frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 27 July 2022
DOI 10.3389/fcomp.2022.891206

OPEN ACCESS

EDITED BY
Antonino Crivello,
National Research Council (CNR), Italy

REVIEWED BY
Jerry Chu-Wei Lin,
Western Norway University of Applied Sciences, Norway
Francesco Furfari,
National Research Council (CNR), Italy

*CORRESPONDENCE
Martin Gjoreski
martin.gjoreski@usi.ch

SPECIALTY SECTION
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

RECEIVED 07 March 2022
ACCEPTED 04 July 2022
PUBLISHED 27 July 2022

CITATION
Gjoreski M, Laporte M and Langheinrich M (2022) Toward privacy-aware federated analytics of cohorts for smart mobility. *Front. Comput. Sci.* 4:891206. doi: 10.3389/fcomp.2022.891206

Toward privacy-aware federated analytics of cohorts for smart mobility

Martin Gjoreski*, Matías Laporte and Marc Langheinrich

Faculty of Informatics, Università della Svizzera italiana (USI), Lugano, Switzerland

Location-based Behavioral Analytics (LBA) holds a great potential for improving the services available in smart cities. Naively implemented, such an approach would track the movements of every citizen and share their location traces with the various smart service providers—similar to today's Web analytics systems that track visitors across the web sites they visit. This study presents a novel privacy-aware approach to location-based federated analytics that removes the need for individuals to share their location traces with a central server. The general approach is to model the behavior of cohorts instead of modeling specific users. Using a federated approach, location data is processed locally on user devices and only shared in anonymized fashion with a server. The server aggregates the data using Secure Multiparty Computation (SMPC) into service-defined cohorts, whose data is then used to provide cohort analytics (e.g., demographics) for the various smart service providers. The approach was evaluated on three real-life datasets with varying dropout rates, i.e., clients not being able to participate in the SMPC rounds. The results show that our approach can privately estimate various cohort demographics (e.g., percentages of male and female visitors) with an error between 0 and 8 percentage points relative to the actual cohort percentages. Furthermore, we experimented with predictive models for estimating these cohort percentages 1-week ahead. Across all three datasets, the best-performing predictive model achieved a Pearson's correlation coefficient above 0.8 (strong correlation), and a Mean Absolute Error (MAE) between 0 and 10 (0 is the minimum and 100 is the maximum). We conclude that privacy-aware LBA can be achieved using existing mobile technologies and federated analytics.

KEYWORDS

federated analytics, privacy, location-based services, mobility modeling, mobile computing

Introduction

Motivation

Location-based Behavioral Analytics (LBA) is widely recognized as being key to providing new services and solutions in many application domains. For example, changes in behavior can be used to recognize impending mental health episodes (Mehrotra et al., 2016), deliver more effective advertising and retail experiences (Krüger et al., 2011), enhance security (Crossler et al., 2013) and shape the provision of urban

services (Mazhar et al., 2016). Within the purely digital domain of the Web, "behavioral analytics" are enabled through the use of cookies that provide the mechanism for tracking individual user interactions. Such user-centric analytics have helped drive forward rapid innovation in the design and deployment of web content—enabling site owners to understand visitor behavior. Google Analytics and similar platforms have also been extended to support mobile analytics that provide data on how users interact with iOS and Android Apps (e.g., install, launch and deletion events). However, while there are numerous application-specific initiatives to measure elements of human behavior (typically using mobile phones), we still do not have viable mechanisms for service providers to track user interactions in the real world without raising significant privacy concerns (Benjamin and Musolesi, 2020).

LBA through user-specific mobility modeling has been studied for a long time. In the past, the focus has been on analyzing frequent mobility patterns using Markov models (Ashbrook and Starner, 2003). Besides the Markov-based approaches, machine learning (ML) approaches have been utilized to predict future movements. The initial ML approaches were based on statistical time- and frequency-based features, extracted from the mobility trajectories (Baumann et al., 2013). The ML algorithms include Decision Trees (Monreale et al., 2009), Support Vector Machines (Bhaskar and La Porta, 2015), Random Forest (Do and Gatica-Perez, 2014), etc. The most recent approaches for mobility modeling replaced the Markov- and feature-based approaches with end-to-end deep learning based on Recurrent Neural Networks (RNNs). For example, DeepMove (Jie et al., 2018), RNN+SAtl (Jun et al., 2019) and Flashback (Dingqi et al., 2020), are deep learning architectures that learn time and location embeddings and apply Recurrent Neural Networks (RNNs) for predicting the next place visited by a user.

A disadvantage of all these methods is the requirement of centralized data i.e., the training data must be available at one place, rendering all these methods questionable with regards to user privacy. After all, the EU General Data Protection Regulation (GDPR) has emphasized the importance of location and movements data by including it in the definition of personal data[1]. One possible solution for the privacy-problem is to use federated ML approaches, where the users' privacy is guaranteed by implementing one simple rule: "No personal data leaves the user's device." In these approaches, each user device acts as a separate computational unit, processing only its local data (e.g., local GPS traces), and sharing globally only the results of the computation—which in federated ML approaches are abstract model updates. An example for such approach is PMF (Jie et al., 2020)—another RNN-based next-place predictor trained using Federated Learning (Andrew et al., 2018). However, a downside of PMF is its requirement of having access to large training datasets and the need for specialized hardware (e.g., GPUs)—just as its centralized cousins DeepMove and RNN+SAtl.

Proposed approach and contributions

Cohort-based modeling, instead of user-specific modeling, is an alternative way of modeling human behavior. By modeling the behavior of cohorts, the user is removed from the focus of the processing pipelines. The personal data from one user is just one data point which is anonymously aggregated within the pool of cohort data, thus location k-anonymity (Sweeney, 2002) is enabled by default (k is the size of the cohort), enabling privacy-aware analytics. Cohort-based mobility modeling has been previously proposed in another context. Jane et al. (2016) analyzed residential trajectories of older men and women born between 1918 and 1947 with regards to socio-historical contexts. In the context of human mobility, such group-based approaches have been proposed to model the distribution of human trajectories between larger areas (e.g., city blocks, cities, countries, and continents). An example of such models is the Exploration and Preferential Return (EPR) model. EPR is a statistical model that does not learn from actual mobility trajectories, but it rather uses equations that depend on parameters such as: waiting time, action selection, exploration phase and return phase. While these statistical models are powerful and widely used for urban mobility planning on a larger scale, their applicability on a smaller scale (e.g., understanding visitors to a specific restaurant or a specific place in town where a smart display has been placed) is limited.

The contribution presented in this study is a novel method for privacy-aware federated analytics of cohorts for smart mobility. The three main characteristics that distinguish our method are: (i) cohort-based modeling instead of user-specific modeling; federated approach, i.e., the privacy-sensitive data stays on the device; and (iii) online learning i.e., the models are continuously refined through time. Other characteristics that define our method: it works both for small datasets (e.g., 80 users) and large datasets (e.g., 65k users); it works both for continuous GPS sensing and on-demand sensing; It does not require specialized hardware (e.g., GPUs); and, it is based on an online-learning approach, where updates are communicated to the system daily. Finally, to the best of our knowledge, this is the first method that provides privacy-aware cohort-based analytics based on mobility data.

The rest of this paper is structured as follows: Section Related work summarizes the existing work in related scientific fields. Section Datasets presents the three experimental datasets used in this study. Section Methods describes the proposed method. Section Experiments presents the experimental setup, the experimental results and example federated analytics of cohorts generated by the proposed method. Section Limitations and future work presents limitations and future improvements.

1 https://eur-lex.europa.eu/eli/reg/2016/679/oj

Section Implications presents the implications this study brings. Finally, Section Conclusion concludes the paper.

Related work

This section first presents a more general overview of the related work on LBA, then focuses on three more specific fields which deal with predictive models for human mobility based on GPS and/or cell data, and the final sub-section presents a summary comparison of the related work methods. The field of next-place prediction offers the most advanced approaches; however, these approaches are focused on modeling the movement of each specific user, are mostly centralized, require specialized hardware for the design and the training of the models (e.g., GPUs) and are quite complex. All these characteristics limit the utility of the next-place predictors for real-life applications. The Federated Learning field solves the problem of data centralization, but the method complexity and the requirement for specialized hardware is still a challenge. The third field deals with the task of counting visitors in a specific place. These approaches are much simpler than the next-pace predictors, e.g., some of them are only based on correlation analysis, and all of them are based on centralized data analytics.

Modeling human behavior for analytics

The ubiquity of mobile phones has significantly changed our understanding of human mobility—and consequently our use of this knowledge in the form of mobility models. Due to the density of today's communication infrastructure (dense urban areas may have cell towers every 200 m), mobile phone operators are able to track a subscriber's path through a city with block-level accuracy. Self-tracked systems that use GPS (e.g., Google Maps) may allow service providers even more fine-grained tracking capabilities. Mobility models allow providers to characterize these traces, e.g., for predicting future activities of individuals to optimize network handovers, or to simulate large-scale network load. Hess et al. (2015) define a mobility model as "a simplified representation of the movement of single or groups of mobile entities in a given context, primarily the spatial environment" (Hess et al., 2015). While mobility models can be entirely synthetic—with start and end points, as well as waypoints, speed, and pauses chosen at random or within a defined set of parameters—a more realistic approach is to base them on actual mobility traces (trace-based models). Several anonymous large-scale traces are available for research, such as the GeoLife dataset (Yu et al., 2010). Data-driven models more accurately reflect actual human behavior, as well as real-world topographies (e.g., streets, bus lines).

