Benchmarking Clustered Federated Learning Algorithms for Next-Point Prediction

Sonal Yadav*, Yacine Belal†, Brent Lagesse*, Afra Mashhadi*
*Computing and Software Systems, University of Washington, Bothell, USA
Emails: sonaly@uw.edu, lagesse@uw.edu, mashhadi@uw.edu
†Computer Science, INSA Lyon, Lyon, France
Email: yacine.belal@insa-lyon.fr

***Abstract*—The collection of spatio-temporal mobility data, especially individual trajectory data from location-based services and smart devices, raises significant privacy concerns. However, it is extremely valuable for policy makers in tasks such as next-point predictions. Federated Learning aims to address these issues by training models locally on edge devices, thus preserving data privacy. However, the heterogeneous nature of individual trajectory data can make it challenging for a single global model to converge effectively in Federated Learning. One approach to overcome this challenge is Clustered Federated Learning (CFL). In this paper, we investigate to what extent CFL algorithms can improve the accuracy of next-point prediction models. We study four state-of-the-art CFL algorithms on two benchmark datasets, namely GeoLife and MDC and compare the performance of these algorithms in terms of accuracy and APR with state-of-the-art personalized FL models. We show that CFL is a viable option for the next-point prediction task and that it can particularly improve the performance of the model for user groups with high and low entropy. We open source a framework that can help the research community benchmark future personalized FL models against CFL algorithms.**

***Index Terms*—Next-Point Prediction, Federated Learning, Clustered Federated Learning**

I. Introduction

The widespread adoption of location-based services, such as mobile phones and smartwatches, has resulted in the generation of vast amounts of spatio-temporal mobility data. This data, which tracks individual movements over time, is crucial for various applications, including traffic management, urban planning, and other types of policy making [1]–[3]. A key subset of this data is trajectory data, which combines time series, spatial information, and highly socially driven human movement patterns. Trajectory data can be divided into two categories: crowd flow data, which captures the movement of large groups, and individual trajectory data, which focuses on predicting the next location of a single user. Individual trajectory data is particularly useful in navigation systems, pandemic management, and other personalized applications.

Traditionally, predictions of the next location of a user have relied on centralized learning models, which analyze mobility traces and historical trends to make predictions. However, the collection and storage of such private mobility data on centralized servers raises significant privacy concerns. Research has shown that even anonymized trajectory data can lead to the re-identification of individuals when combined with other datasets, posing a serious privacy risk [4]. Although deep learning models have been proposed to generate synthetic trajectory data for privacy protection [5], these models often struggle to balance data utility with privacy, potentially leading to memorization and leakage of sensitive information [6].

As an alternative solution, Federated Learning (FL) [7] has been introduced as a paradigm to address some of these privacy concerns. In FL, data remains on client devices and only the model parameters are shared, enhancing the privacy of the data itself. However, this approach faces challenges due to the non-independent and non-identically distributed nature of data between clients [8], leading to poor model convergence and prediction accuracy. To address this problem, personalized next-point prediction FL models have been proposed with great success [9]–[11].

As an alternative solution to the non-IID problem, Clustered Federated Learning (CFL) [12] has emerged, where model performance is improved by grouping clients with similar data distributions and training separate models for each cluster. While CFL has been used primarily in vision tasks, there is limited understanding of how CFL compares to personalized FL approaches in next-point prediction tasks. Ye et al. [13] showed that for Point of Interest (POI) recommendation tasks, CFL outperforms FedAvg [14] and but under-performs compared to PMF [10], a personalized FL model.

Although Ye et al. [13] are the first to perform such a comparison, their analysis is limited to one CFL implementation and makes it difficult to generalize and answer the broader research question of: *Can CFL offer a strong alternative for spatio-temporal predictive tasks (RQ1)? How do CFL compare with personalized FL approaches, namely PMF (RQ2)*? *For what type of users / datasets does CFL show better improvements compared to personalized FL approaches (RQ3)*?

To address these research questions, we compare the performance of four state-of-the-art CFL algorithms on next-point prediction tasks. Specifically, we implement a framework that allows researchers to evaluate IFCA [15], WeCFL [16], CFL [12], FL+HC [17] against PMF [10] and the FedAvg [14] baseline. To facilitate this comparison, we have open-sourced our framework[1] that can support any underlying predictive

[1]https://github.com/YacineBelal/fl-nextpoint-benchmark-DCOSS-IOT-2025

mobility models and any spatio-temporal datasets. In this paper we report the results of our benchmarking analysis on two datasets, GeoLife [18] and MDC [19], in terms of ACC@5 and APR metrics. In addition to reporting overall performance improvements across all users, we also identify which groups of users benefit the most from CFL models. The contributions of this work are as follows:

- We study how CFL algorithms [12], [15]–[17] perform on the next-point prediction task on two datasets [18], [19] and across various metrics. We compare this performance with FedAvg baseline and a personalized FL solution, namely, PMF [10].
- We highlight for which groups of users and what dataset properties CFL algorithms are better suited than personalized FL approaches.
- We open-source our code for the research community in order to enable researchers to baseline their next-point prediction model and approaches against various CFL algorithms on any dataset.

II. Related Works

A. Federated Point-of-Interest Recommendation

Location-based applications, such as POI recommendation, offer numerous benefits, but they also raise significant privacy concerns related to the collection, processing, and sharing of individual mobility data. To account for this, recent research has witnessed the emergence of federated POIs recommendation systems. These systems allow users to train predictors without having clients' data ever leave their premises. For instance, the authors of PREFER [20] proposed a multi-federated server top-K POIs recommendation system. This system was later decentralized in [21]. To gain privacy, [22] proposed to mix both approaches, by learning sensitive model parameters in a decentralized manner while less sensitive ones were learned in a federated manner. In [23], the authors proposed the first federated next point prediction (FNP).

