Enriching Sparse Mobility Information in Call Detail Records

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, Carlos Sarraute

Enriching Sparse Mobility Information in Call Detail Records

Guangshuo Chen[a,b], Sahar Hoteit[c], Aline Carneiro Viana[b], Marco Fiore[d], Carlos Sarraute[e]

[a] *École Polytechnique, Université Paris Saclay, 91128 Palaiseau, France.*
[b] *INRIA Saclay-Île-de-France, Université Paris Saclay, 91120 Palaiseau, France.*
[c] *Laboratoire des Signaux et Systèmes, Université Paris Sud-CNRS-CentraleSupélec, Université Paris-Saclay, 91192 Gif-sur-Yvette, France.*
[d] *CNR-IEIIT, 10129 Torino, Italy.*
[e] *Grandata Labs, 550 15th Street, San Francisco, 94103 California, USA*

Abstract

Call Detail Records (CDR) are an important source of information in the study of diverse aspects of human mobility. The accuracy of mobility information granted by CDR strongly depends on the radio access infrastructure deployment and the frequency of interactions between mobile users and the network. As cellular network deployment is highly irregular and interaction frequencies are typically low, CDR are often characterized by spatial and temporal sparsity, which, in turn, can bias mobility analyses based on such data. In this paper, we precisely address this subject. First, we evaluate the spatial error in CDR, caused by approximating user positions with cell tower locations. Second, we assess the impact of the limited spatial and temporal granularity of CDR on the estimation of standard mobility metrics. Third, we propose novel and effective techniques to reduce temporal sparsity in CDR, by leveraging regularity in human movement patterns.

Keywords: Call Detail Records, spatiotemporal trajectories, data sparsity, cellular networks, mobility, movement inference.

Email addresses: guangshuo.chen@inria.fr (Guangshuo Chen), sahar.hoteit@u-psud.fr (Sahar Hoteit), aline.viana@inria.fr (Aline Carneiro Viana), marco.fiore@ieiit.cnr.it (Marco Fiore), charles@grandata.com (Carlos Sarraute)

1. Introduction

Urbanization challenges the development and sustainability of city infrastructures in a variety of ways, and telecommunications networks are no exception. Understanding human habits becomes essential for managing the available resources in complex smart urban environments. Specifically, a number of network-related functions, such as paging [1], caching [2], dimensioning [3], or network-driven location-based recommending systems [4] have been shown to benefit from insights on subscriber movements. More generally, the investigation of human mobility pattern has attracted a significant attention across disciplines [5–9].

Motivation: Human mobility studies strongly rely on actual human footprints, which are usually provided by spatiotemporal datasets, as a piece of knowledge to investigate human mobility patterns. In this context, using specialized spatiotemporal datasets such as GPS logs seems to be a direct solution, but there is a huge overhead of collecting such a detailed dataset at scale. Hence, Call Detail Records (CDR) have been lately considered the primary source of data for large-scale mobility studies. CDR contain information about *when*, *where* and *how* a mobile phone subscriber generates voice calls and text messages, and are collected by mobile network operators for billing purposes. These records usually cover a large urban population [10], which makes them a practical choice for large scale human mobility analysis.

CDR are able to be regarded as human footprints and thus be used to build human visiting location patterns or measure mobility-related features, *e.g.*, the radius of gyration. Despite significant benefits that the dataset of CDR brings to human mobility analysis regarding its population, an indiscriminate use of such data may raise questions about the validity of the conclusions of the related research efforts. Specifically, CDR have limited accuracy along both the spatial dimension (as the user location is known at the cell sector or base station coverage levels) and the temporal dimension (since a location is recorded only at a time when one performs a voice call or texts a message). Indeed, a cell

(sector) typically spans at least thousands of square meters; even a very active mobile subscriber generate only a few tens of voice or text events per day. Overall, spatiotemporal sparsity exists in CDR, and it is urgent and significant to understand whether and to what extent such a sparsity affects mobility studies.

Existing studies and limitations: Several previous analyses have addressed the validity of mobility studies based on CDR. An influential work [6] observed that using CDR allows to correctly identifying, for each user, popular locations that account for 90% of subscriber's activity. Also, according to the same authors, biases may arise in some cases when measuring individual human mobility features. Besides, other works [11] have explored this topic. [11, 12] discussed biases regarding incompleteness of locations, as CDR usually do not capture every location that one travels through. Nevertheless, another important bias, caused by using cell tower locations instead of actual human positions in CDR, is not yet investigated in the literature.

In the context of constructing reliable human visiting patterns using CDR, it is necessary to fill the spatiotemporal gaps in CDR, or in other words, to complete sparse CDR as much and accurate as possible. For that, the most intuitive solution is to consider that a cell tower location documented in an entry of CDR is available and representative for a period (typically one hour) rather than only at the event's time instance, as discussed in [7, 13]. So far, to the best of our knowledge, no solution in the literature aims at filling the spatiotemporal gaps in CDR or leveraging the unique features of CDR.

Our work and contributions: In this paper, we leverage real human mobility datasets and explore the following research questions. First, we investigate how the spatiotemporal sparsity of CDR affects the accuracy and incompleteness of mobility information by observing actual CDR and cell tower deployments in metropolitan areas. Second, we evaluate the biases caused by such spatiotemporal sparsity in identifying important locations and measuring the span of individual movements. Third, we study whether leveraging instantaneous whereabouts provided by CDR is capable of locating a user continuously in time. This third study is to understand to what degree of completeness CDR

can reconstruct a user's visiting pattern of locations. Our work leads to these major contributions, summarized as follows:

- We assess the capability of sparse CDR of modeling important features of individual visiting patterns. Our results confirm previous findings in the literature.

- We originally study the drawback of representing human whereabouts by cell tower locations in terms of accuracy loss. Our result shows that geographical shifts (distances to actual positions) caused by such representation are below 100 meters in a majority of cases ($50\% - 90\%$ varying across cellular operators), and are mostly (over 99%) below 1 kilometers. This result strongly supports the validity of results obtained by CDR from the literature.

- We implement a number of techniques for CDR completion proposed in the literature and assess their quality in the presence of ground-truth GPS data. Our evaluation sheds light on the quality of the results provided by each approach. We propose original CDR completion solutions and show that they outperform previous proposals, reducing the spatial error in the completed data and shortening the time periods in which no location information is available.

- Particularly, we show how the cell coverage affects the results. Our adaptive solutions for completing CDR utilize the cell coverage and other features that are obtainable via CDR, based on cell boundary (*i.e.*, the considered period for a user to spend on the position associated with each communication activity) identification. Among all the set of tested features, the results reveal the three most significant ones with respect to the estimation of a cell boundary, *i.e.*, the cell size, the radio of gyration, and the time of activity. Leveraging such features, our adaptive solutions outperform the previous approaches in the literature.

The rest of the paper is organized as follows. Related works are introduced

in Sec. 2. In Sec. 3, we present the datasets used in our study. In Sec. 4, we introduce and explore the biases of using CDR for human mobility analysis. In Sec. 5, we discuss the rationale for CDR completion and errors introduced by common literature related approaches. In Sec. 6 and 7, we introduce our contributions on CDR completion allowing improving accuracy, during nighttime and daytime, respectively. Finally, Sec. 8 concludes the paper.

2. Related works

Our work aims at measuring and evaluating possible biases induced by the usage of CDR. Whether and to what extent these biases affect mobility studies have been only partly addressed. In terms of the sparsity of CDR, promising results are obtained when mobility is constrained to transportation networks. Zhang *et al.* [11] find CDR-based individual trajectories to match reference information from public transport data, *i.e.*, GPS logs of taxis and buses, as well as subway transit records. A seminal work in this sense was performed by Ranjan *et al.* [6]. They show that CDR are capable of identifying important locations, and expose the bias of working only on very active CDR users. In our paper, we confirm the observation in [6] and push one step further by considering the spatial bias. Besides, [14] show that using CDR positioning information may lead to a distance error within 1 km compared to a ground-truth collected from 5 users. In our previous study [15], we confirm the observations in [6][14] using a GPS-referenced dataset containing 84 users.

Furthermore, to mitigate the spatiotemporal sparsity in CDR, we explore some mobility completion techniques. The legacy approach in the literature is assuming that the user remains static from some time before and after each communication activity. The span of the static period, named *cell boundary* starting now, is a constant system parameter that is hardly validated [13, 15]. In this paper, we extend previous analyses in [15, 16]; we propose two adaptive approaches to complete user's normal locations and home locations/periods.

3. Datasets

Our analysis requires both coarse-grained and fine-grained geographical data describing the mobility of the same group of users. The former should contain CDR, while the latter data acts as the ground-truth; the latter should provide more detailed mobility information and have a better temporal and spatial granularity. We generate such data from one coarse-grained (CDR) dataset and three fine-grained datasets. Since these datasets do not share the same set of users, we downsample the fine-grained datasets using the inter-event time distribution obtained from the coarse-grained one. This downsampling allows us exploiting the fine-grained datasets to construct multiple ground-truth data as well as coarse-grained data (equivalent CDR), which is then used to investigate the human mobility under different features and granularity. We detail all the used datasets at the remaining of the section.

3.1. Coarse-grained data: CDR dataset

This dataset consists of Call Detail Records (CDR): timestamped and geo-referenced logs on voice calls of serviced customers (mobile subscribers). These logs are generated each time the mobile device connects to the cellular network for making or receiving a voice call. Each record contains the hashed identifiers of the caller and the callee, the call duration in seconds, the timestamp for the call time and the cell tower location to which the device is connected at the beginning. The CDR are collected by a major cellular network operator. They capture the call activities of 1.6 million of users over a consecutive 3-month period in 2015[1] resulting in 681 million CDR in the selected period of study.