A plethora of data-driven mobility models have been suggested. The key differentiation between them lies in the set of features they use and the type of predictions they are designed to make. In the context of this study, which targets methods that can model users in the spatiotemporal domain, a range of models from the literature offer themselves as starting points, e.g., HERMAS (Yiwei et al., 2021), which enables trajectory similarity measurement and user profiling using cellular signaling data; CallSense (Zhihan et al., 2021), which enables a recovery of sparse cellular data for modeling human mobility; a study by Tongqing et al. (2021), which used photo crowdsensing to model human mobility. Furthermore, human mobility modeling often includes modeling transportation modes (e.g., in car, in bus) (Gjoreski et al., 2020).

Note that all these models are based on offline learning, i.e., they ingest a fixed (large) set of mobility traces and use this to train the respective model. In this study, we focus on alternatives to central data collection and instead investigate approaches that support online learning, i.e., incrementally refining mobility models.

Next-place prediction

Next-place prediction is a research field with the goal of predicting where a user will go next, focusing on modeling movements of individual users. A detailed review this field is presented in the survey papers by Schreckenberge et al. (2018) and Massimiliano et al. (2020), with the latter focusing explicitly on deep learning methods.

A sub-task in next-place prediction is the automatic identification of the places (or POIs), especially in datasets collected *via* continuous GPS sensing. A typical approach to identify places from continuous streams of GPS data is to use clustering methods. The identified clusters are then related to specific places where the users spent a certain amount of time (e.g., at least 5 min in a radius of 250 meters). For example, Ashbrook and Starner (2003), used a modified k-means method and Adams et al. (2006) used DBSCAN. In this study, we modified an existing spatiotemporal clustering method available in scikit-mobility (Pappalardo et al., 2019), which is based on the DBSCAN clustering algorithm (Martin et al., 1996).

Once the POIs are well-defined, a variety of mobility modeling approaches can be used. Ashbrook and Starner used first-order Markov model to predict user-specific future movements (Ashbrook and Starner, 2003). Imai et al. (2018) proposed an interesting improvement of Markov-based approaches, where the set of possible places to be visited narrows down as the trip progresses. Their study included GPS data of 1,646 users from commercial services.

Regarding feature-based ML approaches, Baumann et al. (2013) analyzed a variety of spatial (e.g., current location and previous location) and temporal features (e.g., day of the week and weekday/weekend) as possible next-place predictors using data of 37 users collected over a period of 1.5 years.

Nitin et al. (2015) proposed a multi-level approach by predicting the semantics of a place and then the specific place to be visited. The feature set used in these studies include current location, last call, hour of the day, day of the week and applications used. Similarly, Bhaskar and La Porta (2015) used start minute, end minute, normalized start time, and other related features, as input to a Support Vector Machine classifier. Etter et al. (2012) compared a variety of methods, including a majority classifier (35% accuracy), first-order Markov model (44% accuracy), deep belief network (60.7% accuracy), neural network (60.83% accuracy) and gradient boosting trees (57.63% accuracy).

The most recent and advanced next-place predictors are based on end-to-end deep learning methods. ST-RNN (Spatial Temporal Recurrent Neural Networks), DeepMove (Jie et al., 2018), RNN+SAtl (Jun et al., 2019) and Flashback (Dingqi et al., 2020), are all based on RNNs or their variations (e.g., LSTMs or GRUs). DeepMove is an attentional RNN specifically designed to address the problem of sparse trajectories. The method utilizes multi-modal embedding layers to create a dense representation of the spatiotemporal trajectories and user-specific features (thus, it is a user-dependent model). Additionally, the embeddings of the historical trajectories are processed by an attention mechanism to extract mobility patterns, while a GRU processes current trajectories. The output of the multi-modal embedding, the GRU, and the attention mechanism are concatenated and passed to a fully connected layer that provides the final output (next-place prediction). While being state-of-the-art for modeling human mobility, these methods were developed and tested using centralized approaches, using specialized hardware (e.g., GPU for deep learning model), using large training datasets (e.g., 65 thousand users) and they are offline methods.

Federated learning and privacy preservation

Federated Learning (FL) is an iterative technique where each device trains a personalized model on the device itself, and only shares the weights of the trained model, thus protecting the user data (Andrew et al., 2018). The personalized models are then anonymously aggregated by a server to a general model. The general model is then communicated back to the devices again, where each device can either use the general model or perform new updates over the general model using more recent personal data. FL has been used in a variety of domains. Yuanyishu et al. (2022) developed federated BERT—a large-scale language model used for natural language processing. Ittai et al. (2021) applied FL on COVID-19 data from twenty medical institutions to develop a federated model that predicts the future oxygen requirements of symptomatic patients. Usman et al. (2022) explored FL for edge intelligence applied in the domain of customer segmentation using sales data from an online store.

In the domain of FL for human mobility modeling, some privacy concerns related to the next-place predictors can be mitigated using federated next-place predictors (Zipei et al., 2019; Jie et al., 2020). Nevertheless, these methods still require large, labeled datasets including thousands of users and require specialized hardware (GPUs) in the development process (Castro et al., 2022). This is probably the reason why large deep learning models are only evaluated using train-test splits (e.g., train of the first 50% of the data and test on the last 50% of the data), which does not correspond to a real-life usage where a new model is updated every day (online learning).

Counting visitors

Henrikki et al. (2017) analyzed data from Instagram, Twitter, and Flickr, to estimate visitor statistics in 56 national parks in Finland and South Africa in 2014. Hamstead et al. (2018) explored visitation dynamics in New York City parks using Twitter and Flickr data. In both studies, an association between the social media data and the actual visitor counts was found, although the strength of the association depended on the social medial platform, e.g., the models based on Instagram data outperformed the models based on Twitter and Flickr data (Henrikki et al., 2017). In addition, the daily average number of observed visitors in New York City parks had a Pearson's correlation coefficient of 0.58 with the daily average number Flickr users, and a correlation coefficient of 0.76 with the daily average number of Twitter users (Hamstead et al., 2018). Interestingly, their analysis showed that parks with greater areas of green space get fewer visitors, and proportion of minority ethnicity and minority race in the neighborhoods of the parks is also negatively correlated with the number of visitors.

Similarly, Fisher et al. (2018) explored Flickr images and trip reports shared on a hiking forum, for counting visitors at recreational areas in USA. Their analysis showed that correlations between official Forest Service statistics and geo-tagged images ranged between 0.55 and 0.95. For individual trails, monthly visitor counts from on-site measurements were correlated with counts from geo-tagged images (0.79) and trip reports (0.91).

Nathaniel et al. (2020) used smartphone location data in combination with weather data to estimate visitor count for 500 water recreation centers in USA. They tested linear models and Random Forest, with Random Forest showing best results. Takahiro et al. (2020) used cellular data to calculate the economic value of coastal tourism. They analyzed 536 places (beaches) across Japan, but they did not provide ground truth for the analysis. Jung et al. (2020) used cell data to analyze tradeoffs between visitation and biodiversity

for an island park in Korea. Their analysis showed moderate correlations between the cell data and monthly estimates of visitation to several specific locations on the island (Pearson's correlation coefficient of 0.64). Christopher et al. (2019) developed techniques for processing, sampling and calibration that can be applied on cell data for counting vehicles in parks in California. They statistically compared monthly estimates produced by their model and direct counts and found no significant differences.

The methods presented in these studies are rather simple (mostly based on correlation analysis) and yet showed that location-based analytics can be useful in a variety of scenarios, from recreational tourism to counting vehicles in the park. The findings in these studies further motivate the need for cohort-based behavior analytics which can shed a light not only on the visitor counts, but also on the aggregated profile (cohort type) of the visitors, which is much more informative. Additional improvement to the centralized approaches presented in these studies, i.e., location data from all users available at one place, our study presents one step toward privacy-aware federated analytics by creating decentralized models updated daily (online learning).

Related work summary

Table 1 presents a summary comparison between our proposed approach and the related-work methods. From the table, it can be seen that the two main characteristics that distinguish our method are the federated approaches (i.e., user-location data stays on the device), and online learning (i.e., the models are continuously refined through time). Theoretically, online learning should be possible also for the three federated methods (PMF; Zipei et al., 2019; Jie et al., 2020; Castro et al., 2022); however, these methods were only evaluated using static train-test splits (e.g., train of the first 50% of the data and test on the last 50% of the data), which does not correspond to real-life usage. Furthermore, these methods are based on deep learning approaches that require large, labeled datasets including thousands of users and require specialized hardware (GPUs) in the development process. Another important characteristic of our proposed method is that it can work with small datasets (e.g., 80 users in our experimental setup) and with big datasets (e.g., 65k users in our experimental setup). The final important characteristic is that it can work in on-demand sensing scenarios (e.g., Foursquare check-ins) and continuous sensing scenarios. For the continuous sensing scenario, an additional step is required to automatically discover the POIs, i.e., the spatiotemporal clustering that is part of our method. Finally, none of the related methods provide privacy-aware cohort-based analytics.

