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, 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. Tests with real-world datasets show that our solutions can reduce temporal sparsity in CDR by recovering 75% of daytime hours, while retaining a spatial accuracy within 1 km for 95% of the completed data.

Keywords: Call Detail Records, spatiotemporal trajectories, data sparsity,

[*] Corresponding author

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)

cellular networks, mobility, movement inference.

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 movements of mobile network subscribers. 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 as a primary source of data for large-scale mobility studies. CDR contain information about *when*, *where* and *how* a mobile network subscriber generates voice calls and text messages, and are collected by mobile network operators for billing purposes. These records usually cover large populations [10], which makes them a practical choice for performing large-scale human mobility analyses.

CDR can be regarded as footprints of individual mobility and can thus be used to infer visited locations, to learn recurrent movement patterns, and to measure mobility-related features. Despite the significant benefits that CDR bring to human mobility analyses, an indiscriminate use of CDR may question the validity of research conclusions. Indeed, CDR have limited accuracy in the spatial dimension (as the user's location is known at a cell sector or in a base station level) and the temporal dimension (since the device's position is only recorded when it sends or receives a voice call or text message). This is a severe limitation, as a cell (sector) typically spans thousands of square meters at least, and even a very active mobile network subscriber only generates a few tens of voice or text events per day. Overall, CDR are characterized by spatiotemporal sparsity, and understanding whether and to what extent such sparsity affects mobility studies is a critical issue.

Existing studies and limitations: A few previous works have investigated the validity of mobility studies based on CDR. An influential analysis [6] observed that using CDR allows to correctly identify popular locations that account for 90% of each subscriber's activity; however, biases may arise when measuring individual human mobility features. Works such as [6] or the later [11] discussed biases introduced by the incompleteness of positioning information, *i.e.*, the fact that CDR do not capture every location a user has travelled through. Nevertheless, another important bias of CDR, caused by the use of cell tower locations of mobile network subscribers in their footprints instead of their actual positions, has been overlooked in the literature.

Another open research problem is that of completing spatiotemporal gaps in CDR. The most intuitive solution is to consider that the location in an entry of CDR stays representative for a time interval period (*e.g.*, one hour) centered on the actual event timestamp [7, 12]. So far and to the best of our knowledge, no more advanced solution has been proposed in the literature to fill the spatiotemporal gaps in CDR.

Our work and contributions: In this paper, we explore the following research questions. First, we investigate how the spatiotemporal sparsity of CDR affects the accuracy and incompleteness of mobility information, by leveraging CDR and cell tower deployments in metropolitan areas. Second, we evaluate the biases caused by such spatiotemporal sparsity in identifying important locations and measuring individual movements. Third, we study the capability of CDR of locating a user continuously in time, *i.e.*, the degree of completeness of the data. Answering these questions leads to the following main contributions:

- We show that the geographical shifts, caused by the mapping of user locations to cell tower positions, are less than 1 kilometer in the most of cases (*i.e.*, 85%−95% in the entire country or over 99% in the metropolitan areas in France), and the median of the shifts is around $200 - 500$ meters (varying across cellular operators). This result substantiates the validity of many large-scale analyses of human mobility that employ CDR.
- We provide a confirmation of previous findings in the literature regarding the capability of CDR to model individual movement patterns: (1) CDR provides the limited suitability for the assessment of the spread of human mobility and the study of short-term mobility patterns; (2) CDR yield enough details to detect significant locations in users' visiting patterns and to estimate the ranking among such locations.
- We implement different 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 approaches that outperform existing ones, and carry out extensive tests on their performance with substantial real-world datasets collected by mobile network operators and mobility tracing initiatives. Validations against ground-truth movement information of individual users show that, on average, our proposed adaptive techniques can achieve an increased temporal completion of CDR data (75% of daytime hours) and retain significant spatial accuracy (having errors below 1 km in 95% of completed time). Compared with the most common proposal in the literature, our best adaptive approach outperforms by 5% of accuracy and 50% of completion.

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 analyses.

In Sec. 5, we discuss the rationale for CDR completion and errors introduced by common literature related approaches. In Sec. 6 and 7, we describe original CDR completion solutions that achieve improved 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 use of CDR. Understanding whether and to what extent these biases affect human mobility studies is a subject that has been only partly addressed. The early paper by Isaacman [13] unveiled that using CDR as positioning information may lead to a distance error within 1 km compared to ground-truth collected from 5 users. In a seminal work, Ranjan *et al.* [6] showed that CDR are capable of identifying important locations, but they can bias results when more complex mobility metrics are considered; the authors leveraged CDR of very active mobile network subscribers as ground-truth. In our previous study [14], we confirmed these observations using a GPS dataset encompassing 84 users. In the present work, we confirm the observation in [6], and push them one step further by also considering the spatial bias introduced by CDR. For the sake of completeness, we mention that results are instead more promising when mobility is constrained to transportation networks: Zhang *et al.* [11] found 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.

Also relevant to our study are attempts at mitigating the spatiotemporal sparsity of CDR through completion techniques. The legacy approach in the literature consists in assuming that a user remains static from some time before and after each communication activity. The span of the static period, which we will refer to as *temporal cell boundary* hereinafter, is a constant system parameter that is often fairly arbitrary [12, 14]. In this paper, we extend previously proposed solutions [14, 15], and introduce two adaptive approaches to complete subscribers' trajectories inferred from CDR.

3. Datasets

We leverage two types of datasets in our study. *Coarse-grained* datasets are typical CDR data and feature significant spatiotemporal sparsity as well as user locations mapped to cell tower positions. *Fine-grained* datasets describe the mobility of the same user populations in the coarse-grained datasets with a much higher level of details and spatial accuracy. The coarse-grained datasets are treated as CDR in our experiments, while the corresponding fine-grained datasets are used as ground-truth to validate the results.

We have access to one coarse-grained (CDR) and three fine-grained (Internet flow, MACACO, and Geolife) datasets. The CDR and Internet flow datasets share the same set of subscribers, and thus represent a readily usable pair of coarse- and fine-grained datasets. Coarse-grained counterparts of the MACACO and Geolife datasets are instead artificially generated, by downsampling the original fine-grained data. The exact process is detailed in Sec. 3.5.

As a result, we have three pairs of fine- and coarse-grained datasets. The following sections describe each dataset in detail.

3.1. CDR coarse-grained dataset

This dataset consists of actual Call Detail Records (CDR), *i.e.*, time-stamped and geo-referenced logs of network events associated to voice calls placed or received by mobile network subscribers. Specifically, 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 location of the cell tower to which the caller's device is connected to when the call was first started. The CDR are collected by a major cellular network operator. They capture the communication 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.

[1] Due to a non-disclosure agreement with the data owner, we cannot reveal the geographical area or the exact collecting period of this dataset.

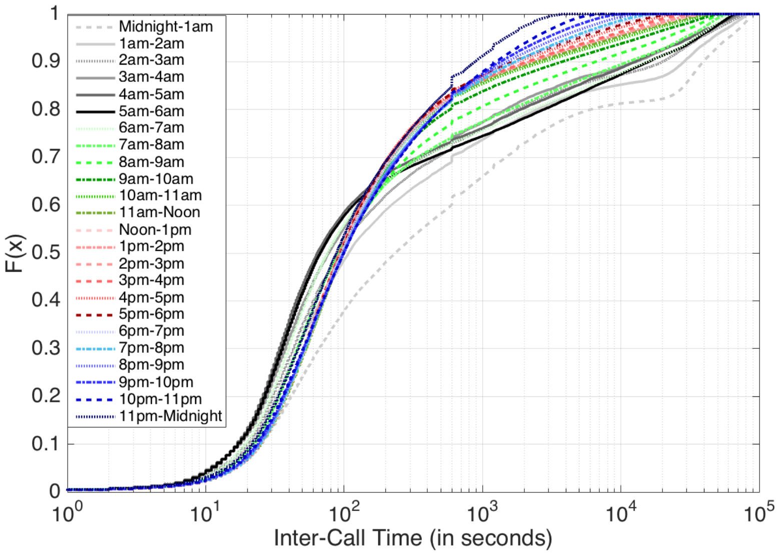
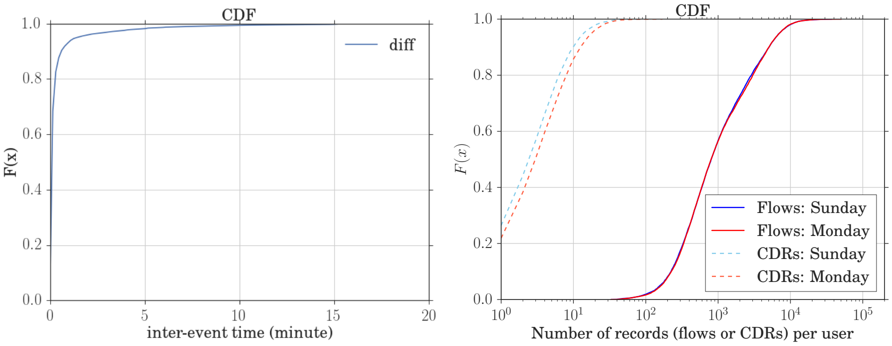


Figure 1: Distributions of the inter-event time in the CDR dataset at different day times.

