 4.1 presents the distribution of users across the three profiles for each dataset, along with the corresponding average numbers of records and unique locations per user trajectory. Figs. 4.3 and 4.4 show the distributions of the five aforementioned features' values for all users, as well as for users in each profile.

Shanghai dataset: As shown in the table, almost half of the population is dominated by users who often alternate between routine and novelty in their mobility patterns, i.e., *Regulars* (49%). The remaining ones are roughly balanced and split between *Scouters* (24%), who are more prone to novelty seeking, and *Routiners* (27%), who have a prevalent routine component. Moreover, even though users in all three profiles tend to have a similar number of records on average, *Scouters* have a much larger number of unique locations – more than twice that of *Routiners*.

Figs. 4.3a and 4.3b show that *Routiners* tend to have the highest values of *Regularity* and *Stationarity*, suggesting that these users are not only more prone to return to previously visited places (Fig. 4.3a) but also tend to stay longer in the same places (Fig. 4.3b). Consistently, they tend to have the lowest values of *Diversity* (Fig. 4.3c) and *Entropy* (Fig. 4.3d). On the other extreme, *Scouters* have the highest values of *Diversity* and *Entropy* (and lowest *Regularity* and *Stationarity*), characterizing them as exploratory users with less predictable mobility trajectories. Finally, we notice that the Shanghai dataset is spatially wide, with some users having a *Radius of gyration* close to 45 km (Fig. 4.3e).

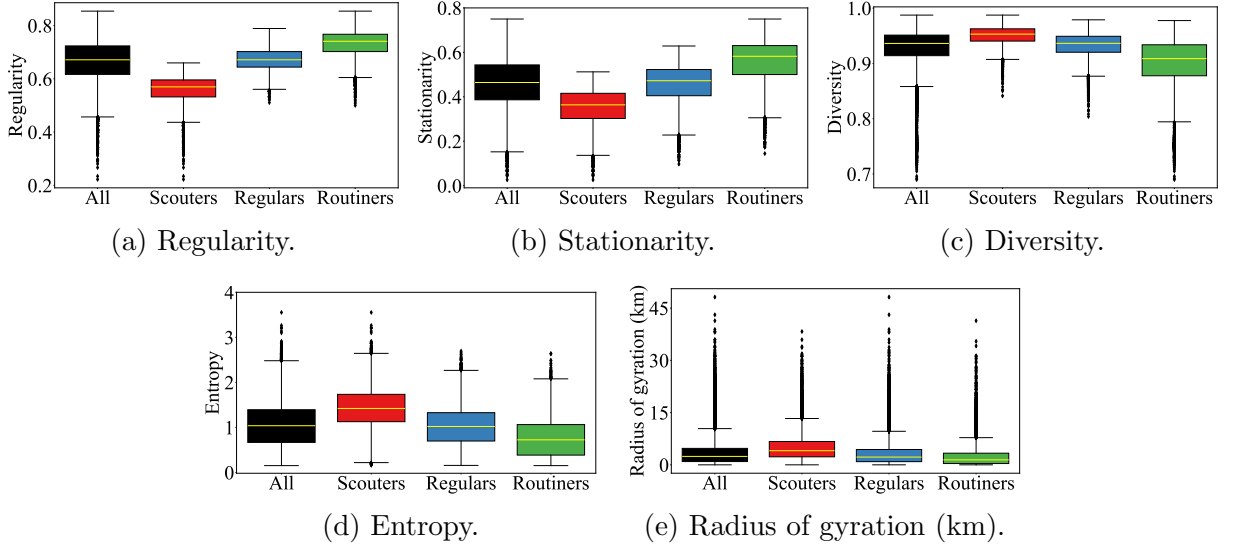


Figure 4.3: Mobility behavior features' analysis for the Shanghai dataset.

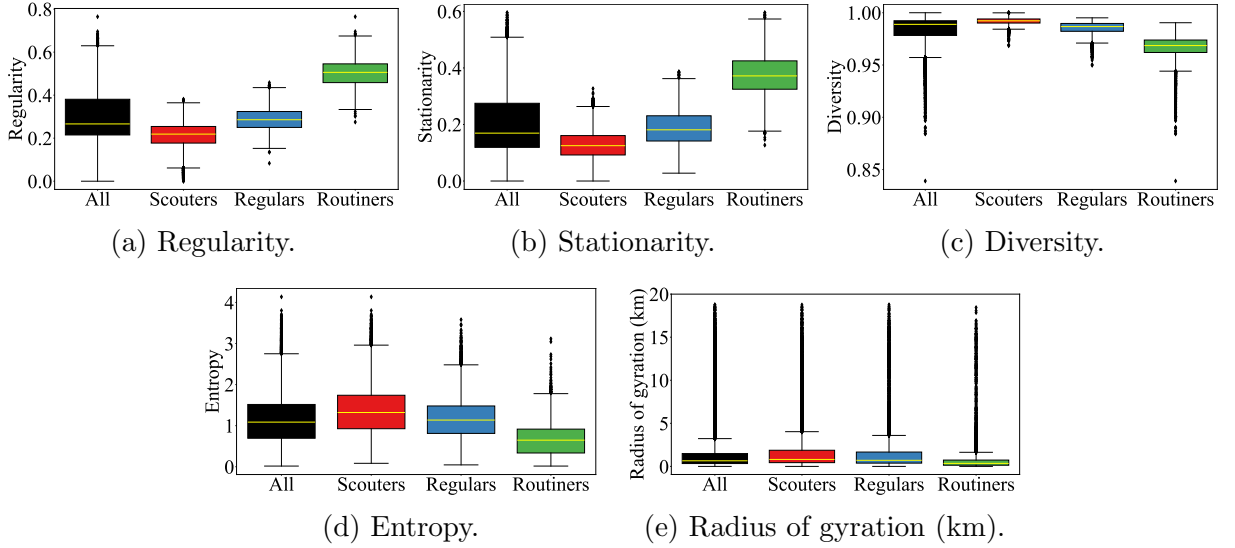


Figure 4.4: Mobility behavior features' analysis for the Shenzhen dataset.

Shenzhen dataset: Unlike the Shanghai dataset, the population in the Shenzhen dataset is dominated by users more prone to explore, i.e., *Scouters* (48%). Also, again, unlike in the Shanghai dataset, the number of records per user trajectory varies greatly across the profiles, with *Routiners* having up to 45% more records on average than *Scouters*. Regarding the number of unique locations, we find that, consistent with the observations made for the whole dataset (Section 4.2), i.e., user trajectories in the Shenzhen dataset are much less diversified. The distributions of *Regularity*, *Stationarity*, *Diversity*, and *Entropy* (Figs. 4.4a, 4.4b, 4.4c, 4.4d) show qualitatively similar patterns for the three profiles in the Shenzhen dataset as observed for the Shanghai one. Finally, compared to the Shanghai dataset, the Shenzhen dataset is much more spatially constrained (which explains the fewer number of unique locations), with *Radius of gyration* always lower than

20 km (Fig. 4.4e).

4.5 Summary

In this chapter, we showed that the mobility behaviors (captured by *Regularity*, *Stationarity*, *Diversity* and *entropy*) of *Scouters*, *Regulars*, and *Routiners* are quite diverse, consistent with their definition, in both Shanghai and Shenzhen. However, the datasets highly differ in terms of the number of unique locations visited, the radius of gyration, and the number of records per trajectory. This discrepancy may facilitate prediction on the Shenzhen dataset, where user trajectories tend to have fewer distinct locations and be more confined, but with increased user history (i.e., data available for training the models). Regarding profiles, the entropy of *Scouters* (cf. *Routiners*) in both datasets is the highest (cf. lowest) among the three profiles, which will make individual mobility prediction more (cf. less) challenging.

Also, the results demonstrate the datasets' diversity in spatial-temporal features and population heterogeneity, i.e., in terms of data and human factors (cf. Section 3). These assets provide the necessary foundation for studying the impact of mobility pattern heterogeneity on FL4iMP from a human behavior perspective. Besides, our analysis reveals the features that contribute to population mobility behavior heterogeneity, presenting three distinct profiles of users. This clear differentiation among profiles and their features offers a robust basis for challenging mobility prediction and most importantly, for interpreting the impact of mobility heterogeneity (i.e., non-IID data) on FL4iMP results.

