Filling the Gaps: On the Completion of Sparse Call Detail Records for Mobility Analysis

Sahar Hoteit, Guangshuo Chen, Aline Viana
INRIA Saclay
1 Rue Honoré d'Estienne d'Orves,
91120 Palaiseau, France
name.surname@inria.fr

Marco Fiore
CNR - IEIIT
Corso Duca degli Abruzzi 24
10129 Torino, Italy
marco.fiore@ieiit.cnr.it

ABSTRACT

Call Detail Records (CDRs) have been widely used in the last decades for studying different aspects of human mobility. The accuracy of CDRs strongly depends on the user-network interaction frequency: hence, the temporal and spatial sparsity that typically characterize CDR can introduce a bias in the mobility analysis. In this paper, we evaluate the bias induced by the use of CDRs for inferring important locations of mobile subscribers, as well as their complete trajectories. Besides, we propose a novel technique for estimating real human trajectories from sparse CDRs. Compared to previous solutions in the literature, our proposed technique reduces the error between real and estimated human trajectories and at the same time shortens the temporal period where users' locations remain undefined.

CCS Concepts

•**Networks** → **Network monitoring;**

Keywords

Human trajectory, important locations; movement inference.

1. INTRODUCTION

A deep understanding of human-mobility patterns can yield interesting insights into a variety of important societal issues, such as urban planning [1], road traffic monitoring and forecasting [2], or disease spreading and containment [3]. Moreover, from a networking point of view, cloud and content delivery networks [4], paging operations in cellular networks [5], cognitive network functions [6], as well as location-based recommender systems [7, 8] can all get a great benefit from quantitative and qualitative knowledge of users' mobility patterns.

In this context, specialized spatiotemporal datasets such as GPS logs have shown significant potential as a knowledge base about human mobility patterns. However, due to the overhead of collecting such detailed datasets at scale, many researchers have been urged to explore other data sources as potential proxies for human mobility.

Among these, Call Detail Records (CDRs) are a very promising option. Originally collected by mobile network operators for billing purposes, CDRs contain timestamped and geo-referenced logs on each voice call or texting activity of every serviced customer [9]. They thus convey fairly detailed information about the movements of very large populations, typically comprising millions of subscribers.

The digital footprints generated by mobile phone users have rapidly emerged as a primary source of knowledge about human mobility. The analysis of CDRs has allowed revealing, for instance, the spatial recurrence and temporal periodicity of the movement patterns of people, who show a strong tendency to return to previously visited locations [10, 11]. This entails a high predictability potential for human mobility [12]. Similarly, significant places in our lives (e.g. home, work, shopping- or hobby-related locations) are easily inferred from CDRs [13]. This information can be further used to help guide policies intended to reduce people carbon footprint [14]. Other relevant examples of applications of CDR analyses include the detection and modeling of aggregate mobility flows at large scales [15], the characterization of individual movement patterns [16], or the computation of origin-destination matrices in urban areas [17]. A comprehensive survey of studies that leverage CDRs for mobility analysis is included in [9].

Despite the significant benefits that it brings to human mobility analysis at scale, an indiscriminate use of CDR may raise questions on the validity of the conclusions of the related research efforts. Specifically, CDRs 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 the user location is recorded only if and when he performs a voice call or texts a message). Indeed, cell (sectors) typically span at least thousands of square meters, and very active mobile subscribers generate a few tens of voice or text events per day: Overall, this leads to spatiotemporal sparsity in the CDR data.

The question of whether and to what extent such a sparsity affects mobility studies has been only partly addressed. Promising results are obtained when mobility is constrained to transportation networks. Zhang *et al.* [2] 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. Asgari *et al.* [18] prove that CDRs are sufficient to reconstruct with fair accuracy the multi-modal trips of individual users. In this case, GPS data from a small set of ten users, for which CDR information was also available, is leveraged as a ground truth.

Conclusions are instead less clear when it comes to gen-

CHANTS'16, October 03-07, 2016, New York City, NY, USA

DOI: http://dx.doi.org/10.1145/2979683.2979685

eral human movement patterns. Isaacman *et al.* [19] report a spatial error in the order of 1 km when comparing CDR positions with ground-truth reference logged by five volunteers when they initiated some mobile phone activity. This error can be neglected in some applications (e.g., investigations of inter-city mobility), but may impair others (e.g., identification of the preferred shops of a precise subscriber). Ranjan *et al.* [20] show that CDR allow nto correctly identify, for each user, popular locations that account for 90% of the subscriber's activity. However, according to the same authors, plain CDRs do not allow inferring transient locations or measures of the geographical spread of the users' mobility. These considerations are the result of a comparison of CDR-based analyses with equivalent studies of logs of data traffic –which tend to have much higher frequency than CDR, but are also more difficult and expensive to collect [9]– generated by the same mobile user population.