B. Data Heterogeneous Federated Next-Point Prediction

Non-independent and identically distributed Data (non-IID) is a general open problem in the federated framework [24], which rises when clients have dissimilar data distributions. This often induces a phenomenon called client drift [25], where the models updates of clients have varying directions, which leads the model to diverge [26]. Trajectory data is considered as prime candidate for non-iid data [27], due to the number of factors that can determine the mobility of a user (*e.g.*, personal preferences, time, weather, and etc.). FNP remains highly challenging. To tackle this issue, there are mainly two family of works: i) Personalized FL and ii) Clustered FL (CFL).

Personalized FL. In this category, works often incorporate personalized layers (*i.e.*, locally trained) to the clients' models to capture their personal preferences/conditions. In this context, [28] proposed to leverage attention mechanisms and few-shot learning to the FNP task. However, their approach suffers from communication complexity and struggles with erratic human mobility patterns. In PMF [10], several layers are added to the FL model and are trained purely on a client-level. Associated with a group-based sampling technique, this approach showed promising results. In [11], Wang et al. presented STLPF, which employed self-attention layers and collaborative training without relying on a global model. [29] tested Flashback models [30], known to mitigate model forgetting phenomenon, to keep track of the local patterns learned by the clients' models. Another approach to personalized FNP is to leverage meta-learning techniques [31], to train the model simultaneously on a global task (*i.e.*, learning to predict a next POI), and learning the specific preferences for each client. Several works have taken this direction [32], [33]. For instance, [33] adopt the model-agnostic meta-learning method, which enables clients' to be collaboratively train models that can quickly adapt to their local distributions. These approaches usually require computing the hessian of the loss function, which is computationally expensive.

Clustered FL. CFL is based on the idea of grouping clients with similar data distributions and training distinct (partial) models for each cluster. This is achieved by analyzing the similarity of the clients' model updates during training. This approach ensures that models are better tailored to the specific data characteristics of each cluster, resulting in higher utility for individual clients. Depending on the clustering algorithm, there are many existing approaches in CFL [13], [15]–[17] (See Section III-C for details). While all of these works are considered state-of-the-art in the CFL community, their efficiency on the FNP task has rarely been quantified. In fact, CPF-POI [13] is, to the best of our knowledge, the only work that draws such an investigation. However, it has not investigated the correlation between the level of unpredictability of a client, hence, their likely dissimilarity with other clients, and the benefit of the underlying clustering schema. Answering this question is one of the main objectives of our work.

III. Methodology and Analysis

In this section, we formulate the problem within the standard federated learning framework before introducing our approach to analyzing the impact of heterogeneity in the next POI prediction task and the various methods we use for benchmarking.

A. Problem Definition

Let $[k] = \{1, 2, \ldots, k\}$ for any positive integer k. We consider a standard federated setting. We denote the set of clients as $\mathcal{N} = [n]$ and the set of POIs as $\mathcal{M} = [m]$. Each client i holds a private dataset $\mathcal{D}_i = (X_i, Y_i)$ where X_i represents the current POI and possible contextual features (*e.g.*, previous POIs, weather, ...etc.) and Y_i represents the next POI. The clients leverage the union of their local datasets, denoted $\mathcal{D} = \cup_{i \in \mathcal{N}} \mathcal{D}_i$, to optimize a model that predicts the next POI in a privacy-friendly manner (*i.e.*, without sharing their data). Let us denote the model parameters by $\Theta \in \mathbb{R}^d$, where d is the parameter space dimension. The learning objective in this framework can often be formulated as a

minimization problem where the goal is to find the optimal set of parameters Θ^* such that:

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} f(\Theta; \mathcal{D}_i) \tag{1}$$

where f is some cost function.

Problem 1 is solved following gradient descent in $[T]$ rounds. Specifically, at each round $t \in [T]$, (1) the server first broadcasts Θ_t to the clients. (2) Each client i locally updates the model on D_i by doing gradient steps, thus, obtaining a locally optimized model $\Theta_t^{(i)}$. (3) Finally, the server receives and aggregates the local models $\{\Theta_t^{(i)}\}_{i \in \mathcal{N}}$. In this work, we leverage FedAvg [14], where each local model is weighted by the number of samples it was trained on. This process is repeated until Θ^* is found (or a good approximation of it).

B. Analyzing Data Heterogeneity in Human Mobility

In this paper, we focus on two benchmark human mobility datasets, namely, GeoLife [18] and the Mobile Data Challenge (MDC) [19]. The GeoLife dataset contains GPS trajectory data from 182 users, spanning four years. MDC is provided by the Idiap Research Institute and comprises GPS data from 185 users over one year in Switzerland.

TABLE I
SPARSITY ON MDC AND GEOLIFE.

Dataset	#Users	#POIs	Sparsity
MDC	163	293	76.32%
GeoLife	124	848	90.53%

Within the general federated framework, we are interested in data heterogeneous environments, that is, environments where the data distribution between clients significantly differs. Indeed, this is one of the main characteristics of human mobility datasets where i) clients can visit largely different locations, which causes a label distribution skew, *i.e.*, $(\forall (i,j) \in \mathcal{N}^2 : P(Y_i) \neq P(Y_j))$ and ii) in similar contextual conditions (*e.g.*, time of the day, ...etc.), clients can have different mobility behaviors, *i.e.*, $(\forall (i,j) \in \mathcal{N}^2 : P(Y_i|X_i) \neq P(Y_j|X_j))$. To quantify this heterogeneity and study its impact on the performance of models, we propose to correlate the sparsity of mobility datasets with client-level entropy. Sparsity is characterized by a low number of clients visiting each POI. As for entropy, it describes the level of unpredictability of the movements of clients. We argue that higher levels of entropy, conjoined with sparsity, correspond to a high level of data heterogeneity. The rationale behind this is the following: high entropy clients necessarily visit a reasonable amount of POIs, however, as the dataset is sparse, this implies a low cardinality of POI intersection between clients, hence more heterogeneity.