In the next chapter, we evaluate the impact of each factor presented in Chapter 3, using the two datasets detailed in this chapter.

Chapter 5

Evaluation

In this chapter, we dive into our evaluation of the impact of the various factors introduced in Chapter 3 on FL4iMP. Using the two datasets detailed in Chapter 4, we leverage the user profiles identified and characterized therein to capture human and data-related factors. We begin by outlining our experimental setup and explaining how we accounted for algorithm and architecture factors (Section 5.1). We then present our key findings, initially focusing on the human and data-related factors (Section 5.2), and subsequently on the factors related to the FL algorithms and architecture (Sections 5.3, 5.4 and 5.5).

5.1 Experimental methodology

FL architecture: To establish a unified environment for studying all identified factors, we used *Flower* (Beutel et al. [9]), a state-of-the-art framework for FL. We configured *Flower* to distribute all users in the input dataset uniformly across all clients. We assume all clients participate in every round. This approach aims to minimize variability from diverse client selections, allowing us to focus on the number of clients (i.e., local models) as the primary architectural factor.

Local learning algorithms: Regarding the algorithm used to learn the local models, we chose to focus on two FL4iMP solutions based on more traditional and less costly neural network algorithms (Cunha et al. [15]): *FL-Flashback* and *GRU-Spatial* FL models (Ezequiel et al. [20]). These solutions rely on two distinct neural networks, namely RNN and GRU, respectively, as algorithm to learn the local models. We used the implementations of both *FL-Flashback* and *GRU-Spatial* developed in *Flower* and made available by the authors of Ezequiel et al. [20].

Aggregation algorithm: We consider the *FedAvg* (McMahan et al. [42]) and the *FedProx* (Li et al. [36]) aggregation strategies to build a global model, both available in *Flower*. Similarly to the *FedProx* discussion in Li et al. [36], our analysis using the two datasets showed that the hyper-parameter $\mu = 1$ produces the best results, while

smaller values closer to 0 yield similar results to *FedAvg*. Higher μ values (e.g., 1000) also produced results comparable to those for $\mu = 1$. Thus, we set μ to 1. Unless otherwise noted, *FedAvg* was considered as the aggregation algorithm.

Parameterization: The hyper-parameters for the *FL-Flashback* and *GRU-Spatial* local models were set following their original configuration in Ezequiel et al. [20]. Notably, the input trajectory length was set to 20. The number of epochs in the local model and the communication rounds between clients and remote servers were set to 10 (unless otherwise noted). Batch size, embedding size, and learning rate were set to 128, 10, and 0.001, respectively.

Hardware setup: Our experiments were executed in the following machine’s configuration: NVIDIA GPU GeForce RTX 3090 (24,576 MB of VRAM), AMD Ryzen 5 5600X processor (6 cores / 12 threads) @ 3.7 GHz and 64 GB RAM (4x 16 GB DDR4) @ 2133 MT/s. As previously observed in Ezequiel et al. [20] and Proteasa et al. [47], GPU memory (VRAM) quickly becomes a bottleneck as the number of clients increases. Indeed, prior studies were restricted to at most 4 (Ezequiel et al. [20]) or 10 (Proteasa et al. [47]) clients for their scenarios. We were able to reach as many as 17 clients in ours.

Evaluation scenarios: To isolate and assess the impact of the studied factors as much as possible, we build several different evaluation scenarios where all users either have the same profile (cf. Section 4.3), representing an IID data, or have mixed profiles, representing a non-IID case. We also varied the number of clients and the number of users per client. These scenarios will be introduced in the following sections, along with the corresponding results.

Training and test sets: For each scenario, defined by a set of users, we split the data into training and test sets as follows. We take the initial 80% of the records of all user trajectories as training set, which is used to learn the prediction model. The final 20% of the records of all trajectories are used as test set to evaluate it. We ran each scenario 5 times with random initial model weights.

Evaluation metrics: Our evaluation considers metrics of both prediction *effectiveness* and *efficiency*.

To assess effectiveness, we follow prior work (Feng et al. [23], Fan et al. [21], Li et al. [35], Wang et al. [62], Ezequiel et al. [20], Zhang et al. [69], Park et al. [46]) and measure the number of times the correct location is among the top- k predicted locations (for $k = 1$ and 5), referred to as *accuracy in the top- k* or $Acc@k$. In addition to being reported by prior work, $Acc@5$ is a relevant metric in the context of recommendation systems (Ko et al. [33]).

For efficiency, we measure the consumption of energy (in watts) and VRAM (in megabytes) used by the clients on the GPU, as well as the training (per communication round) and total execution times (in seconds). While the amount of data impacts all efficiency metrics, energy and VRAM usages are particularly sensitive to the complexity

of users’ mobility behaviors (i.e., of the processed data). We developed a Python script that measures the energy (wattage) and the used GPU VRAM every second. Efficiency metrics are aggregated across all clients, whereas $Acc@k$ is measured for each client and then aggregated. We report **average** results computed across 5 runs along with 95% confidence intervals for all metrics, i.e., total energy consumption has to consider the evaluation total execution time.

5.2 Impact of human and data factors

We first focus on human and data factors, guided by the discussions in Section 4. To assess the impact of user mobility behavior on FL4iMP, we build various scenarios as described in Table 5.1, sampling users from both the Shanghai and Shenzhen datasets. To avoid overwhelming the discussion with excessive results, we concentrate on setups using *GRU-Spatial* as the local algorithm and *FedAvg* as the aggregation algorithm, with four clients. Yet, the qualitative conclusions remain similar across different configurations. Additionally, the input data is uniformly distributed across all four clients. Lets first consider the scenarios in rows 1-7 of Table 5.1, which, to control for population size, consists of samples of 6,000 users with diverse mobility patterns. For such scenarios, the average FL4iMP results are shown in Tables 5.2 and 5.3 for the Shanghai and Shenzhen datasets, respectively.

Scenario	Description
1-Random Scouters (6K)	6,000 users randomly selected from all Scouters.
2-Random Regulars (6K)	6,000 users randomly selected from all Regulars.
3-Random Routiners (6K)	6,000 users randomly selected from all Routiners.
4-Super-Scouters (6K)	6,000 users with the highest values of entropy, being those with the hardest to predict mobility patterns.
5-Super-Routiners (6K)	6,000 users with lowest entropy (easiest to predict).
6-Equal mix (6K)	6,000 users, with 2,000 users that are randomly sampled from each profile.
7-Original mix (6K)	6,000 users randomly sampled with the distribution of user profiles kept the same as in the original data.
8-Original mix (20K)	20,000 users randomly sampled with the distribution of user profiles kept the same as in the original data.
9-Original mix (all)	Full original dataset.

Table 5.1: **Scenarios for human and data factors analysis.** Profile distributions of Scenarios 7-9: For Shanghai: 24% Scouters, 49% Regulars and 27% Routiners. For Shenzhen: 48% Scouters, 28% Regulars and 24% Routiners.

Effectiveness (tables 5.2a and 5.3a): Results vary greatly across the three user profiles (rows 1-3 in the tables) and even more for the two extreme scenarios of *4-Super Scouters (6K)* and *4-Super Routiners (6K)* (rows 4-5). In the Shanghai dataset, the average *Acc@1* reaches as high as 0.805 for *5-Super Routiners (6K)*, i.e., those with the easiest-to-predict mobility patterns, and as low as 0.343 for the hardest-to-predict *4-Super Scouters (6K)*. The gap is even larger in the Shenzhen dataset, due to an intricate combination of more extreme values of *Regularity*, *Stationarity*, and *Diversity* compared to those of Shanghai (see Figs. 4.3-4.4), and a smaller number of unique locations visited (see Table 4.1). As argued in Section 4.4, the Shenzhen dataset is much more spatially constrained, which naturally leads to simpler mobility patterns that are easier to learn and predict.

Scenario	Acc@1	Acc@5
1-Random Scouters	0.459 $\pm$ 0.00	0.651 $\pm$ 0.00
2-Random Regulars	0.542 $\pm$ 0.00	0.731 $\pm$ 0.00
3-Random Routiners	0.628 $\pm$ 0.00	0.780 $\pm$ 0.00
4-Super Scouters	0.343 $\pm$ 0.00	0.578 $\pm$ 0.00
5-Super Routiners	0.805 $\pm$ 0.00	0.863 $\pm$ 0.00
6-Equal mix	0.534 $\pm$ 0.00	0.710 $\pm$ 0.00
7-Original mix	0.542 $\pm$ 0.00	0.724 $\pm$ 0.00

(a) Accuracy.