Data completion is an interesting approach to mitigate the spatiotemporal sparsity in CDRs, and thus the problems the latter raises in terms of dependability of the results of CDR analyses. Data completion consists in filling the spatiotemporal gaps in CDRs, so as to derive subscribers' trajectories that are as comprehensive as possible.

In this paper, we contribute to the effort on reliable completion of CDR data in a number of ways.

- We introduce an original dataset of GPS logs, which we leverage as a ground truth for our analysis. Unlike previous datasets, our GPS data features regular high-frequency position sampling, and covers the movements of 84 users worldwide for more than 18 months. We also propose an original technique to subsample GPS data in time, so as to mimic sparse CDRs. To that end, we leverage real-world large-scale data from a mobile network operator. Details are provided in Sec. 2.
- We assess the capability of sparse CDRs of modelling important features of individual trajectories. Our results confirm previous findings in the literature. Details are provided in Sec. 4.
- We implement a number of techniques for CDR data completion proposed in the literature, and assess their quality in presence of ground-truth GPS data. This evaluation sheds light on the quality of the results provided by each approach. In addition, we propose original CDR data completion solutions, and show that they outperform previous proposals, reducing the spatial error in the completed data and shortening the time periods where no location information is available. Details are provided in Sec. 5.

Conclusions and perspectives of our works are finally discussed in Sec. 6.

2. DATASETS

Our study requires two datasets. The first provides fine-grained information on the spatiotemporal trajectory of individuals, and is used as our ground truth. The second is instead sparse so as to mimic CDR data. These two datasets are presented in Sec. 2.1 and Sec. 2.2, respectively.

2.1 MACACOApp GPS data

This dataset is obtained through an Android mobile phone application, MACACOApp[1], developed in the context of the EU CHIST-ERA MACACO project [21]. The application collects data related to the user's digital activities (e.g., the mobile services he uses, the uplink/downlink traffic he generates, the type of network connectivity he leverages, etc.), and, more importantly to our work, includes GPS locations. These are logged with fixed periodicity, at every 5 minutes: we remark that this sampling approach is different from those employed by popular GPS tracking projects, such as MIT Reality Mining [22] or Microsoft GeoLife [8], where users' positions are sampled irregularly and often present very large gaps between subsequent data records. With respect to such previous efforts, the regular sampling of MACACOApp data grants a much neater and comprehensive vision of the user's movement patterns.

The MACACOApp 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. The geographical map of the users considered in our study is shown in Fig. 1, where dots refer to the center of mass (i.e., barycenter) of each user's positions during every day of activity.

[1]Available at https://macaco.inria.fr/macacoapp/.

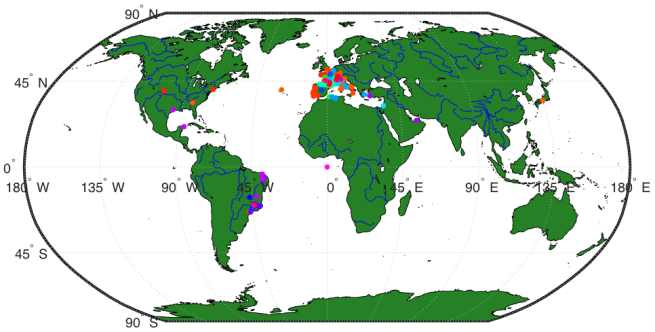


Figure 1: Geographical map of the MACACOApp users considered in our study. Each dot refers to the center of mass of the positions of one user in a specific day.

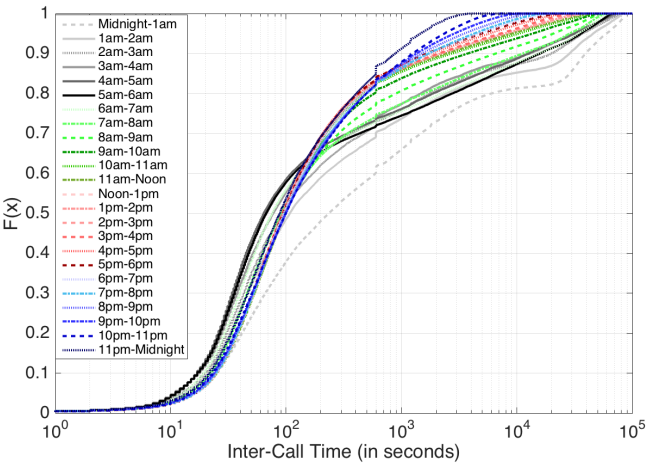


Figure 2: CDF of the inter-event time in CDRs, per hour.