Concretely, we measure the sparsity by reporting the quantity $1 - \frac{|\mathcal{D}|}{n \times m}$. Table I illustrates this phenomenon for MDC and Geolife. We note that both datasets have a high level of sparsity and in particular GeoLife dataset is highly sparse. For the client-level entropy, we leverage the standard Shannon entropy [34], defined as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \tag{2}$$

where $P(x)$ is the probability of POI x being visited. In practice, we compute the entropy for each client individually.

Furthermore, to study the correlation between the entropy and the utility of the investigated works, we categorize clients into low, medium, and high entropy groups using the interquartile range (IQR) of entropy values. Clients with entropy values below Q1 were categorized as low entropy, those between Q1 and Q3 as medium entropy, and those above Q3 as high.

Figure 1 shows the Cumulative Distribution Funtion (CDF) of client entropy as well as a Kernel Density Etimate (KDE) of the number of unique POIs visited per user, for the GeoLife and MDC datasets respectively. In both datasets, a wide range of entropy values is observed, reflecting diverse movement patterns across clients. For GeoLife, most clients exhibit entropy values between 1.5 and 4.5, indicating substantial variability in their movement behaviors. The MDC dataset shows a slightly narrower distribution, with entropy values primarily ranging from 2.5 to 4.

To complement this temporal analysis, the KDE plot illustrates the distribution of spatial diversity across users by capturing how many distinct POIs each user visits. The x-axis represents the number of unique POIs visited, while the y-axis indicates the density of users exhibiting that behavior. GeoLife displays a broader and more skewed distribution, with a notable presence of users visiting a large number of locations. In contrast, MDC users are more concentrated in a narrower range, suggesting less exploratory movement.

Together, the entropy CDF and POI-count KDE highlight the heterogeneous nature of human mobility datasets, where both sparsity and entropy play significant roles in modeling challenges. Low entropy generally corresponds to predictable, routine activities, such as daily commutes or frequent visits to the same locations. In contrast, high entropy represents more irregular or diverse movement patterns, where individuals engage in unpredictable activities or visit a wide variety of locations. When combined with high sparsity, this variability amplifies data heterogeneity, posing substantial challenges for predicting next-point movements.

C. Benchmark Methods

The standard federated setting aims to find the solution to Problem 1. However, due to data heterogeneity, this model can be sub-optimal for a potentially significant number of clients. One popular approach to accommodate this is CFL. This line of works aims to find a set of clusters $\mathcal{C} = [k]$ and incorporate each client i in the cluster that best matches some predefined criterion. We consider the following four CFL works:

- **IFCA [15]**. In this work, each user assigns itself to the cluster that minimizes its local cost function on a local holdout set. This can be formulated as:

$$\forall i \in \mathcal{N} : \Theta_c = \underset{\Theta}{\operatorname{argmin}}\, f(\Theta, \hat{\mathcal{D}}_i) \implies i \in c \tag{3}$$

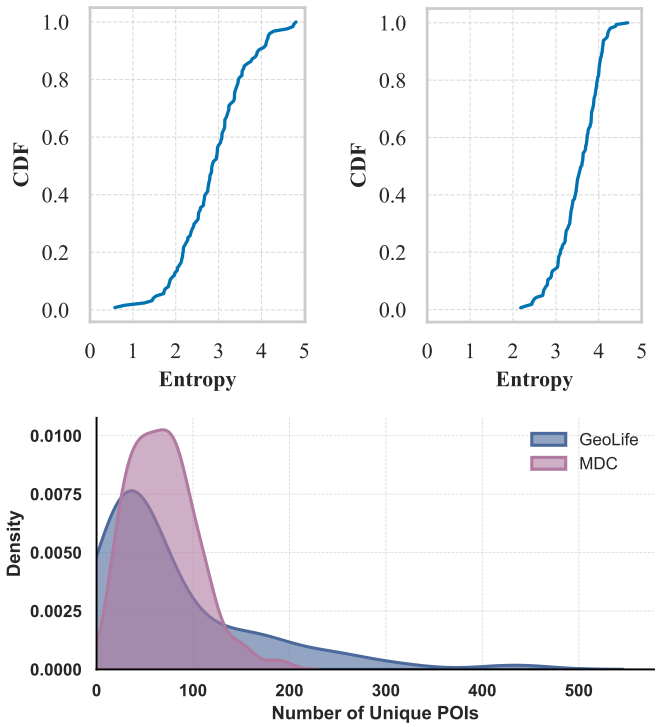

Fig. 1. CDF of User Entropy for GeoLife (left) and MDC (right).

Where $\hat{\mathcal{D}}_i$ is the holdout dataset of client i.

- **WeCFL [16]**. In this work, a k-means clustering algorithm is performed directly on the clients' models to build $\mathcal{C}$. It is worth noting that in both IFCA and WeCFL, the aggregation step is performed inter-cluster.
- **CFL [12]**. In this work, clients are recursively partitioned into clusters based on the cosine similarity of their model updates. After each round, the server checks if the average norm of the updates within a cluster is below ϵ_1 and the maximum norm exceeds ϵ_2. If both conditions are met, the cluster is split into two. This can be expressed as:

$$\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\|\Delta w_i\| < \epsilon_1 \quad \text{and} \quad \max_{i\in\mathcal{C}}\|\Delta w_i\| > \epsilon_2 \tag{4}$$

where Δw_i represents the weight update of client i in cluster $\mathcal{C}$. After clustering, FedAvg is applied within each cluster to improve personalization for clients with similar data distributions.