Scenario	Energy (W)	VRAM (MB)
1-Random Scouters	201.5 $\pm$ 1.35	6845 $\pm$ 3
2-Random Regulars	200.8 $\pm$ 0.86	6439 $\pm$ 1
3-Random Routiners	200.2 $\pm$ 0.76	6402 $\pm$ 1
4-Super Scouters	202.3 $\pm$ 1.03	6558 $\pm$ 2
5-Super Routiners	196.5 $\pm$ 0.76	6226 $\pm$ 2
6-Equal mix	200.2 $\pm$ 0.65	6766 $\pm$ 3
7-Original mix	199.9 $\pm$ 1.14	6479 $\pm$ 1

(b) Resource.

Scenario	Train (s)	Total
1-Random Scouters	15.5 $\pm$ 0.12	516.6 $\pm$ 0.89 (08m 36s)
2-Random Regulars	15.0 $\pm$ 0.04	511.4 $\pm$ 0.70 (08m 31s)
3-Random Routiners	14.9 $\pm$ 0.09	510.6 $\pm$ 0.70 (08m 30s)
4-Super Scouters	15.6 $\pm$ 0.00	517.6 $\pm$ 0.43 (08m 37s)
5-Super Routiners	14.2 $\pm$ 0.04	504.2 $\pm$ 0.66 (08m 24s)
6-Equal mix	15.4 $\pm$ 0.11	516.0 $\pm$ 0.96 (08m 36s)
7-Original mix	14.9 $\pm$ 0.12	510.0 $\pm$ 1.11 (08m 30s)

(c) Time.

Table 5.2: **Impact of the user mobility behaviors for the *Shanghai* dataset.** Average (with 95% C.I.) results for GRU-Spatial, FedAvg, 6,000 users and 4 clients.

Indeed, the results for the two mixed scenarios (rows 6-7 of Tables 5.2 and 5.3), which hardly reflect those for specific profiles (especially *Routiners* and *Scouters*), illustrate how average results for a heterogeneous user population, as reported in prior studies, can be misleading concerning the performance perceived by individual users (or user profiles). Such results vary somewhat depending on the distribution of profiles, being lower for cases where *Scouters*, who typically exhibit much more complex mobility patterns, are a smaller fraction (e.g., *6-Equal mix (6K)* in Shanghai).

Scenario	Acc@1	Acc@5
1-Random Scouters	0.504 $\pm$ 0.00	0.874 $\pm$ 0.00
2-Random Regulars	0.560 $\pm$ 0.00	0.906 $\pm$ 0.00
3-Random Routiners	0.730 $\pm$ 0.00	0.973 $\pm$ 0.00
4-Super Scouters	0.297 $\pm$ 0.00	0.745 $\pm$ 0.00
5-Super Routiners	0.867 $\pm$ 0.00	0.988 $\pm$ 0.00
6-Equal mix	0.619 $\pm$ 0.00	0.929 $\pm$ 0.00
7-Original mix	0.592 $\pm$ 0.00	0.917 $\pm$ 0.00

(a) Accuracy.

Scenario	Energy (W)	VRAM (MB)
1-Random Scouters	181.6 $\pm$ 0.62	5663 $\pm$ 2
2-Random Regulars	181.6 $\pm$ 0.82	5599 $\pm$ 2
3-Random Routiners	181.5 $\pm$ 0.59	5600 $\pm$ 1
4-Super Scouters	181.3 $\pm$ 0.92	5600 $\pm$ 1
5-Super Routiners	181.3 $\pm$ 1.34	5599 $\pm$ 1
6-Equal mix	180.9 $\pm$ 0.71	5665 $\pm$ 2
7-Original mix	180.8 $\pm$ 1.04	5599 $\pm$ 1

(b) Resource.

Scenario	Train (s)	Total
1-Random Scouters	10.0 $\pm$ 0.04	511.2 $\pm$ 1.02 (08m 31s)
2-Random Regulars	10.5 $\pm$ 0.09	532.2 $\pm$ 1.02 (08m 52s)
3-Random Routiners	14.9 $\pm$ 0.11	831.2 $\pm$ 1.70 (13m 51s)
4-Super Scouters	11.2 $\pm$ 0.13	570.6 $\pm$ 1.42 (09m 30s)
5-Super Routiners	16.3 $\pm$ 0.07	811.2 $\pm$ 1.02 (13m 31s)
6-Equal mix	11.8 $\pm$ 0.12	617.4 $\pm$ 0.89 (10m 17s)
7-Original mix	11.2 $\pm$ 0.11	579.6 $\pm$ 1.42 (09m 39s)

(c) Time.

Table 5.3: **Impact of the user mobility behaviors for the *Shenzhen* dataset.** Average (with 95% C.I.) results for GRU-Spatial, FedAvg, 6,000 users and 4 clients.

Efficiency (tables 5.2b, 5.2c, 5.3b and 5.3c): Most results are quite similar across all scenarios for either dataset. Exceptions are the training and total execution times in the Shenzhen dataset, which are significantly longer for the *3-Random* and *5-Super Routiners (6K)* (row 3 and 5). This is due to the much larger number of records per

trajectory for those users ($\approx 45\%$ higher in Table 4.1). In contrast, the number of records per trajectory in Shanghai is roughly the same for all profiles, leading to similar execution times in all phases. VRAM and energy usage are somewhat lower for *5-Super Routiners (6K)* in Shanghai, possibly due to a combination of profile simplicity in pattern prediction, fewer unique locations, and a roughly similar number of records per user trajectory (see Table 4.1).

Population size: We now evaluate the impact of increasing the population size. We do so by considering the last two scenarios of Table 5.1, where the profile distribution is maintained as in the original data, but the sample is increased to 20,000 users (*8-Original mix (20K)*) and the full dataset (*9-Original mix (all)*). Tables 5.4 and 5.5 present the results for these scenarios for both datasets. These results should be compared to those in row 7 of Tables 5.2 and 5.3 (i.e., *7-Original mix (6K)*), for which the population is kept at 6,000 users.

Scenario	Acc@1	Acc@5
8-Original mix (20K)	0.629 ± 0.00	0.916 ± 0.00
9-Original mix (all)	0.656 ± 0.00	0.955 ± 0.00

(a) Accuracy.

Scenario	Energy (W)	VRAM (MB)
8-Original mix (20K)	206.3 ± 0.51	6600 ± 1
9-Original mix (all)	209.2 ± 0.45	6687 ± 1

(b) Resource.

Scenario	Train (s)	Total
8-Original mix (20K)	52.8 ± 0.10	1811.2 ± 1.61 (30m 11s)
9-Original mix (all)	157.4 ± 0.16	5319.4 ± 7.83 (1h 28m 39s)

(c) Time.

Table 5.4: **Impact of the user population size for the *Shanghai* dataset.**

Average (with 95% C.I.) results for GRU-Spatial, FedAvg, 20,000 and 58,471 users (respectively) and 4 clients.

Since the number of clients is fixed at 4, a larger population size results in more data (i.e., user trajectories) available for training the local models, thereby enhancing each client’s training process. We observe that the execution times naturally increase with population size in both datasets as more data leads to more processing. Interestingly, the increases are only marginal regarding energy and VRAM consumption, suggesting the strong correlation of these two metrics with data complexity rather than data size. Indeed, since the profile distribution is the same in all scenarios of Tables 5.4 and 5.5, as well as in the last scenario of Tables 5.2 and 5.3, the increase of population size (from 6k, to 20k, and to 58k (Shanghai) / 42k (Shenzhen)) does not significantly increase the complexity in learning and predicting. Still, learning greatly benefits from more data,

Scenario	Acc@1	Acc@5
8-Original mix (20K)	0.601 $\pm$ 0.00	0.939 $\pm$ 0.00
9-Original mix (all)	0.607 $\pm$ 0.00	0.942 $\pm$ 0.00

(a) Accuracy.