We carry out a preliminary analysis of the CDR dataset, by extracting the experimental statistical distributions of the inter-event time (*i.e.*, the time between consecutive events). These distributions will be later leveraged in Sec. 3.5 to downsample the fine-grained datasets. The resulting cumulative distribution functions (CDF) are shown, for different hours of the day, in Fig. 1. We observe that a majority of events occur at a temporal distance of a few minutes, but a non-negligible amount of events are spaced by hours. This observation confirms results in the literature on the burstiness of human digital communication activities, with rapidly occurring events separated by long periods of inactivity [16]. The curves in Fig. 1 allow 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. Internet flow fine-grained dataset

This dataset is composed of mobile Internet session records, termed *flows* in the following. These records are generated and stored by the operator every time a mobile device establishes a TCP/UDP session for certain services



(a)

(b)

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

(*i.e.*, Facebook, Google Services, WhatsApp *etc*). 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 the location of the cell tower handling the session. The dataset refers to two-day period consisting of a Sunday and a Monday in 2015. In each day, the data covers a constant time interval, *i.e.*, from 10 am to 6 pm.

The flows in the Internet flow dataset have a considerably higher time granularity than the original CDR. Namely, at least one flow (*i.e.*, one location) is provided within every 20 minutes, for all users. 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 (solid lines) and CDR (dashed lines) for each user appearing in both datasets: the number of events per user in the Internet flow case is more than two orders of magnitude larger than that observed in the CDR case. We conclude that the Internet flows represent a suitable fine-grained dataset that can be associated to the coarse-grained CDR dataset.

Tab. 1 summarizes the number of users in the Internet flow dataset. In particular, the over 10K and 14K subscribers recorded on Sunday and Monday,

Table 1: Overview of the Internet Flow Dataset

Day of the week	Users	Rare CDR users	Frequent CDR users
Sunday	$10,856$	$6,154$	$4,702$
Monday	$14,353$	$7,215$	$7,138$

respectively, are separated into two similarly sized categories based on their CDR as follows:

- *Rare CDR users* are not very active in placing or receiving voice calls and thus have limited records in the CDR dataset. As in [7], we use the threshold of 0.5 event/hour below which the user is considered to belong to this category.
- *Frequent CDR users* are more active callers or callees and have more than 0.5 event/hour in the CDR dataset.

This distinction will be leveraged later on in our performance evaluation.

3.3. MACACO fine-grained dataset

This dataset is obtained through an Android mobile phone application, MACACOApp[2], developed in the context of the EU CHIST-ERA MACACO project [17]. The application collects data related to the user's digital activities such as used mobile services, generated uplink/downlink traffic, available network connectivity, and visited GPS locations. These activities are logged with a fixed periodicity of 5 minutes. We remark that this sampling approach differs from those employed by popular GPS tracking projects, such as MIT Reality Mining [18] or GeoLife [19], where users' positions are sometimes irregularly sampled. With respect to such previous efforts, the regular sampling in MACACO data grants a neater and more comprehensive overview of a user's movement patterns. The MACACO data covers 84 users who have stayed in 6

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

different countries and travel worldwide. The data collection spans 18 months approximately, from July 10, 2014, to February 4, 2016.

3.4. Geolife fine-grained dataset

This is the latest version of the Geolife dataset [19], which provides timestamped GPS locations of 182 individuals, mostly in Beijing [19]. The dataset spans a three-year time period, from April 2007 to August 2012. Unfortunately, the Geolife dataset is often characterized by large temporal gaps between subsequent data records. As a result, not all users present a number of locations or mobility level sufficient to our analysis. We thus select users given the criteria that the entropy rate of each individual's data points falls below the theoretical maximum entropy rate, which are used in [20] to select the Geolife users for analyzing individual human mobility.

3.5. Generating coarse-grained equivalents for MACACO and Geolife

We do not have access to CDR datasets for users in the MACACO nor Geolife datasets. We thus generate CDR-equivalent coarse-grained datasets, by leveraging the experimental distributions of the inter-event time in the CDR dataset (shown in Fig. 1, *cf.* Sec. 3.1). Specifically, we downsample the MACACO and Geolife datasets so that the inter-event times match those in the experimental distributions. Therefore, we first randomly choose one GPS record of the user as the seed CDR entry. We then randomly choose an inter-event time value from the distribution for the corresponding hour of the day, and use such interval to sample the second GPS record for the same user, mimicking a new CDR entry. We repeat this operation through the whole fine-grained trajectories of all users, and obtain datasets of downsampled GPS records that follow the actual inter-event time distributions of CDR.

Note that tailoring the inter-event distribution on a specific hour of the day allows taking into account the daily variability of CDR sampling. Also, upon downsampling, we filter out users who have an insufficient number of records, *i.e.*, users with less than 30 records per day on average or less than 3 days

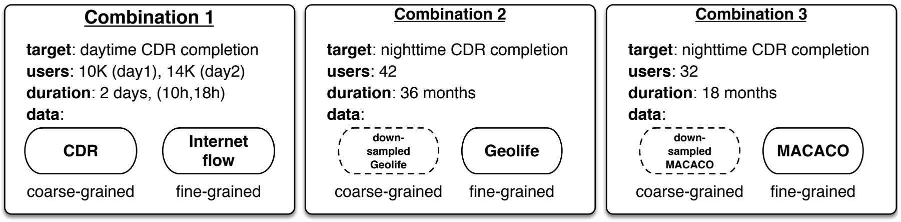


Figure 3: Combinations of corresponding coarse- and fine-grained datasets.

of activity. The final CDR-like coarse-grained versions of the MACACO and Geolife datasets contain 32 and 42 users, respectively.

3.6. Summary

By matching or downsampling the original data, we obtain three combinations of coarse-grained and fine-grained datasets for the same sets of users. Fig. 3 outlines them.

An important remark is that, as already mentioned in Sec. 3.2, the Internet flow dataset only covers working hours, from 10 am to 6 pm. As a result, the first data combination is well suited to the investigation of CDR completion during daytime. The relevant analysis is presented in Sec. 7.

The second and third data combinations, issued from the MACACO and Geolife datasets, cover instead all times. We thus employ them to overcome the limitations of the CDR and Internet flow pair, and to study CDR completion during night hours. Details are provided in Sec. 6.

4. Biases in CDR-based mobility analyses

Before delving into CDR completion, we present an updated analysis of the suitability of CDR data for the characterization of human mobility. Indeed, as anticipated in Sec. 1, CDR are typically sparse in space and time, which may affect the validity of results obtained from CDR mining.

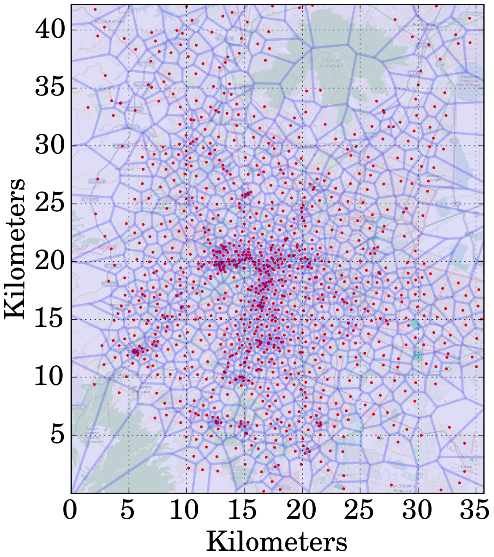
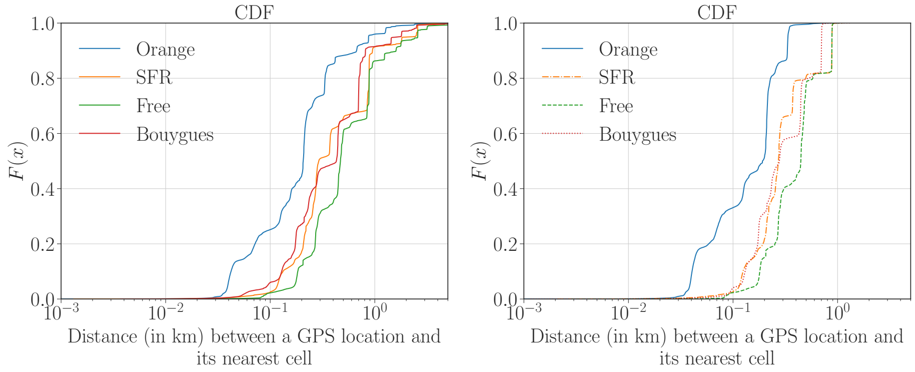


Figure 4: Deployment of cell towers in the target metropolitan area. Purple dots represent the base stations, whose coverage is approximated by a Voronoi tessellation.