- **FL+HC [17]**. In this work, clients initially perform FedAvg for several rounds. Then, **FL+HC [17]** applies agglomerative clustering to group clients based on the distance between their model parameters:

$$d(w_i, w_j) = \|w_i - w_j\| \tag{5}$$

where w_i and w_j are the model parameters of clients i and j. Once clustered, FedAvg is applied within each cluster. Clustering continues until a distance threshold is met or a fixed number of clusters is reached.

- **PMF [10]**. This work aims at building more personalized models, which allows them to be more resilient to data non-iidness. In practice, the authors propose to augment the recommendation model with a personal adaptor, which consists in one or several additional transformations. Specifically, let H_g be the hidden state coming before the output layer in the learning model. PMF suggests adding a trainable personal bias vector v_p, st. $H_p = H_g * \sigma(v_p)$, with H_p being the new personalized hidden state and σ the sigmoid function. Evidently, v_p is locally fine-tuned for each client and not shared with the FL server. Finally, It is worth noting that PMF advocates for the usage of differential privacy [35] to generate noisy data on which POI embeddings are trained, which we do not consider in this work. The reason behind this is that our model does not contain such highly sensitive parameters. As such, the PMF implementation we investigate can be seen as a more utility oriented one.

IV. Evaluation

A. Next-Point Prediction Model

In centralized settings, where all users' trajectories are available, Recurrent Neural Network (RNN)-based approaches, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are widely used for trajectory and predictive tasks [36], [37]. Building on these works, we implemented a next-point prediction model utilizing an LSTM architecture. The model takes POI sequences of length 10 as input. The architecture comprises a single LSTM layer with 100 hidden units, followed by a fully connected layer of size m, which outputs the logits of the unique POIs in the dataset. The optimization process leverages a cross-entropy loss. Table II summarizes the hyper-parameter setting for our evaluation.

TABLE II
Federated Learning hyper-parameters.

Hyper-parameter	Value
Context window size	10
Learning rate	0.001
Batch size	1
Epochs	2
FL rounds	40

B. Data Preprocessing

To prepare for the next-point prediction tasks, we set the context length to 10 for the GeoLife and MDC datasets. This refers to the number of previous points (locations) used to predict the next point in a user's trajectory, ensuring that each sequence fed into the model contains sufficient historical information for accurate predictions. As part of this process, any clients or labels with fewer than 10 trajectory points were removed, as they did not meet the minimum context length requirement for the model. Additionally, each user ID in the datasets was treated as a unique client within the federated learning setup, preserving the decentralized nature of the experiments.

C. Evaluation Metrics

To measure the predictive performance of the models, we employed the following metrics:

- **Accuracy@K:** Measures the fraction of times the correct next point is predicted within the top-K predicted points. It is calculated as:

$$\text{Accuracy@K} = \frac{1}{N}\sum_{i=1}^{N} 1(y_i \in \hat{y}_{i,k})_{k \leq K} \tag{6}$$

  where N is the number of predictions, y_i is the true next location, $\hat{y}_{i,k}$ is next location predicted at rank k and 1 is the indicator function.
  In our experiments, we compute Accuracy@1 and Accuracy@5.
- **Average Percentile Rank (APR):** Evaluates the ranking performance of predictions by calculating the average percentile rank for the correct next location:

$$\text{APR} = \frac{1}{N}\sum_{i=1}^{N}(\frac{\text{rank}(y_i)}{C} \times 100) \tag{7}$$

  where $\text{rank}(y_i)$ represents the rank of the true label among the predicted rankings, and C is the total number of classes. Concretely, the APR ranges between 0 and 100, where smaller values indicate a tendency of the model to recommend true POIs higher in the ranking list. The lower the APR value, the better the underlying model is at making predictions.

D. Experimental Setup

The experiments were conducted on a server with 2 Intel Xeon Silver 4210 CPUs (40 cores total), 4 NVIDIA GeForce RTX 2080 Ti GPUs, 187 GB of RAM, and 15 TB of HDD storage. The system ran Linux (kernel version 4.18) with CUDA 12.3, and the code was implemented using Python with PyTorch and the Flower Federated Learning framework [38]. Each experiment utilized one GPU and took 11 hours on average to train for 40 rounds.

V. Results

A. RQ1: Does CFL generalize to the FNP task?

In this section, we investigate the ability of different CFL methods to adapt to the FNP task, considering the previously mentioned characteristics that makes it particularly difficult. To this end, we first look at the average performance metrics across all clients for each dataset with standard deviation noted in the parenthesis, which can be found in Table III. We observe that three out of the four CFL solutions perform similarly or worse than standalone FedAvg. For instance, on Geolife, CFL, FL+HC and WeCFL obtain an accuracy@5 of 24%, 23% and 15%, respectively, compared to 24% for FedAvg. These numbers are further exacerbated in MDC, due to the sparsity of the dataset. In parallel, IFCA seems quite efficient for this task on both datasets, achieving up to 44% on Geolife, for instance. We attribute the difference in performance between IFCA and other CFL methods to the notion of *mode connectivity* [39], [40]. In essence, this notion dictates that models with similar performance (*i.e.*, local optima), are often permutation functions of each other. As such, these models can differ distance-wise while still being equally performing for similar data distributions. While CFL methods that are based on parameter-wise clustering such as FL+HC and WeCFL do not detect these models; a solution that this loss-based as IFCA can detect them and correctly put them in the same cluster. **A critical insight gained from this result, is that for a task such as NFP, which requires non-convex loss functions, clustering algorithms have to be a level of abstraction beyond the model parameters. For instance, considering approaches such as spectral clustering, or taking into account additional meta-data (*e.g.*, category of POIs) can be an interesting research direction.**