Scenario	Energy (W)	VRAM (MB)
8-Original mix (20K)	185.0 $\pm$ 0.66	5609 $\pm$ 1
9-Original mix (all)	185.0 $\pm$ 0.67	5672 $\pm$ 1

(b) Resource.

Scenario	Train (s)	Total
8-Original mix (20K)	37.8 $\pm$ 0.13	2004.6 $\pm$ 1.89 (33m 24s)
9-Original mix (all)	80.9 $\pm$ 0.23	4286.4 $\pm$ 8.03 (1h 11m 26s)

(c) Time.

Table 5.5: **Impact of the user population size for the *Shenzhen* dataset.**

Average (with 95% C.I.) results for GRU-Spatial, FedAvg, 20,000 and 42,266 users (respectively) and 4 clients.

justifying the slight increase in energy and VRAM. When comparing the two datasets, a much higher number of *unique* locations in Shanghai (see Table 4.1), naturally leads to more complex mobility patterns compared to Shenzhen, what is reflected in lower energy and VRAM consumption in this latter dataset.

Summary: FL4iMP performance is a multi-faceted problem where the final result depends, in non-trivial ways, on the combination of human behavior and data properties. In particular it can be very sensitive to user mobility patterns, highlighting the need to take the inherent heterogeneity of human behavior into account when analyzing alternative solutions.

In terms of accuracy, our results present higher accuracy values when compared to the literature, given the usage of CDR data instead of LBSN data (as discussed in Section 3.2). Also, the presented *Acc@1* values (~ 0.6 in both Shanghai and Shenzhen datasets, i.e., *9-Original mix (all)*) are reasonable for applications that are not mission/safety critical, i.e., applications that do not have critical efficiency requirements, e.g., recommendation systems (Ko et al. [33]).

5.3 Impact of number of clients

We now assess the impact of a key architectural parameter, namely, the number of clients. To that end, we keep the population size fixed at 6,000 users, i.e., *7-Original mix (6K)* scenario (see Table 5.1), and use *GRU-Spatial* and *FedAvg* as the local learning algorithm and aggregation strategy, respectively. We varied the number of clients from 2 to the maximum possible in our setup (before running out of VRAM), which was 15 for the Shanghai and 17 for the Shenzhen datasets, respectively. Results are shown in Tables 5.6 and 5.7 (results for 4 clients are the same as those in Tables 5.2 and 5.3 for the same scenario).

Since the population is kept fixed, increasing the number of clients results in a reduction of the amount of data available for learning the local models. As shown, accuracy drops significantly as the number of clients increases, especially for the Shanghai dataset, which has trajectories with more distinct locations. The reduced data available to each client challenges the effective learning of the more complex mobility patterns.

Resource consumption naturally increases with the number of clients due to contention on the shared hardware. Conversely, execution time, especially for local model training, decreases as the amount of data decreases. Interestingly, we see a change of pattern from 10 to 15 clients. This occurs because, in this case, the number of CPU threads available (12) is exceeded. The CPU is responsible for loading the data into the GPU. Thus, resource contention arises as the number of CPU threads becomes scarce to feed all clients, leading to extra delays. Energy consumption, in turn, reduces as the GPU may occasionally become idle, waiting for the CPU. Such idle periods also contribute to increasing execution times.

5.4 Impact of the local algorithm

All the results discussed so far were obtained by using *GRU-Spatial* as the learning algorithm at the clients. We now focus on this specific factor and compare *GRU-Spatial* against *FL-Flashback*. In this analysis, we consider a population of 6,000 users and 4 clients and, once again, use *FedAvg* to aggregate learned local models into a global model.

As discussed in Section 2.2.3, these algorithms' accuracy and training times were analyzed using LBSN datasets in Ezequiel et al. [20]. The reported results suggest that in terms of accuracy, there is no clear winner. Yet, unlike *GRU-Spatial*, *FL-Flashback*

Number of clients	Acc@1	Acc@5
2 clients	0.614 $\pm$ 0.00	0.864 $\pm$ 0.01
4 clients	0.542 $\pm$ 0.00	0.724 $\pm$ 0.00
10 clients	0.391 $\pm$ 0.00	0.539 $\pm$ 0.00
15 clients	0.306 $\pm$ 0.00	0.441 $\pm$ 0.00

(a) Accuracy.

Number of clients	Energy (W)	VRAM (MB)
2 clients	191.0 $\pm$ 0.88	3254 $\pm$ 1
4 clients	199.9 $\pm$ 1.14	6479 $\pm$ 1
10 clients	205.4 $\pm$ 0.28	16161 $\pm$ 3
15 clients	201.6 $\pm$ 0.82	24225 $\pm$ 7

(b) Resource.

Number of clients	Train (s)	Total
2 clients	18.0 $\pm$ 0.07	571.2 $\pm$ 2.03 (09m 31s)
4 clients	14.9 $\pm$ 0.12	510.0 $\pm$ 1.11 (08m 30s)
10 clients	14.1 $\pm$ 0.04	473.0 $\pm$ 0.55 (07m 53s)
15 clients	13.7 $\pm$ 0.14	511.2 $\pm$ 0.35 (08m 31s)

(c) Time.

Table 5.6: **Impact of the number of clients for the *Shanghai* dataset.**

Average (with 95% C.I.) results for GRU-Spatial, FedAvg, Scenario 7-Original mix (6K).

captures both temporal and spatial contexts, which yields a much larger model, thus requires much longer training times.

Our results are shown in Tables 5.8 and 5.9. For ease of comparison, the *GRU-Spatial* results are repeated from Tables 5.2 and 5.3, which correspond to the same scenarios. In terms of accuracy, *GRU-Spatial* consistently outperforms *FL-Flashback* in almost all cases. This superior performance can be attributed to the gating mechanisms within every GRU cell, which allow the model to learn which data in a sequence is important to retain and which data should be discarded or ignored (Luca et al. [40]). However, the performance gap between the algorithms varies depending on the complexity of the mobility profiles in the input data. For example, for the *4-Super Scouters (6K)* scenario, which consists of users with the harder-to-predict profile, the improvements in *Acc@1* are around 29% and 21% for the Shanghai and Shenzhen datasets, respectively. For the simpler *5-Super Routiners (6K)* scenario, *GRU-Spatial* yields a 94% improvement in *Acc@1* over *FL-Flashback* in Shanghai. The only case where both algorithms deliver similar performance (with a slight advantage to *FL-Flashback*) is the *5-Super Routiners (6K)* in the Shenzhen dataset. Users in this scenario exhibit very simple (i.e., easy-to-predict) mobility trajectories, which are much easier than the same profile in Shanghai. Indeed, the *5-Super Routiners (6K)* scenario in Shenzhen has only 540 distinct locations, compared to

Number of clients	Acc@1	Acc@5
2 clients	0.593 $\pm$ 0.00	0.935 $\pm$ 0.00
4 clients	0.592 $\pm$ 0.00	0.917 $\pm$ 0.00
10 clients	0.584 $\pm$ 0.00	0.843 $\pm$ 0.00
15 clients	0.570 $\pm$ 0.00	0.813 $\pm$ 0.00
17 clients	0.563 $\pm$ 0.00	0.802 $\pm$ 0.00

(a) Accuracy.

Number of clients	Energy (W)	VRAM (MB)
2 clients	172.3 $\pm$ 1.11	2812 $\pm$ 1
4 clients	180.8 $\pm$ 1.04	5599 $\pm$ 1
10 clients	183.7 $\pm$ 0.83	13964 $\pm$ 4
15 clients	182.5 $\pm$ 1.07	20931 $\pm$ 3
17 clients	181.2 $\pm$ 0.44	23714 $\pm$ 1

(b) Resource.

Number of clients	Train (s)	Total
2 clients	16.8 $\pm$ 0.12	665.4 $\pm$ 1.80 (11m 05s)
4 clients	11.2 $\pm$ 0.11	579.6 $\pm$ 1.42 (09m 39s)
10 clients	9.6 $\pm$ 0.04	559.8 $\pm$ 0.35 (09m 19s)
15 clients	10.6 $\pm$ 0.09	555.2 $\pm$ 0.66 (09m 15s)
17 clients	10.9 $\pm$ 0.14	571.2 $\pm$ 0.66 (09m 31s)

(c) Time.