4.1. Cell tower locations

In most CDR datasets, the position information 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 CDR entry. Such a shift may impact 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 local deployment of base stations. Fig. 4 shows the deployment of cell towers in the metropolitan area where our CDR dataset was collected. The presence of cell towers is far from uniform, with a higher density in downtown areas where a cell tower covers an approximately 2 km^2 area on average: in these cases, the cell coverage grants a fair granularity in the localization of mobile network subscribers. The same may not be true for cells in the city outskirts, which cover areas of several tens of km^2.

We evaluate how the cell deployment can bias human mobility studies. To



(a) Entire country

(b) Metropolitan areas

Figure 5: Distributions of the distances to the nearest cell tower (shifts), for $718,987$ GPS locations in the MACACO data of users in (a) the whole area and (b) major metropolitan areas (Paris Region, Lyon, Toulouse) in France.

this end, we perform a quantitative analysis of spatial shifts introduced by CDR positioning information by leveraging GPS logs in the MACACO dataset. Our focus on the MACACO dataset is 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 their 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 the Geolife users are mainly from.

We first extract $718,987$ GPS locations in the mainland of France[3] from the MACACO dataset. Among these locations, 74% are collected from the major metropolitan areas in France, including Paris Region, Lyon, and Toulouse. We then extract cell tower locations of the four major cellular network operators in France (*i.e.*, Orange, SFR, Free, and Bouygues) from the open government data [21].

Fig. 5(a) is the CDF of the distance between each GPS location in the

[3] The study focuses on the area in the latitude and longitude ranges of $(43.005, 49.554)$ and $(-1.318, 5.999)$, respectively.

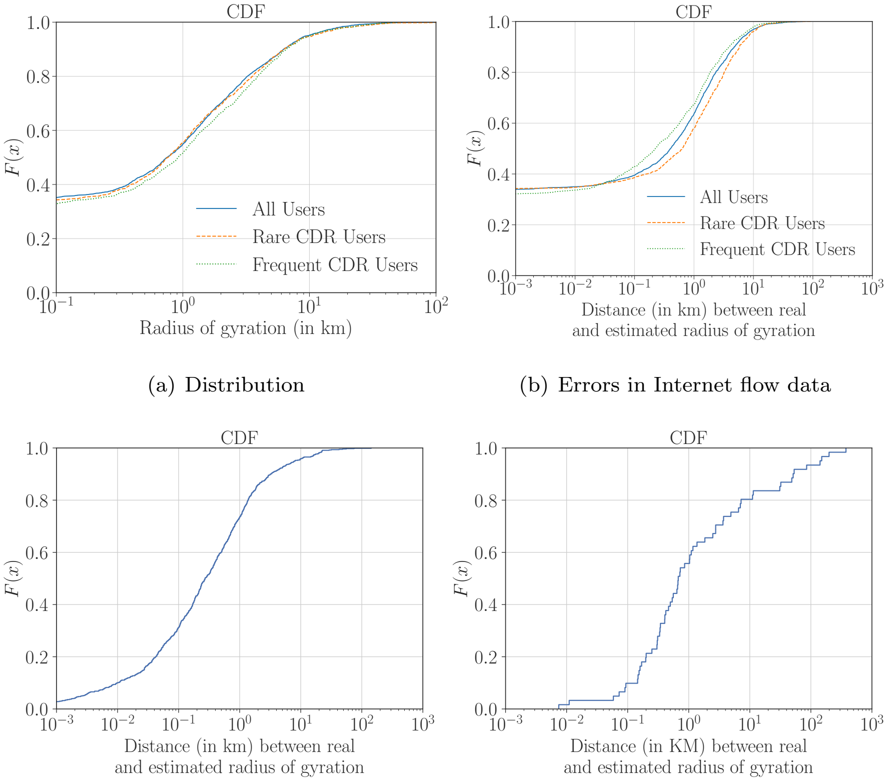
MACACO dataset and its nearest cell tower. We observe that most of the locations have a distance below 1 km when shifting to their nearest cells (*i.e.*, 95% for Orange, 91% for SFR, 86% for Free, and 91% for Bouygues). Nevertheless, when we focus on the metropolitan areas as shown in Fig. 5(b), almost all the shifts (*i.e.*, over 99%) are below 1 km and all the operators have their median shifts around $200 - 500$ meters. This indicates that the shifts above 1 km are all observed in rural areas. Still, most of the shifts are higher than 100 meters, indicating the presence of some bias of using cell tower locations. We stress that these values provide an upper bound to the positioning error incurred by CDR, as mobile network subscribers may be associated to antennas that are not the nearest ones, due to the signal propagation phenomena or load balancing policies enacted by the operator.

Still, the level of accuracy in Fig. 5, although far from that obtained from GPS logs, is largely sufficient for a variety of metropolitan-level or inter-city mobility analyses. For instance, it was shown that a spatial resolution of $2-7$ km is sufficient to track the vast majority of mobility flows in a large dual-pole metropolitan region [22].

4.2. Human movement span

We then examine whether mining CDR data is a suitable means for measuring the geographical span of movement of individuals. For that, we compute for each user u in the set of study $\mathcal{U}$ the *radius of gyration*, *i.e.*, the deviation of the user's positions to their centroid. Formally, $R_g^u = \sqrt{\frac{1}{n}\sum_{i=1}^{n}||\mathbf{r}_i^u - \mathbf{r}_{\text{centroid}}^u||_{\text{geo}}^2}$, 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 a popular measure used in human mobility studies [3, 5, 7, 23]. An individual who repeatedly moves among several fixed nearby locations still yields a small radius of gyration, even if she may total a large traveled distance.

We are able to compute both *estimated* (due to the temporal sparsity of the actual or the equivalent CDR data) and *real* (due to the finer granularity in the ground-truth provided by the Internet flow, MACACO, and Geolife datasets)



(a) Distribution

(b) Errors in Internet flow data

(c) Errors in MACACO data

(d) Errors in Geolife data

Figure 6: (a) CDF of the radius of gyration of two categories (Rare and Frequent) of CDR users in the Internet flow dataset. (b)(c)(d) CDF of the distance between the real and the estimated radius of gyration from CDR over the users of the (b) Internet flow, (c) MACACO, and (d) Geolife datasets.

radius of gyration for each user. Fig. 6 summarizes the results.

Let us first consider the users of the Internet flow dataset and their radii of gyration. Three curves denote different cases: *all*, *rare*, and *frequent* CDR users (*cf.* Sec. 3.2). The associated radius of gyration CDF are portrayed in Fig. 6(a). 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 if they are rare CDR users, *i.e.*, have a limited number of mobile communication activities.

When considering the error between real and estimated radius of gyration, in Fig. 6(b) for the CDR and Internet flow datasets, and in Fig. 6(c) and 6(d) for the MACACO or Geolife datasets, respectively, we observe the following:

- The distribution of large errors is similar in all cases, and outlines a decent accuracy of the coarse-grained CDR or CDR-like datasets. For approximately 90% of the Internet flow users, 95% of the MACACO users and 70% of the Geolife users, the errors between the real and the estimated radius of gyration are less than 5 km. The higher errors 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.

- A more accurate radius of gyration can be obtained for the CDR users who are especially active: 92% of the frequent CDR users have their errors lower than 5 km, while the percentage decreases to 86% for the rare CDR users.

- When considering small errors, the distributions tend to differ, with far lower errors in the case of CDR than MACACO or Geolife. This is in fact an artifact of considering cell tower locations as the ground-truth user positions in the fine-grained Internet flow dataset (*cf.* Sec. 4.1). In the more accurate GPS data of MACACO and Geolife, around 30% and 10% of the users enjoy their errors lower than 100 meters, while around 35% of the users in the CDR dataset have errors below 1 meter.

Overall, these results confirm the previous findings on the limited suitability of CDR for the assessment of the spread of human mobility [6]. They also unveil how different datasets can affect the data reliability at diverse scales.