2.2 Sparse CDR data

In order to derive the second dataset, containing sparse spatiotemporal information and mimicking CDRs, we downsample the MACACOApp data. This is an inevitable step, as we do not have access to mobile network operator CDRs

for the MACACOApp users.

In order to downsample the GPS data in a realistic way, we leverage real-world CDRs collected by a major cellular operator in Mexico. The dataset contains voice call and texting logs of approximately 221,000 subscribers in Mexico City, during a 3-month period from October 1, 2014 to December 31, 2014. We extract from such data experimental statistical distributions of the inter-event time. The corresponding cumulative distribution functions (CDF) for different hours of the day are shown in Fig. 2. It is worth mentioning that for two consecutive events happening at different time slots, the corresponding inter-event time is attributed to the time slot of the first event. We observe that the majority of events occur at a distance of a few minutes. However, a non-negligible amount of events are spaced by hours. These results confirm 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 [23]. The curves in the figure 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. We downsample the MACACOApp data according to the distributions in Fig. 2. 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: More precisely, we select only users with more than 30 positions per day and more than 3 days of activity and we remove all the weekends. This results in equivalent CDR data for a total of 32 users. Hereafter, we will refer to this as CDR data.

3. BIASES IN CALL DETAIL RECORDS

Here, we compare the ground-truth GPS data and the equivalent sparse CDRs, in terms of the results these datasets yield when they are employed for human mobility analysis. We perform three tests, aimed at understanding the dependability of CDRs for the characterization of *(i)* the home and work locations of users, in Sec. 4.1, *(ii)* their daily span of movement, in Sec. 4.2, and *(iii)* complete trajectories, including transient locations, in Sec. 4.3.

3.1 Home and work locations

The identification of significant places where people live and work is often an important first step toward the characterization of human mobility. To capture the home and work locations of users, we first separate both ground-truth and CDR data into two time windows, mapping to work time (9 am to 5 pm) and night time (10 pm to 7 am). The places where the majority of work time records occur are considered as a proxy of work locations, whereas the equivalent records at night time are a proxy of home locations [24].

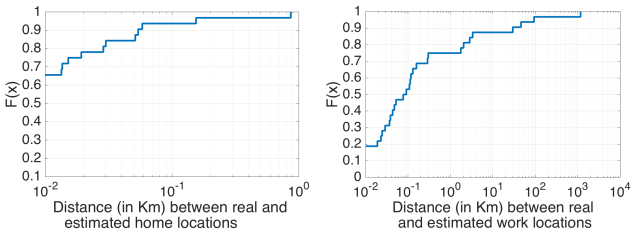
Formally, let us consider a user $u \in \mathcal{U}$, where $\mathcal{U}$ is the observed population. The spatiotemporal trajectory of u is a sequence of samples $\{(\ell_u^0, t_u^0), (\ell_u^1, t_u^1), \ldots, (\ell_u^n, t_u^n)\}$, where the i-th sample (ℓ_u^i, t_u^i) denotes the location ℓ_u^i where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

$$\ell_u^H = mode(\ell_u^i \mid t_u^i \in t^H), \quad (1)$$

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

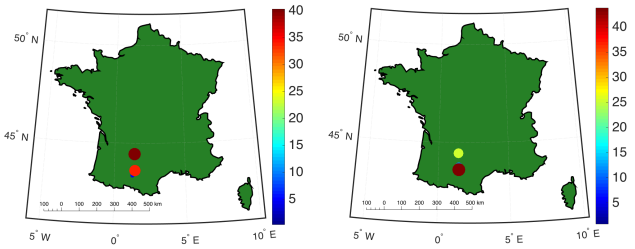
$$\ell_u^W = mode(\ell_u^i \mid t_u^i \in t^W), \quad (2)$$

where t^W is the work time interval.



(a) Home locations (b) Work locations

Figure 3: CDF of the spatial error (in km) between (a) home and (b) work locations estimated from GPS and CDR data.