Table 5.7: **Impact of the number of clients for the *Shenzhen* dataset.**

Average (with 95% C.I.) results for GRU-Spatial, FedAvg, Scenario 7-Original mix (6K).

3,919 in Shanghai. This suggests that the Shenzhen sample has much more redundancy, making it easier for both algorithms to perform reasonably well.

Regarding training times, our results agree with those reported in Ezequiel et al. [20]: *GRU-Spatial* has much faster training times, being also 5-times smaller (i.e., having about 5-times less trainable parameters) than Flashback model. Interestingly, despite having more parameters, we observe that *FL-Flashback* consistently consumes less energy and VRAM compared to *GRU-Spatial*. The reason for this lower resource usage might be due to the differences in how the two models handle computational resources. *Flashback*, although having more parameters, appears to be designed in a way that is more computationally efficient, possibly optimizing its operations to reduce resource usage. In contrast, GRU-Spatial, with its efficient architecture and specialized gating mechanisms, lead to higher computational demands to manage and process the data. These gating mechanisms allow GRU-Spatial to effectively learn which parts of the input data are important, improving prediction accuracy. However, this process is computationally intensive, leading to higher energy consumption and VRAM usage during training and inference. Thus, while GRU-Spatial is more accurate and fast, the trade-off is higher in

Scenario	Local algorithm	Acc@1	Acc@5
4-Super Scouters	FL-Flashback	0.266 ± 0.00	0.453 ± 0.00
	GRU-Spatial	0.343 ± 0.00	0.578 ± 0.00
5-Super Routiners	FL-Flashback	0.415 ± 0.00	0.607 ± 0.00
	GRU-Spatial	0.805 ± 0.00	0.863 ± 0.00

(a) Accuracy.

Scenario	Local algorithm	Energy (W)	VRAM (MB)
4-Super Scouters	FL-Flashback	178.5 ± 0.88	6249 ± 1
	GRU-Spatial	202.3 ± 1.03	6558 ± 2
5-Super Routiners	FL-Flashback	176.3 ± 1.30	5991 ± 2
	GRU-Spatial	196.5 ± 0.76	6226 ± 2

(b) Resource.

Scenario	Local Algorithm	Train (s)	Total
4-Super Scts.	FL-Flashback	76.0 ± 0.37	1136.0 ± 4.00 (18m 56s)
	GRU-Spatial	15.6 ± 0.00	517.6 ± 0.43 (08m 37s)
5-Super Rots.	FL-Flashback	76.3 ± 0.16	1137.0 ± 1.47 (18m 57s)
	GRU-Spatial	14.2 ± 0.04	504.2 ± 0.66 (08m 24s)

(c) Time.

Table 5.8: Impact of the local algorithm for the *Shanghai* dataset. Average (with 95% C.I.) results for FedAvg, 6,000 users and 4 clients.

Scenario	Local algorithm	Acc@1	Acc@5
4-Super Scouters	FL-Flashback	0.246 ± 0.00	0.564 ± 0.00
	GRU-Spatial	0.297 ± 0.00	0.745 ± 0.00
5-Super Routiners	FL-Flashback	0.872 ± 0.00	0.951 ± 0.00
	GRU-Spatial	0.867 ± 0.00	0.988 ± 0.00

(a) Accuracy.

Scenario	Local algorithm	Energy (W)	VRAM (MB)
4-Super Scouters	FL-Flashback	172.2 ± 1.04	5506 ± 0
	GRU-Spatial	181.3 ± 0.92	5600 ± 1
5-Super Routiners	FL-Flashback	171.4 ± 0.50	5505 ± 1
	GRU-Spatial	181.3 ± 1.34	5599 ± 1

(b) Resource.

Scenario	Local Algorithm	Train (s)	Total
4-Super Scts.	FL-Flashback	69.0 ± 0.40	1167.6 ± 4.06 (19m 27s)
	GRU-Spatial	11.2 ± 0.13	570.6 ± 1.42 (09m 30s)
5-Super Rots.	FL-Flashback	100.1 ± 0.38	1676.8 ± 4.09 (27m 56s)
	GRU-Spatial	16.3 ± 0.07	811.2 ± 1.02 (13m 31s)

(c) Time.

Table 5.9: Impact of the local algorithm for the *Shenzhen* dataset. Average (with 95% C.I.) results for FedAvg, 6,000 users and 4 clients.

resource consumption compared to Flashback.

Summary: Our results on the impact of the number of clients and the learning algorithm agree with prior work on a qualitative level. Yet, we here show, once again, the need to investigate FL4iMP performance separately for diverse mobility profiles, as performance gaps exhibit striking differences. Whereas for some mobility profiles, changing the learning algorithm has a strong impact (e.g., *5-Super Routiners (6K)* in Shanghai), in scenarios of more complex patterns, the impact is more modest (e.g., *4-Super Scouters (6K)* in both datasets). In cases of very simple mobility patterns, the impact may be marginal (e.g., *Super Routiners* in Shenzhen). Yet, the resource usage results of the two learning algorithms suggest that there is a balance between achieving high prediction accuracy and the amount of computational resources required, which is an important consideration for practical applications.

5.5 Impact of data distribution across clients

We here evaluate IID vs. non-IID users distributions across a variety of scenarios. Table 5.10 presents each scenario evaluated. Each scenario contains 6,000 users, distributed across 3 clients. The scenario *10-Mixed IID (6K)* contains a mix of profiled users (about 30%-37% *Scouters* and 32%-35% *Routiners*) on each client, representing the traditional approach of combining users with distinct patterns in the same client (Ezequiel et al. [20], Proteasa et al. [47]). In the *11-Mixed non-IID (6K)* scenario, the **exact same users** of the previous scenario are now divided between clients given their mobility profile, where: Client 1 has 2,000 (*Super*) *Scouters* with the highest values of entropy; Client 2 has 2,000 Random Regulars randomly selected from the 6,000 in Table 5.1; and Client 3 has 2,000 (*Super*) *Routiners* with the lowest values of entropy. Next, the scenarios *12-Only Scouters IID (6K)* and *13-Only Routiners IID (6K)*, are IID scenarios with only Scouters or Routiners, divided by their entropy. Tables 5.11 and 5.12 reports accuracy results on the scenarios *10-Mixed IID (6K)* and *11-Mixed non-IID (6K)*, with GRU-Spatial and both FedAvg and FedProx.

Combining users with different mobility profiles (*10-Mixed IID (6K)*) leads to misleading accuracy results that are not aligned with the entropy results of profiles shown in Figs. 4.3 and 4.4 for the two datasets. *Super Scouters*, for instance, reach the highest entropy in both datasets, indicating how troublesome their predictability is, i.e., how difficult it is to correctly predict their location visits compared to *Super Routiners* users. However, the *10-Mixed-IID (6K)* scenario hides the natural prediction uncertainty of *Super Scouters*'s mobility by reporting average *Acc@1* values that are higher than 0.572 in

Client 1 (2K)	Client 2 (2K)	Client 3 (2K)
Mixed (Super Scouters, Random Regulars, Super Routiners)	Mixed (Super Scouters, Random Regulars, Super Routiners)	Mixed (Super Scouters, Random Regulars, Super Routiners)
(a) 10-Mixed IID (6K).		
Client 1 (2K)	Client 2 (2K)	Client 3 (2K)
Super Scouters (top 2K in entropy)	Random Regulars (2K Regulars)	Super Routiners (bottom 2K in entropy)
(b) 11-Mixed non-IID (6K).		
Client 1 (2K)	Client 2 (2K)	Client 3 (2K)
Super Scouters (top 2K in entropy)	Moderate Scouters (top 2K to 4K in entropy)	Light Scouters (top 4K to 6K in entropy)
(c) 12-Only Scouters IID (6K).		
Client 1 (2K)	Client 2 (2K)	Client 3 (2K)
Super Routiners (bottom 2K in entropy)	Moderate Routiners (bot. 2K to 4K in entropy)	Light Routiners (bot. 4K to 6K in entropy)
(d) 13-Only Routiners IID (6K).		