4.3. Missing locations

Due to spatiotemporal sparsity, the mobility information provided by CDR is usually incomplete. We investigate the phenomenon in the case of users in the CDR dataset, and plot in Fig. 7(a) the ratio r_{N_L} of unique locations detected from CDR (N_L^{CDR}) to those from the ground-truth (N_L^{Flow}), *i.e.*, Internet flow data, as

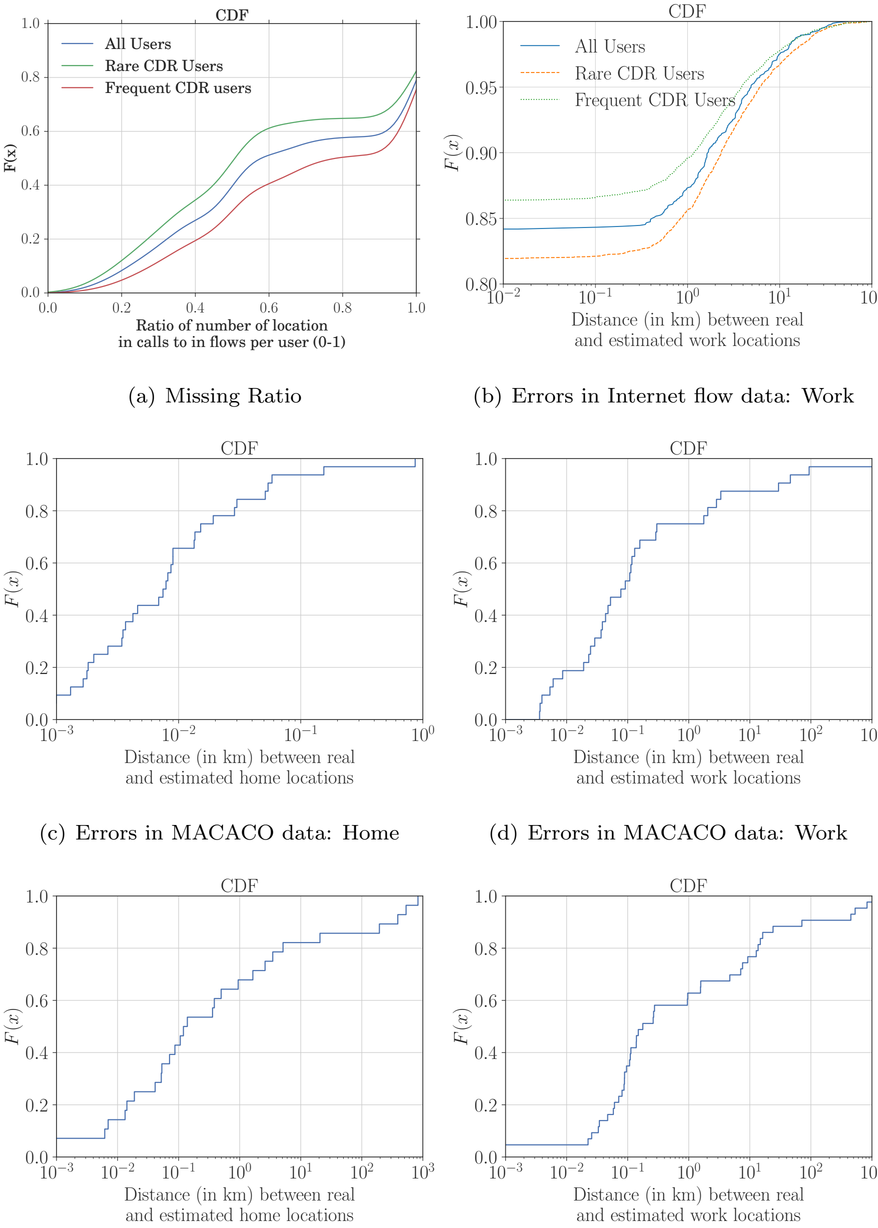
$$r_{N_L} = N_L^{\text{CDR}} / N_L^{\text{Flow}}. \tag{1}$$

We notice that 42% in the population of study (*i.e.*, all users) have their r_{N_L} higher than 0.8. For these users, 80% of their unique visited locations appear in the CDR data. The percentage of all users having this criterion is slightly higher for the frequent CDR users (50%) and lower for the rare CDR users (37%). These results confirm 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 generally regarded as an important step in the characterization of human mobility. Here, we focus on 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-like and ground-truth datasets. For each user, the places where the majority of work time records occur are considered a proxy for work locations; the equivalent records at night time are considered a proxy for home locations [24]. It is worth noting that, as the Internet flow dataset covers only $(10am, 6pm)$, we only infer work locations for this dataset.

Formally, let us consider a user u from the user set. The visiting pattern of the user u is a sequence of samples $\{(\ell_u^1, t_u^1), \ldots, (\ell_u^n, t_u^n)\}$, where the i-th sample (ℓ_u^i, t_u^i) denotes the location ℓ_u^i where the user u is recorded at time t_u^i. The



(a) Missing Ratio

(b) Errors in Internet 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 7: (a) CDF of the radio r_{N_L} of the number of locations in each user's coarse-grained trajectory to the one in her fine-grained trajectory. (b)(c)(d)(e)(f) CDF of the distances between each user's real and estimated important locations located by her CDR and ground-truth: (b) work locations over the Internet flow users; (c) home and (d) work locations over the MACACO users; (e) home and (f) work locations over the Geolife users.

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 (2) and (3) to determine home and work locations and then evaluate the accuracy of the CDR-based significant locations by measuring the geographical distance that separates them from the equivalent locations estimated via the corresponding fine-grained ground-truth datasets.

The results are shown in Fig. 7(b)-(f) as the CDF of the spatial error in the position of home and work places for different user groups for the three datasets. We 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. 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. 7(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 positions extracted from the GPS data. A close behavior can be noticed from the Internet flow and Geolife users, as shown in Fig. 7(b) and Fig. 7(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. 7(b), the errors between the real and the estimated significant locations are null for approximately 85% of all users, indicating that the use of the coarse-grained dataset is fairly sufficient for inferring these significant locations.
- The errors are 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.
- There is only a slight difference in the distribution of the errors associated with work locations between the rare and the frequent CDR users as shown in Fig. 7(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 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 [6], 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. Current approaches to 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, although transient and less important places visited may be lost in CDR data, capturing most of one's historical locations is not impossible. The good news is that, even in those cases, the error induced by CDR is relatively small. A major issue remains that CDR only provide instantaneous information about user's locations at a few time instants over a whole day. Overcoming the problem would help the already significant efforts in mobility analyses with CDR [10], allowing the exploration of scales much larger than those enabled by GPS datasets.

Temporal *CDR completion* aims at filling the time gaps in CDR, by estimating users' locations in between their mobile communication activities. Several strategies for CDR completion have been proposed to date. In this section, we introduce and discuss the two most popular solutions adopted in the literature.

5.1. Baseline `static` *solution*

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.* [25] to compute subscriber's presence in mobile traffic meta-data 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.*, the higher is the number of CDR, the 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.

5.2. Baseline `stop-by` *solution*

Building on in-depth studies proving individuals to stay most of the time in the vicinity of their voice call places [26], Jo *et al.* [27] 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.

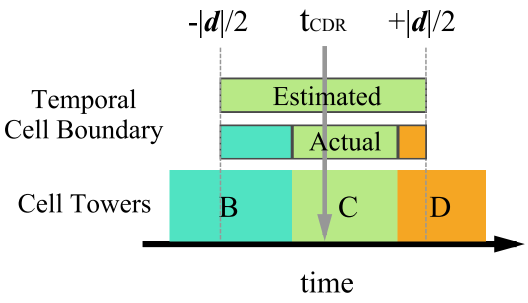


Figure 8: An example of a temporal cell boundary in the `stop-by` approach: A period $(t_{\text{CDR}} - |\mathbf{d}|/2, t_{\text{CDR}} + |\mathbf{d}|/2)$ is given as a temporal cell boundary at the cell C attached with a CDR entry at time t_{CDR}. In this temporal cell boundary, the user is assumed to be at the cell C, while actually she moves from the cell B to D: this leads to a spatial error.