We extract the experimental statistical distributions of the inter-event time (*i.e.*, the time between consecutive events) from the CDR dataset. The corresponding cumulative distribution functions (CDF) for different hours of the day are shown in Fig. 1. For two consecutive events happening at different

[1]According to the privacy requirement of the data owner, we cannot point out the area as well as the collecting period of this dataset.

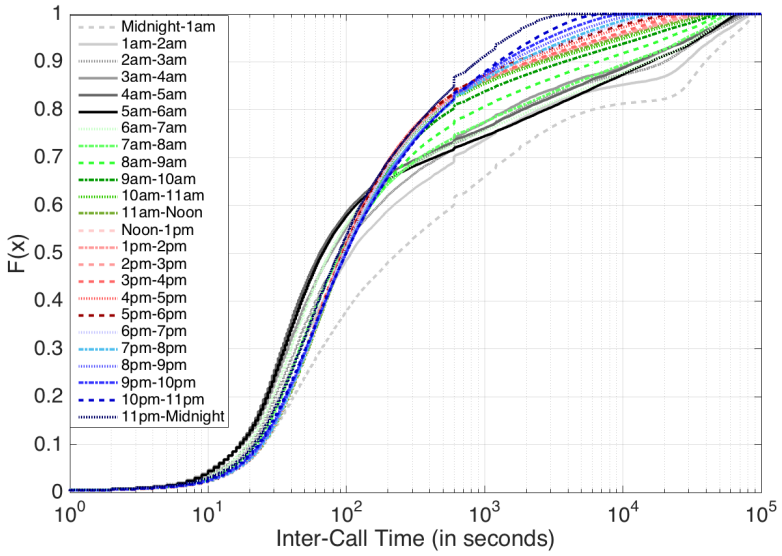


Figure 1: CDF of the inter-event(voice call) time in CDR collected on an hourly basis.

time slots, the corresponding inter-event time is attributed to the time slot of the first event. We observe that a majority of events occur at a distance of a few minutes, but a non-negligible amount of events are spaced by hours. This observation confirms literature studies on the timing issue of many human activities, which are characterized by bursts of rapidly occurring events separated by long periods of inactivity [17]. The curves in Fig. 1 tell apart the distributions observed during different hours of the day: this allows appreciating the longer inter-event times during low-activity hours (*e.g.*, midnight to 6 am) that become progressively shorter during the day.

3.2. Fine-grained data: Internet flow, Macaco GPS and Geolife GPS datasets

We leverage three datasets that act as the *actual* ground-truth data of the CDR introduced above. First, the Internet flow dataset contains locations logged to data records generated by a part population of and collected in a sub-period of the CDR dataset. Hence, this dataset is used directly as one ground-truth, but it only covers daytime and not nighttime. This dataset is further presented in Sec. 3.2.1. Second, both MACACO and Geolife datasets,

Table 1: Overview of the Internet Flow Dataset

Date	Users	Rare CDR users	Frequent CDR users
2015-07-19 (Sunday)	10,856	6,154	4,702
2015-07-20 (Monday)	14,353	7,215	7,138

covering all time of the day, are utilized to overcome the time period limit of the Internet flow dataset. They are presented in Sec. 3.2.2 and Sec. 3.2.3.

3.2.1. Internet flow data

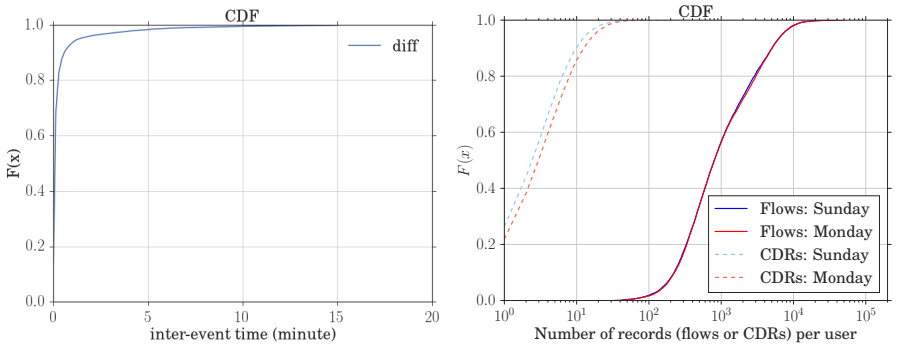
The Internet flow data is composed of Internet data records, named hereafter *flows*. These records are obtained every time a mobile device establishes a TCP/UDP session for some services (*e.g.*, Facebook, Google, and WhatsApp). Each flow entry contains the hashed device identifier, the type of service, the volume of exchanged upload and download data, the timestamps denoting the starting and ending time of the session, and more importantly, the location of cell tower handling the session.

The Internet flow dataset is collected from users appearing in the CDR dataset by the same cellular network operator. Thus, for a set of users, we have both CDR and flow data during the period of study given at the Internet flow dataset. Nevertheless, this observing period merely covers the $(10am, 6pm)$ time interval (prevailing working hours) during two consecutive days, as shown in Table. 1.

The users in the Internet flow dataset are categorized according to their actual CDR as:

- *Rare CDR users*: those who in the CDR dataset are not very active in making or receiving voice calls, and sending or receiving SMS/MMS. As in [7], we use the threshold of 0.5 CDR/hour below which the user is considered to belong to this category.

- *Frequent CDR users*: those who in the CDR dataset are comparatively active. They have more than 0.5 CDR/hour.



(a) (b)

Figure 2: (a) CDF of the inter-event time in the ground-truth dataset; (b) CDF of the number of records (flows or CDR) per user in a weekend and a weekday.

In the two days of study, the Internet flow dataset has a considerable temporal granularity, which means having at least one flow (*i.e.*, one location) in every 20 minutes. In particular, the statistical distribution of the per-user inter-flow time is shown in Fig. 2(a). We note that in 98% of cases, the inter-event time is less than 5 minutes, and in less than 1% of cases, the inter-event time is higher than 10 minutes. We also plot in Fig. 2(b) the CDF of the number of flows (as solid lines) and CDR (as dashed lines) per user for comparison. The results tell that this fine-grained dataset brings richer information about user visiting patterns of locations because the user tends to have far more flows seen in this dataset than CDRs seen in the coarse-grained dataset.

High temporal granularity supports the use of visiting location patterns in the Internet flow dataset as the ground-truth. Nevertheless, the flow data covers daytime only. To have ground-truth covering more time, we introduce the other two fine-grained datasets thereafter.

3.2.2. MACACO GPS data

This dataset is obtained through an Android mobile phone application, MACACOApp[2], developed in the context of the EU CHIST-ERA MACACO project [18]. The application collects data related to the user's digital activities such as mobile services she uses, uplink/downlink traffic she generates, network connectivity she leverages, and the visited GPS locations. These activities are logged with fixed periodicity at every 5 minutes. We remark that this sampling approach differs from those employed by popular GPS tracking projects, such as MIT Reality Mining [19] or GeoLife [20], where users' positions are sampled irregularly as previously explained. With respect to such previous efforts, the regular sampling of MACACO data grants a much neater and comprehensive vision of the user's movement patterns. The MACACO data cover 84 users who live in 6 different countries and travel worldwide. The data collection spans 18 months approximately, from July 10, 2014, to February 4, 2016.

3.2.3. Geolife GPS data

To perform a comparative study, we also use the latest version of Geolife dataset [20] as a GPS-based mobility source. This dataset provides geolocalized and time-stamped points from 182 people during a three-year span (from April 2007 to August 2012), mostly in Beijing [20]. In this dataset, every GPS trajectory comes from an ordered sequence of time-stamped points, each of which contains the information of latitude, longitude, and altitude. The main issue in this dataset is that very often it presents huge gaps between subsequent data records. Hence, some users in this dataset have an insufficient number of locations or do not move at all during the day. To overcome these issues, we select users that have active records and are not static during the whole observation period, *i.e.*, 43 users in total[3].

[2]Available at `https://macaco.inria.fr/MACACOApp/`.

[3]The Geolife users are used in our study with the following identifiers: 3-5,10,12,14,17,20-22,24,26,30,35,45,48,51,63,65,66,72,78,80,93,96,102,103,107,109,110,115-117,119-121,125-127,133,143,153,173.

3.3. Equivalent CDR: from MACACO and Geolife GPS data

Both MACACO and Geolife datasets serve as fine-grained ground-truth. To have coarse- and fine-grained data sharing the same set of users, we separately downsample MACACO and Geolife to generate coarse-grained data mimicking CDR. The downsampling is an inevitable step, as we do not have access to the mobile network operator for MACACO or Geolife users. To downsample the GPS logs realistically, we leverage statistical distributions of the inter-event time among the CDR: we downsample the two datasets according to the inter-event distributions that are shown in Fig. 1. This allows taking into account the differences emerging across day hours. Also, upon subsampling, we select only users having a sufficient number of records during weekdays: we eliminate all weekends and select only users with more than 30 positions per day and more than 3 days of activity. The selection results in those referred as *equivalent CDR* hereafter, for a total of 32 MACACO and 43 Geolife users.

4. Biases of using CDR

This section presents major characteristics of CDR and biases induced by using CDR into analyzing human mobility. As previously explained, CDR are sparse in time and space, and such sparsity impacts the validity of results obtained from CDR in the form of biases. This section reveal some of these biases.