(a) GPS-based (b) CDR-based

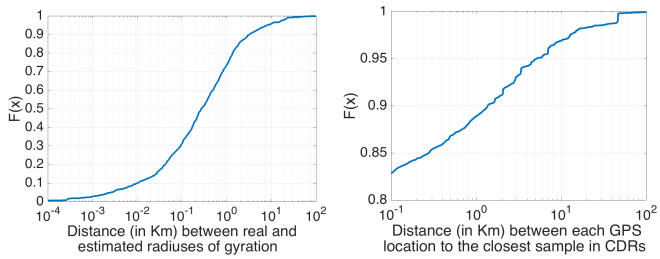
Figure 4: Example of popular locations during work hours, along with their percentage of occurrence (color bars) inferred from (a) GPS and (b) CDR data.

We use the definitions in (4) and (5) above to determine the subscribers' home and work locations, using both GPS and CDR data. We then evaluate the accuracy of the CDR-based home and work locations by measuring the geographical distance that separates them from the equivalent locations estimated via the GPS ground-truth data.

The results are displayed in Fig. 6a and Fig. 6b, showing the CDF of the spatial error in the position of home and work places, respectively. We can clearly observe the following.

- The errors related to home locations are fairly small. They are below 1 km for all users, and within 100 m for 94% of them. The latter figure is easily one order of magnitude smaller than the coverage radius of medium-sized cell sectors. In other words, the positioning error due to the temporal sparsity of CDRs seems negligible when compared to that induced by mapping the subscriber's location to the base station position.

- The errors associated with work locations are sensibly higher than those measured for home locations. While 75% of users have an error of less than 300 m, the work places of a significant portion of individuals (around 12% of the total) are identified at a distance higher than 10 km from the position extracted from GPS data. Further investigations show that these large errors typically occur for users who do not seem to have a stable work location, and might be working in different places depending on, e.g., the day of the week. As an example, Fig. 7 shows for both GPS and CDR data, the same most popular locations at which one specific user appears during the work hours, along with their percentage of occurrence during such an interval (cf. color bars). For both data, two important



(a) Radius of gyration (b) All samples

Figure 5: CDF of the spatial error (in km) between the (a) radiuses of gyration and (a) all samples from the GPS and CDR data.

places emerge, in different cities at around 100 km of distance in Southern France. The overall work-time presence of the user is close between the two locations in the ground-truth case (40% versus 33%) and the location tagged as the effective user' work location is the one given by the highest presence (cf. dark red location in Fig. 7a). On the other side, the reduced amount of data in CDRs reverses (26% versus 44%) the more reliable ranking obtained with GPS data, and consequently, tags a different location as the effective user's work location (cf. dark red location in Fig. 7b). Still, situations like the one in Fig. 7 prove that, although the granularity of CDRs is sufficient to detect popular locations for individual subscribers, granularity improvement is required to better infer effective locations.

Overall, these results confirm previous findings [20], and further prove that CDRs yield enough detail to detect effective locations in users' movement patterns. However, they also point out that they may lead to incorrect estimations in the ranking among such locations.

3.2 Span of movement

As a second test, we study whether CDRs can be used to determine the geographical span of the movement of individual users. To that end, we employ the radius of gyration as a relevant metric. For a user $u \in \mathcal{U}$, the radius of gyration r_u is computed as

$$r_u = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ell_u^i - \ell_u^*)}, \tag{3}$$

where ℓ_u^* is the center of mass of all locations recorded in the spatiotemporal trajectory of u, i.e., $\ell_u^* = \frac{1}{n}\sum_{i=1}^{n}\ell_u^i$.

We compute the radius of gyration from the daily data of each user, using both GPS and CDR records. Fig. 8a shows the CDF of the error obtained when comparing the results. The span of movement according to CDRs entails a small error, lower than 100 m, only in 30% of cases. The errors is instead larger than 1 km for 26% of the users. These numbers are far from those obtained in the case of home locations, and corroborate previous findings on the limited suitability of CDRs for the assessment of the spread of subscribers' mobility [20].

3.3 Complete trajectories

Finally, we compare the full GPS and CDR trajectories. To that end, we compute the geographical distance between each GPS sample and the CDR sample that is the closest in time. The resulting CDF is shown in Fig. 8b. For 83% of GPS points, the error in the equivalent CDRs is minimal, i.e., 100 m or less: This is consistent with the well-known behavior of many individuals who tend to be fairly static and spend most long periods of time at a same location [11, 25].