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 be not always suitable. This solution lacks flexibility in dealing with various human mobility behaviors. As exemplified in Fig. 8, a single CDR is observed at time t_{CDR} at cell C. Following the `stop-by` solution, the user is considered to be 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 *temporal cell boundary*. In the example of Fig. 8, this temporal 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 temporal cell boundary. To have a quantitative manner of such an error, we define the spatial error of a temporal 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 $c_t^{(\text{real})}$ and her estimated cell location $c^{(\text{CDR})}$, during the time period $\mathbf{d}$. The interpretation of the spatial error is straightforward, as follows:

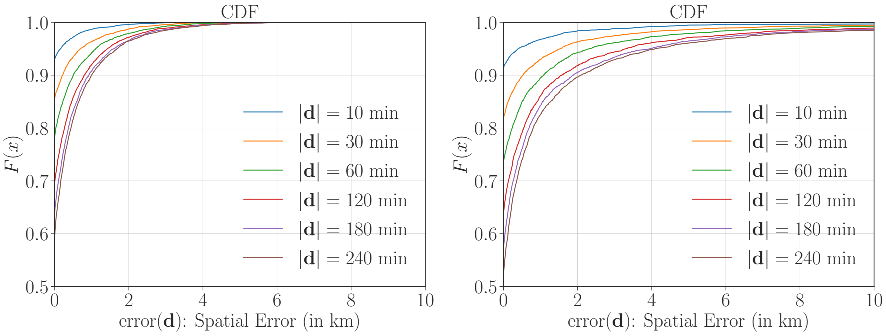
- When $\text{error}(\mathbf{d}) = 0$, it means that the user stays at the cell $c^{(\text{CDR})}$ during the whole temporal cell boundary. Still, the estimation of $\mathbf{d}$ may be conservative, since a larger $|\mathbf{d}|$ could be more adapted in this case.
- When $\text{error}(\mathbf{d}) > 0$, it means that the temporal cell boundary is oversized: the user in fact, moves to other cells in the corresponding time period. Thus, a smaller $|\mathbf{d}|$ should be used for the cell.

Due to the relevance of this parameter on the model performance, in the following we evaluate the impact of $|\mathbf{d}|$ on the spatial error.

5.3. Impact of parametrization on `stop-by` *accuracy*

We evaluate the performance of the `stop-by` approach, by considering the CDR and ground-truth Internet flow datasets (*cf.* Sec. 3). CDR are used to generate temporal cell boundaries, while locations in the fine-grained data of flows are adopted as actual locations and are used to compute the spatial errors. We consider a comprehensive range of values $|\mathbf{d}| = \{10, 30, 60, 120, 180, 240\}$ minutes for the `stop-by` parameters.

Fig. 9(a) and 9(b) show the CDF of the spatial error of temporal cell boundary on Monday and Sunday, respectively. We observe that $\text{error}(\mathbf{d}) = 0$ 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 is a strong indicator that users tend to remain in cell coverage areas for long intervals around their instant locations recorded by CDR. It is also true that many users are simply static, *i.e.*, only appear at one single location in their Internet flows, and, consequently have an associated radius of gyration $R_g^u = 0$: this behavior accounts for approximately 35% and 40% on Monday and Sunday, respectively. The high percentage of temporal cell boundaries with $\text{error}(\mathbf{d}) = 0$ in Fig. 9 may be due to these static users, since they will not entail any spatial



(a) CDF of Spatial Error: Monday

(b) CDF of Spatial Error: Sunday

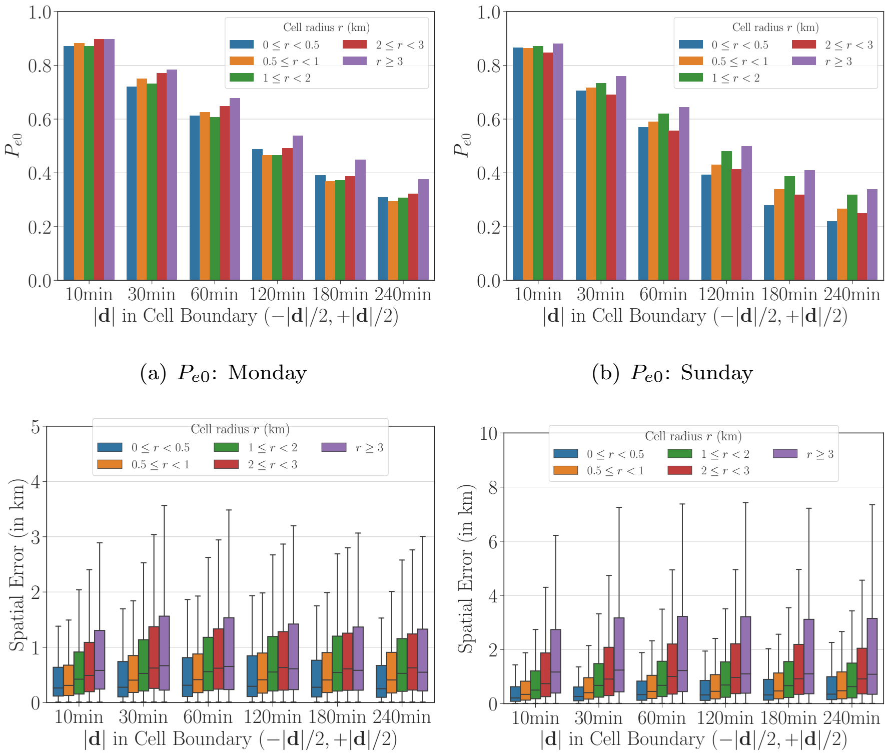
Figure 9: CDF of the spatial error of temporal cell boundaries of CDR generated by the `stop-by` solution over two groups of the users in the Internet flow dataset on (a) Monday and (b) Sunday.

error, under any $|\mathbf{d}|$. To account for this aspect, we exclude the static users in the following, and only consider the *mobile* users, *i.e.*, ones having $R_g^u > 0$.

An interesting consideration is that the spatial error incurred by the `stop-by` approach is not uniform across cells. Intuitively, a cell tower covering a larger area is expected to determine longer user dwelling times and hence better estimates with `stop-by`. We thus compute for each cell its coverage as the *cell radius*: specifically, we assume a homogeneous propagation environment and an isotropic radiation of power in all directions at each cell tower, and roughly estimate each cell radius as that of the smallest circle encompassing the Voronoi polygon of the cell tower. We remark that this approach yields overlapping coverage at temporal cell boundaries, which reflects what happens in real-world deployments. In the target area under study, shown in Fig. 4, 70% of the cells have radii within 3 km, and the median radius is approximately 1 km.

We can now evaluate the probability of having a temporal cell boundary with a null spatial error, as $P_{e0} = \mathbf{Pr}\{\text{error}(\mathbf{d}) = 0\}$. Fig. 10(a) and 10(b) present the probabilities P_{e0} grouped by the cell radius, when applying varying sizes of temporal cell boundary on the days of study. We notice the following.

- The probability P_{e0} decreases with the increasing period marked by $|\mathbf{d}|$,



(a) P_{e0}: Monday

(b) P_{e0}: Sunday

(c) Error: Monday

(d) Error: Sunday

Figure 10: Spatial errors of temporal 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 temporal cell boundary $(-|\mathbf{d}|, |\mathbf{d}|)$, where $|\mathbf{d}| \in \{10, 30, 60, 120, 180, 240\}$ minutes, under several groups of cell radius on (a) Monday and (b) Sunday; (c)(d) Box plot of non-zero spatial errors, grouped by the cell radius and the time period of temporal 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.

indicating that using a large period on the temporal cell boundary increases the chances of generating some spatial errors. 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 to around 0.6).

- The probability P_{e0} correlates 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, handovers are frequent for users moving among small cells and less so for users traveling across large cells.

The results support the idea that there is a strong correlation between the temporal cell boundary and the cell coverage. Nevertheless, since CDR are usually sparse in time, using a small temporal cell boundary could only cover an insignificant amount of cell visiting time, while using a big temporal 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 $\text{error}(\mathbf{d})$) in Fig. 10(c) and 10(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 $c^{(\text{CDR})}$.

5.4. Key insights

Overall, we assert that temporal 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, giving rise to spatial errors, when increasing $|\mathbf{d}|$. Hence, the trade-off between the completion and the accuracy should be carefully considered when completing CDR using temporal cell boundaries. Using a constant $|\mathbf{d}|$ over all users as in the `stop-by` solution is unlikely to be an appropriate approach.

Building on these considerations, we propose enhancements to the `stop-by` and `static` solutions in the remainder of the paper. The data completion strategies introduced in the following leverage common trends in human mobility, in terms of (1) attachment to a specific location during night periods, and (2) a tendency to stay for some time in the vicinity of locations where digital activities take place. In particular, we tell apart strategies for CDR completion at night time and daytime: Sec. 6 presents nighttime completion strategies inferring the home location of users; Sec. 7 introduces our adaptive temporal cell boundary strategies leveraging human mobility regularity during daytime.

6. Identifying temporal home boundaries

The main goal of our strategies for CDR completion during nighttime is to infer temporal boundaries where users are located, with a high probability, at their home locations. We refer to this problem as the identification of the user's *temporal home boundary*. Gaps in CDR occurring within the home boundary of each user are then filled with the identified home location. The rationale for this approach stems from our previous observations that CDR allow identifying the home location of individuals with high accuracy.

6.1. Proposed solutions

We extend the `stop-by` solution (*cf.* Sec. 5.2) in the following ways. Note that all techniques below assume that the home location is the user's most active location during some night time interval $\mathbf{h}$, and that CDR not in $\mathbf{h}$ are completed via legacy `stop-by`.