Table 5.10: Scenarios to analyze IID and non-IID data distribution across clients.

both datasets and equal to mixed users from *Super Routiners* or *Regulars* profiles. This occurs because combining users and simply reporting the resulting average neglects the hard/easy predictability of each user. In the *11-Mixed non-IID (6K)* scenario, we notice the discrepancy between the hardest to predict *Super Scouters (2K)* users of Client 1 (i.e., with $Acc@1$ of 0.221 in Shanghai and 0.213 in Shenzhen) and the easiest to predict *Super Routiners (2K)* users of client 3 (i.e., with $Acc@1$ of 0.797 in Shanghai and 0.986 in Shenzhen). This discrepancy ultimately reduces the resulting average $Acc@1$ to 0.519 in Shanghai and 0.583 in Shenzhen, highlighting the performance obfuscation across clients of the *10-Mixed IID (6K)* scenario. Furthermore, we notice that *FedProx* produced slightly better results for most non-IID clients than *FedAvg*, due to its regularization described in Chapter 3.3.

Next, we separately compare scenarios with either *Super Scouters* or *Super Routiners* users. Our goal is to assess if combining similar mobility patterns across clients can improve accuracy or convergence speed. Fig. 5.1 presents how such *Super* users perform in scenarios having mixed profiled users (i.e., Clients 1 and 3 of *11-Mixed non-IID (6K)* scenario) compared to scenarios only having similar profiles (i.e., Client 1 in *12-Only Scouters IID (6K)* or in *13-Only Routiners IID (6K)*). As depicted, separating clients by profile produce better results, i.e., faster convergence, specially for clients harder to predict, i.e., *Scouters*. Furthermore, we observe $Acc@1$ values of *Super Scouters* in the *12-Only*

Scenario	10-Mixed IID		11-Mixed non-IID	
Client	Acc@1	Acc@5	Acc@1	Acc@5
1	0.586 $\pm$ 0.01	0.764 $\pm$ 0.00	0.221 $\pm$ 0.01	0.443 $\pm$ 0.01
2	0.572 $\pm$ 0.00	0.753 $\pm$ 0.01	0.540 $\pm$ 0.01	0.764 $\pm$ 0.01
3	0.591 $\pm$ 0.00	0.763 $\pm$ 0.00	0.797 $\pm$ 0.01	0.886 $\pm$ 0.01
Avg.	0.583 $\pm$ 0.00	0.760 $\pm$ 0.00	0.519 $\pm$ 0.01	0.698 $\pm$ 0.01

(a) FedAvg.

Scenario	10-Mixed IID		11-Mixed non-IID	
Client	Acc@1	Acc@5	Acc@1	Acc@5
1	0.583 $\pm$ 0.01	0.764 $\pm$ 0.00	0.226 $\pm$ 0.00	0.446 $\pm$ 0.00
2	0.573 $\pm$ 0.01	0.750 $\pm$ 0.00	0.541 $\pm$ 0.00	0.768 $\pm$ 0.01
3	0.589 $\pm$ 0.00	0.760 $\pm$ 0.01	0.806 $\pm$ 0.00	0.889 $\pm$ 0.00
Avg.	0.581 $\pm$ 0.00	0.758 $\pm$ 0.00	0.524 $\pm$ 0.00	0.701 $\pm$ 0.00

(b) FedProx.

Table 5.11: **Impact of the data distribution on accuracy results for the *Shanghai* dataset.** Average (with 95% C.I.) results for GRU-Spatial, 6,000 users and 3 clients.

Scenario	10-Mixed IID		11-Mixed non-IID	
Client	Acc@1	Acc@5	Acc@1	Acc@5
1	0.631 $\pm$ 0.00	0.865 $\pm$ 0.00	0.213 $\pm$ 0.00	0.579 $\pm$ 0.00
2	0.636 $\pm$ 0.00	0.861 $\pm$ 0.00	0.549 $\pm$ 0.00	0.905 $\pm$ 0.00
3	0.603 $\pm$ 0.00	0.836 $\pm$ 0.00	0.986 $\pm$ 0.00	0.999 $\pm$ 0.00
Avg.	0.623 $\pm$ 0.00	0.854 $\pm$ 0.00	0.583 $\pm$ 0.00	0.828 $\pm$ 0.00

(a) FedAvg.

Scenario	10-Mixed IID		11-Mixed non-IID	
Client	Acc@1	Acc@5	Acc@1	Acc@5
1	0.631 $\pm$ 0.00	0.865 $\pm$ 0.00	0.211 $\pm$ 0.00	0.576 $\pm$ 0.00
2	0.636 $\pm$ 0.00	0.861 $\pm$ 0.00	0.550 $\pm$ 0.00	0.904 $\pm$ 0.00
3	0.604 $\pm$ 0.00	0.836 $\pm$ 0.00	0.987 $\pm$ 0.00	0.999 $\pm$ 0.00
Avg.	0.624 $\pm$ 0.00	0.854 $\pm$ 0.00	0.582 $\pm$ 0.00	0.826 $\pm$ 0.00

(b) FedProx.

Table 5.12: **Impact of the data distribution on accuracy results for the *Shenzhen* dataset.** Average (with 95% C.I.) results for GRU-Spatial, 6,000 users and 3 clients.

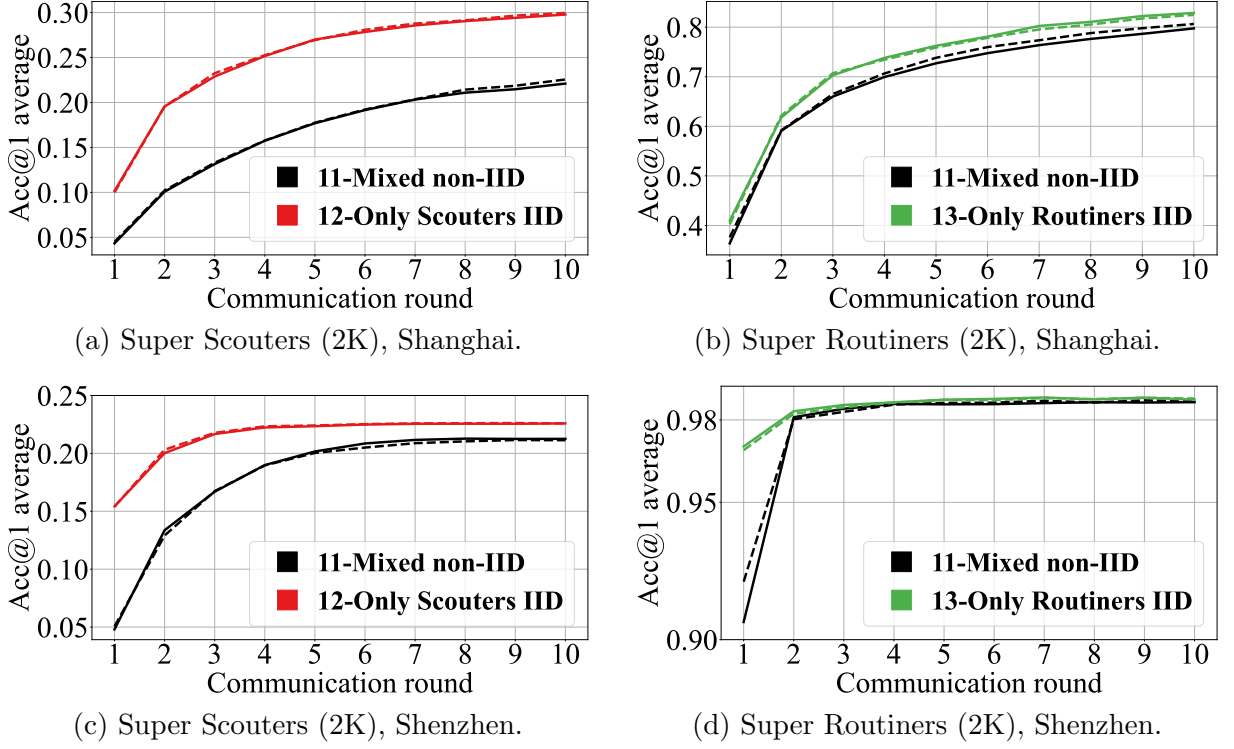


Figure 5.1: Impact of data distribution on $Acc@1$ and convergence in scenarios of Table 5.10. FedAvg (Solid line). FedProx (Dashed line).

Scouts IID (6K) scenario are higher than the one of *Super Scouts* in the (*10-Mixed non-IID* (6K)) scenario. This results correlates with the aggregation algorithm, that can generate a global model more helpful for the clients having closer mobility patterns.