However, for around 11% of samples, the information in CDRs is highly erroneous, with spatial displacements of at least 1 km, with peaks of several tens of km. These errors can be imputed to periods of significant mobility of subscribers (corresponding to transition periods, e.g., commuting or traveling), during which sparse CDR data cannot track positions reliably. Thus, the results confirm that CDRs are not suited to the analysis of transient movement patterns of individuals [20].

4. BIASES IN CALL DETAIL RECORDS

Here, we compare the ground-truth GPS data and the equivalent sparse CDRs, in terms of the results these datasets yield when they are employed for human mobility analysis. We perform three tests, aimed at understanding the dependability of CDRs for the characterization of *(i)* the home and work locations of users, in Sec. 4.1, *(ii)* their daily span of movement, in Sec. 4.2, and *(iii)* complete trajectories, including transient locations, in Sec. 4.3.

4.1 Home and work locations

The identification of significant places where people live and work is often an important first step toward the characterization of human mobility. To capture the home and work locations of users, we first separate both ground-truth and CDR data into two time windows, mapping to work time (9 am to 5 pm) and night time (10 pm to 7 am). The places where the majority of work time records occur are considered as a proxy of work locations, whereas the equivalent records at night time are a proxy of home locations [24].

Formally, let us consider a user $u \in \mathcal{U}$, where $\mathcal{U}$ is the observed population. The spatiotemporal trajectory of 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 u is then defined as the most frequent location during night time:

$$\ell_u^H = mode(\ell_u^i \mid t_u^i \in t^H), \tag{4}$$

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

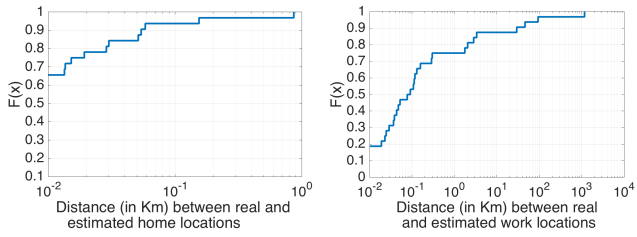
$$\ell_u^W = mode(\ell_u^i \mid t_u^i \in t^W), \tag{5}$$

where t^W is the work time interval.

We use the definitions in (4) and (5) above to determine the subscribers' home and work locations, using both GPS and CDR data. We then evaluate the accuracy of the CDR-based home and work locations by measuring the geographical distance that separates them from the equivalent locations estimated via the GPS ground-truth data.

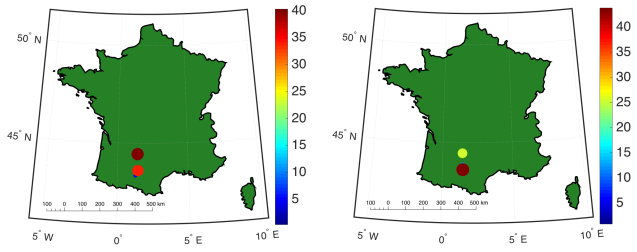
The results are displayed in Fig. 6a and Fig. 6b, showing the CDF of the spatial error in the position of home and work places, respectively. We can clearly observe the following.

- The errors related to home locations are fairly small. They are below 1 km for all users, and within 100 m for 94% of them. The latter figure is easily one order of magnitude smaller than the coverage radius of



(a) Home locations (b) Work locations

Figure 6: CDF of the spatial error (in km) between (a) home and (b) work locations estimated from GPS and CDR data.

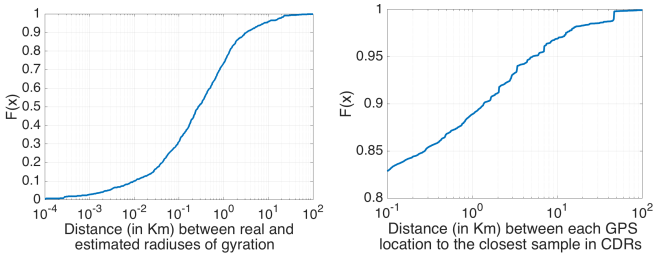


(a) GPS-based (b) CDR-based

Figure 7: Example of popular locations during work hours, along with their percentage of occurrence (color bars) inferred from (a) GPS and (b) CDR data.

medium-sized cell sectors. In other words, the positioning error due to the temporal sparsity of CDRs seems negligible when compared to that induced by mapping the subscriber's location to the base station position.