Datasets

We used three separate datasets in this study: the Breadcrumbs dataset (Arielle et al., 2019), which is collected by 80 smartphone users via continuous GPS sensing; the Foursquare dataset, which is generated by 65 thousand Foursquare users; and the Gowalla dataset which is generated by 319 thousand users. The main difference among the datasets, besides their size, is that Breadcrumbs is collected with a sampling frequency close to 1 Hz, whereas the data in Foursquare and in Gowalla is collected *via* "on-demand" sensing (check-ins).

Breadcrumbs dataset

The Breadcrumbs dataset was introduced by Arielle et al. (2019) in 2019. The dataset was collected by 80 smartphone users, mainly in the city of Lausanne (Switzerland), for a period of 94 days. It contains data from continuous GPS sensing, user demographic information, POIs (e.g., longitude and latitude for each POI) and user-supplied semantic labels for the POIs (e.g., university, home, restaurant, etc.). The motivation behind the data collection campaign was to advance the research in fields such as next-place prediction, trajectory prediction, privacy preserving location-based services, and supervised and unsupervised detection of points of interests.

To define cohorts of users, we used the demographic information available in the dataset. We defined threes cohorts based on the university the users attended, i.e., UNIL: 44 (55%); EPFL: 29 (36%); Other: 7 (9%). Besides these, other cohorts can be defined based on the demographic information, including age, gender and nationality.

To segment the continuous GPS data, spatiotemporal clustering algorithms are typically used to detect stop-points, i.e., places (defined by a radius size) where the users spent a certain amount of time (e.g., 5 min). In this study, we modified an existing spatiotemporal clustering method available in scikit-mobility (Pappalardo et al., 2019), which is based on the DBSCAN clustering algorithm (Martin et al., 1996). We refer to the existing spatiotemporal clustering method as ST-DBSCAN, and we refer to the modified method as Hierarchical Spatiotemporal DBSCAN (HST-DBSCAN). More details about the clustering methods are presented in Section Smartphone users, and an experimental comparison between ST-DBSCAN and HST-DBSCAN is presented with the experimental results (Section POI detection for continuous sensing).

Figure 1 presents the relative frequency of the semantic labels in the Breadcrumbs dataset obtained after using HST-DBSCAN. For example, the most frequent semantic label is Home (with a frequency close to 40%). The second and the third most frequent semantic labels represent the two main universities in Lausanne: the Swiss Institute of

TABLE 1 Comparison between the proposed approach and the related work.

Methods	Type	Federated	Online learning	Needs GPUs	Small data, $n <$ 1,000	Big data	Continuous sensing	On-demand sensing
Ashbrook and Starner (2003), Etter et al. (2012), Baumann et al. (2013), Bhaskar and La Porta (2015), Nitin et al. (2015), Imai et al. (2018)	Next-place predictors, classical machine learning	No	No	No	Yes	No	Yes	No
DeepMove (Jie et al., 2018), RNN+SAtl (Jun et al., 2019), Flashback (Dingqi et al., 2020)	Next-place predictors, deep learning	No	No	Yes	No	Yes	No	No
PMF (Zipei et al., 2019; Jie et al., 2020; Castro et al., 2022)	Next-place predictors, deep learning	Yes	No	Yes	No	Yes	Yes	Yes
Fisher et al. (2018), Christopher et al. (2019), Jung et al. (2020), Nathaniel et al. (2020), and Takahiro et al. (2020)	Counting visitors, majority use correlation analysis	No	No	No	Yes	Yes	Yes	Yes
Proposed approach	Cohort-based predictors	Yes	Yes	No	Yes	Yes	Yes	Yes

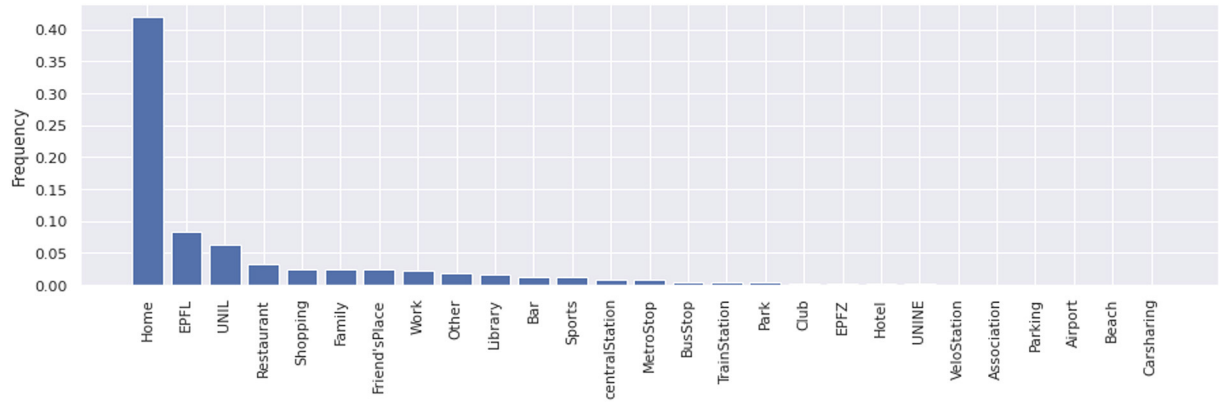

FIGURE 1
Semantic labels in Breadcrumbs and their relative frequency, obtained after HST-DBSCAN clustering.

Technology (EPFL) and the University of Lausanne (UNIL). Not surprisingly, the majority of the dataset was collected by EPFL and UNIL students.

Foursquare dataset

The Foursquare dataset (Yang et al., 2016), is a widely used dataset for the evaluation of location-based methods. The specific version of the dataset used in our study contains 18 months (April 2012 to September 2013) of global-scale check-in data collected from Foursquare. This dataset also contains user profiles, including the gender, the number of friends, and the number of followers the users have. The specific cohorts analyzed in this study were based on the user gender, which have the following distribution: Female: 25,061 (11%); Male: 168,327 (72%); Unknown: 41,207 (17%).

Figure 2 presents the 100 most frequently visited places in the Foursquare dataset. The y-axis presents the relative frequency (scaled by the overall number of check-ins in the dataset) and on the x-axis is the category. Besides the category, by using the longitude and latitude for these places one can also get more details about each of them. For example, the first two places are in the center of Istanbul (Turkey), the third place is a bridge also in Istanbul, the fourth one is a train station in Japan (Tokyo Station), and the fifth one is another train station in Japan (Shinjuku Station).

Gowalla dataset

This dataset was collected from Gowalla, a location-based social network, which had more than 600,000 users since November 2010 and was acquired by Facebook in December 2011. The dataset authors used the Gowalla APIs to collect the user profiles, user friendship, location profiles, and check-ins (Liu et al., 2014). The released dataset contains 36 million check-ins made by 319,063 users in 2.8 million locations. The locations in Gowalla are grouped into 7 main categories, i.e., community, entertainment, food, nightlife, outdoors, shopping and travel.

We defined four cohorts based on the number of friends each user has: (1) Number of friends in [0–25th percentile]; (2) Number of friends in [25th percentile to 50th percentile]; (3) Number of friends in [50th percentile to 75th percentile]; and (4) Number of friends above the 75th percentile. The 25th, 50th, and 75th percentiles were calculated using the overall dataset.

Methods

For a given list of POIs (e.g., a specific restaurant defined by GPS coordinates), our proposed method estimates the cohort percentages (e.g., male and females, students and workers, young and old, etc.) of the visitors in a privacy-aware manner. The estimations can be summarizations of historical visits, but also predictions about the future (e.g., week-ahead predictions). The estimations are performed using aggregated and anonymized historical data. The proposed method is depicted in Figure 3.

Smartphone users

At the beginning, the system shares a list of system-defined POIs (defined *via* GPS coordinates) for which cohort-based statistics is being collected. Next, the participants can select to which (system-defined) cohort they belong (e.g., their gender, occupation, or other demographics). For cohorts that require general statistics, e.g., what is the participant's rank compared to the rest of the users of the system, the system sends to all participants decision rules based on

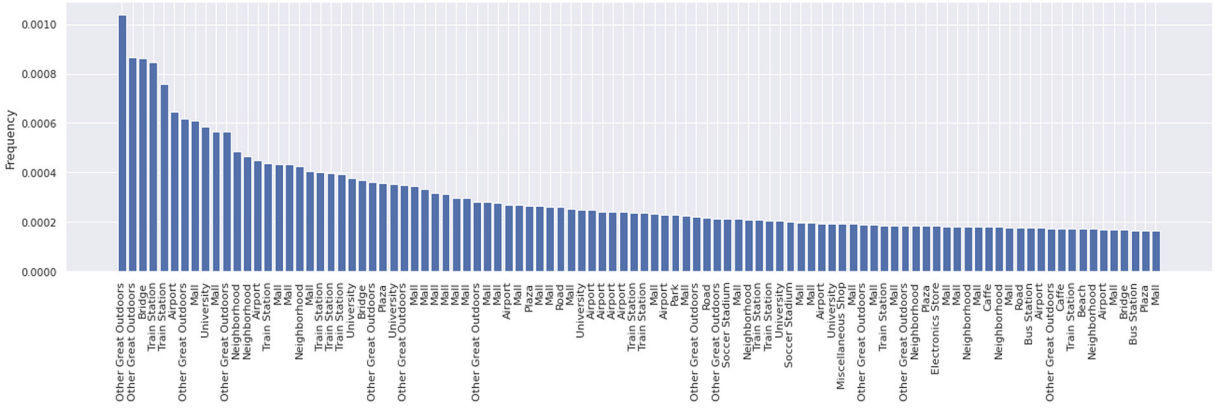

FIGURE 2
Top 100 places ranked by the number of check-ins scaled by the overall number of check-ins in the Foursquare dataset.