4.1. Using cell tower location

In most CDR datasets, the location property in the CDR entry is actually represented by the cell tower location handling the corresponding communication. Hence, a shift from the user's actual location to the cell tower location always exists in every entry. Such a shift leads to a bias that impacts the accuracy of individual mobility measurements. Usually, CDR are collected in metropolitan areas. In this case, the precision of human locations provided by CDR is related to the deployment of base stations in the area. Fig. 3 shows the deployment of cell towers in a metropolitan area in a Latin American country,

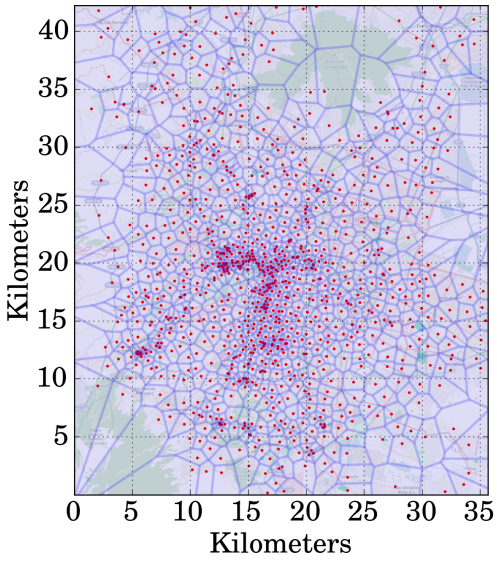


Figure 3: Deployment of cell towers in a metropolitan area in a latin American country. Red dots represent the base stations and the Voronoi tessellation approximates the coverage of each cell.

in which the cell tower locations are collected from approximately 1 million of CDR. We clearly observe that these cell towers are fairly dense in the center observing area, where a cell tower covers a 2km^2 area on average: the density grants a fair granularity in the localization of mobile subscribers. Intuitively, such density should lead to small shifts. We give a quantitative analysis of shifts across various cell towers, and evaluate the bias to human mobility studies.

We leverage GPS logs in the MACACO dataset in this specific analysis, due to two reasons: *(i)* the Internet flow and CDR datasets lack GPS information of visited locations or only provide cellular-level information of visited locations of the users; *(ii)* no available reliable source allows the extraction of cell tower information (*i.e.*, coordinates or covered area of deployed cell towers) in the area of Beijing that Geolife users are mainly from. In particular, we extract $718,987$ GPS locations in a major area of France[4] from the MACACO dataset. We

[4]The study focuses on the area in the latitude and longitude ranges of $(43.005, 49.554)$ and

then extract cell tower locations of the three major cellular network operators in France from the OpenCellid [21]. The operators, marked as Operator A, B and C, have the MCC code 208 (*i.e.*, that stands for France) and their own MNC codes (Operator A: $20, 21, 88, 215$; Operator B: $1, 2, 91, 95$; Operator C: $9, 10, 11, 13, 100, 200$) in the identifier of their cells.

Fig. 4 is the CDF of the distance between each GPS location in the MACACO dataset and its nearest cell tower extracted from OpenCellid [21]. We observe that over 99% of the locations have a distance below 1 km when shifting to their nearest cells. More precisely, only 0.65%, 0.24% and 0.27% of the locations have shifts of more than 1 km to the nearest cells for the three operators, respectively. This result indicates that a user's time sequence of cell tower locations is able to capture user's actual footprints in a degree of few hundreds of meters; such accuracy is quite enough in various scenarios[5]. It is worth noting that such precision is still far below that obtained from civilian GPS locations (*i.e.*, the precision of GPS locations is around a few or tens of meters).

4.2. Span of human movement

The second study consists examines whether CDR can be adapted for measuring the geographical span of movement of subscribers. For that, we consider the *radius of gyration* parameter computed for each user $u \in \mathcal{U}$ (the user set of study), defined as the deviation of user's positions to its centroid position, *i.e.*, $R_g^u = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{r}_i^u - \mathbf{r}_{\text{centroid}}^u||^2}$, where $\mathbf{r}_{\text{centroid}}^u$ is the center of mass of locations of the user u, *i.e.*, $\mathbf{r}_{\text{centroid}}^u = \frac{1}{n}\sum_{i=1}^{n} \mathbf{r}_i^u$. This metric reflects how widely the subscribers move, and is hardly affected by the actual distance traveled, *e.g.*, having a subscriber repeatedly moving among fixed locations does not increase or decrease the radius of gyration, but moving to a new location does. Hence, the radius of gyration is often measured in human mobility studies [3, 5, 7, 22].

$(-1.318, 5.999)$, respectively.

[5]To be strictly accurate, our result shows an upper bound to the error of using cellular-level locations incurred by CDR, as the user could not always be assigned to the nearest antenna due to the networking optimization of cellular operator.

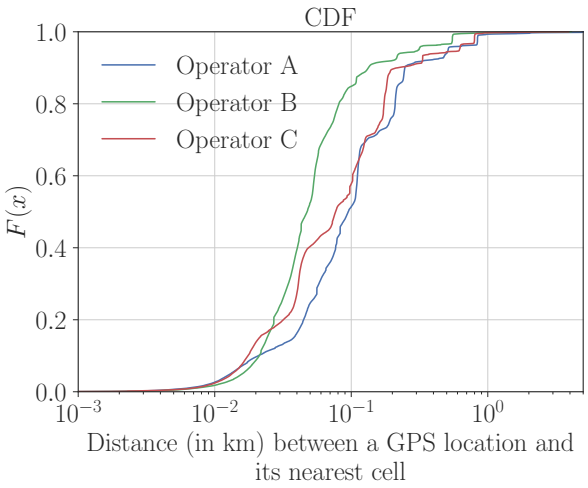
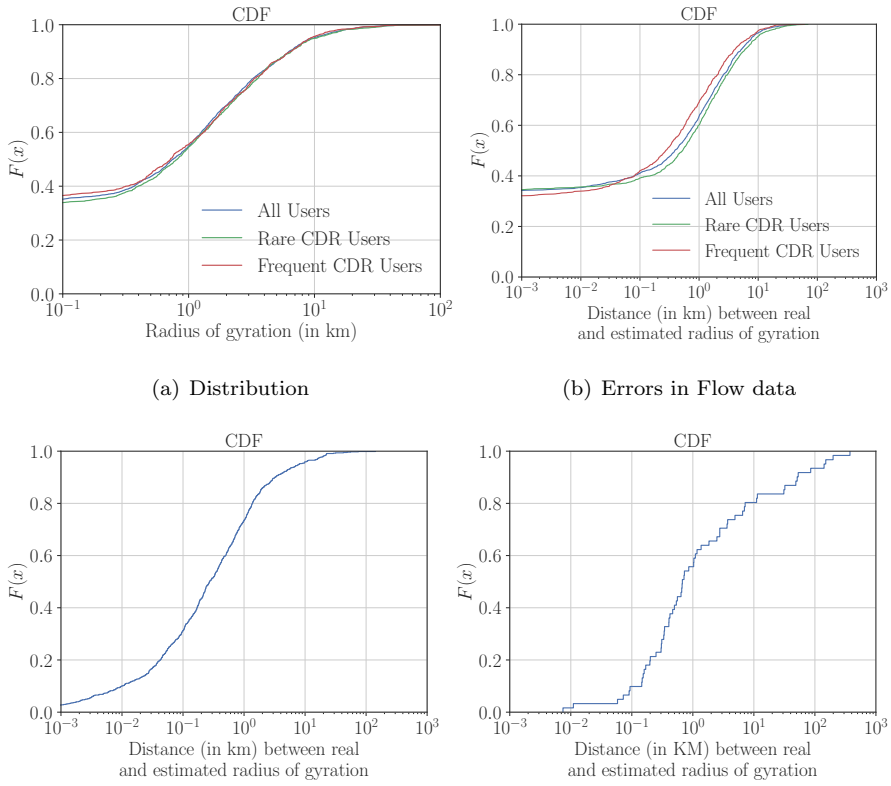


Figure 4: CDF of the distance to the nearest cell tower location across 718,987 GPS-based locations, collected from the users of the MACACO data in France.

We compute this metric for all users in each dataset introduced in Sec. 3. The obtained values represent the *estimated* (due to the temporal sparsity of the actual or the equivalent CDR) and the *real* (due to the finer granularity in the ground-truth provided by the Internet flow, MACACO, and Geolife datasets) radius of gyration.

Let us first consider the users of the Internet flow dataset and their radii of gyration. These users are regarded as three categories: *i.e.*, all users, rare CDR users, and frequent CDR users (as described in Section 3.2.1). Fig. 5(a) shows the CDF of the distributions of the radius of gyration over the three categories of users. The three distributions are quite similar, indicating that one can get a reliable distribution of R_g^u from a certain number of users even whatever activeness they show in the cellular network. Such distribution is often used to illustrate human mobility patterns over a large population [3, 5]. In fact, the number of CDR per user is not a primary concern, as it hardly impacts the obtained distribution.

We then compute the error between the real and the estimated radius of gyration: Fig. 5(b) shows the error when comparing the CDR with the Internet



(a) Distribution

(b) Errors in Flow data

(c) Errors in Macaco data

(d) Errors in Geolife data

Figure 5: (a) CDF of the radius of gyration over two categories (Rare and Frequent) of CDR users in the Internet flow dataset. (b)(c)(d) CDF of the distance between the real and estimated radius of gyration from CDR over the users of (b) Internet flow dataset, (c) MACACO dataset and (d) Geolife dataset.

flows, while Fig. 5(c) and 5(d) show the error for MACACO or Geolife datasets and their respective equivalent CDR, respectively. We notice the following:

- Intuitively, a more accurate radius of gyration can be obtained by taking into consideration more locations visited by the user: we notice that 92% of frequent CDR users have an error lower than 5 km, while the percentage decreases to 86% for rare CDR users.

- The span of movement according to CDR entails a small error. For approximately 90% of the Internet flow users, 95% of the MACACO users and 70% of the Geolife users, the error between the real and the estimated radius of gyration is less than 5 km. The higher error obtained from Geolife dataset may be interpreted by the irregular sampling in the original data and the presence of very large gaps between consecutive logs.