- The errors associated with work locations are sensibly higher than those measured for home locations. While 75% of users have an error of less than 300 m, the work places of a significant portion of individuals (around 12% of the total) are identified at a distance higher than 10 km from the position extracted from GPS data. Further investigations show that these large errors typically occur for users who do not seem to have a stable work location, and might be working in different places depending on, e.g., the day of the week. As an example, Fig. 7 shows for both GPS and CDR data, the same most popular locations at which one specific user appears during the work hours, along with their percentage of occurrence during such an interval (cf. color bars). For both data, two important places emerge, in different cities at around 100 km of distance in Southern France. The overall work-time presence of the user is close between the two locations in the ground-truth case (40% versus 33%) and the location tagged as the effective user' work location is the one given by the highest presence (cf. dark red location in Fig. 7a). On the other side, the reduced amount of data in CDRs reverses (26% versus 44%) the more reliable ranking obtained with GPS data, and consequently, tags a different location as the effective user's work location (cf. dark red location in Fig. 7b). Still, situations like the one in Fig. 7 prove that, although the granularity of CDRs is sufficient to detect popular locations for individual subscribers, granularity improvement is required to better infer effective locations.



(a) Radius of gyration (b) All samples

Figure 8: CDF of the spatial error (in km) between the (a) radiuses of gyration and (a) all samples from the GPS and CDR data.

Overall, these results confirm previous findings [20], and further prove that CDRs yield enough detail to detect effective locations in users' movement patterns. However, they also point out that they may lead to incorrect estimations in the ranking among such locations.

4.2 Span of movement

As a second test, we study whether CDRs can be used to determine the geographical span of the movement of individual users. To that end, we employ the radius of gyration as a relevant metric. For a user $u \in \mathcal{U}$, the radius of gyration r_u is computed as

$$r_u = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ell_u^i - \ell_u^*)}, \qquad (6)$$

where ℓ_u^* is the center of mass of all locations recorded in the spatiotemporal trajectory of u, i.e., $\ell_u^* = \frac{1}{n}\sum_{i=1}^{n}\ell_u^i$.

We compute the radius of gyration from the daily data of each user, using both GPS and CDR records. Fig. 8a shows the CDF of the error obtained when comparing the results. The span of movement according to CDRs entails a small error, lower than 100 m, only in 30% of cases. The errors is instead larger than 1 km for 26% of the users. These numbers are far from those obtained in the case of home locations, and corroborate previous findings on the limited suitability of CDRs for the assessment of the spread of subscribers' mobility [20].

4.3 Complete trajectories

Finally, we compare the full GPS and CDR trajectories. To that end, we compute the geographical distance between each GPS sample and the CDR sample that is the closest in time. The resulting CDF is shown in Fig. 8b. For 83% of GPS points, the error in the equivalent CDRs is minimal, i.e., 100 m or less: This is consistent with the well-known behavior of many individuals who tend to be fairly static and spend most long periods of time at a same location [11, 25].

However, for around 11% of samples, the information in CDRs is highly erroneous, with spatial displacements of at least 1 km, with peaks of several tens of km. These errors can be imputed to periods of significant mobility of subscribers (corresponding to transition periods, e.g., com-

muting or traveling), during which sparse CDR data cannot track positions reliably. Thus, the results confirm that CDRs are not suited to the analysis of transient movement patterns of individuals [20].

5. CALL DETAIL RECORD COMPLETION

As mentioned in Sec. 1, data completion aims at filling the gaps in CDR data, so as to estimate users' positions also in between voice calls or texting activities. Several attempts at data completion have been made to date. A simple solution is to hypothesize that a user remains static at the same location where he is last seen in the CDR data. This methodology is adopted, e.g., by Khodabandelou *et al.* [26] to compute subscribers' presence in mobile traffic metadata used for population density estimation. We will refer to this approach as the `static` solution in the following, and we will use it as a basic benchmark for more advances techniques.

A second approach stems from a continuous-movement model: users continuously move between consecutive samples in CDR data, without stopping at all. An in-depth study has been carried out by Hoteit *et al.* [27], who find that the inter-sample interpolation needs to be adapted on a per-user basis in order to achieve the best accuracy. Specifically, the trajectory of users having a small radius of gyration can be reconstructed using the linear interpolation between consecutive samples. Instead, the trajectory of users having a high radius of gyration is well approximated using a cubic interpolation (i.e., a third-degree spline that interpolates the function by a cubic polynomial using values of the function and its derivatives at the ends of each subinterval). The threshold between small and large radiuses of gyration is set at 32 km. We will refer to this as the `continuous` approach in our comparative performance evaluation.