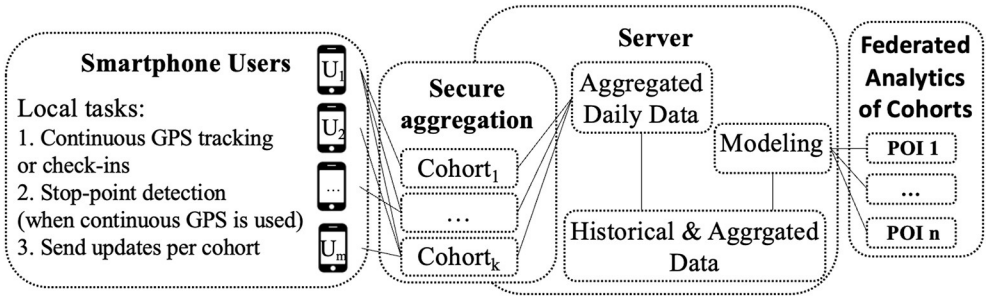

FIGURE 3
Proposed method for privacy-aware federated analytics of cohorts for smart mobility.

which the participant's cohort is defined. These rules can be as simple as, if you travel more than n km a day, you are in the top 25% of the participants. Furthermore, the participants can opt-out both from the cohorts and the POIs for which they do not want to share their data. This allows participants to inspect and control the data collection practices (Langheinrich, 2002).

Next, the participants should enable continuous GPS sensing locally on their device, or use manual check-ins. In the case of manual check-ins, the data is already segmented on POIs. In the other case, the continuous GPS data need to be segmented on POIs using spatiotemporal clustering method running locally on the device. For this task, we initially used the spatiotemporal clustering method based on DBSCAN (ST-DBSCAN) as implemented in scikit-mobility No Matches Found, which worked well in general, but it also produced clusters with big radii, mostly because ST-DBSCAN does not allow us to specify the maximum radius of a cluster. The algorithm does allow one to specify the maximum distance between any two points within the same cluster, which is different from a cluster radius. We thus implemented HST-DBSCAN. HST-DBSCAN utilizes ST-DBSCAN as a base method but goes one step further by ensuring that each cluster has a radius smaller than a predefined threshold. That threshold was set to 250 meters in this study because we wanted to focus on specific places (e.g., library, restaurant) and not regions. This is done by splitting clusters with bigger radii into clusters with smaller radii, which is a process that requires running ST-DBSCAN several times. This process produces a hierarchy of clusters and in this work, we used the clusters at the lowest level of the hierarchy. The radius of the clusters R was calculated as the maximum geodesic distance G between the cluster's median (defined with median longitude and median latitude) and any GPS point x within that cluster. G is the shortest distance on the surface of an ellipsoidal model of the earth (Charles, 2013).

$$R = \max\left(G\left(\left[Lon_{median},\ Lat_{median}\right], [Lon_x,\ Lat_x]\right)\right)$$

An experimental comparison between ST-DBSCAN and HST-DBSCAN is presented with the experimental results (Section POI detection for continuous sensing).

Server—secure aggregation and modeling

The goal of this step is to count the number of visitors per POI that belong to a specific cohort. This information is available locally on the participant's devices, i.e., we know when a specific participant visited a specific POI. Using secure multiparty computation, we can compute sums of the visitors for each POI. The specific protocol used in this study is "Practical Secure Aggregation" (Bonawitz et al., 2017) proposed by Google and implemented in the machine-learning platform TensorFlow Federated. The protocol was specifically designed for securely computing sums of vectors, and has a constant number of rounds, low communication overhead, robustness to failures, and requires only one server with limited trust. The server routes messages between the participants and calculates the final sum. From the final calculation, the server only learns the results, and it cannot learn the number of devices that participated in the computation. Formal privacy guarantees of the protocol are presented in the original study of the protocol (Bonawitz et al., 2017). This protocol has been also utilized for generating city-level location heatmaps (e.g., Manhattan) in a privacy-aware manner on (Eugene et al., 2021)—which is different than our study as we focus on fine-grained analysis based on smartphone sensing and online learning.

Dropout rate is an important characteristic that influences the computed sums. For example, at certain days, some devices may not be available for participating during the secure aggregation. This would cause an under-estimation of the number of visitors. That is why our method focuses on cohort percentage, and not the actual numbers of the visitors. For example, if we have 100 male participants and 50 female participants at a dropout rate of 10%, the secure aggregation will

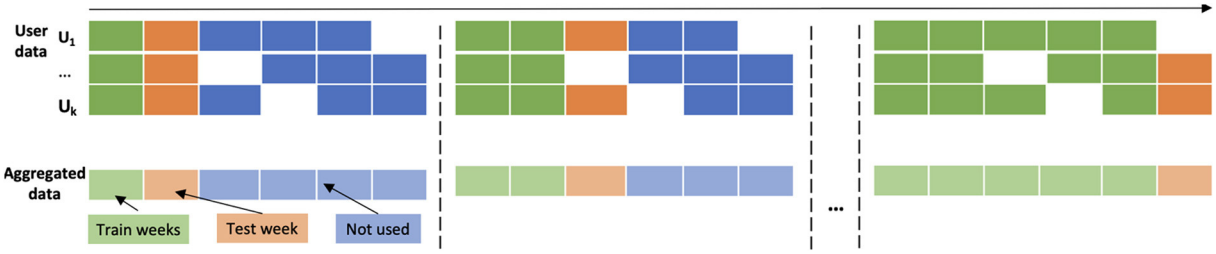

FIGURE 4
Evaluation technique used for evaluating week-ahead predictive models.

count 90 out of 100 male participants and 45 out of 50 female participants. This is under the assumption that the dropout rate is constant for the whole system across the related cohorts (10% in this example). Thus, the cohort percentages estimated with a 10% dropout rate would be equal to the cohort percentages estimated under perfect conditions (0% dropout rate). To put this example into numbers, see Equation (1):

$$females\ at\ 10\%\ dropout = \frac{45}{45+90} = 33\% = \frac{50}{50+100} = females\ at\ 0\%\ dropout \quad (1)$$

To estimate the actual error induced by the dropout rate, we performed experiments with varying dropout rate between 0 and 50% (see Section Dropout rate and error rate).

The cohort percentages are computed on a daily basis and are stored locally on the server for further modeling. The modeling can be simple summarizations of historical visits, but also predictions about the future. For example, we experimented with predictive models that estimate average cohort percentages 1-week ahead (see Section Next-week predictions).

Experiments

This section presents the study's experimental results. The first subsection presents the results from the automatic detection of POIs in the Breadcrumbs dataset. For the Foursquare and the Gowalla datasets, we performed experiments with the top 100 POIs, ranked by the number of overall check-ins. After the POI detection, all datasets were in a similar format, i.e., user, user cohort, place (POI) and time of visits.

In the second set of experiments (Section Dropout rate and error rate), we evaluated the relation between the dropout rate and the error for estimating the cohort percentages. For quantifying the error, we used Mean Absolute Error (MAE):

$$MAE\ (in\ percentage\ points) = \frac{1}{N}\sum_{1}^{N} |Acual\ Cohort\ Percentage - Estimated\ Cohort\ Percentage|$$

where N is the number of POIs in the specific dataset.

In the third set of experiments (Section Next-week predictions), we explored the possibility to predict the average cohort percentages 1-week ahead for each POI. For these experiments we tested four simple predictive models:

1. Autoregressive Integrated Moving Average (ARIMA)—POI-specific and Cohort-specific ARIMA models, i.e., the models were fitted for each POI and each cohort separately. Furthermore, the method's main parameters, auto-regressive order (p), the degree of differencing (d), and the moving-average order (q), were tuned for each model specifically using Akaike information criterion.
2. Overall mean predictor—POI-specific and Cohort-specific models that output the mean cohort percentage calculated from the historical data.
3. 3-weeks mean predictor—POI-specific and Cohort-specific models that output the mean cohort percentage calculated from the past 3 weeks.
4. 1-week mean predictor—POI-specific and Cohort-specific models that output the mean cohort percentage calculated from the past 1 weeks.

To evaluate the predictive models, we used an approach that represents a real-life, online usage:

- We used weekly evaluation depicted in Figure 4. In the N^{th} iteration, the models are fitted using the data of the previous N weeks and they are tested using the data of the (N^{th}+1) week. Thus, the models are using the overall historical data to predict next-week's average cohort

percentages. The same procedure was repeated for the three datasets.
- The models were fitted using noisy data, i.e., at each week we used a random dropout rate uniformly sampled from the interval [0–50%].
- The models' predictions were evaluated against the actual next-week cohort percentages.

Finally, Section Analytics of cohorts presents examples for federated analytics of cohorts.