- Besides, the span of movement is lower than 100 meters, only in 30% and 10% of the Macaco and Geolife users, respectively. Also, the error larger than 10 km occurs in 5% and 20% of the two user sets, respectively.

These results confirm the previous findings on the limited suitability of CDR for the assessment of the spread of human mobility [6].

4.3. Missing locations

Due to the spatiotemporal sparsity, the mobility information provided by CDR is usually incomplete. We study the users of the Internet flow dataset and plot in Fig. 6(a) the ratio r_{N_L} of unique locations detected from CDR (N_L^{CDR}) to those from the ground-truth (N_L^{Flow}):

$$r_{N_L} = N_L^{\mathrm{CDR}} / N_L^{\mathrm{Flow}}. \tag{1}$$

We notice that 42% in the population of study (*i.e.*, all users) have their r_{N_L} higher than 80%. For these user, only 80% of the unique visited locations have already appeared in their CDR. The percentage of all the users having this criterion is slightly higher for the frequent CDR users (around 50%) and lower

for the rare CDR users (37%). These results confirm the benefit of adding a criteria to ensure a better completeness of mobility information, on the minimum number of CDR that a user has. Also, we can reconfirm that using CDR to study very short-term mobility patterns is not a good idea due to the high temporal sparsity and the lack of locations in CDR.

4.4. Important locations

The identification of significant places where people live and work is often an important preliminary step towards characterizing human mobility. To capture user's home and work locations, we separate the period of study into two time windows, mapping to work time (9 am to 5 pm) and night time (10 pm to 7 am) for both CDR and ground-truth. The places, where the majority of work time records occur, are considered a proxy of work locations; the equivalent records at night time are considered a proxy of home locations [23]. It is worth mentioning that, as the Internet flow dataset covers only $(10am, 6pm)$, we only infer from this dataset the work location.

Formally, let us consider a user $u \in \mathcal{U}$ from the user set of study. The visiting pattern of the user u is a sequence of samples $\left\{(\ell_u^0, t_u^0), (\ell_u^1, t_u^1), \ldots, (\ell_u^n, t_u^n)\right\}$, where the i-th sample (ℓ_u^i, t_u^i) denotes the location ℓ_u^i where user u is recorded at time t_u^i. The home location ℓ_u^H of the user u is then defined as the most frequent location during night time:

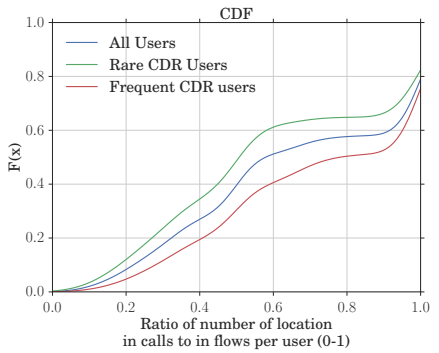
$$\ell_u^H = mode(\ell_u^i \mid t_u^i \in t^H), \tag{2}$$

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

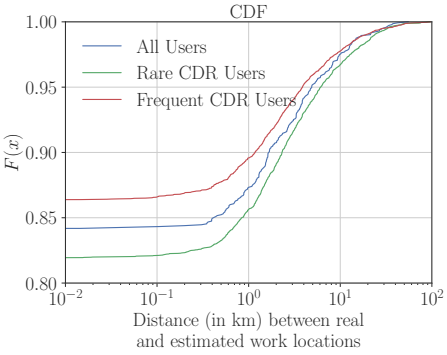
$$\ell_u^W = mode(\ell_u^i \mid t_u^i \in t^W), \tag{3}$$

where t^W is the work time interval.

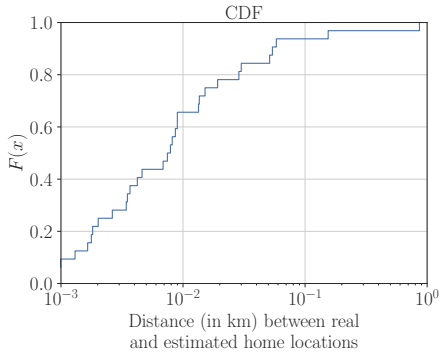
We use the definitions in Equation. (2) and (3) to determine home and work locations and then evaluate the accuracy of the CDR-based significant locations



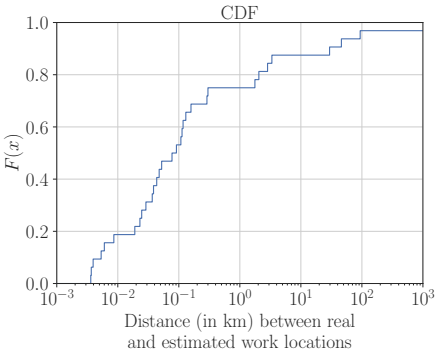
(a) Missing Ratio



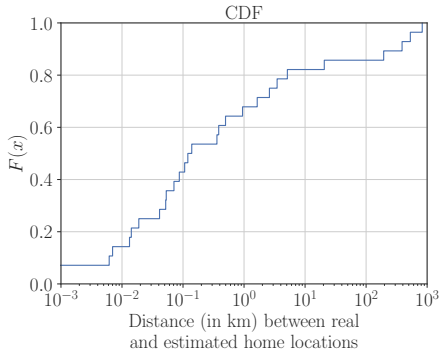
(b) Errors in Flow data: Work



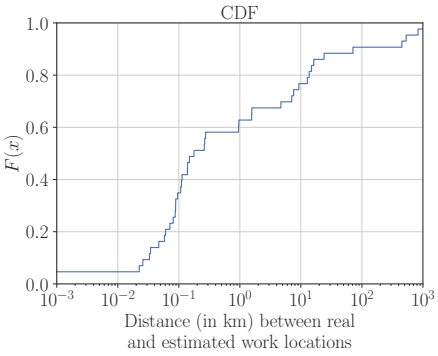
(c) Errors in Macaco data: Home



(d) Errors in Macaco data: Work



(e) Errors in Geolife data: Home



(f) Errors in Geolife data: Work

Figure 6: (a) CDF of the radio $r_N L$ of the number of location in each user's coarse-grained trajectory to the one in her fine-grained trajectory. (b)(c)(d)(e)(f) CDF of the distance between each user's real and estimated important locations located by her CDR and ground-truth: (b) work locations over the flow users; (c) home and (d) work locations over the MACACO user; (e) home and (f) work locations over the Geolife user.

by measuring the geographical distance that separates them from the equivalent locations estimated via the GPS ground-truth data.

The results are shown in Fig. 6(b)-(f) as the CDF of the spatial error in the position of home and work places for different user groups and for the three datasets. We can clearly observe the following.

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

- For both MACACO and Geolife users, The errors associated with work locations are sensibly higher than those measured for home locations. For instance, as shown in Fig. 6(d), while 75% of the MACACO users have an error of less than 300 meters, the work places of a significant portion of individuals (around 12%) are identified at a distance higher than 10 km from the position extracted from the GPS data. A close behavior can be noticed from the Internet flow and and Geolife users, as shown in Fig. 6(b) and Fig. 6(f). These large errors typically occur for users who do not seem to have a stable work location and may be working in different places depending on, *e.g.*, the time of day.

- The errors are significantly reduced when using cell tower locations as in the Internet flow dataset instead of actual GPS positions as in the MACACO or Geolife datasets. For the Internet flow users in Fig. 6(b), the error between the real and the estimated significant locations is null for approximately 85% of all the users, indicating that the usage of the coarse-grained dataset is fairly sufficient for inferring these significant locations.

- The error is non-null for the remaining Internet flow users (15%). Among them, 10% have relatively small errors (less than 5 km), while 5% have

errors larger than 5 km.

- Besides, there is only a slight difference in the distribution of the errors associated with work locations between rare CDR and frequent CDR users as shown in Fig. 6(b). The reason is that, most of CDR are generated in significant locations, and hence the most frequent location obtained from CDR of a user is likely to be be her actual work location during daytime. Still, it is relatively difficult to capture actual location frequencies if a user has only a few of CDR. Hence the rare CDR users have higher errors.

Overall, these results confirm previous findings [12], and further prove that CDR yield enough details to detect significant locations in users' visiting patterns. Besides, the results reveal a small possibility of incorrect estimation in the ranking among such locations.

5. Rationale for CDR completion

The previous results confirm the quality of mobility information inferred from CDR, regarding the span of user's movement and significant locations. They also indicate that some biases are present: specifically, as transient and less important places visited may lose, capturing one's entire history of locations is almost impossible. The good news is that, even in those cases, the error induced by CDR is relatively small, meaning that CDR are capable of locating users. The bad aspect is that common approaches designed for locating users via their GPS data are hardly applicable on CDR, as the data suffers from its temporal sparsity and usually does not have a stable data rate. Indeed, CDR only provide instantaneous information about user's locations at a few time instants over a whole day. Nevertheless, CDR still have come into wide use in the literature [10], because it is much easier to obtain CDR than to collect GPS surveys at a large scale.

As a matter of fact, *CDR temporal completion* aims at tackling that problem by filling the gaps in CDR, so to estimate users' locations between consecutive

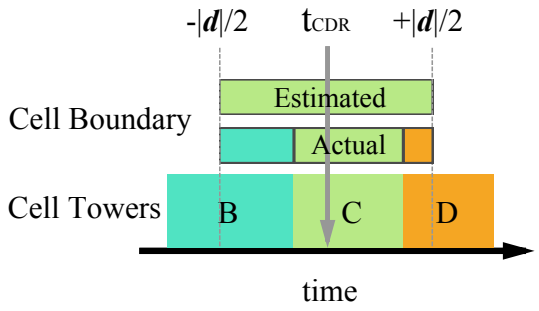


Figure 7: An example of a cell boundary in the `stop-by` approach: A period $(t_{\text{CDR}} - |\mathbf{d}|/2, t_{\text{CDR}} + |\mathbf{d}|/2)$ is given as a cell boundary at the cell C attached with a CDR entry at time t_{CDR}. In this cell boundary, the user is assumed to be at the cell C, while actually she moves from the cell B to D: this leads to a spatial error.

activities of voice calls or text messages. Several attempts at data completion have been made to date. In this section, we introduce and discuss two basic solutions used in the literature.