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

- The `stop-by-flexhome` strategy refines the previous approach by exploiting the diversity in the habits of individuals. The fixed night time temporal home 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 the most probable interval of time $\mathbf{h}_{\text{flex}}^{(u)} \subseteq \mathbf{h}$ during which the user is at her home location. Then, as for `stop-by-home`, the user will be considered at her home location to fill gaps in her CDR data during $\mathbf{h}_{\text{flex}}^{(u)}$.

- The `stop-by-spothome` strategy augments the previous technique by accounting for positioning errors that can derive (1) from users 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 $\mathbf{h}_{\text{flex}}^{(u)}$ is not identified, and if she is last seen at no more than 1 km from her home location, she is considered to be at her home location.

We compare the above strategies with the `static` and the legacy `stop-by` solution introduced in Sec. 5, assuming $|\mathbf{d}| = 60$ min. Our evaluation considers 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 *completion*, *i.e.*, the percent of the time during which the position of a user is determined. Note that the `static` solution (*cf.* Sec. 5) provides user locations at all times, but this is not true for `stop-by` or the derived techniques above. In this case, the CDR is completed only for a portion of the total period of study, and the users' whereabouts remain unknown in the remaining time.

6.2. Accuracy and completion results

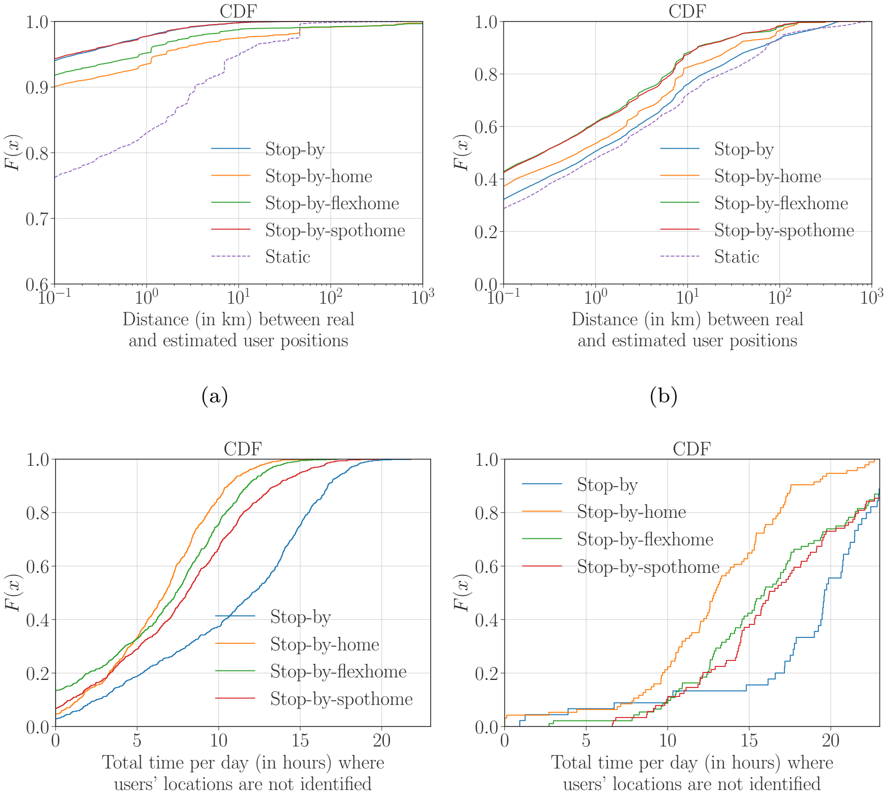
We first compute the geographical distance between the positions in the GPS records in MACACO and Geolife and those in their equivalent CDR-like coarse-grained datasets. These strategies are not designed to provide positioning information at all times expect the `static` solution, hence distances are only measured for GPS samples whose timestamps fall in the time periods for which completed data is available.

Fig. 11(a) and 11(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.

These results confirm that a model where the user remains static for a limited temporal interval around each measurement timestamp is fairly reliable when it comes to accuracy of the completed data. They also support previous observations on the quite static behavior of mobile network subscribers [26]. More importantly, the information of home locations can be successfully included in such models, by accounting for the specificity of each user's habits overnight.

The `stop-by` and derived solutions do not provide full completion by design. Fig. 11(c) and 11(d) show the CDF of the hours per day during which a user



(a)

(b)

(c)

(d)

Figure 11: 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 completion of completed data over the (c) MACACO and (d) Geolife data.

cannot be localized by such solutions, for individuals in the MACACO and Geolife CDR-like datasets, respectively. The completion performance is in fact 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 the plots, we assert that the more irregular sampling of the Geolife dataset translates into larger time gaps and smaller completion. Interestingly, the `stop-by` approach yields the worst result for both datasets, with unknown user positions in 12 and 19 hours per day in the median cases. Our proposed refinements to the `stop-by` solution increase the completion by inferring missing user positions overnight, when the CDR sampling is reduced. The improvement is significant, with a median gain over the basic `stop-by` solution of $4-5$ hours for MACACO dataset and $3-7$ hours for the Geolife dataset.

Overall, the combination of the results in Fig. 11 indicates the `stop-by-spothome` solution as that achieving the best combination of high accuracy and fair completion, among the different completion techniques considered.

7. Identifying temporal cell boundaries

We now consider the possibility of completing CDR during daytime. Our strategy is based again on inferring temporal boundaries of users. However, unlike what has been done with nighttime periods in Sec. 6, here we leverage the communication context of human mobility habits and extend the time span of the position associated with each communication activity to so-called *temporal cell boundaries.*

7.1. Factors impacting temporal cell boundaries

Hereafter, we aim to answer the following question: *how to choose a proper and adaptive period for a temporal cell boundary instead of a static fixed-to-all period?* To answer the question, we need to understand the correlation between the routine behavior of users in terms of mobile communications and their movement patterns. For this, we first study how human behavior factors that can be extracted from CDR may affect daytime temporal cell boundaries. We categorize factors in three classes, *i.e.*, event-related, long-term behavior, and location-related, as detailed next. Then, we leverage them to design novel approaches to estimate temporal cell boundaries.

7.1.1. Event-related factors

We include in this class the meta-data contained in records of common CDR datasets. They include the activity `time`, `type` (*i.e.*, voice call or text message), and `duration`[4]. Intuitively, these factors have direct effects on temporal 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`, a long phone call may imply that the user is static, while a single text message may indicate that the user is on the move. Besides, these factors are commonly found in and easily extracted from any common CDR entries.

7.1.2. Long-term behavior factors

This class includes factors describing users' activities over extended time intervals. They are the radius of gyration (`URg`), the number of unique visited locations (`ULoc`), and the number of active days during which at least one event is recorded (`UDAY`). These factors characterize a user by giving indications of *(i)* her long-term mobility and *(ii)* her habit on generating calls and text messages, which may be indirectly related to her temporal cell boundaries. For each user, these factors are computed from our CDR dataset (*cf.* Sec. 3.1) by aggregating data during the whole 3-month period of study.

7.1.3. Location-related factors

Factors in this class relate to positioning information. The first factor is the cell radius (`CR`), which we already proved to be affecting the reliability of CDR completion schemes in Sec. 5. The other location-related factors take account for the relevance that different places have for each user's activities. The intuition is

[4] We set the duration text messages to 0 seconds.

that individuals spend long time periods at their important places. Specifically, we explore it by applying the algorithm presented by Isaacman *et al.* [28], which determines prominent locations where the user usually spends a large amount of time or visits frequently.

The algorithm applies Hartigan's clustering [29] on visited cell locations of users in CDR and use logistic regression to estimate a location's importance to the user from 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 km) 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. The weights assigned to locations are the fractions of days in which they are visited over the whole observing period. The clustering process 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 [28]. It is worth noting that we cannot reproduce the exact model in [28], since the used ground-truth is not publicly available. However, we can still use the same factors for our objective, *i.e.*, identifying temporal cell boundaries.

7.2. Supervised temporal cell boundary estimation

So far, we have introduced human behavior factors that might be directly or indirectly related to temporal cell boundaries. In order to use them for our purpose, we need a reliable model linking them to actual temporal cell boundaries. In the following we introduce two approaches to do so, both based on supervised machine learning.

7.2.1. Symmetric and asymmetric temporal cell boundaries

We define two kinds of temporal cell boundaries: symmetric and asymmetric. Given a CDR entry at time t, determining its temporal cell boundary means to expand the instantaneous time t to a time interval $\mathbf{d}$, during which the user is assumed to remain within coverage of the same cell. For a symmetric temporal cell boundary, this period is generated from a CDR-based parameter $d^{\pm}$ as $\mathbf{d} = (t - d^{\pm}, t + d^{\pm})$, *i.e.*, it is symmetric with respect to the CDR time t. Instead, the period of an asymmetric temporal cell boundary is generated from two independent parameters d^{+} and d^{-} as $\mathbf{d} = (t - d^{-}, t + d^{+})$.