Summary: Mixed users distribution across clients significantly impact the performance of FL4iMP models, even when the same number of users per client is considered. As shown, mixing users of dissimilar mobility patterns in the same clients obscure performance differences that emerge for users with diverse mobility profiles (i.e., the *Mixed IID* case), and separating them by their patterns (i.e., the *Mixed non-IID* case) highlights the hidden heterogeneity. Finally, combining users with similar behaviors (i.e., the *Only* case) facilitates the aggregation process, leading to better results, specially for harder to predict users.

5.6 Summary

In this chapter, we made an extensive evaluation that shows that FL4iMP is a multi-faceted problem, that in terms of accuracy is very sensitive to user mobility behaviors, given each user predictability. However, this factor showed little variation in terms

of execution time, which is controlled by the amount of data that needs to be processed, either by the number of records or the user population size. Moreover, resource usage quickly bottlenecks with an increased number of clients, however, a data silos architecture mitigates this issue, by allowing for an increased population size (which presents little resource usage impact) if the number of clients is kept fixed. Furthermore, this architecture showed that increasing the population size beyond 6,000 users did not enhanced accuracy significantly, given that this smaller population size can feed the local algorithm with enough unique mobility patterns, that are only repeated once the population increases.

In the next chapter, we make a summary of our results and also discuss possible next steps and future directions for research on the topic of the Federated Learning framework for individual mobility prediction.

Chapter 6

Conclusions and future directions

The last chapter of our work summarizes what was discussed throughout the thesis (Section 6.1), and proposes future directions and extensions for upcoming works (Section 6.2).

This thesis was centered around a main goal, which aimed at filling the existing FL literature gap on individual mobility prediction, that lacked a *comprehensive understanding of the factors that most impact FL prediction results*, including aspects ranging from human behavior to solution design.

To do so, the key research question driving our study was: “*How do various factors associated with the FL4iMP environment, from the human nature of its participants to architectural and algorithmic design decisions, impact its performance?*” Specifically, we aimed to address: “*How does FL4iMP perform for people with very diverse mobility patterns?*”.

6.1 Conclusions

To answer the aforementioned research questions, in Chapter 2 we first conducted an extensive literature review, pinpointing strengths and shortcomings of previous work. Next, in Chapter 3 we categorized and detailed a group of key factors impacting FL4iMP, dividing them by the *human* mobility behavior, *data* properties, FL *architecture* and *algorithms*. Subsequently, in Chapter 4, we described how we processed and characterized two distinct datasets, in order to lead to a comprehensive evaluation in Chapter 5.

We started answering our research questions in Section 5.2, where we first showed that FL4iMP performance is very sensitive to user mobility patterns and data properties. We showed that a great discrepancy of accuracy results can be found in the same dataset, even with the same model parameters, if the inherent heterogeneity of human behavior is not taken into account.

Our discussion continued in Section 5.3, where we shifted the focus to a key ar-

chitectural parameter, the number of clients. In this part, we found that accuracy drops significantly as the number of clients increases, given that the amount of data available for learning the local models is reduced. Furthermore, resource consumption naturally increases with the number of clients due to contention on the shared hardware.

We then carry the discussion to the algorithm side, and in Section 5.4 highlighted the balance between achieving high prediction accuracy and the amount of computational resources required, which is fundamental for practical applications.

Finally, in Section 5.5 we ultimately answered our research questions, showing that mixing users of dissimilar mobility patterns in the same clients obscure performance differences that emerge for users with diverse mobility profiles, and separating them by their patterns the hidden heterogeneity. Also, combining users with similar behaviors facilitates the aggregation process, leading to better results, specially for harder to predict users.

In a nutshell, our results showed that FL4iMP is a multi-faceted problem, where the factor with greatest impact on performance depends on the perspective being analyzed: (i) for accuracy: user mobility behaviors; (ii) for resource usage: number of clients; (iii) for execution time: population size. Furthermore, the population size has the lowest impact on accuracy and resource usage, whereas the user mobility behaviors have the lowest impact on execution time. We highlighted the fact that disregarding the pinpointed factors impacting FL4iMP, leads to poor and misleading performance results, obfuscating differences that emerge for users with different mobility patterns. Our findings recommend combining users with similar behaviors to facilitate the aggregation process, improve performance, and resources usability.

6.2 Future directions

In this section, we present possible extensions to this work, that can further empower the research on FL4iMP. We divide the future work following the factors categorized of Chapter 3.

6.2.1 Human and data factors

As presented throughout this thesis, the driving factor for better accuracy results rely on the individuals' mobility behaviors and data characteristics.

As seen in Section 5.2, there exists a great accuracy discrepancy between *Super Scouters* and *Super Routiners* in both datasets. However, this discrepancy differs in *Shanghai*, where the $Acc@1$ of *Super Scouters* was 0.343, and of *Super Routiners* was 0.805, where in *Shenzhen* the results were 0.297 and 0.867 respectively.

Therefore, a robust and complete analysis of the results has not only to consider the individuals different mobility behaviors, but also the data characteristics, as done in Chapter 4.

One possible extension to the human and data factors analysis, is to study the individuals' profile behaviors throughout different time windows. For example, an individual might have a strict routine during weekdays, but an exploratory behavior on weekends. Also, during vacation periods, individuals might opt to rest, staying longer periods at home, or to explore new cities and places.

Therefore, a possible extension is to consider the individuals behaviors throughout different periods, e.g., during weekdays, weekends and vacation periods, and utilizing this information for further improving the individuals' modeling and consequently achieving better effectiveness and efficiency results.

6.2.2 Algorithm and architecture factors

As discussed in Sections 5.3, 5.4, and 5.5, algorithm and architecture factors have a strong correlation.

In Section 5.3, we showed that increasing the number of clients (architecture factor) greatly challenged the local algorithms. Then, in Section 5.4, we showed that the local algorithm of choice can have a more robust neural network architecture that can lead to improved results. Finally, in Section 5.5, we showed that depending on how the same set of individuals is distributed among clients, different results can be obtained, with some architecture choices leading to misleading result interpretations.

Naturally, aside from proposing algorithms that better model the individuals' routine (as discussed in the previous section), Section 5.4 shows that, when considering users with simpler mobility patterns (e.g., *Routiners*), a more simplistic algorithm can be adopted for this type of users, where a more complex algorithm will be suitable to harder

to predict users (e.g., *Scouters*). By doing so, not only similar accuracy results can be obtained (as presented in the *Super Routiners* scenario in *Shenzhen*), but also resources (computation, time and energy) can be saved.

Furthermore, as presented in Section 5.5, accuracy results improve when the considered individuals are from the same profile, i.e., when individuals on each client present similar mobility behaviors. Therefore, coupled with the algorithm selection for each mobility profile, and architecture with, for example, three global models, one per unique profile, can be considered as a next step, in order to achieve better effectiveness and efficiency results.

Bibliography

- [1] Charu C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319944622.
- [2] Licia Amichi, Aline Carneiro Viana, Mark Crovella, and Antonio AF Loureiro. Understanding individuals' proclivity for novelty seeking. In *Proceedings of the 28th international conference on advances in geographic information systems*, pages 314–324, 2020.
- [3] Licia Amichi, Aline Carneiro Viana, Mark Crovella, and Antonio AF Loureiro. From movement purpose to perceptive spatial mobility prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, pages 500–511, 2021.
- [4] Licia Amichi, Aline Viana Carneiro, Mark Crovella, and Antonio Loureiro. Revealing an inherently limiting factor in human mobility prediction. *IEEE Transactions on Emerging Topics in Computing*, 2022.
- [5] Fereshteh Asgari, Vincent Gauthier, and Monique Becker. A survey on human mobility and its applications. *arXiv preprint arXiv:1307.0814*, 2013.
- [6] 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):1247–1254, 2020.
- [7] James Barrett, Phil Legg, Jim Smith, Tim Barnes, and Charles Boyle. Trends and Challenges in Data Analytics and Machine Learning Using Call Detail Records in Telecom Systems, 2024.
- [8] Yacine Belal, Sonia Ben Mokhtar, Hamed Haddadi, Jaron Wang, and Afra Mashhadi. Survey of federated learning models for spatial-temporal mobility applications. *arXiv preprint arXiv:2305.05257*, 2023.
- [9] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchu Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. Flower: A Friendly Federated Learning Research Framework. *arXiv preprint arXiv:2007.14390*, 2020.