POI detection for continuous sensing

To define POIs in the Breadcrumbs dataset, i.e., joint places where the users spend most of their time, we used the HST-DBSCAN clustering algorithm. We focused on joint POIs and not on personal POIs because joint POIs allow for privacy-aware cohort-based analytics, and because modeling of personal POIs may single-out users *via* their unique POIs (e.g., home location). The algorithm used the following parameters: 5 min waiting time (stop-points with a smaller duration were disregarded) and 250 meters maximum cluster radius. Figure 5 presents a comparison between the clusters generated with ST-DBSCAN and HST-DBSCAN. From the boxplots we see that HST-DBSCAN generates clusters with smaller radius, which should lead to "cleaner" POIs.

From the generated clusters, we selected only those that had more than 5 visitors per day on average. This led to 8 POIs depicted in Figure 6. These POIs will serve as example places for which the method can generate federated analytics. In a real-life scenario, such places could be predefined with specific GPS coordinates (e.g., central train station in a specific city, a specific restaurant, etc.), so a centralized clustering approach would not be required. However, even with predefined POIs, HST-DBSCAN may be needed to detect stop-points on the user device when continuous sensing is used.

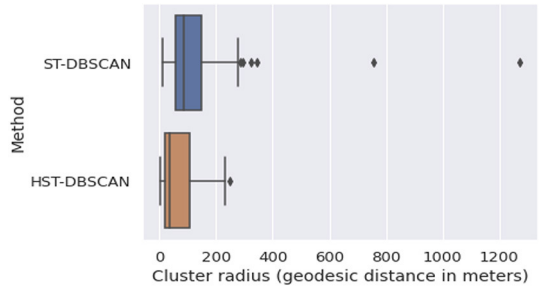

FIGURE 5
Cluster radius—ST-DBSCAN vs. HST-DBSCAN.

Dropout rate and error rate

The results of these experiments are presented in Figure 7. The figures in the left column depict the relation between the dropout rate and the MAE (in percentage points). The histograms in the right column depict the distribution of the cohort percentages calculated from the experimental data. A general observation for all datasets is that a bigger dropout rate causes a bigger MAE, which is expected. However, this relationship is sublinear, i.e., a dropout rate 50% does not cause a MAE of 50 percentage points, but rather the average MAE in most of the experiments is below 8 percentage points. For the Gowalla dataset (top left figure), the largest error of 5 percentage points can be seen for the green cohort "# Friends percentile in [50–75]." This is also the largest cohort, as it can be seen from the Gowalla histogram (top right figure). Similar relation can be seen also for the Breadcrumbs dataset, where the larger cohorts (EPFL and UNIL) have a larger MAE. This relation is expected because for these cohorts the predictive models work on a larger scale. For a cohort that is represented with <10% of the overall population, it is normal that the MAE (expressed in percentage points) is much lower than 10 percentage points. For the Foursquare dataset, we present MAE only for one cohort because the error for the second is equal to the presented one (note that here we have only two cohorts). The MAE ranges between 0 and 8 percentage points, depending on the dropout rate.

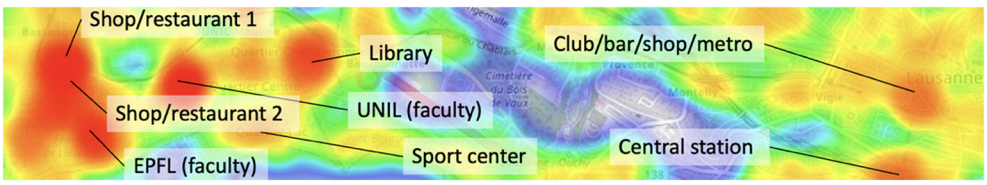

FIGURE 6
Heatmap of the GPS data available in the dataset and location of the POIs analyzed in this study.

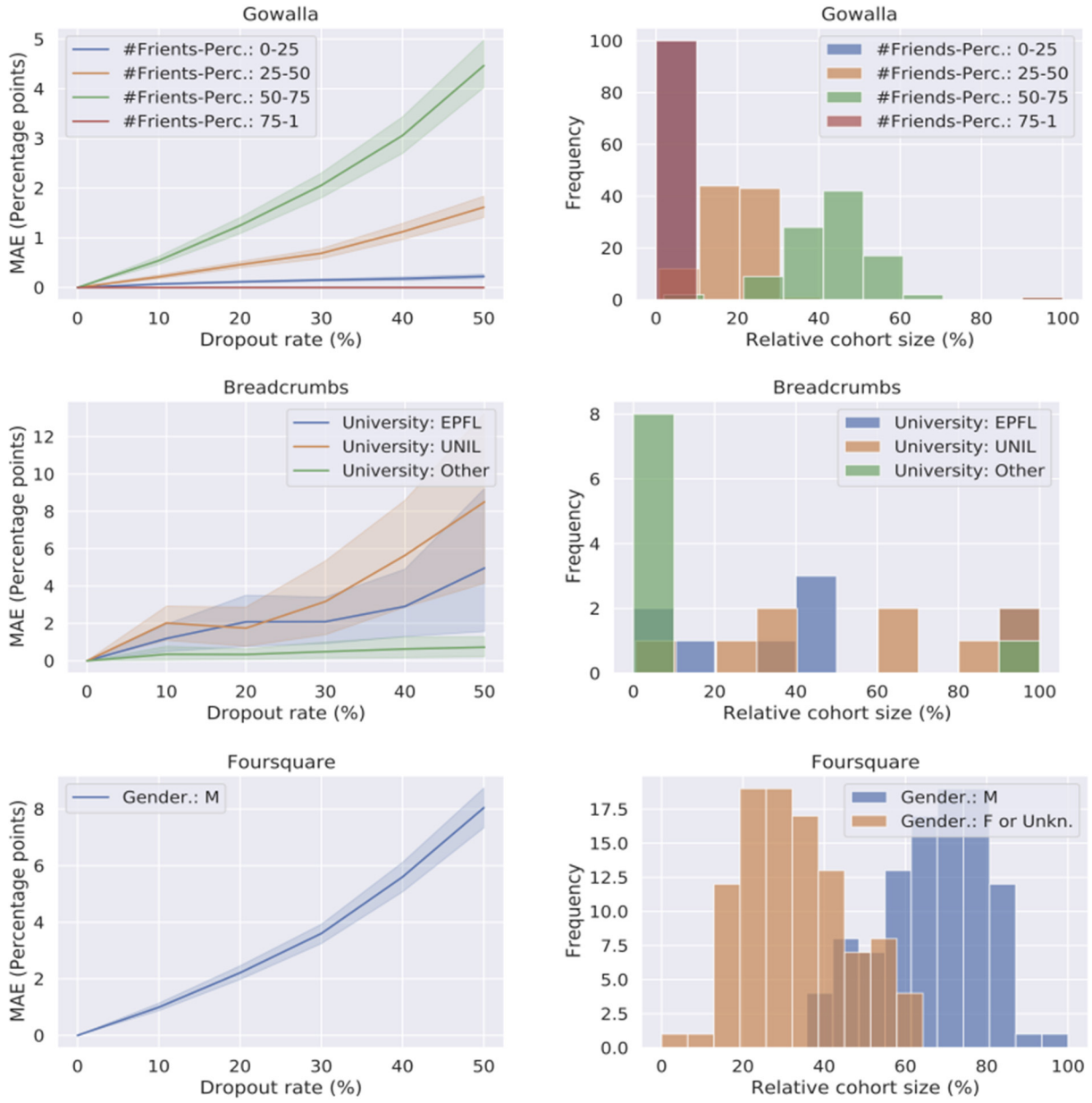

FIGURE 7
The three figures in the left column represent the relation between dropout rate (x-axis) and the corresponding Mean Absolute Error (y-axis) averaged across all POIs in each dataset. The shaded areas represent 95% confidence intervals. The three figures in the right column represent the distribution of the relative cohort size calculated over all POIs in one dataset. Each row represents one dataset (from top to bottom: Gowalla, Breadcrumbs and Foursquare).

Next-week predictions

The results of these experiments are presented in Table 2. The table contains average values and standard deviations for the Pearson's Correlation Coefficients (PPC) and MAE (in percentage points), calculated for the four predictive models, MO (overall mean predictor), M3 (3-weeks mean predictor), M1 (1-week mean predictor) and A (ARIMA). For most of the cohorts (6 out of 9), the best predictor is the M1 predictor which has a PCC above 0.8 (strong correlation). For one cohort, the PCC achieved by the M1 predictor is 0.71 (Gowalla—# Friends: Perc. 25–50). The other two cohorts, which represent special cases (Gowalla—"# Friends: Perc. 0–25" and Gowalla—"# Friends: Perc. 75–100"), the PCC is low because these cohorts are rare for the analyzed POIs and thus are hard to model. The cohort distribution can be seen in Figure 7 (top-left histogram).

TABLE 2 Pearson's Correlation Coefficient (PCC) and Mean Absolute Error (MAE) for predicting next-week's average cohort ratios.