B. RQ2: How do CFL compare with PMF?

In this section, we aim to compare the performance of CFL algorithms with Personalized FL approaches for FNP task. From Table III, **we observe that the best CFL approach, namely, IFCA, is systematically competitive with PMF on accuracy metrics**. It slightly outperforms it on Geolife for rank=5 where it obtains 48% compared to the 44% of IFCA, while IFCA outperforms on MDC (17% versus 14% of accuracy@5). **However, the APR metric demonstrates a clear superiority of IFCA over PMF on both datasets (8.03% versus 13.41% for Geolife and 18.39% versus 40.23% on MDC)**. This result reveals that while IFCA might not systematically correctly predict test POIs in the top rank (*i.e.*, false negatives), it still often ranks them higher compared to POIs not belonging to the test (*i.e.*, true negatives). In contrast, PMF seems to struggle when facing more challenging test POIs (*e.g.*, not seen in the training set), as it tends to rank them significantly lower, which is penalized by the APR. This intuition is confirmed by the growing in gap in APR on MDC, where sparsity is higher, and such PMF is more likely to encounter challenging test POIs. **At its core, these results indicate higher generalization capabilities for IFCA over PMF, which can be attributed to the fact that IFCA does a *global personalization*, by leveraging models of other (similar) clients, while PMF optimizes the loss functions strictly on a local basis.**

C. RQ3: How does CFL compare to PMF on a group-level performance?

In this section, we attempt to confirm previously drawn insight by investigating the performance of users grouped in the three degrees of entropy: High, Medium and Low. To this end, we calculate improvement gains in comparison to FedAvg for each strategy and for each metric. Tables IV, and V report the percentage of clients that have gained improvements (positive delta denoted by $\uparrow$) or decline in performance (negative delta denoted by $\downarrow$) in comparison to FedAvg. In addition to the delta metrics, we also present the Cumulative Distribution Function (CDF) of Accuracy@5 for low and medium entropy groups in the MDC dataset, shown in Figure 5. The CDF plots

TABLE III
COMPARISON OF VARIOUS ALGORITHMS ON GEOLIFE AND MDC DATASETS REPORTED IN MEAN AND (STANDARD DEVIATION).

	Geolife			MDC		
	Acc@1	*Acc@5*	*APR*	*Acc@1*	*Acc@5*	*APR*
FedAvg [14]	0.08 (0.15)	0.24 (0.24)	13.82 (13.42)	0.003 (0.02)	0.04 (0.13)	17.49 (8.96)
CFL [12]	0.09 (0.15)	0.24 (0.25)	14.06 (13.86)	0.004 (0.02)	0.02 (0.09)	18.78 (9.25)
FL+HC [17]	0.09 (0.15)	0.23 (0.24)	13.83 (13.68)	0.004 (0.03)	0.01 (0.06)	18.53 (8.74)
WeCFL [16]	0.08 (0.16)	0.15 (0.26)	31.63 (16.22)	0.003 (0.03)	0.04 (0.11)	34.41 (10.48)
IFCA [15]	0.22 (0.23)	0.44 (0.29)	8.03 (7.98)	0.02 (0.07)	0.17 (0.21)	18.39 (9.17)
PMF [10]	0.22 (0.23)	0.48 (0.28)	13.41 (13.80)	0.06 (0.15)	0.14 (0.18)	40.23 (13.12)

show how accuracy is distributed across clients, helping us visualize the proportion of clients reaching different accuracy levels.

TABLE IV
PERCENTAGE OF CLIENTS WITH POSITIVE (↑) AND NEGATIVE (↓) DELTA ACCURACY@5. THE HIGHEST ACCURACY GAIN COMPARED TO FEDAVG PER GROUP PER DATASET ARE UNDERLINED.

Datasets	Strategy	Low	Med.	High
GeoLife	**CFL**	↑9.68% ↓6.45%	↑11.29% ↓9.68%	↑25.81% ↓16.13%
	FL+HC	↑12.90% ↓22.58%	↑12.90% ↓19.35%	↑25.81% ↓12.90%
	IFCA	↑58.1% ↓12.9%	↑59.7% ↓19.4%	↑77.4% ↓16.1%
	WeCFL	↑35.5% ↓22.6%	↑41.9% ↓16.1%	↑45.2% ↓19.7%
	PMF	↑<u>83.87</u>% ↓9.68%	↑<u>67.74</u>% ↓20.97%	↑<u>83.87</u>% ↓12.90%
MDC	**CFL**	↑4.88% ↓43.90%	↑9.88% ↓30.86%	↑21.95% ↓31.71%
	FL+HC	↑2.44% ↓41.46%	↑7.41% ↓37.04%	↑14.63% ↓31.71%
	IFCA	↑48.8% ↓29.3%	↑80.3% ↓12.4%	↑77.8% ↓14.6%
	WeCFL	↑19.5% ↓32.1%	↑32.1% ↓28.4%	↑51.2% ↓14.6%
	PMF	↑<u>63.41</u>% ↓17.07%	↑<u>81.48</u>% ↓3.70%	↑<u>90.24</u>% ↓2.44%

a) High Entropy Group: This group contains users with the most unpredictable mobility traces, hence, the expectation is that these users are the most likely to benefit from CFL/Personalization approaches. This expectation is reasonably confirmed in Table IV, where high entropy groups benefit more often than others groups overall. However, there are clear disparities between different strategies and datasets. **On Geolife, we observe positive accuracy@5 improvement for all strategies, albeit more pronounced for PMF (83.47%), followed by IFCA (77.4%) and WeCFL (45.2%).** This observation translates to the MDC dataset for these three strategies, while the other approaches seem to be halted by the sparsity of the dataset. To confirm these results, we put the lens on the distribution of accuracy@5 and APR, in Figure 2 on the Geolife dataset. This further demonstrates two key phenomenons:i) **While PMF does showcase the best accuracy@5**, there is a non-negligible proportion of clients which are better under IFCA (∼15%) in the long lower tail. That is, **IFCA manages to improve a significant proportion of the worst performing clients**. Moreover, both approaches (IFCA and PMF) seem to be quite competitive on APR. **All in all, considering both accuracy and APR, as well as the delta improvement, PMF seems to have an edge over IFCA and WeCFL for higher entropy users. The rationale behind this result is that higher entropy clients often have sufficient data, which allows them to fully utilize PMF.**

TABLE V
PERCENTAGE OF CLIENTS WITH POSITIVE (↑) AND NEGATIVE (↓) DELTA APR. THE HIGHEST ACCURACY GAIN COMPARED TO FEDAVG PER GROUP PER DATASET ARE UNDERLINED.