We design `sym-adaptive` and `asym-adaptive` approaches, both of which receive a CDR entry as input and return an estimate of its associated temporal cell boundary. More precisely, the factors discussed in Sec. 7.1 are extracted for each user and CDR record, and converted to an input vector $\mathbf{x}$, under the following rules: *(i)* the categorical factor `type` is converted to two binary features by one-hot encoding[5]; *(ii)* the `time` is converted to the distances (in seconds) separating it from $10am$ and from $6pm$[6]; *(iii)* the other factors are used as plain scalar values. Given a CDR entry and its input vector $\mathbf{x}$, we have the following approaches:

- The `sym-adaptive` approach contains one model that accepts the input vector and predicts the parameter $d^{\pm}$ as a symmetric estimation of the corresponding temporal cell boundary, *i.e.*, $d^{\pm} = F_{\text{sym}}(\mathbf{x})$.

[5] Used to deal with the unbalanced occurrence of the types.

[6] Daytime interval covered by the used dataset (*cf.* Sec. 3.2).

- The `asym-adaptive` approach contains two models that separately predict the parameters d^+ and d^- as a joint asymmetric estimation of the corresponding temporal cell boundary, *i.e.*, $d^+ = F^+_{\text{asym}}(\mathbf{x})$ and $d^- = F^-_{\text{asym}}(\mathbf{x})$.

We use supervised machine learning techniques to build the models. It is worth noting that the user identifier is not in the input vector $\mathbf{x}$ because we do not want to train models that bound themselves to any particular user. This gives our models better flexibility and ensures higher potential for applying the trained model into other mobile phone datasets where the same factors can be derived.

7.2.2. Estimating temporal cell boundaries via supervised learning

We detail our methodology and results, by *(i)* formalizing the optimization problems that capture our goal, *(ii)* discussing how they can be addressed via supervised machine learning, and *(iii)* presenting a complete experimental evaluation.

Optimization problems. All the models are generalized from a training set $\mathcal{X}$ consisting of CDR entries (as input vectors) and their real temporal cell boundaries (which are originally asymmetric), *i.e.*, $\mathcal{X} = \{(\mathbf{x}_i, d_i^+, d_i^-)\}$.

To build the `asym-adaptive` approach, the objective is to find two separate approximations, as $F^+_{\text{asym}}(\mathbf{x})$ and $F^-_{\text{asym}}(\mathbf{x})$, to functions $F^+(\mathbf{x})$ and $F^-(\mathbf{x})$ that respectively minimize the expected values of two losses $L(d^+, F^+(\mathbf{x}))$ and $L(d^-, F^-(\mathbf{x}))$, *i.e.*,

$$F^+_{\text{asym}}(\mathbf{x}) = \quad \arg\min_{F^+} \ \mathbb{E}_{d^+,\mathbf{x}}[L(d^+, F^+(\mathbf{x}))], \tag{5}$$

$$F^-_{\text{asym}}(\mathbf{x}) = \quad \arg\min_{F^-} \ \mathbb{E}_{d^-,\mathbf{x}}[L(d^-, F^-(\mathbf{x}))], \tag{6}$$

where L is the squared error loss function, *i.e.*, $L(x, y) = \frac{1}{2}(x - y)^2$.

To build the `sym-adaptive` approach, a modified training set $\mathcal{X}^{\pm} = \{(\mathbf{x}_i, d_i^{\pm})\}$ is firstly generated from the original $\mathcal{X}$ by applying $d_i^{\pm} = \min\{d_i^+, d_i^-\}$ on each real asymmetric temporal cell boundary. Then, as our objective, we need to find an approximation $F_{\text{sym}}(\mathbf{x})$ to a function $F^{\pm}(\mathbf{x})$ that minimizes the expected value of the loss $L(d^{\pm}, F^{\pm}(\mathbf{x}))$, *i.e.*,

$$F_{\text{sym}}(\mathbf{x}) = \arg\min_{F^{\pm}} \mathbb{E}_{d^{\pm},\mathbf{x}}[L(d^{\pm}, F^{\pm}(\mathbf{x}))]. \tag{7}$$

Learning technique. In order to compute the approximations, we utilize a typical supervised machine learning technique, *i.e.*, Gradient Boosted Regression Trees (GBRT) [30, 31]. Although several supervised learning techniques can be adopted, we pick the GBRT technique because *(i)* it is a well-understood approach with thoroughly-tested implementations, *(ii)* it has advantages over alternative techniques, in terms of predicative power, training speed, and flexibility to accommodate heterogeneous input (which is our case) [32], and *(iii)* it returns quantitative measures about the contribution of each factor to the overall approximation [30].

In the GBRT technique, an approximation function is the weighted sum of an ensemble of regression trees. Each tree divides the input space (*i.e.*, the vector $\mathbf{x}$ of factors) 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 that of each single regression tree [31]. The ensemble is initialized with a single-leaf tree (*i.e.*, a constant value). During each iteration, a new regression tree is added to the ensemble by minimizing the loss function via gradient descent. An algorithm of the GBRT technique for building the approximation of the function F_{sym} in the `sym-adaptive` approach is given in Alg. 1. In the algorithm, the function `FitRegrTree` is used to build a *regression tree* based on the input and the gradients of the function in the last iteration, of which we refer the reader to [31, Chapter 9.2.2] for the detail. Two important tuning parameters are in the algorithm, *i.e.*, the number of iterations M (*i.e.*, the number of regression trees to be added to the ensemble) and the learning rate ν (*i.e.*, the level of contribution expected by a new regression tree), which we determine via cross validation and discuss later. In the `asym-adaptive` approach, the same algorithm is used except that the training set $\mathcal{X}^{\pm}$ is replaced by $\mathcal{X}$.

Algorithm 1: GBRT algorithm [31, Algorithm 10.3] for finding the approximation function $F_{\text{sym}}(\mathbf{x})$ in the `sym-adaptive` approach

```
function GBRT($\mathcal{X}^{\pm}, M, \nu$);
Input  : $\mathcal{X}^{\pm}$ - training set, $M$ - number of iterations, $\nu$ - learning rate
Output: $F_{\text{sym}}(\mathbf{x})$ - symmetric temporal cell boundary estimation
        function
$F_0^{\pm}(\mathbf{x}) = \arg\min_{\gamma} \sum_i L(d_i^{\pm}, \gamma)$;
for $m \leftarrow 1$ to $M$ do
    for $(d_i^{\pm}, \mathbf{x}_i) \in \mathcal{X}^{\pm}$ do
        $g_i = -\frac{\partial L(d_i^{\pm}, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)}$;
    end
    $\mathcal{G} = \{(g_i, \mathbf{x}_i)\}$;
    $h_m(\mathbf{x}) = \texttt{FitRegrTree}(\mathcal{G})$;
    $\rho_m = \arg\min_{\rho} \sum_i L(d_i^{\pm}, F_{m-1}(\mathbf{x}_i) + \nu \cdot \rho \cdot h_m(\mathbf{x}_i))$;
    $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \rho_m \cdot h_m(\mathbf{x})$;
end
return $F_M(\mathbf{x})$;
```

Experiments. The first step is to build the training sets. For that, we randomly select 50% of the users from the two available days (*i.e.*, a Monday and a Sunday) in the Internet flow dataset (*cf.* Sec. 3.2). In particular, from the CDR and Internet flow datasets, we first extract for each CDR entry of these selected users its corresponding input vector $\mathbf{x}$ as well as the parameters d^+, d^- of its real temporal cell boundary. We then build the two training sets $\mathcal{X}$ and $\mathcal{X}^{\pm}$.