5.1. Basic solution: `Static`

A simple solution is to hypothesize that a user remains static at the same location where she is last seen in her CDR. This methodology is adopted, *e.g.*, by Khodabandelou *et al.* [24] to compute subscriber's presence in mobile traffic metadata used for population density estimation. We will refer to this approach as the `static` solution and will use it as a basic benchmark for more advanced techniques. It is worth noting that this solution has no spatiotemporal flexibility; its performance only depends on the number of CDR a user generates in the period of study: *i.e.*, higher is the number of CDR, lower will be the spatial error in the completed data by the `static` solution. In other words, there is no space (configurable setting or initial parameter) for customizing this solution to obtain better accuracy. Next, we introduce the `stop-by` basic solution and analyze its performance with respect to its setting.

5.2. Basic solution: `Stop-by`

Building on in-depth studies proving individuals to stay most of the time in the vicinity of their voice call places [25], Jo *et al.* [26] assume that users can be found at the locations where they generate some digital activities for an hour-long interval centered at the time when each activity is recorded. If the time between consecutive activities is shorter than one-hour, the inter-event interval is equally split between the two locations where the bounding events occur. This solution will be denoted as `stop-by` in the remaining sections.

The drawback of the `stop-by` is that it uses a constant hour-long interval for all calls as well as users in CDR, which may not always be suitable. This solution lacks flexibility in dealing with various human mobility behaviors. As exemplified in Fig. 7, a single CDR is observed at time t_{CDR} at cell C. Following the `stop-by` solution, the user is consider stable at this cell C during the period $\mathbf{d} = (t_{\text{CDR}} - |\mathbf{d}|/2, t_{\text{CDR}} + |\mathbf{d}|/2)$, while in fact the user has moved to two other cell towers during this period. We call the period estimated from an instant CDR entry, a *cell boundary*. In the example of Fig. 7, this cell boundary is overestimated.

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

Intuitively, a spatial error occurs if the user moves to other different cells during the cell boundary. To have a quantitive manner of such an error, we define the spatial error of a cell boundary with a period $\mathbf{d}$ as follows:

$$\text{error}(\mathbf{d}) = \frac{1}{|\mathbf{d}|} \int_{\mathbf{d}} \left\| c^{(\text{CDR})} - c_t^{(\text{real})} \right\|_{\text{geo}} \mathrm{d}t. \tag{4}$$

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

- When error($\mathbf{d}$) = 0, it means that the user stays at the cell $c^{(\mathrm{CDR})}$ during the whole cell boundary. Still, the estimation of $\mathbf{d}$ may be conservative, since a larger $|\mathbf{d}|$ could be more adapted in this case.

- When error($\mathbf{d}$) > 0, it means that the cell boundary is over-sized: the user in fact, moves to other cells in the corresponding time period. Thus, a smaller $|\mathbf{d}|$ would be more adapted.

Next section shows the evaluation of the impact of varying $|\mathbf{d}|$ on the spatial error, when the `stop-by` solution is used as CDR completion strategy.

5.2.1. `Stop-by` accuracy

We evaluate the performance in terms of the spatial error between CDR and Internet flows of the users introduced in Sec. 3. CDR are used to generate cell boundaries. Locations in Internet flows are adopted as actual locations for measuring spatial errors. The evaluation is performed in the two days of study (a Monday and a Sunday). On each day, we apply the `stop-by` solution on CDR over $|\mathbf{d}|$ = 10/30/60/120/180/240 minutes, and then measure spatial errors on cell boundaries generated.

We plot in Fig. 8(a) and (b) the CDF of the spatial error of cell boundary on Monday and Sunday, respectively. We observe error($\mathbf{d}$) = 0 for 80% of CDR on Monday (cf. 75% on Sunday) when $|\mathbf{d}|$ = 60 minutes and for 60% of CDR on Monday (cf. 53% on Sunday) when $|\mathbf{d}|$ = 240 minutes. This result strongly supports that the users remain in the cell coverage temporally around their CDR activities. Yet on Monday approximately 35% (cf. 40% on Sunday) of the observed users are *stable*, *i.e.*, each of them has only one location observed in the Internet flows and consequently, $R_g^u = 0$. The high percent of cell boundaries with error($\mathbf{d}$) = 0 shown in Fig. 8 may credit to these stable users, since any $|\mathbf{d}|$ for these users will be conservative: no spatial error occurs at all. To further analyze the spatial error, we exclude these users from the next evaluation, where only *mobile* ($R_g^u > 0$) users are involved.

An interesting issue regarding the spatial error is the coverage of each cell

tower. Intuitively, when a cell covers a bigger area, the user is expected to stay inside the cell for a longer time. We estimate for each cell its coverage as the *cell radius*. Since we have no knowledge of the actual cell coverage, we assume a homogeneous propagation environment and an isotropic radiation of power in all directions at each cell tower; we roughly estimate each cell's radius using a composition of Voronoi cells extracted from CDR covering the area. A cell tower's coverage is considered the smallest circle centered at the cell tower location. The radius of that circle is the largest distance between such location and the Voronoi polygon contour. The resulting circle covers the Voronoi polygon entirely and yields overlapping coverage at cell boundaries. In the area of our study (shown in Fig. 3), 70% of the cells have radii within 3 km, and the median radius is approximately 1 km.

Hereby, we evaluate the probability of having a cell boundary with a null spatial error, as $P_{e0} = \mathbf{Pr}\{\text{error}(\mathbf{d}) = 0\}$. Fig. 9(a) and Fig. 9(b) present the probabilities P_{e0} grouped by the cell radius, when applying varying sizes of cell boundary on the days of study. We notice the following.

- The probability P_{e0} decreases with the increasing period marked by $|\mathbf{d}|$, indicating that using a large period on the cell boundary increases the chances of having a spatial error. For instance, for $|\mathbf{d}| = 30$ minutes, the probability of having a null spatial error is around 0.7 depending on the date and on the cell radius. When a larger $|\mathbf{d}|$ is used, the probability significantly increases (*e.g.*, for $|\mathbf{d}| = 60$ minutes, the probability P_{e0} reduces and is around 0.6).

- The probability P_{e0} increases positively with the cell radius r. This trend is seen on both Monday and Sunday (except some cases), indicating that the cell size has an impact on the time interval during which the user stays within the cell coverage. Intuitively, for a moving user, if small cells surround her, quite often may handovers occur; if big cells do, less often may handovers occur.

The results support the idea that there is a strong correlation between the

cell boundary and the cell coverage. Nevertheless, since CDR are usually sparse in time, using a small cell boundary could only cover an insignificant amount of cell visiting time, while using a big cell boundary increases the risk of having a non-null spatial error. To investigate this trade-off, we plot the variation of the statistical distribution of the spatial errors after excluding the null errors (*i.e.*, keeping only cases with non-null error($\mathbf{d}$)) in Fig. 9(c) and Fig. 9(d). We observe that:

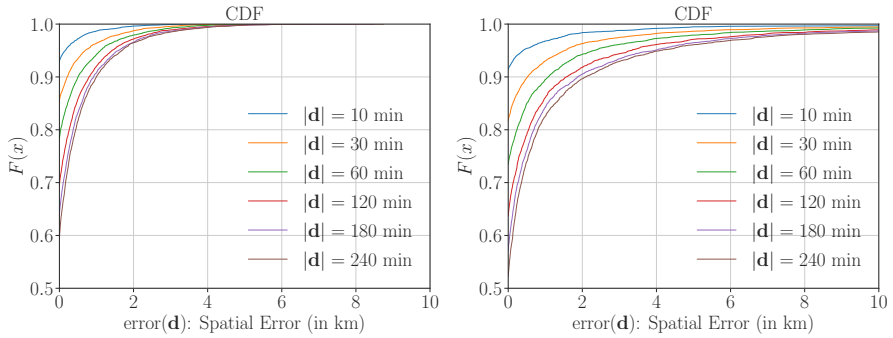
- The spatial error varies widely: it goes from less than 1 km to very huge values (up to 3.6 km on Monday and to 7.5 km on Sunday). Hence, for some users, the `stop-by` solution is unsuitable for reconstructing visiting patterns due to the presence of such high spatial errors.

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

5.3. Major outcomes

Overall, we assert that cell boundary estimates user's locations with a high accuracy when $|\mathbf{d}|$ is small. This validates the previous finding that users usually stay in proximity of call locations for certain time. The accuracy reduces significantly (spatial error arise), when increasing $|\mathbf{d}|$. Hence, the trade-off between the time coverage and the accuracy should be carefully considered when completing CDR using cell boundaries. Using a constant $|\mathbf{d}|$ over all users as well as CDR, as in the `stop-by` solution, makes hard achieving a good coverage vs. accuracy trade-off.