Datasets	Strategy	Low	Med.	High
GeoLife	**CFL**	↑35.48% ↓61.29%	↑45.16% ↓54.84%	↑54.84% ↓45.16%
	FL+HC	↑41.94% ↓54.84%	↑50.00% ↓50.00%	↑74.19% ↓25.81%
	IFCA	↑<u>83.9</u>% ↓12.9%	↑45.2% ↓51.6%	↑<u>77.4</u>% ↓16.1%
	WeCFL	↑45.2% ↓51.6%	↑19.7% ↓0.0%	↑25.8% ↓1.6%
	PMF	↑51.61% ↓45.16%	↑<u>59.68</u>% ↓40.32%	↑74.19% ↓25.81%
MDC	**CFL**	↑14.63% ↓82.93%	↑7.41% ↓92.59%	↑9.76% ↓90.24%
	FL+HC	↑12.20% ↓87.80%	↑11.11% ↓88.89%	↑12.20% ↓87.80%
	IFCA	↑56.1% ↓43.9%	↑49.4% ↓13.7%	↑<u>80.5</u>% ↓19.5%
	WeCFL	↑14.6% ↓53.7%	↑34.2% ↓0.0%	↑21.0% ↓4.9%
	PMF	↑<u>90.24</u>% ↓9.76%	↑<u>95.06</u>% ↓4.94%	↑78.05% ↓21.95%

b) Medium Entropy Group: From Table IV, we observe that both CFL and FL+HC fail to show improvement in accuracy@5, especially on MDC (*e.g.*, 0% for FL+HC). For this group, IFCA and WeCFL seem to show an improvement. However, it simultaneously comes with a considerable proportion of users that experience a decline. For instance, in the GeoLife dataset, IFCA 59.7% users for the accuracy@5 while it declines 19.4%. Similar results can seen in the MDC dataset. In comparison, PMF appears to (almost) systematically im-

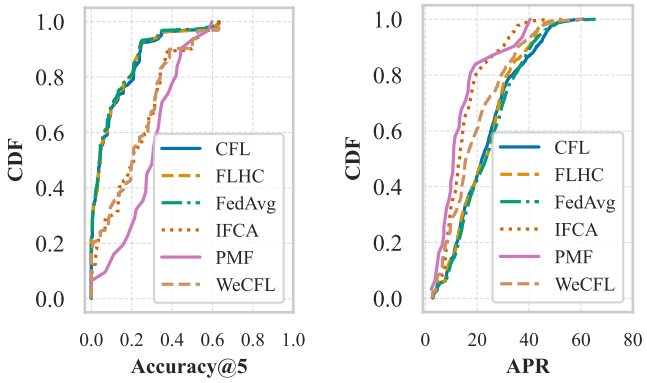

Fig. 2. Accuracy@5 and APR for High Entropy Group on Geolife.

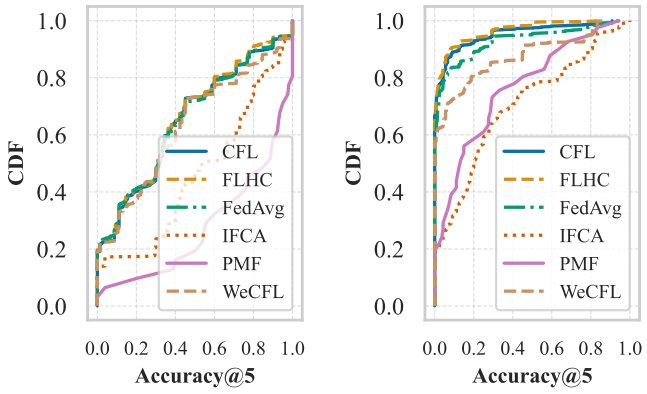

Fig. 3. Accuracy@5 for Low Entropy Group Low on Geolife (left) and MDC (right).

proves the users in this group. For instance, it increases the accuracy@5 of 81.48% while only declining 3.7% users. A similar observation can be for the APR (See Table V) where we note an improvement of up to 95.06% users. Although this might disfavour CFL approaches compared to PMF for medium entropy users, a closer examination of the CDF distribution of the metrics shows otherwise (See Figure 5 for CDF of Accuracy@5 and APR on MDC). **Indeed, we note that IFCA is better in accuracy than PMF while all CFL approaches and even FedAvg outperform PMF on APR**. This indicates that while PMF manages to improve a significant number of users, it does not do so by a large margin. In other terms, CFL works, and especially IFCA and WeCFL improves a smaller proportions of users with larger margins. This is likely to be a consequence of the difficulty to cluster some medium entropy clients, which are not sufficiently unpredictable (*e.g.*, high entropy group) nor sufficiently predictable (*e.g.*, clients visiting popular POIs) to find themselves in an ideal cluster. However, when an ideal cluster is found, users benefit largely from CFL. **To conclude, accounting for users that cannot be clustered ideally could strengthen CFL.**