Cohort	**# Friends: Perc. 0–25**				**# Friends: Perc. 25–50**				**#Friends: Perc. 50–75**				**# Friends: Perc. 75–100**			
Gowalla																
Model	A	MO	M1	M3	A	MO	M1	M3	A	MO	M1	M3	A	MO	M1	M3
PCC avg.	0.29	0.35	0.25	0.33	0.75	0.66	0.71	0.76	0.83	0.63	0.82	0.84	/	/	/	/
PCC std.	0.2	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	/	/	/	/
MAE avg.	2	2	2	2	6	12	7	6	9	23	10	9	0.1	0.1	0.1	0.1
MAE std.	1	1	1	1	2	4	2	2	2	5	3	2	0.1	0.1	0.1	0.1
Cohort	**University: EPFL**				**University: UNIL**				**University: other**							
Breadcrumbs																
Model	A	MO	M1	M3	A	MO	M1	M3	A	MO	M1	M3				
PCC avg.	0.64	0.38	0.83	0.38	0.77	0.05	0.88	0.32	0.55	0.2	0.88	0.52				
PCC std.	0.3	0.2	0.1	0.2	0.2	0.2	0.1	0.4	0.4	0.5	0.1	0.3				
MAE avg.	7	11	4	7	7	14	3	9	1	2	0.1	1				
MAE std.	5	7	3	4	4	7	3	6	1	2	1	2				
Cohort	**Gender: male**				**Gender: female or unknown**											
Foursquare																
Model	A	MO	M1	M3	A	MO	M1	M3								
PCC avg.	0.6	0.12	0.81	0.55	0.6	0.12	0.81	0.55								
PCC std.	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1								
MAE avg.	9	14	6	11	9	14	6	11								
MAE std.	3	3	1	2	3	3	1	2								

Predictive models: A, Arima; MO, Overall Mean; M3, 3-weeks mean; M1, 1-week mean.

Because these cohorts are rare, the MAE scores achieved by the same predictor are also low (2 and 0.1).

Regarding the overall MAE scores, the M1 predictor is better than the M3 predictor, and the two are also better than the MO predictor. This confirms that online learning approach, where the models are updated weekly (or even daily), is the most suitable approach for this task. Furthermore, the ARIMA predictors performed similarly as the M3 predictors (in some cases even better), but worse than the M1 predictor. The ARIMA models were fitted using the overall historical data, which may be one cause for the underperforming.

The weekly MAE scores for the two best performing predictors (M1 and ARIMA) are presented in Figure 8. The MAE scores are averaged across all cohorts and all POIs in each dataset. From the weekly MAE curves, it can be seen that the ARIMA predictor performed slightly better than the M1 predictor for the Gowalla dataset, but also performed slightly worse for the other two dataset. Furthermore, the Foursquare dataset seem to be more complex for modeling since the MAE curves exhibit higher irregularities.

Analytics of cohorts

Figure 9 presents an example cohort analytics that could be provided to the operators of a specific place (in this case a shop and a restaurant nearby). From the analytics, the owners of the place can see the ratio of "Swiss" vs. "Other" nationality visitors for each weekday. More specifically, we see that during most working days (Tuesday through Friday to be more precise), the ratio of "Swiss" to "Other" visitors is close to 50/50. However, during the other days most of the visitors are non-Swiss.

Limitations and future work

The cohorts used in the experiments were based on the descriptive information available in the experimental datasets. Besides such pre-defined cohorts, automatic detection of cohorts could be exploited in the future. For example, one assumption could be that similar participants visit similar places, thus cohorts could be automatically detected by clustering the descriptions of the places the participants visit. Such clustering could be performed in privacy-preserving manner, e.g., using Google's approach for Federated Learning of cohorts (FloC) (Google Research Ads., 2020). The idea in FloC is that the participants do not send their private data to a centralized server, but rather the server sends to the participants a trained clustering model, which is then executed on the participants' devices to determine to which cohort each participant belongs. Another approach would be grouping participants with respect to their privacy-preferences, as recently

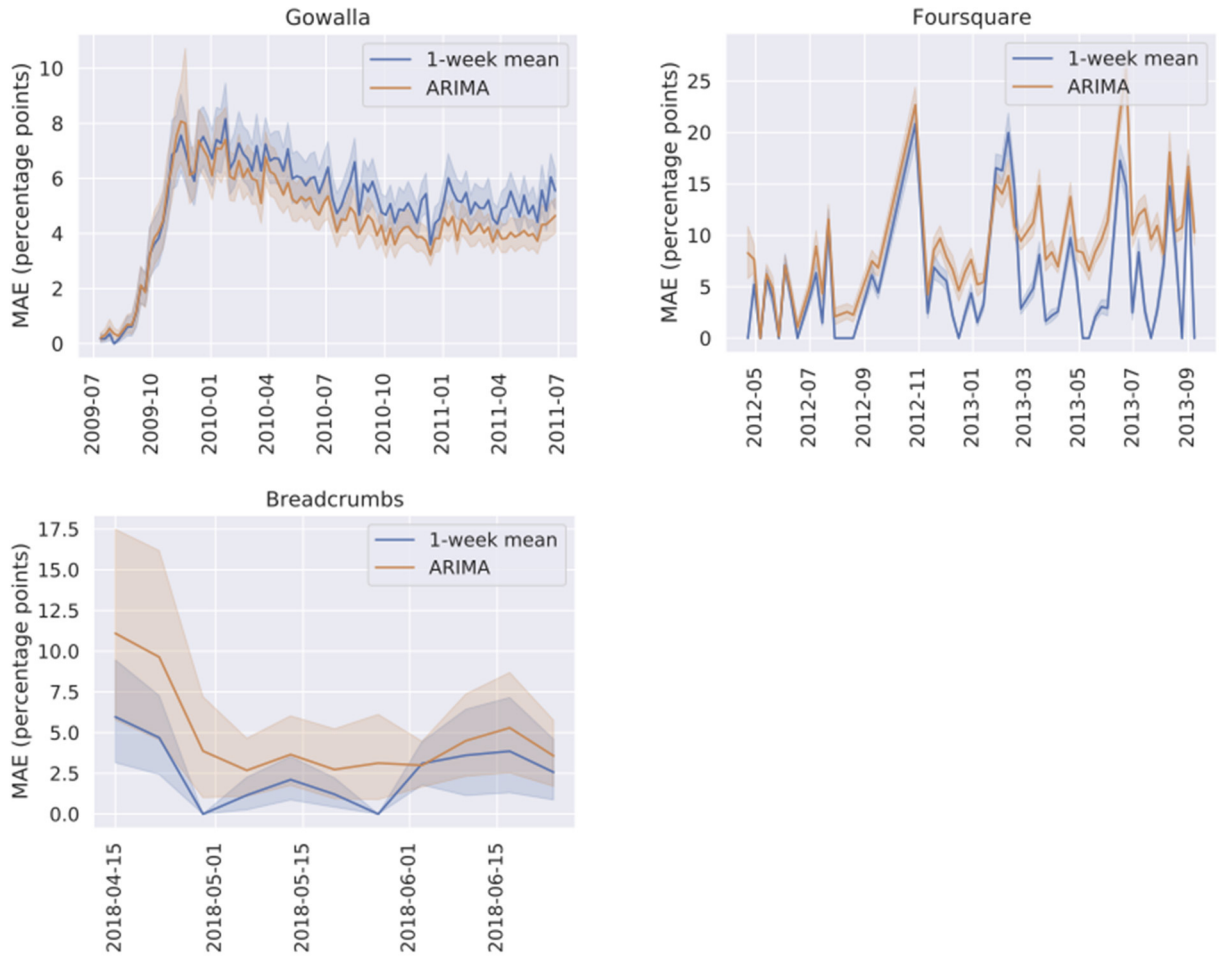

FIGURE 8
Weekly Mean Absolute Error (MAE in percentage points) for predicting next-week's cohort ratios in each of the three datasets (Gowalla, Foursquare and Breadcrumbs) using ARIMA predictor and simple 1-week mean predictor. The MAE numbers are averaged for all POIs in one dataset. The shaded areas represent 95% confidence intervals.

proposed by Tongqing et al. (2021). However, while this technique enables k-anonymity, it would create cohorts that are not intuitively explainable.

The federated analytics of cohorts that the method provides for each place depends on historical data available for that specific place. Thus, finding a representative sample for each place may be one of the biggest challenges for applying the presented method in real-life scenarios. One possible solution is to utilize historical location data that participants already store, e.g., as part of the Google Maps timeline. Nevertheless, monetary or similar incentives should be provided to the data-sharing participants as the awareness about the value of such data increases (Goldfarb and Tucker, 2012). Therefore, a balance between participants and businesses must be reached, where participants knowingly agree to share their data with individual businesses, and businesses provide some value to the participants in return. Some companies even tried to develop this idea of data-exchange transparency by providing a decentralized data marketplace, where citizens could sell their data, and businesses acquire it. For location data in particular, other companies were able to take advantage of the social needs of users through gamification, like Gowalla and Foursquare. In other cases, the benefits for participants are more obvious and direct. Location applications included by default in smartphones, like Apple Maps, Google Maps, and Bing Maps, provide a wealth of up-to-date information about different POIs (like shops, restaurants, museums, public venues, and more), allowing application users to make informed decisions about places to visit and things to do. Maps applications produce real and tangible value to their users in exchange for their location data. As our approach can be considered as a layer on top of those applications, participants would not need to collect any additional data to what they already do, but they would benefit from additional monetary or similar incentives.