As an enhancement to the `stop-by` and `static` solutions, the data completion strategies introduced in the following sections leverage human nature in terms of (1) their attachment to a specific location in night periods or (2) their



(a) CDF of Spatial Error: Monday

(b) CDF of Spatial Error: Sunday

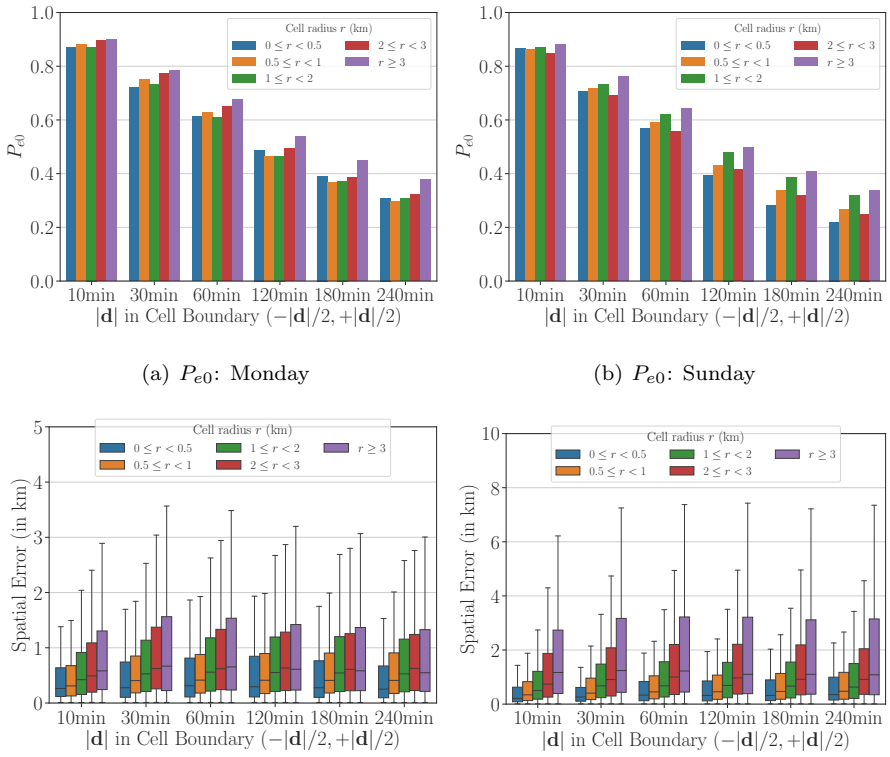
Figure 8: CDF of the spatial error of cell boundaries of CDR generated by the `stop-by` solution over two groups of the users in the Internet flow dataset on (a) Monday and (b) Sunday.

tendency to stay for some time interval in the vicinity of the location where the digital activity takes place. In particular, we classify our strategies into nighttime and daytime completion: Sec. 6 presents nighttime completion strategies inferring the home location of users; Sec. 7 introduces our adaptive cell boundary strategies leveraging the human mobility tendency during the daytime.

6. Identifying home boundaries

This section discusses the CDR completion during nighttime. The main goal of the introduced strategies is first, to infer temporal boundaries where users are located, with a high probability, at their home location, *i.e.*, to identify their *home boundary*. Gaps in CDR occurring at home boundary of users are then filled with the identified home location (cell). Overall, leveraging that CDR allow identifying the home location of individuals with high accuracy, the following strategies extend the `stop-by` solution (presented in Sec. 5.2).

- The `stop-by-home` strategy adds fixed temporal home boundaries in the `stop-by` technique. If a user's location is unknown during the night time interval $\mathbf{h} = (10pm, 7am)$, due to the absence of CDR in that period, the user will be considered at her home location during $\mathbf{h}$. Note that the home



(a) P_{e0}: Monday

(b) P_{e0}: Sunday

(c) Error: Monday

(d) Error: Sunday

Figure 9: Spatial error of cell boundaries of CDR generated by the `stop-by` solution over users with their $R_g > 0$: (a)(b) the probability (P_{e0}) of having a non-error cell boundary $(-|\mathbf{d}|, |\mathbf{d}|)$, where $|\mathbf{d}|$ is 10/30/60/120/180/240 minutes, under several groups of cell radius on (a) Monday and (b) Sunday; (c)(d) Boxplot of the spatial error grouped by the cell radius and the time period of cell boundary on (c) Monday and (d) Sunday. Each box denotes the median and $25^{th} - 75^{th}$ percentiles and the whiskers denote $5^{th} - 95^{th}$ percentiles.

location is the user's most active location during $\mathbf{h}$, and CDR not in $\mathbf{h}$ is completed via the `stop-by`, so do the other extended strategies.

- The `stop-by-flexhome` strategy refines the previous approach by exploiting the diversity in the habits of individuals. In this technique, the fixed night time temporal boundaries are relaxed and become flexible, which allows adapting them on a per-user basis. Specifically, instead of considering $\mathbf{h} = (10pm, 7am)$ as the fixed home boundaries for all users, we compute for each user $u \in \mathcal{U}$ the most probable interval of time $\mathbf{h}^{(u)}_{\text{flex}} \subseteq \mathbf{h}$ during which the user is at her home location. Then, as for `stop-by-home`, the user will be considered at her home location if gaps are identified at her CDR during $\mathbf{h}^{(u)}_{\text{flex}}$.

- The `stop-by-spothome` strategy augments the previous technique by accounting for positioning errors that can derive from users (1) who are far from home during some nights, or (2) from ping-pong effects in the association to base stations when the user is within their overlapping coverage region. In this approach, if a user's location during $\mathbf{h}^{(u)}_{\text{flex}}$ is not identified, and if she is last seen at no more than 1 kilometer from her home location, she is considered to be at her home location.

As we hereby focus on the completion during nighttime, we use $|\mathbf{d}| = 60$ min for the `stop-by` part of all the strategies in this section. More advanced completing strategies for daytime completion are discussed in Sec. 7.

In the following, we compare the above strategies with the `static` and the pure `stop-by` solution introduced in Sec. 5, with respect to two dual perspectives. The first is *accuracy*, *i.e.*, the spatial error between mobility metrics computed from ground-truth GPS data and from CDR completed with the different techniques above. The second is *coverage*, *i.e.*, the percent of the time during which the CDR completion technique can determine the position of a user. Indeed, the `static` solution discussed in Sec. 5 provide user's locations at all times, but this is not true for the `stop-by` or our derived techniques. In this

case, the CDR is completed only for a portion of the total period of study, and user's whereabouts remain unknown in the remaining time.

6.1. Coverage and accuracy

We compute the geographical distance between MACACO and Geolife GPS samples and their equivalent CDR, when the following completion strategies are used: `static`, `stop-by`, and derived solutions introduced above. These strategies are not designed to provide positioning information at all times expect the `static` solution. As this limitation is already evaluated by the coverage metric, we need to ensure a fair comparison of accuracy. To that end, distances are only measured for GPS samples whose timestamp fall in the time periods for which completed data is available.

Fig. 10(a) and 10(b) summarize the results of our comparative evaluation of accuracy, and allow drawing the following main conclusions.

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

We conclude that those solutions, based on a model where the user remains static for a limited temporal interval around each measurement time, are clear winners when it comes to accuracy of the completed data. This result supports previous observations on the mostly static behavior of mobile subscribers [25]. Moreover, the information of home locations can be successfully included in such models, by accounting for specificities in each user's habits at night.

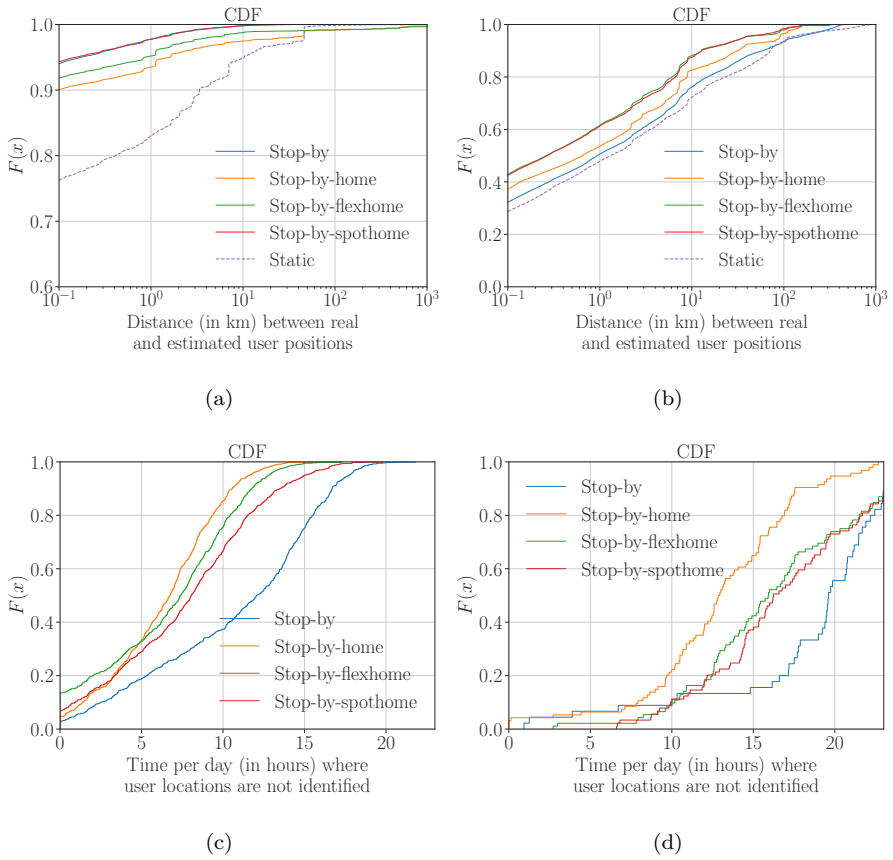


Figure 10: CDF of the spatial error (in km) between samples from the GPS and completed data over the (a) MACACO and (b) Geolife data. CDF of the temporal coverage of completed data over the (c) MACACO and (d) Geolife data.