c) Low Entropy Group: Table IV shows that low entropy users sensibly improve their accuracy@5 under PMF with 83,87% and 63.41% of on Geolife and MDC, respectively. It is followed closely by IFCA, which experiences 58.1% and 48% improvement for these datasets. The disparity in improvement, compared to the relative proximity in improvement noticed for the high entropy group or the average performance can be attributed to two factors: i) First, the nature of low entropy users. Indeed, as these users are more predictable, the model is less likely to encounter test POIs that were not seen in the training. **This makes locally personalized models such as PMF ideal for low-entropy users. Nevertheless, users must still possess sufficient data to locally personalize these models**, which leads to the second point. ii) The improvement quantified the proportion of users but not the margin of improvement for individual, which as seen for previous groups can incorporate a bias. To account for this, we illustrate in Figure 3 the accuracy@5 CDF for the low entropy group. **Interestingly, we note that IFCA is much more competitive with PMF on Geolife and outperforms it on MDC.** Similar patterns can be observed in the APR results illustrated in Table V. However, we note that the the gap in improvements between PMF and IFCA is less significant. In fact, IFCA provides the best improvement for lower-entropy on Geolife with up to 83.9% compared to 51.61 for PMF. As explained in RQ2, this stems mainly from the generalization abilities of IFCA compared to PMF, which cannot be captured by the accuracy@5. Moreover, we note more significant improvements for other CFL methods on the APR, suggesting that these latter show interesting generalization abilities too, in spite of their parameter-wise clustering limits. To better observe this phenomenon, we illustrate in Figure 4 the CDF of APR for the low entropy group. The latter results not only demonstrate the superiority of IFCA on APR, but also reveal that the improvement in most other CFL methods is more evenly distributed between clients compared to PMF, as with lower average improvement, these methods have often better individual, especially for the long tail users (users having APR between 0 and 20). In fact, PMF appears to perform the worst in this experiment. **All in all, considering both accuracy and APR, IFCA appears to be the most promising approach, followed by WeCFL. This result shows that CFL can perform well even for more predictable users, especially in highly sparse environments.**

VI. Discussion

Our analysis reveals that CFL approaches are promising directions for FL paradigm but with most benefit for users in the high and low entropy groups, where they consistently provide large improvements in accuracy and APR metrics. Indeed, future work on augmenting CFL approaches such that they optimize for parity in accuracy for all user groups may include integration of re-reinforcement learning into the clustering strategy. Across different CFL approaches, we see variations in performance, with IFCA outperforming the others. Furthermore, IFCA outperforms personalized FL approach, PMF in terms of APR, making it a viable choice for next point

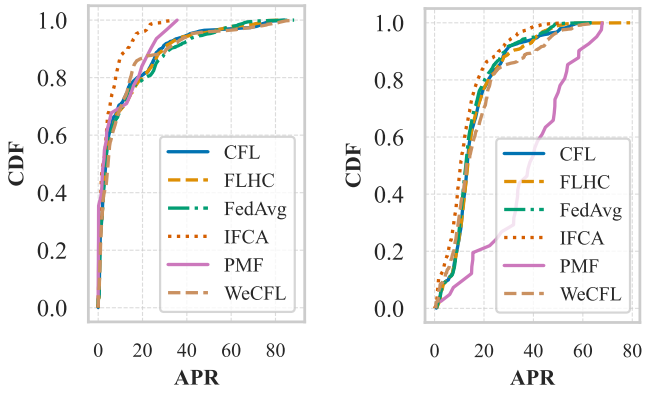

Fig. 4. Comparison of APR for Low Entropy Group on Geolife (left) and MDC (right).

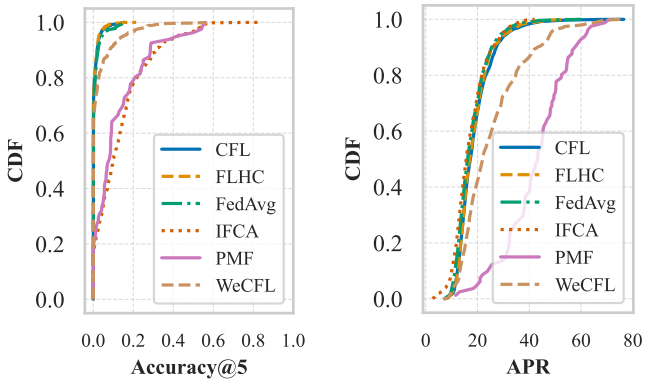

Fig. 5. Accuracy@5 and APR CDF for Medium Entropy Group on MDC.

prediction tasks. Our results also show that IFCA is more robust and generalizable across user groups and datasets.

It is worth noting that in this work we studied a simple underlying prediction model that is *light-weight for smartphone deployment,* as opposed to those more complex ones such as GRU [36] and DeepMove [37]. The choice of this model was primarily for us to be able to observe the impact of CFL on a low-resource predictive model. Our results have shown that even with such a basic model, CFL approaches can offer comparative results to PMF. We believe a more complex model with additional layer could enhance the performance gain of the CFLs even further, albeit with increasing computational complexity on the user's device. To this end, our open-source framework is designed to allow researchers easily evaluate their model and baseline against CFL approaches.

In releasing our framework, we take a big step in encouraging the research community to contribute their models in order to enable standardization and benchmarking next point prediction approaches. Our framework which is based on Flower [38], allows for integration of any model be it in Pytorch or TensorFlow, and its agile functionality allows researchers to use any spatial temporal datasets.

VII. Conclusion

Our study evaluates Clustered Federated Learning (CFL) for next-point prediction with non-IID trajectory data, comparing four state-of-the-art CFL methods against personalized FL (PMF) on the GeoLife and MDC datasets. We find that CFL, particularly IFCA, performs strongly for high- and low-entropy user groups, improving both accuracy and APR while demonstrating superior generalization in sparse data environments where PMF tends to overfit. These results establish CFL as a competitive alternative to personalized FL, with IFCA emerging as the most effective approach. To support further research, we open-source our benchmarking framework for evaluating CFL and FL models on spatiotemporal datasets.