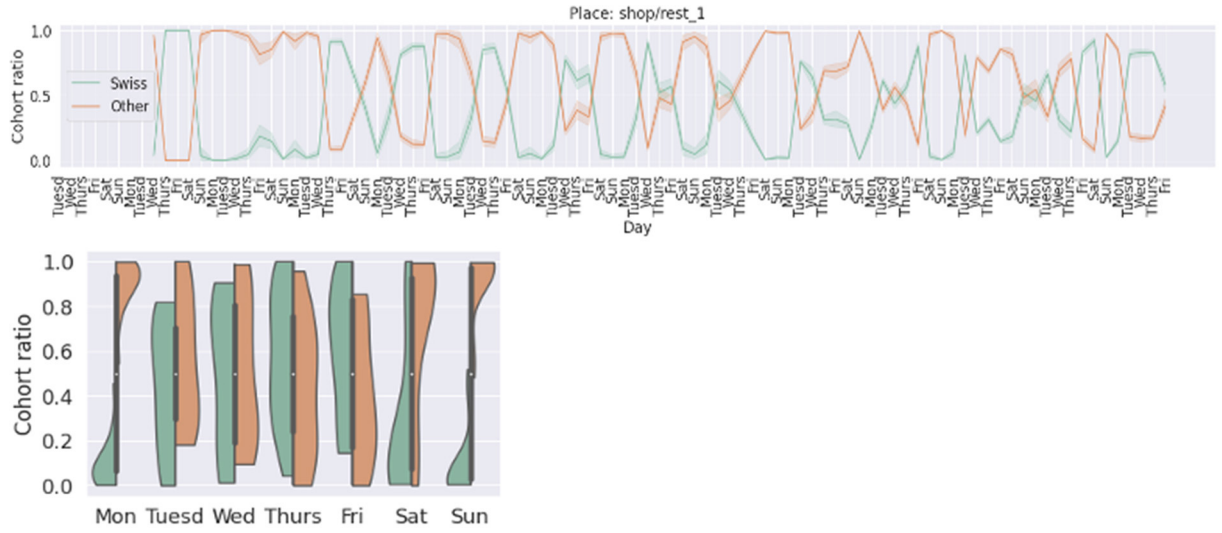

FIGURE 9
Example cohort (nationality) analytics provided to place operators (analytics users). Y-axis: aggregated cohort ratio. X-axis top figure: day of the week through the 3 months study. X-axis left figure: split-violin plots for each day of the week. Green: Swiss nationality. Orange: Other nationality. Breadcrumbs dataset.

Implications

In this study we evaluate the proposed method using GPS data acquired from smartphones. However, the proposed method is general, and it can be applied to any type of data that provides information about user's proximity to a specific Point Of Interest (POI). A successful real-world implementation of the method would enable IoT service providers to develop innovative and efficient services while preserving the privacy of the users. Example use-cases include:

- Information provision—Future smart environments are likely to be saturated with networks of digital displays (e.g., digital signs) (Davies et al., 2012) that can be used for a wide range of applications and will provide an important new form of communication (Elhart et al., 2017). The current state-of-the-art in understanding how users interact with these displays is to equip each display with a camera that can record basic demographics on the audience in front of the display at any point in time. This demographic data can be combined with a record of the content shown to provide basic audience statistics. Such approach is both privacy intrusive and limited to the characteristics that can be recognized using video analysis. On the other hand, our proposed method offers similar functionalities tracked in a privacy-aware manner (e.g., cohort-based analytics instead of person-based analytics) and in addition it holds the potential to provide enriched LBA since the users may be willing to provide more characteristics for such privacy-aware cohort-based analytics. For example, knowing the type of viewers through the cohort-based analytics may help provide insights into viewer behaviors (e.g., which groups of viewers are more likely to purchase items based on the behavior of the viewers from the same cohort).
- Environmental exposure—There is currently significant research interest in deploying environmental sensors in urban environments to monitor factors such as temperature, noise, and air quality (Hart and Martinez, 2006). In contrast to classic environmental monitoring in which sensors are a scarce resource, emerging systems feature large numbers of sensors that are widely deployed (Rawat et al., 2014). While such sensors can provide precise measurements of environmental conditions in a particular location, they are unable to capture the exposure of cohorts as they move around the urban environment. For example, while sensors can provide detailed maps of pollution hot-spots within a city, estimating the exposure of cohorts requires an understanding of their movement within the city. To address this shortcoming, researchers have explored the use of personal, wearable environmental sensors (Oscar and Labrador, 2012). However, it is unrealistic to expect citizens to carry and maintain such sensors for any significant period. The proposed cohort-based analytics would enable the production of reports from a fixed sensor base that captures the exposure of a specific cohort of users to environmental factors, helping to understand risk profiles for different user groups and the effectiveness of interventions such as pollution-aware routing (Jarjour et al., 2013).

- Cohort-based energy profiling—Modern buildings include a wide range of monitoring systems that can report on factors such as occupancy levels, temperature, and energy consumption on a room-level granularity. Such systems enable organizations to carefully monitor energy use in terms of buildings. In most of the cases, the existing insights target specific buildings or energy meters. The ability to provide LBA that report at cohort level would allow a fundamentally different understanding of how energy is "spent" to support different activities and user groups.

Conclusion

This study presented a novel privacy-aware method for federated analytics of cohorts for smart mobility. The experimental results confirmed that the method works both for small datasets (the Breadcrumbs dataset) and large datasets (the Foursquare and the Gowalla datasets), and it works both for continuous GPS sensing and on-demand sensing (check-ins). The method is easy to implement as it does not require specialized hardware (e.g., GPUs). Furthermore, the method is based on an online learning approach, where updates are communicated daily. This allows to track distribution shifts in the data, which solves one more problem that many methods have in dynamic environments.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Gowalla (https://www.yongliu.org/datasets); Foursquare (https://sites.google.com/site/yangdingqi/home/foursquare-dataset); Breadcrumbs (https://github.com/doplab/breadcrumbsDB).

Author contributions

MG: investigation, conceptualization, methodology, validation, visualization, software, and writing—review and editing. MLap: conceptualization, formal analysis, investigation, and writing—review and editing. MLan: conceptualization, methodology, writing—review and editing, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by the Swiss National Science Foundation, project 200021_182109 (BASE: Behavioral Analytics for Smart Environments).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adams, B., Phung, D. Q., and Venkatesh, S. (2006). "Extraction of social context and application to personal multimedia exploration." in *Proceedings of the 14th ACM International Conference on Multimedia* (New York, NY: ACM), 987–996. doi: 10.1145/1180639.1180857

Andrew, H., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., et al. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

Arielle, M., Kulkarni, V., Ghiringhelli, P. A., Chapuis, B., Huguenin, K., et al. (2019). "Breadcrumbs: a rich mobility dataset with point-of-interest annotations," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 508–511.

Ashbrook, D., and Starner, T. (2003). (2003). Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiquit. Comput.* 7, 275–286. doi: 10.1007/s00779-003-0240-0

Baumann, P., Kleiminger, W., and Santini, S. (2013). "The influence of temporal and spatial features on the performance of next-place prediction algorithms," in *Proceedings of the 2013ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 449–458. doi: 10.1145/2493432.2493467

Benjamin, B., and Musolesi, M. (2020). Where you go matters: a study on the privacy implications of continuous location tracking. *Proc. ACM Interact. Mobile Wearable Ubiquit. Technol.* 4, 1–32. doi: 10.1145/3432699

Bhaskar, P., and La Porta, T. (2015). "Spatial and temporal considerations in next place predictions," in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (Hong Kong: IEEE), 390–395.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., et al. (2017). "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17, 1175–91* (Dallas, TX: ACM Press). doi: 10.1145/3133956.3133982

Castro, J. E., Gjoreski, M., and Langheinrich, M. (2022). Federated learning for privacy-aware human mobility modeling. *Front. Artif. Intell.* 5, 867046. doi: 10.3389/frai.2022.867046

Charles, K. F. (2013). Algorithms for geodesics. *J. Geodesy* 87, 43–55. doi: 10.1007/s00190-012-0578-z

Christopher, M., Mitrovich, M., D'Antonio, A., and Sisneros-Kidd, A. (2019). Using mobile device data to estimate visitation in Parks and protected areas: an

example from the nature reserve of orange County, California. *J. Park Recreat. Admi.* 37, 92–109. doi: 10.18666/JPRA-2019-9899

Crossler, R. E., Johnston, A. C., Lowry, P. B., Hu, Q., Warkentin, M., Baskerville, R., et al. (2013). Future directions for behavioral information security research. *Comput. Security* 32, 90–101. doi: 10.1016/j.cose.2012.09.010

Davies, N., Langheinrich, M., Jose, R., and Schmidt, A. (2012). Open display networks: a communications medium for the 21st century. *Computer* 45, 58–64. doi: 10.1109/MC.2012.114

Dingqi, Y., Fankhauser, B., Rosso, P., and Cudre-Mauroux, P. (2020). "Location prediction over sparse user mobility traces using RNNs: flashback in hidden states!" in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2184–2190.