The `stop-by` and derived solutions provide incomplete temporal coverage by design. Fig. 10(c) and 10(d) show the CDF of the hours per day during which a user's position cannot be identified by such solutions for the users in both MACACO and Geolife datasets, respectively. We can notice from these figures that the coverage performance is very heterogeneous across users, for all solutions: it can range from one hour per day for some individuals up to 23 hours per day for other subscribers. By comparing both figures, we assert that the irregular sampling of the Geolife dataset, which occurs often, is translated directly to higher time gaps (*i.e.*, time per day where user locations are not identified) by the different techniques. Moreover, for both datasets, the `stop-by` approach yields the worst result, with an unknown user position of 12 hours per day and 19 hours per day in the median case, for both datasets, respectively. The refinements of the `stop-by` solution increase the coverage as expected, since the derived approaches aim at defining the users' positions overnight, when actual CDR are absent. The improvement is significant, with a median gain of $4-5$ hours for MACACO dataset and a gain of $3-7$ hours for the Geolife users dataset over the basic `stop-by` solution.

Overall, the combination of the results in Fig. 10 indicates that among the different completion techniques, the `stop-by-spothome` solution achieves the best combination of high accuracy and fair coverage.

7. Identifying cell boundaries

This section discusses the CDR completion during the daytime. Still, we infer temporal boundaries of users. Nevertheless, differing from the nighttime case in Sec. 6, we leverage context of human mobility habits while communicating instead of regular or personal nighttime behavior. We target the completion for common CDR during daytime by extending the time span of the position associated with each communication activity to *cell boundaries*.

7.1. Factors impacting cell boundaries

The `stop-by` solution comes from that a user stays within cell coverage for a certain time before and after a communication activity takes place, and converts CDR to cell boundaries by tackling this stable time to one hour. As introduced in Sec. 5, this solution leads to a spatial bias that relates to coverage of the cell (as its radius) while communicating, and consequently, needs improvement.

Hereafter, we aim to answer the following question: how to choose a proper and adaptive period for a cell boundary instead of an inflexibly fixed-to-all period? To answer the question, we need to understand the correlation between the routine behavior of users, in terms of voice/text communication and their mobility. For this, we first investigate human behavior factors that can be extracted from CDR. These factors may be used to determine cell boundaries. We classify these factors describing human behavior in three classes, named event-related, long-term behavior, and location-related factors, described hereafter. Then, we use them to design our approaches for estimating cell boundaries.

7.1.1. Event-related factors

These are the metadata of CDR, including the activity's `time`, `type` (call/message), and `duration`[6]. Intuitively, these factors have direct effects on cell boundaries. For instance, in terms of `time`, a user may stay within a fixed cell during her whole working period. In terms of `type` and `duration`, one making a long phone call may stay stable for a long time, but one sending a short text message may be on the move. Besides, these factors are easily extracted from any CDR, since they are necessary to describe events.

7.1.2. Long-term behavior factors

This kind of factors describes a user in terms of her long-term behaviors. They are the radius of gyration (`URg`) of a user, the number of a user's locations (`ULoc`) appearing in the observing period, and the number of a user's active days

[6]For this attribute, the duration of a text message is set to 0 second.

(UDAY). These factors characterize a user by giving senses of *(i)* her long-term mobility and *(ii)* her habit on generating calls and text messages. Intuitively, they are not directly related to estimating cell boundaries. We choose them to seek opportunities to build an indirect correlation between cell boundaries and long-term human behaviors. For each user, the factors are computed from the CDR dataset by aggregating her CDR during the whole 3-month period of study.

7.1.3. Location-related factors

The first factor in this category is related to the cell coverage, *i.e.*, the average call radius (CR), shown being capable of contributing to determining "boundaries" of human movement already in Sec. 5.

The remaining location-related factors describe the location where the activity happens regarding its importance to the user. We select them into design given the intuition that a user may stay a huge amount of time in her "important" places. For this, we apply the algorithm presented by Isaacman *et al.* [27]. Their algorithm is designed to determine prominent locations where the user usually spends a large amount of time and/or visits frequently.

The algorithm applies Hartigan's clustering [28] on visited cell locations of users in CDR and then, use logistic regression to estimate a location's importance to the user using factors extracted from the cluster that the location belongs to. To start with, the cluster approach chooses the cell tower from the first CDR and makes it the first cluster. Then, it recursively checks all cell towers in the remaining CDR. If a cell tower is within the distance threshold (we use 1 kilometer) to the centroid of a certain cluster, the cell tower is added to the cluster, and the centroid of the cluster is moved to the weighted average of the locations of all the cell towers in the cluster. Weights of locations are numbers of days that they appear in the observing period. The clustering finishes once all cell towers are assigned to clusters.

Once clusters are defined, the importance of each cluster is identified according to the following observable factors: *(i)* the number of days on which any cell

tower in the cluster was contacted (`CDay`); *(ii)* the number of days that elapse between the first and the last contact with any location in the cluster (`CDur`); *(iii)* the sum of the number of days cell towers in the cluster were contacted (`CTDay`); *(iv)* the number of cell towers inside the cluster (`CTower`); *(v)* the distance from the registered location of the activity to the centroid of the cluster (`CDist`).

These factors derived from a cluster correlate with the time that the user spends in the cluster's locations, as shown by Isaacman *et al.* via their logistic regression model [27]. It is worth mentioning we can not reproduce the model in [27], since the used ground-truth is not publicly available. Thus, we decided to use those factors but to build our cell boundary approaches.

7.2. Supervised cell boundary estimation

So far, we have introduced human behavior factors that might be directly or indirectly related to cell boundaries. Recall the question of determining a cell boundary properly. For that, we need a model linking the setting of cell boundaries with human behavior factors. For this, in the following, we introduce two approaches according to different models built via supervised machine learning. Both approaches are able to adaptively build cell boundaries from the aforementioned behavior factors of communicating events and have far more flexibility than the `stop-by` solution does.

7.2.1. Formalizing cell boundaries

We define two kinds of cell boundaries: symmetric and asymmetric boundaries. Given a CDR entry at time t, generating its cell boundary means to expand the instance time t to a period $\mathbf{d}$ assuming that the user stays in the cell during the whole period. For a symmetric cell boundary, this period is generated from a CDR-based parameter t_{sym} as $\mathbf{d} = (t - t_{sym}, t + t_{sym})$, and is symmetric to the CDR time t. Similarly, the period of an asymmetric cell boundary is generated from two independent parameters t^{+}_{asym} and t^{-}_{asym} as $\mathbf{d} = (t - t^{-}_{asym}, t + t^{+}_{asym})$.

We design the `sym-adaptive` and `asym-adaptive` approaches by modeling the estimation of a cell boundary to two regression problems corresponding to the symmetric and asymmetric cases, respectively. Given a CDR entry, its corresponding factors are extracted and converted to an input vector $\mathbf{x}$, given the rules that: *(i)* the categorial factor `type` is converted to two binary features by one-hot encoding; *(ii)* the `time` is converted to the distances (in seconds) separating it from $10am$ and from $6pm$; *(iii)* the other factors are used as scalar values they are. For the `sym-adaptive` approach, the parameter t_{sym} is modeled from a function $t_{sym} = f_{sym}(\mathbf{x})$. For the `asym-adaptive` approach, the parameters t^+_{asym} and t^-_{asym} are modeled from two independent functions $t^+_{asym} = f^+_{asym}(\mathbf{x})$ and $t^-_{asym} = f^-_{asym}(\mathbf{x})$.

7.2.2. Estimating cell boundaries via supervised learning

Our goal is to predict t_{sym}, t^+_{asym} and t^-_{asym} as accurate as possible to their actual values in a cell boundary. For that we utilize Gradient Boosted Regression Trees (GBRT) [29] to estimate all the functions in the two approaches. GBRT has advantages in terms of flexibility for heterogeneous features, good predictive power, and training speed. A GBRT model is an ensemble of regression trees with limited depth. In the model, each tree divides the input space into disjoint regions and predicts a constant value in each region. The GBRT technique combines the predictive power of all regression trees having a weak predicting performance by making a joint predictor: it is proved that the performance of such a joint predictor is better than of each single regression tree. The ensemble has one tree at the beginning. During each iteration, a new regression tree is added to the ensemble by minimizing the loss function via gradient descent.

We train three GBRT predictors for f_{sym}, f^+_{asym} and f^-_{asym} using 50% of users, randomly selected from the two available days (*i.e.*, a Monday and a Sunday) of the CDR and Internet flow datasets, as in Sec. 5.2.1. Cell boundaries are then generated from the CDR. Actual locations are extracted for verification from flows (used as the ground-truth). Finally, we extract the input vector of factors $\mathbf{x}$ as well as the parameters t_{sym}, t^+_{asym}, and t^-_{asym} of the cell boundary,

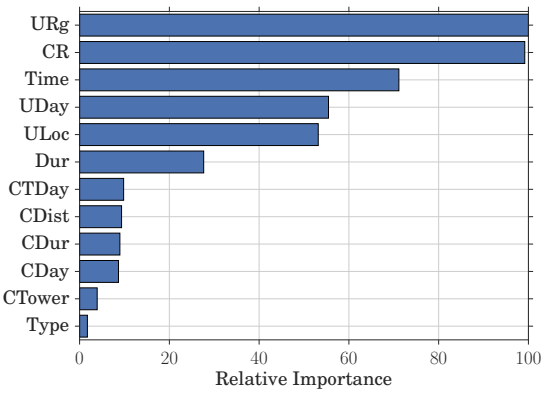


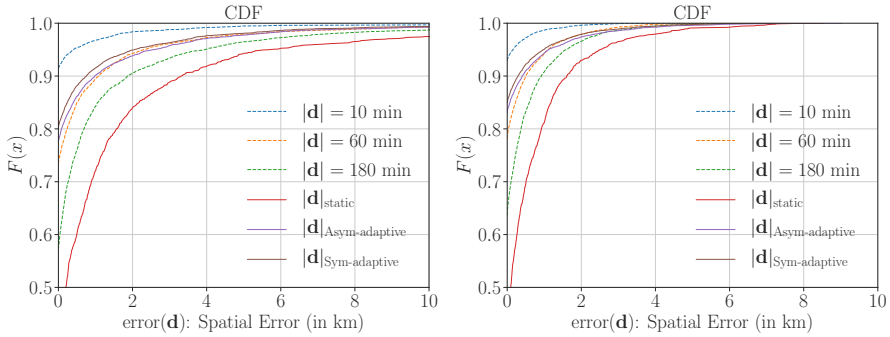
Figure 11: Relative Importance of features in determining accurate cell boundaries.

from the CDR and its ground-truth. The predictors use the Huber loss function. Fig. 11 shows the relative importance of factors with respect to the estimation of a cell boundary on the training of GBRT predictors. The importance indicates the degree of a feature contributing to the construction of the regression trees. This figure allows us drawing the following main conclusions, valid for both approaches.