The second step is to build the approximation functions (*i.e.*, F_{asym}^+, F_{asym}^-, and F_{sym}) from the training sets. For that, we have to first tune the M and ν parameters of Alg. 1 of each approximation function. To this end, we use a 3-fold cross-validation to select the number of iterations M from the candidate set $\{100, 500, 1000, 2000, \cdots, 10000\}$ and the learning rate ν from the candidate set $\{0.1, 0.2, \cdots, 1\}$. In particular, we divide the training set $\mathcal{X}$ (or $\mathcal{X}^{\pm}$) into

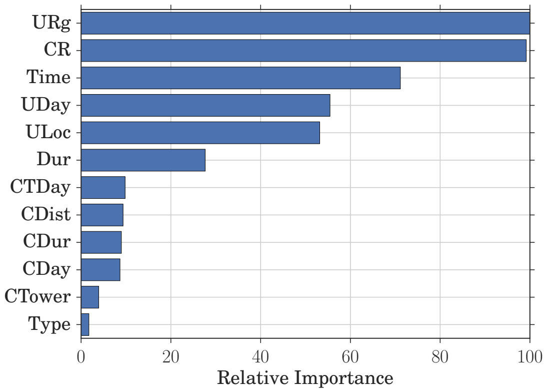
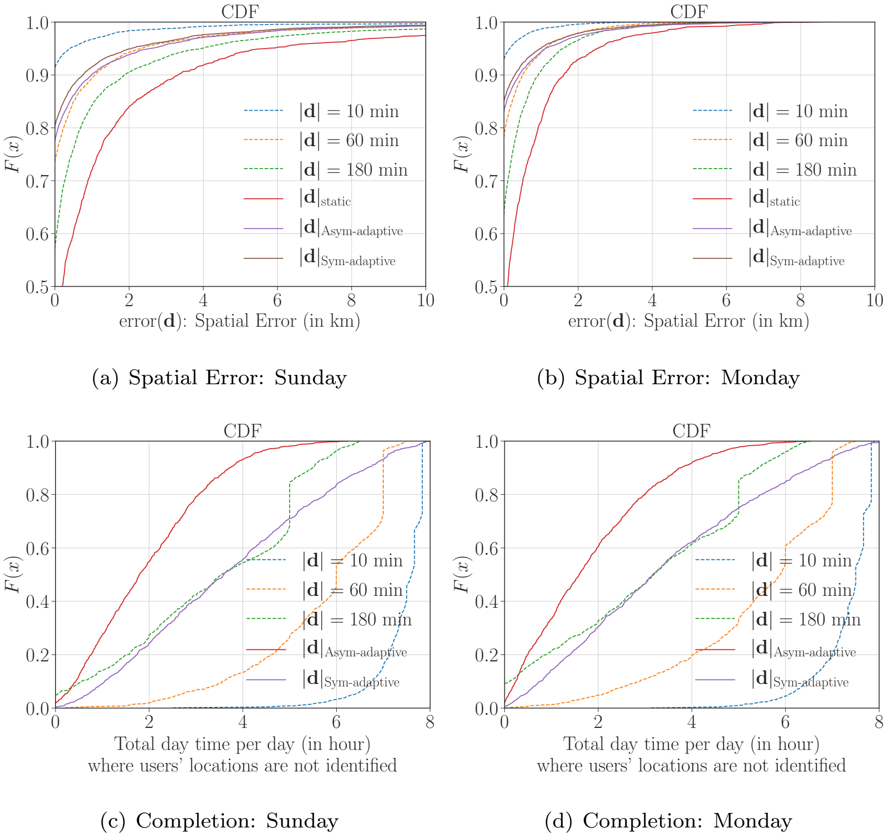


Figure 12: Relative Importance of features in determining accurate temporal cell boundaries.

equal-sized three subsets. For each combination of M and ν, we train the model corresponding to each approximation function based on one subset and validate it on the other two subsets. We repeat this operation three times with each of the subsets used as training data. We select as our actual parameters the M and ν values that achieve the lowest loss in the cross-validation. Finally, we use the training sets $\mathcal{X}$ and $\mathcal{X}^{\pm}$ and the tuning parameters that we select to build the functions F^{+}_{asym}, F^{-}_{asym}, and F_{sym} corresponding to the `asym-adaptive` and `sym-adaptive` approaches.

Fig. 12 shows the relative importance of factors with respect to the estimation of a temporal cell boundary in the training procedure of the GBRT technique. For each factor, its importance is computed as a relative value of the sum of its corresponding importance in all the three approximations. 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.

- The three most important factors are the timestamp of the activity, the cell radius, and the radius of gyration. This indicates that the time spent by a user within coverage of a same cell mainly depends on the cell size, the precise time when the activity occurred, and the user's long-term mobility.



(a) Spatial Error: Sunday

(b) Spatial Error: Monday

(c) Completion: Sunday

(d) Completion: Monday

Figure 13: CDF of the spatial errors of temporal cell boundaries computed on (a) Sunday and (b) Monday; CDF of the completion of completed data on (c) Sunday and (d) Monday, across the `stop-by`, `static`, `sym-adaptive`, and `asym-adaptive` approaches.

- Surprisingly, the activity's `type` is the least relevant factor, indicating that knowing whether a user generates a call or a message is useless in determining a temporal cell boundary.

7.3. Accuracy and completion results

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. For the two `sym-adaptive` and `asym-adaptive` approaches, we build two testing sets from the CDR entries of the remaining users. We then let them generate adaptive symmetric and asymmetric temporal cell boundaries using the input vectors in the testing sets. Besides, we let the `stop-by` approach generate temporal cell boundaries using $|\mathbf{d}| = \{10, 60, 180\}$ minutes. As in Sec. 6, we make a comparative study by evaluating the solutions regarding *accuracy* and *completion*, where the accuracy is measured by evaluating the *spatial error* in (4) (*cf.* 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 completion simultaneously.

Fig. 13(a) and 13(b) display the distribution of the spatial errors over all temporal cell boundaries. Our results confirm that the spatial error increases as t_d becomes larger when using the `stop-by` approach. More importantly, the two adaptive approaches perform slightly better than the `stop-by` approach does with its most common setting ($|\mathbf{d}| = 60$ minutes) in terms of the spatial error. As expected, the `static` solution has the worst performance, similarly to what observed in the case of home boundaries using the MACACO and Geolife datasets.

Fig. 13(c) and 13(d) plot the distribution of the completion per users over all approaches except `static` (of which the completed data always covers the whole period). The x-axis of the figures has 8 hours because the Internet flow dataset only covers an eight-hour day time, *i.e.*, $(10am, 6pm)$. We remark that both our adaptive approaches score a significant performance improvement in terms of completion: the amount of time during which users' locations stay unidentified is substantially reduced with respect to the legacy `stop-by` approach. On average, only approximately 2 hours (25% of the period of study) of the user's day time remains unidentified after applying the `asym-adaptive` approach, while 3 hours remains unidentified after using the `sym-adaptive` and `stop-by` ($|\mathbf{d}| = 180$ minutes) approaches. The `stop-by` approach with its most common setting ($|\mathbf{d}| = 60$ minutes) has the same degree of accuracy as the adaptive approaches but has a far less degree of completion (*i.e.*, a median of 6 unidentified hours).

Overall, these results highlight a clear advantage provided by adaptive approaches for CDR completion based on supervised learning. Consequently, the adaptive approaches achieve a slightly better performance in terms of accuracy but have a far better performance in terms of completion. The `asym-adaptive` approach has an obvious advantage than the competitors: it completes 75% of the day hours with a fairly good accuracy.

8. Conclusion

In this paper, we leveraged real-world CDR and GPS datasets to characterize the bias induced by the use of CDR for the study of human mobility, and evaluated CDR completion techniques to reduce some of the emerging limitations of this type of data. Our results confirm previous findings on the sparsity of CDR, and, more importantly, provide a first comprehensive investigation of techniques for CDR completion. In this context, we propose solutions that *(i)* dynamically extend the time intervals spent by users at locations where they are pinpointed by the CDR data during daytime, and *(ii)* sensibly place users at their home locations during nighttime. Extensive tests with heterogeneous real-world datasets prove that our approaches can achieve excellent combinations of accuracy and completion. On average, for daytime hours, our approaches can complete 75% of the time in which 95% have errors below 1 km; for nighttime hours, our refinements of the legacy solution have a performance gain of 4-5 or 3-7 hours on two datasets regarding completion and up to 10% of a performance gain regarding accuracy. Particularly, compared with the most common proposal in the literature, our best adaptive approach outperforms by 5% of accuracy and 50% of completion.

Acknowledgment

This work is supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO and is performed in the context of the EMBRACE Associated Team of Inria.

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?, SIGMOBILE Mob. Comput. Commun. Rev. 16 (3) (2012) 33–44. `doi:10.1145/2412096.2412101`.

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

[13] 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.

[14] 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`.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] G. Smith, R. Wieser, J. Goulding, D. Barrack, A refined limit on the predictability of human mobility, in: Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on, 2014, pp. 88–94. doi:10.1109/PerCom.2014.6813948.

[21] France Open Data, https://www.data.gouv.fr/fr/datasets/.

[22] M. Coscia, S. Rinzivillo, F. Giannotti, D. Pedreschi, Optimal spatial resolution for the analysis of human mobility, in: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2012, pp. 248–252. doi:10.1109/ASONAM.2012.50.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

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

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

[31] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Vol. 1, Springer series in statistics New York, 2001.

[32] Scikit-learn, Ensemble methods, http://scikit-learn.org/stable/modules/ensemble.html, accessed: 2017-12-20.