- [10] David E Boyce and Huw CWL Williams. *Forecasting urban travel: Past, present and future*. Edward Elgar Publishing, 2015.
- [11] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.
- [12] Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304, 2015.
- [13] Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1): 1–24, 2019.
- [14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [15] Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M Almeida, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481, 2021.
- [16] Andrea Cuttone, Sune Lehmann, and Marta C González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7:1–17, 2018.
- [17] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.
- [18] João Paulo Esper, Aline Carneiro Viana, and Jussara M. Almeida. Beauty or Beast: Human Behavioral Insights and Learning Power of Federated Mobility Prediction. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '24, page 325–337, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711077. doi: 10.1145/3678717.3691323. URL <https://doi.org/10.1145/3678717.3691323>.
- [19] João Paulo Esper, Aline Carneiro Viana, and Jussara M Almeida. Impact of Mobility Patterns on Federated Learning applied to Human Mobility Prediction. In *NetMob 2023*, 2023.

- [20] Castro Elizondo Jose Ezequiel, Martin Gjoreski, and Marc Langheinrich. Federated Learning for Privacy-Aware Human Mobility Modeling. *Frontiers in Artificial Intelligence*, 5, 2022.
- [21] Zipei Fan, Xuan Song, Renhe Jiang, Qunjun Chen, and Ryosuke Shibasaki. Decentralized attention-based personalized human mobility prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–26, 2020.
- [22] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*, pages 1459–1468, 2018.
- [23] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–21, 2020.
- [24] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the first workshop on measurement, privacy, and mobility*, pages 1–6, 2012.
- [25] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.
- [26] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- [27] Andrea Hess, Karin Anna Hummel, Wilfried N Gansterer, and Günter Haring. Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys (CSUR)*, 48(3):1–39, 2015.
- [28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [29] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.
- [30] Ruei-Hau Hsu, Yi-Cheng Wang, Chun-I Fan, Bo Sun, Tao Ban, Takeshi Takahashi, Ting-Wei Wu, and Shang-Wei Kao. A privacy-preserving federated learning system for android malware detection based on edge computing. In *2020 15th Asia Joint Conference on Information Security (AsiaJCIS)*, pages 128–136. IEEE, 2020.

- [31] Zongyuan Huang, Shengyuan Xu, Menghan Wang, Hansi Wu, Yanyan Xu, and Yao-hui Jin. Human mobility prediction with causal and spatial-constrained multi-task network. *EPJ Data Science*, 13(1):22, 2024.
- [32] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, pages 1–9, 2013.
- [33] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.
- [34] Junghye Lee, Jimeng Sun, Fei Wang, Shuang Wang, Chi-Hyuck Jun, Xiaoqian Jiang, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR medical informatics*, 6(2):e7744, 2018.
- [35] Anliang Li, Shuang Wang, Wenzhu Li, Shengnan Liu, and Siyuan Zhang. Predicting human mobility with federated learning. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 441–444, 2020.
- [36] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
- [37] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Learning on Non-IID Data Silos: An Experimental Study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.
- [38] Kuang-Yao Lin and Wei-Ren Huang. Using federated learning on malware classification. In *2020 22nd international conference on advanced communication technology (ICACT)*, pages 585–589. IEEE, 2020.
- [39] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [40] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1):1–44, 2021.
- [41] Wesley Mathew, Ruben Raposo, and Bruno Martins. Predicting future locations with hidden Markov models. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 911–918, 2012.

- [42] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [43] Eduardo Mucceli, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and José Ignacio Alvarez-Hamelin. On the Regularity of Human Mobility. *Pervasive and Mobile Computing*, 2016. URL <https://inria.hal.science/hal-01367825>.
- [44] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org> . <https://www.openstreetmap.org>, 2017.
- [45] Luca Pappalardo, Filippo Simini, Gianni Barlacchi, and Roberto Pellungrini. scikit-mobility: A Python Library for the Analysis, Generation, and Risk Assessment of Mobility Data. *Journal of Statistical Software*, 103(1):1–38, 2022. doi: 10.18637/jss.v103.i04. URL <https://www.jstatsoft.org/index.php/jss/article/view/v103i04>.
- [46] Chung Park, Taekyoon Choi, Taesan Kim, Mincheol Cho, Junui Hong, Minsung Choi, and Jaegul Choo. FedGeo: Privacy-Preserving User Next Location Prediction with Federated Learning. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–10, 2023.
- [47] Vlad-Alexandru Proteasa, Radu-Ioan Ciobanu, Ciprian Dobre, and Radu-Corneliu Marin. Federated Learning for Human Mobility. In *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pages 780–785. IEEE, 2023.
- [48] Pian Qi, Diletta Chiaro, Antonella Guzzo, Michele Ianni, Giancarlo Fortino, and Francesco Piccialli. Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, 150:272–293, 2024.
- [49] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.
- [50] 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.
- [51] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation, 1985.
- [52] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

- [53] Thiago H. Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio A. F. Loureiro, and Daniele Quercia. Urban Computing Leveraging Location-Based Social Network Data: a Survey. *ACM Computing Surveys*, March 2019. doi: 10.1145/3301284.
- [54] Gavin Smith, Romain Wieser, James Goulding, and Duncan Barrack. A refined limit on the predictability of human mobility. In *2014 IEEE international conference on pervasive computing and communications (PerCom)*, pages 88–94. IEEE, 2014.
- [55] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [56] Craig Stedman. What are data silos and what problems do they cause? <https://www.techtarget.com/searchdatamanagement/definition/data-silo>, 2021. [Accessed on 07-May-2024].
- [57] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.
- [58] Douglas Do Couto Teixeira, Aline Carneiro Viana, Mário S Alvim, and Jussara M Almeida. Deciphering predictability limits in human mobility. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 52–61, 2019.
- [59] Douglas Do Couto Teixeira, Jussara M Almeida, and Aline Carneiro Viana. On estimating the predictability of human mobility: the role of routine. *EPJ Data Science*, 10(1):49, 2021.
- [60] Douglas Do Couto Teixeira, Aline Carneiro Viana, Jussara M Almeida, and Mrio S Alvim. The impact of stationarity, regularity, and context on the predictability of individual human mobility. *ACM Transactions on Spatial Algorithms and Systems*, 7(4):1–24, 2021.
- [61] Jinzhong Wang, Xiangjie Kong, Feng Xia, and Lijun Sun. Urban human mobility: Data-driven modeling and prediction. *ACM SIGKDD explorations newsletter*, 21(1): 1–19, 2019.
- [62] Shuang Wang, Bowei Wang, Shuai Yao, Jiangqin Qu, and Yuezheng Pan. Location prediction with personalized federated learning. *Soft Computing*, pages 1–12, 2022.
- [63] Chenhao Xu, Youyang Qu, Yong Xiang, and Longxiang Gao. Asynchronous federated learning on heterogeneous devices: A survey. *Computer Science Review*, 50:100595, 2023.

- [64] Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudre-Mauroux. Location prediction over sparse user mobility traces using rnns. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2184–2190, 2020.
- [65] Khong-Lim Yap, Yung-Wey Chong, and Weixia Liu. Enhanced handover mechanism using mobility prediction in wireless networks. *PloS one*, 15(1):e0227982, 2020.
- [66] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. UrbanCPS: a cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*, pages 238–247, 2015.
- [67] Dongxiang Zhang, Long Guo, Liqiang Nie, Jie Shao, Sai Wu, and Heng Tao Shen. Targeted advertising in public transportation systems with quantitative evaluation. *ACM transactions on information systems (TOIS)*, 35(3):1–29, 2017.
- [68] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and A Salman Avestimehr. Federated learning for the internet of things: Applications, challenges, and opportunities. *IEEE Internet of Things Magazine*, 5(1):24–29, 2022.
- [69] Xiao Zhang, Qilin Wang, Ziming Ye, Haochao Ying, and Dongxiao Yu. Federated representation learning with data heterogeneity for human mobility prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [70] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-IID data: A survey. *Neurocomputing*, 465:371–390, 2021.