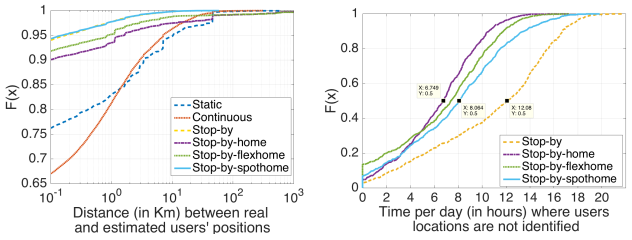
Building on in-depth studies proving individuals to stay in the vicinity of their voice call places most of the time [25], Jo *et al.* [28] assume that users can be found at the location where they generate some digital activity for a hour-long interval centered at the time when the activity is recorded. If the time between consecutive CDR events is shorter than one-hour, the inter-event interval is equally split between the two locations where the bounding events occur. This solution is denoted as `stop-by` in the remainder of the paper.

In addition to those above, we introduce several new techniques, which represent refinements of the `stop-by` solution. They are as follows.

- The `stop-by-home` technique leverages the fact that CDR data allow identifying the home location of individuals with high accuracy. It thus extends the solution in [28] by adding fixed temporal home boundaries. In other words, if a user's location is unknown during the night time interval t^H, due to the absence of CDR samples in that period, the user will be considered at his home location throughout t^H.

- The `stop-by-flexhome` technique 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 t^H as the fixed boundaries for all users, we compute for each user $u \in \mathcal{U}$ the most probable interval of time $t_u^H \subseteq t^H$ during which the user is at his home location.

- The `stop-by-spothome` technique augments the previous technique by accounting for positioning errors that can derive from users who are far from home during some nights, or 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 t_u^H is not identified and he was last seen at no more than 1 km from his home location, he is moved to his home location.



(a) Accuracy (b) Coverage

Figure 9: CDF of (a) the spatial error (in km) between samples from the GPS and completed CDR data, and (b) the temporal coverage of completed CDR data.

In the following, we compare all these techniques from two dual perspectives. The first is that of *accuracy*, i.e., the spatial error between mobility metrics computed from ground-truth GPS data and from CDRs completed with the different techniques above. The second is *coverage*, i.e., the percent of time during which the data completion technique can determine the position of a user. Indeed, the `static` and `continuous` approaches provide some user position at all times, but this is not true for `stop-by` and our derived techniques. In this case, the CDR data is completed only for a portion of the total time, and the location of the user remains unknown in the remaining time.

Accuracy and coverage are discussed in Sec. 5.1 and Sec. 5.2, respectively.

5.1 Completed data accuracy

We compute the geographical distance between each GPS sample and its time equivalent (i.e., real or estimated) CDR sample. As we pointed out above, some data completion solutions are not designed to provide positioning information at all times. 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 computed for GPS samples whose timestamp fall in the time periods for which completed CDR data is available.

Fig. 9a summarizes the results of our comparative evaluation of accuracy, and allows drawing the following main conclusions.

- The `static` and `continuous` techniques provide very poor accuracy. More precisely, the `static` approach yields results that are typically worse than –and at best equivalent to– those obtained by the direct comparison of the GPS data and original CDRs in Fig. 8b. The `continuous` approach slightly reduces the emergence of large errors, but it also significantly reduces negligible errors below 100 m. Overall, both techniques yield an error of 3 km or more for around 10% of spatiotemporal samples.

- The `stop-by-home` and `stop-by-flexhome` techniques largely improve the data precision, with an error that

is lower than 100 m in 90-92% of cases. However, they introduce some very large errors, above 50 km, mainly due to situations where the user is travelling and is very far from his actual home location overnight: Forcing the user's position to his home location in such conditions causes significant inaccuracy.

- The `stop-by` and `stop-by-spothome` techniques have nearly identical performance, as the respective curves overlap. The result is very good in both cases, with about 95% of samples that lie within 100 m of the ground-truth position, and only 1% that yield an error larger than 3 km.

We conclude that 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, home location information can be successfully included in such type of models, by accounting for specificities in each user's habits at night.

5.2 Completed data coverage

The `static` and `continuous` techniques provide full coverage by design. However, this is not the case for the `stop-by` and derived solutions. Fig. 9b shows the CDF of the hours per day during which a user's position cannot be identified by such solutions. The coverage performance is very heterogeneous across users, for all solutions: It can range between one hour per day for some individuals up to 20 hours per day for other subscribers. In this case, the `stop-by` technique yields the worst result, with an unknown user position 12 hours per day in the median case. The refinements of the same approach increase the coverage: this is expected, since these approaches aim at defining the users' positions overnight, when actual CDR samples are absent. The improvement is significant, with a median gain of 4-5 hours over the basic `stop-by`.