- We notice the three most important factors: the timestamp of the activity, the cell radius, and the radius of gyration. This indicates that for a cell, how long a user stays inside it mainly depends on its size, the precise time the activity occurred, and the user's long-term mobility.
- Surprisingly, the activity's `type` is the most pointless factor, indicating that knowing whether a user generates a call or a message is useless in determining a cell boundary.

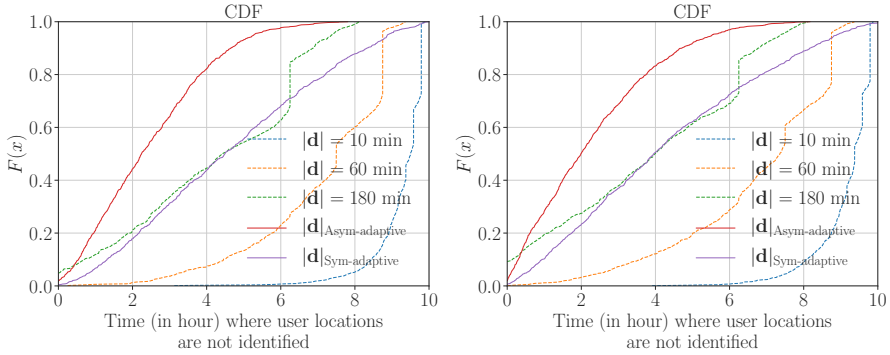
7.3. Coverage and accuracy

We compare our two trained approaches with the `stop-by` and `static` approaches using the CDR from the remaining 50% of the randomly-selected users. We let the `stop-by` approach generate fixed-period cell boundaries using $|\mathbf{d}| = 10/60/180$ minutes. As in Sec. 6, we make a comparative study by



(a) Spatial Error: Sunday

(b) Spatial Error: Monday



(c) Coverage: Sunday

(d) Coverage: Monday

Figure 12: CDF of the spatial error of cell boundaries computed on (a) Sunday and (b) Monday; CDF of the temporal coverage of completed data on (c) Sunday and (d) Monday, across the `stop-by`, `static`, `sym-adaptive`, and `asym-adaptive` approaches.

evaluating the solutions regarding *accuracy* and *coverage*, where the accuracy is measured by evaluating the *spatial error* (Equation. 4) introduced in Sec. 5. Recall that a good data completion approach should cover the observing period as much and precise as possible, *i.e.*, satisfying high accuracy and coverage simultaneously.

Fig. 12(a)(b) plots the distribution of the spatial error over all cell boundaries. Our results confirms that the spatial error increases as t_d becomes larger when using the `stop-by` approach. More importantly, the performance of the two adaptive approaches is nearly as good as the `stop-by` approach with $|\mathbf{d}| = 60$ minutes in terms of the spatial error. As expected, the `static` solution has the worst performance, as we observe in home boundaries using the MACACO and Geolife datasets.

Fig. 12(c)(d) plots the distribution of the temporal coverage per users over all approaches except `static` (which always covers the whole period). Since the Internet flow dataset only covers $(10am, 6pm)$, the x-axis is from 0 to 10 hours. We notice that both adaptive approaches show a splendid performance of temporal coverage: users have far less time with unidentified locations in the data completed by the adaptive techniques than by the `stop-by` approach. The asym-adaptive approach even outperforms the sym-adaptive approach.

Overall, we see a clear advantage of the adaptive approaches over the basic solutions in Fig. 12. The `sym-adpative` approach achieves a fairly good combination of accuracy and temporal coverage: it can complete more time, while it can still ensure accuracy. As to the asym-adaptive approach, it performs better in terms of coverage with losing a small degree of accuracy, compared with the sym-adaptive approach.

8. Conclusion

We leveraged novel datasets of GPS logs and CDR to characterize the bias induced by the use of CDR for the study of human mobility, and evaluate data completion techniques to reduce such a bias. Our results confirm previous

findings of the limitations imposed by the sparsity of CDR and provide the first investigation of techniques for CDR completion. Specifically, we show solutions that *(i)* extend for a limited amount of time the stays of users at known locations, and *(ii)* place users at their home locations with a grain of salt can achieve good accuracy and fair coverage. Such novel approaches outperform previous proposals in the literature.

References

[1] H. Zang, J. C. Bolot, Mining call and mobility data to improve paging efficiency in cellular networks, in: MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking, ACM, New York, New York, USA, 2007, pp. 123–134.

[2] K. Y. Lai, Z. Tari, P. Bertok, Supporting user mobility through cache relocation, Mobile Information Systems 1 (4) (2005) 275–307.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, S. R. Das, Understanding traffic dynamics in cellular data networks, in: INFOCOM, 2011 Proceedings IEEE, IEEE, 2011, pp. 882–890.

[4] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the 18th international conference on World wide web, ACM, 2009, pp. 791–800.

[5] M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.

[6] G. Ranjan, H. Zang, Z.-L. Zhang, J. Bolot, Are call detail records biased for sampling human mobility?, ACM SIGMOBILE Mobile Computing and Communications Review 16 (3) (2012) 33–44.

[7] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of Predictability in Human Mobility, Science 327 (5968) (2010) 1018–1021.

[8] C. Iovan, A.-M. O. Raimond, T. Couronné, Z. Smoreda, Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies., AGILE Conf. (Chapter 14) (2013) 247–265.

[9] M. Ficek, L. Kencl, Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model., IEEE INFOCOM 2012 (2012) 469–477.

[10] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, Large-scale Mobile Traffic Analysis: a Survey, IEEE Communications Surveys & Tutorials PP (99) (2015) 1–1.

[11] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, T. He, Exploring human mobility with multi-source data at extremely large metropolitan scales, in: Proc. of MobiCom, New York, USA, 2014. doi:10.1145/2639108.2639116.

[12] G. Ranjan, H. Zang, Z.-L. Zhang, J. Bolot, Are call detail records biased for sampling human mobility?, SIGMOBILE Mob. Comput. Commun. Rev. 16 (3) (2012) 33–44. doi:10.1145/2412096.2412101.

[13] H. H. Jo, M. Karsai, J. Karikoski, K. Kaski, Spatiotemporal correlations of handset-based service usages, EPJ Data Science 1 (2012) 1–18.

[14] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, Ranges of human mobility in los angeles and new york, in: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, IEEE, 2011, pp. 88–93.

[15] S. Hoteit, G. Chen, A. Viana, M. Fiore, Filling the gaps: On the completion of sparse call detail records for mobility analysis, in: Proceedings of the Eleventh ACM Workshop on Challenged Networks, CHANTS '16, ACM, New York, NY, USA, 2016, pp. 45–50. doi:10.1145/2979683.2979685.

[16] G. Chen, A. C. Viana, C. Sarraute, Towards an adaptive completion of sparse call detail records for mobility analysis, in: Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on, IEEE, 2017, pp. 302–305.

[17] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, Nature 435 (2005) 207.

[18] EU CHIST-ERA Mobile context-Adaptive CAching for COntent-centric networking (MACACO) project, https://macaco.inria.fr/.

[19] N. Eagle, A. (Sandy) Pentland, Reality mining: Sensing complex social systems, Personal Ubiquitous Comput. 10 (4) (2006) 255–268. doi:10.1007/s00779-005-0046-3.

[20] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the World Wide Web Conference, New York, NY, USA, 2009.

[21] Opencellid, http://wiki.opencellid.org/wiki/FAQ.

[22] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, G. Pujolle, Estimating human trajectories and hotspots through mobile phone data, Computer Networks 64 (2014) 296–307.

[23] S. Phithakkitnukoon, Z. Smoreda, P. Olivier, Socio-geography of human mobility: A study using longitudinal mobile phone data, PLoS ONE 7 (6) (2012) 1–9. doi:10.1371/journal.pone.0039253.

[24] G. Khodabandelou, V. Gauthier, M. El-Yacoubi, M. Fiore, Population estimation from mobile network traffic metadata, in: IEEE World of Wireless Mobile and Multimedia Networks (WoWMoM), 2016.

[25] M. Ficek, L. Kencl, Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model, in: INFOCOM,

2012 Proceedings IEEE, 2012, pp. 469–477. doi:10.1109/INFCOM.2012.6195786.

[26] H.-H. Jo, M. Karsai, J. Karikoski, K. Kaski, Spatiotemporal correlations of handset-based service usages, EPJ Data Science 1 (2012) 1–18. doi:10.1140/epjds10.

[27] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, Identifying important places in peopleś lives from cellular network data, in: Pervasive computing, Springer, 2011, pp. 133–151.

[28] J. A. Hartigan, Clustering, Annual review of biophysics and bioengineering 2 (1) (1973) 81–102.

[29] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.