VIII. Acknowledgment

This work was generously supported by the United States National Science Foundation (NSF) award IIS-2304213 and OISE-1853953.

References

[1] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," 2021.

[2] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] D. L. Ferreira, B. A. Nunes, C. A. V. Campos, and K. Obraczka, "A deep learning approach for identifying user communities based on geographical preferences and its applications to urban and environmental planning," *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, vol. 6, no. 3, pp. 1–24, 2020.

[4] A. Graser, A. Jalali, J. Lampert, A. Weißenfeld, and K. Janowicz, "Mobilitydl: a review of deep learning from trajectory data," *GeoInformatica*, pp. 1–33, 2024.

[5] Q. Long, H. Wang, T. Li, L. Huang, K. Wang, Q. Wu, G. Li, Y. Liang, L. Yu, and Y. Li, "Practical synthetic human trajectories generation based on variational point processes," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2023, pp. 4561–4571.

[6] X. Kong, Q. Chen, M. Hou, H. Wang, and F. Xia, "Mobility trajectory generation: a survey," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 3057–3098, 2023.

[7] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[8] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, 2013.

[9] C. E. J. Ezequiel, M. Gjoreski, and M. Langheinrich, "Federated learning for privacy-aware human mobility modeling," *Frontiers in Artificial Intelligence*, vol. 5, p. 867046, 2022.

[10] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, "Pmf: A privacy-preserving human mobility prediction framework via federated learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–21, 2020.

[11] S. Wang, B. Wang, S. Yao, J. Qu, and Y. Pan, "Location prediction with personalized federated learning," *Soft Computing*, pp. 1–12, 2022.

[12] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3710–3722, 2020.

[13] Z. Ye, X. Zhang, X. Chen, H. Xiong, and D. Yu, "Adaptive clustering based personalized federated learning framework for next poi recommendation with location noise," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, p. 1273.

[15] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

[16] J. Ma, G. Long, T. Zhou, J. Jiang, and C. Zhang, "On the convergence of clustered federated learning," *arXiv preprint arXiv:2202.06187*, 2022.

[17] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[18] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 312–321.

[19] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen *et al.*, "The mobile data challenge: Big data for mobile computing research," in *Pervasive computing*, 2012.

[20] Y. Guo, F. Liu, Z. Cai, H. Zeng, L. Chen, T. Zhou, and N. Xiao, "Prefer: Point-of-interest recommendation with efficiency and privacy-preservation via federated edge learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–25, 2021.

[21] Y. Belal, A. Bellet, S. B. Mokhtar, and V. Nitu, "Pepper: Empowering user-centric recommender systems over gossip learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–27, 2022.

[22] C. Chen, J. Zhou, B. Wu, W. Fang, L. Wang, Y. Qi, and X. Zheng, "Practical privacy preserving poi recommendation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–20, 2020.

[23] A. Li, S. Wang, W. Li, S. Liu, and S. Zhang, "Predicting human mobility with federated learning," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 441–444.

[24] S. Vahidian, M. Morafah, C. Chen, M. Shah, and B. Lin, "Rethinking data heterogeneity in federated learning: Introducing a new notion and standard benchmarks," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 3, pp. 1386–1397, 2023.

[25] Y. Shi, Y. Zhang, Y. Xiao, and L. Niu, "Optimization strategies for client drift in federated learning: A review," *Procedia Computer Science*, vol. 214, pp. 1168–1173, 2022.

[26] J. H. Faghmous and V. Kumar, "Spatio-temporal data mining for climate data: Advances, challenges, and opportunities," *Data mining and knowledge discovery for big data: Methodologies, challenge and opportunities*, pp. 83–116, 2014.

[27] Y. Belal, S. Ben Mokhtar, H. Haddadi, J. Wang, and A. Mashhadi, "Survey of federated learning models for spatial-temporal mobility applications," *ACM Transactions on Spatial Algorithms and Systems*, 2024.

[28] Z. Fan, X. Song, R. Jiang, Q. Chen, and R. Shibasaki, "Decentralized attention-based personalized human mobility prediction," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, pp. 1–26, 2019.

[29] J. E. C. Elizondo, M. Gjoreski, and M. Langheinrich, "Federated learning for privacy-aware human mobility modeling," *Frontiers in Artificial Intelligence*, vol. 5, p. 867046, 2022.

[30] X. Rao, L. Chen, Y. Liu, S. Shang, B. Yao, and P. Han, "Graph-flashback network for next location recommendation," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 1463–1471.

[31] Y. Huang, Y. Liang, and L. Huang, "Provable generalization of over-parameterized meta-learning trained with sgd," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 563–16 576, 2022.

[32] S. Zhang, J. Guo, C. Liu, Z. Li, and R. Li, "Next point-of-interest recommendation for cold-start users with spatial-temporal meta-learning," in *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*. IEEE, 2023, pp. 80–87.

[33] Y. Cui, H. Sun, Y. Zhao, H. Yin, and K. Zheng, "Sequential-knowledge-aware next poi recommendation: A meta-learning approach," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 2, pp. 1–22, 2021.

[34] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[35] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[36] D. Liao, W. Liu, Y. Zhong, J. Li, and G. Wang, "Predicting activity and location with multi-task context aware recurrent neural network." in *IJCAI*, 2018, pp. 3435–3441.

[37] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "Deepmove: Predicting human mobility with attentional recurrent networks," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1459–1468.

[38] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[39] Z. Zhou, Y. Yang, X. Yang, J. Yan, and W. Hu, "Going beyond linear mode connectivity: The layerwise linear feature connectivity," *Advances in Neural Information Processing Systems*, vol. 36, pp. 60 853–60 877, 2023.

[40] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," *Advances in neural information processing systems*, vol. 31, 2018.