Overall, the combination of the results in Fig. 9a and Fig. 9b indicate that the `stop-by-spothome` solution achieves the best combination of high accuracy (97% of completed CDR samples within 600 m of the actual user's location, exaclty as in the `stop-by` case) and fair coverage (84% of the users being assigned a position half of the time or more, against the 50% scored by the `stop-by` technique).

6. CONCLUSION

We leveraged novel datasets of GPS data and reconstructed CDRs in order to characterize the bias induced by the use of CDRs for the study of human mobility, and evaluate data completion techniques to reduce such a bias. Our results confirm previous findings about the limitations imposed by the sparsity of CDRs, and provide a first clear ranking of techniques for CDR data completion. Specifically, we show that a solution that *(i)* extends for a limited amount of time the stays of users at known locations, and *(ii)* places users at their home locations with a grain of salt can achieve good accuracy and fair coverage. Such a novel approach outperforms previous proposals in the literature.

Acknowledgment

The authors would like to thank GranData for providing the data used for the experiments. This work was supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO.

7. REFERENCES

[1] M. De Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri, "The death and life of great italian cities: A mobile phone data perspective," in *Proc. of the 25th International Conference on World Wide Web*, 2016.

[2] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proc. of MobiCom*, (New York, USA), 2014.

[3] A. Lima, M. D. Domenico, V. Pejovic, and M. Musolesi, "Disease containment strategies based on mobility and information dissemination," *Nature Scientific Reports*, vol. 5, no. 10650, 2015.

[4] T. Leighton, "Improving performance on the internet," *Commun. ACM*, vol. 52, pp. 44–51, Feb. 2009.

[5] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, MobiCom '07, (New York, NY, USA), pp. 123–134, ACM, 2007.

[6] A. Furno, D. Naboulsi, R. Stanica, and M. Fiore, "Mobile demand profiling for cellular cognitive networking," *IEEE Transactions on Mobile Computing*, vol. to appear, 2016.

[7] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang, "Equip tourists with knowledge mined from travelogues," in *Proceedings of the World Wide Web International Conference*, (New York, NY, USA), 2010.

[8] Y. Zheng, L. Zhang, X. Xie, and 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.

[9] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[10] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, June 2008.

[11] E. Mucceli, A. C. Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin, "On the regularity of human mobility," *To appear in Pervasive and Mobile computing (PMC) Journal, Elsevier*, 2016.

[12] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[13] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, (Berlin, Heidelberg), pp. 133–151, Springer-Verlag, 2011.

[14] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris, "Assessing the potential of ride-sharing using mobile and social data: A tale of four cities," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, (New York, NY, USA), pp. 201–211, ACM, 2014.

[15] B. Csáji, A. Browet, V. Traag, J.-C. Delvenne, E. Huens, P. V. Dooren, Z. Smoreda, and V. D.

Blondel, "Exploring the mobility of mobile phone users," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 6, pp. 1459 – 1473, 2013.

[16] A. Hess, I. Marsh, and D. Gillblad, "Exploring communication and mobility behavior of 3g network users and its temporal consistency," in *2015 IEEE International Conference on Communications (ICC)*, June 2015.

[17] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frías-Martínez, and J. J. Ramasco, "Cross-checking different sources of mobility information," *PLoS ONE*, vol. 9, no. 8, 2014.

[18] "Ct-mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network," *Computer Communications, Special Issue on Mobile Traffic Analytics*, 2016.

[19] S. Isaacman, R. Becker, R. CÃąceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Ranges of human mobility in los angeles and new york," in *Proc. of PERCOM*, 2011.

[20] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 16, pp. 33–44, Dec. 2012.

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

[22] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, pp. 255–268, Mar. 2006.

[23] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, p. 207, 2005.

[24] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PLoS ONE*, vol. 7, pp. 1–9, 06 2012.

[25] M. Ficek and L. Kencl, "Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model," in *INFOCOM, 2012 Proceedings IEEE*, pp. 469–477, March 2012.

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

[27] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296 – 307, 2014.

[28] H.-H. Jo, M. Karsai, J. Karikoski, and K. Kaski, "Spatiotemporal correlations of handset-based service usages," *EPJ Data Science*, vol. 1, pp. 1–18, 2012.