Do, T. M. T., and Gatica-Perez, D. (2014). Where and what: using smartphones to predict next locations and applications in daily life. *Pervasive Mob. Comput.* 12, 79–91. doi: 10.1016/j.pmcj.2013.03.006

Elhart, I., Mikusz, M., Gomez Mora, C., Langheinrich, M., and Davies, N. (2017). "Audience monitor: an open source tool for tracking audience mobility in front of pervasive displays," in *Proceedings of the 6th ACM International Symposium on Pervasive Displays*, 1–8. doi: 10.1145/3078810.3078823

Etter, V., Kafsi, M., and Kazemi, E. (2012). "Been there, done that: what your mobility traces reveal about your behavior," in *Mobile Data Challenge Workshop* (Newcastle, UK).

Eugene, B., Kairouz, P., Mellem, S., Gascón, A., Bonawitz, K., Estrin, D., et al. (2021). Towards sparse federated analytics: location heatmaps under distributed differential privacy with secure aggregation" *arXiv preprint arXiv:2111.02356*.

Fisher, D. M., Wood, S. A., White, E. M., Blahna, D. J., Lange, S., Weinberg, A., et al. (2018). Recreational use in dispersed public lands measured using social media data and on-site counts. *J. Environ. Manage.* 222, 465–474. doi: 10.1016/j.jenvman.2018.05.045

Gjoreski, M., Janko, V., Slapničar, G., Mlakar, M., Reščič, N., Bizjak, J., et al. (2020). Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors. *Inform. Fusion* 62, 47–62. doi: 10.1016/j.inffus.2020.04.004

Goldfarb, A., and Tucker, C. (2012). Shifts in privacy concerns. *Am. Econ. Rev.* 102, 349–353. doi: 10.1257/aer.102.3.349

Google Research and Ads. (2020). *Evaluation of Cohort Algorithms for the FLoC API*. Available online at: https://github.com/google/ads-privacy/tree/master/proposals/FLoC (accessed June 15, 2022).

Hamstead, Z., Fisher, A. D., Ilieva, R. T., Wood, S. A., McPhearson, T., Kremer, P., et al. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Comput. Environ. Urban Syst.* 72, 38–50. doi: 10.1016/j.compenvurbsys.2018.01.007

Hart, J. K., and Martinez, K. (2006). Environmental sensor networks: a revolution in the earth system science? *Earth Sci. Rev.* 78, 177–191. doi: 10.1016/j.earscirev.2006.05.001

Henrikki, T., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., et al. (2017). Instagram, Flickr, or Twitter: assessing the usability of social media data for visitor monitoring in protected areas. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-18007-4

Hess, A., Hummel, K. A., Wilfried, N. G., and Haring, G. (2015). Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys* 48, 1–39. doi: 10.1145/2840722

Imai, R., Tsubouchi, K., Konishi, T., and Shimosaka, M. (2018). Early destination prediction with spatio-temporal user behavior patterns. *Proc. ACM Interact. Mob. Wearable Ubiquitous Comput.* 1, 1–19. doi: 10.1145/3161197

Ittai, D., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* 27, 10, 1735–1743. doi: 10.1038/s41591-021-01506-3

Jane, F., Sage, J., Stone, J., and Vlachantoni, A. (2016). Residential mobility across the life course: continuity and change across three cohorts in Britain. *Adv. Life Course Res.* 30, 111–123. doi: 10.1016/j.alcr.2016.06.001

Jarjour, S., Jerrett, M., Westerdahl, D., de Nazelle, A., Hanning, C., Daly, L., et al. (2013). Cyclist route choice, traffic-related air pollution, and lung function: a scripted exposure study. *Environ. Health* 12, 1–12. doi: 10.1186/1476-069X-12-14

Jie, F., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., et al. (2018). "Deepmove: predicting human mobility with attentional recurrent networks," in *Proceedings of the 2018 World Wide Web Conference*, 1459–1468.

Jie, F., Rong, C., Sun, F., Guo, D., and Li, Y. (2020). PMF: a privacy-preserving human mobility prediction framework via federated learning. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 4, 1–21. doi: 10.1145/3381006

Jun, Z., He, X., Tang, H., and Wen, J. (2019). "A next location predicting approach based on a recurrent neural network and self-attention," in *International Conference on Collaborative Computing: Networking, Applications and Worksharing* (Cham: Springer), 309–322. doi: 10.1007/978-3-030-30146-0_21

Jung, K. Y., Kun Lee, D., and Kim, C. K. (2020). (2020). Spatial tradeoff between biodiversity and nature-based tourism: considering mobile phone-driven visitation pattern. *Glob. Ecol. Conserv.* 21, e00899. doi: 10.1016/j.gecco.2019.e00899

Krüger, A., Johannes, S., and Patrick, O. (2011). How computing will change the face of retail. *Computer* 44, 84–87. doi: 10.1109/MC.2011.112

Langheinrich, M. (2002). "A privacy awareness system for ubiquitous computing environments," in *UbiComp 2002: Ubiquitous Computing: 4th International Conference G{ö}teborg, Sweden, September 29 – October 1, 2002*, eds G. Borriello, L. E. Holmquist (Berlin, Heidelberg: Springer Berlin Heidelberg), 237–245. doi: 10.1007/3-540-45809-3_19

Liu, Y., Wei, W., Sun, A., and Miao, C. (2014). "Exploiting geographical neighborhood characteristics for location recommendation," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*, 739–748. doi: 10.1145/2661829.2662002

Martin, E., Kriegel, H.-., P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In kdd* 96, 226–231.

Massimiliano, L., Barlacchi, G., Lepri, B., and Pappalardo, L. (2020). Deep learning for human mobility: a survey on data and models. *arXiv preprint arXiv:2012.02825*.

Mazhar, R. M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Comput. Netw.* 101, 63–80. doi: 10.1016/j.comnet.2015.12.023

Mehrotra, A., Hendley, R., and Musolesi, M. (2016). "Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 1132–1138. doi: 10.1145/2968219.2968299

Monreale, A., Pinelli, F., Trasarti, R., and Giannotti, F. (2009). "Wherenext: a location predictor on trajectory pattern mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 637–646. doi: 10.1145/1557019.1557091

Nathaniel, M. H., Atkinson, S. F., Mulvaney, K. K., Mazzotta, M. J., and Bousquin, J. (2020). Using data derived from cellular phone locations to estimate visitation to natural areas: an application to water recreation in New England, USA. *PLoS ONE* 15, e0231863. doi: 10.1371/journal.pone.0231863

Nitin, B., Kidiyoor, G. V., Varun, S. K., Kalambur, S., Sitaram, D., Kollengode, C., et al. (2015). "Predicting the next move: determining mobile user location using semantic information," *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2359–2365.

Oscar, L., and Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE Communic. Surv. Tutor.* 15, 1192–1209. doi: 10.1109/SURV.2012.110112.00192

Pappalardo, L., Simini, F., Barlacchi, G., and Pellungrini, R. (2019). scikit-mobility: a Python library for the analysis, generation and risk assessment of mobility data. *arXiv preprint arXiv:1907.07062*.

Rawat, P., Singh, K. D., Chaouchi, H., and Bonnin, J. M. (2014). Wireless sensor networks: a survey on recent developments and potential synergies. *J. Supercomput.* 68, 1–48. doi: 10.1007/s11227-013-1021-9

Schreckenberge, C., Beckmann, S., and Bartelt, C. (2018). "Next place prediction: a systematic literature review," in *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*, 37–45. doi: 10.1145/3283590.3283596

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* 10, 557–570. doi: 10.1142/S0218488502001648

Takahiro, K., Uryu, S., Yamano, H., Tsuge, T., Yamakita, T., Shirayama, Y., et al. (2020). Mobile phone network data reveal nationwide economic value of coastal tourism under climate change. *Tourism Manag.* 77, 104010. doi: 10.1016/j.tourman.2019.104010

Tongqing, Z., Cai, Z., and Liu, F. (2021). The crowd wisdom for location privacy of crowdsensing photos: spear or shield? *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 5, 1–23. doi: 10.1145/3478106

Usman, A., Srivastava, G., and Lin, J. C.-W. (2022). Reliable customer analysis using federated learning and exploring deep-attention edge intelligence. *Fut. Gen. Comput. Syst.* 127, 70–79. doi: 10.1016/j.future.2021.08.028

Yang, D., Zhang, D., and Qu, B. (2016). Participatory cultural mapping based on collective behavior data in location based social networks. *ACM Trans. Intell. Syst. Technol.* 7, 1–23. doi: 10.1145/2814575

Yiwei, S., Jiang, D., Liu, Y., Qin, Z., Tan, C., Zhang, D., et al. (2021). HERMAS: a human mobility embedding framework with large-scale cellular signaling data. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 5, 1–21. doi: 10.1145/3478108

Yu, Z., Xie, X., and Ma, W. (2010). Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* 33, 32–39.

Yuanyishu, T., Wan, Y., Lyu, L., Yao, D., Jin, H., Sun, L., et al. (2022). FedBERT: when federated learning meets pre-training. *ACM Trans. Intell. Syst. Technol.* doi: 10.1145/3510033

Zhihan, F., Yang, Y., Yang, G., Xian, Y., Zhang, F., Zhang, D., et al. (2021). CellSense: human mobility recovery via cellular network data enhancement. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 5, 1–22. doi: 10.1145/3478087

Zipei, F., Song, X., Jiang, R., Chen, Q., and Shibasaki, R. (2019). Decentralized attention-based personalized human mobility prediction. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 3, 1–26. doi: 10.1145/3369830