Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis

Sahar Hoteit, Guangshuo Chen, Aline C Viana, Marco C Fiore

Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis

Sahar Hoteit[1], Guangshuo Chen[2,3], Aline C. Viana[3] and Marco Fiore[4]

[1]*ISEP, France* [2]*École Polytechnique, Université Paris-Saclay, France*
[3]*INRIA, Université Paris-Saclay, France* [4]*CNR - IEIIT, Italy*

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.

Keywords: Human trajectory; important locations; movement inference; call detail record

1 Introduction

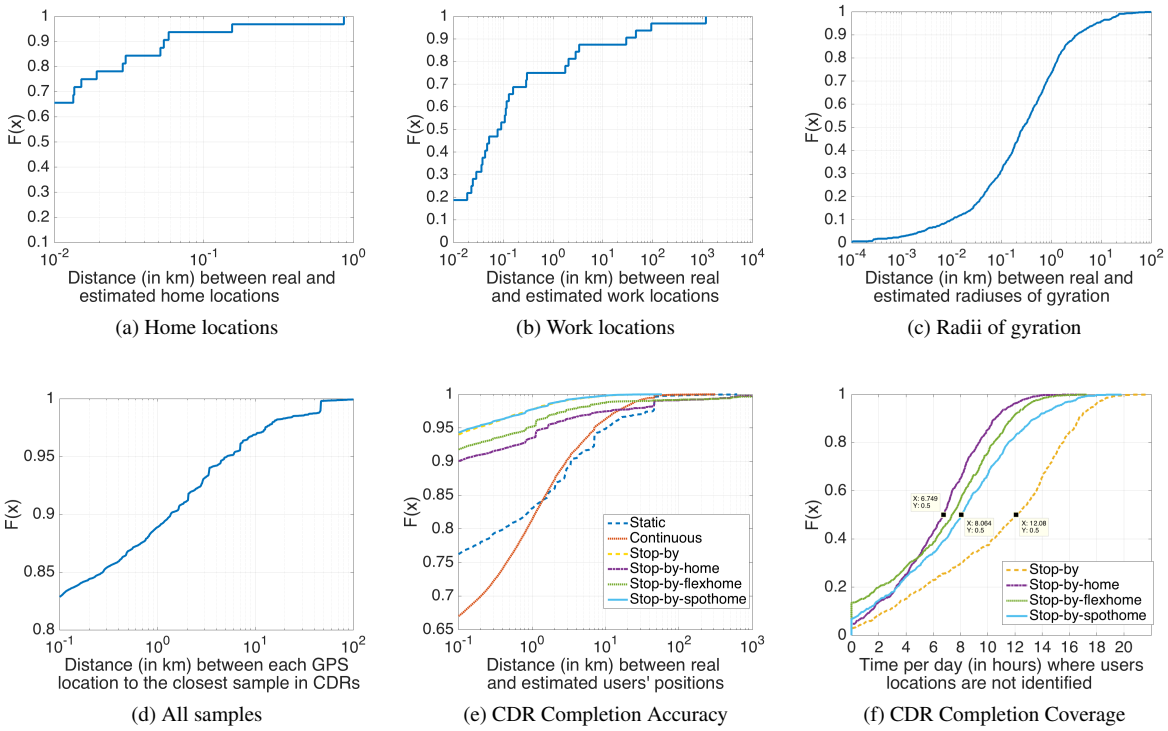
A deep understanding of human-mobility patterns can yield interesting insights into a variety of important societal and networking issues. In this context, Call Detail Records (CDRs) have rapidly emerged as a primary source of knowledge about human mobility, they contain timestamped and geo-referenced logs on each voice call or texting activity of every serviced customer [NFRS16]. The analysis of CDRs has revealed, 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 [GHB08]. Similarly, significant places in our lives (e.g. home, work, shopping- or hobby-related locations) are easily inferred from CDRs [IBC$^+$]

However, the sparsity of CDRs often has an adverse impact on the dependability of study results. Due to the bursty and irregular nature of the communication activities they capture, CDRs are habitually sparse in time, as user's locations may be not recorded with a stable and consistent frequency, and also sparse in space, as locations are known at the cell sector or base station coverage levels. The question of whether and to what extent such a sparsity affects mobility studies has been only partly addressed. For instance, the comparison of CDR-based analyses with equivalent studies of logs of data traffic that have much higher frequency than CDR has shown that CDRs allow to correctly identify, for each user, popular locations that account for 90% of the subscriber's activity. However, the CDRs do not allow inferring transient locations or measures of the geographical spread of the users' mobility [RZZB12]. We refer the reader to [HCVF16] for a more elaborate version of the state of the art.

Data completion that consists of filling the spatiotemporal gaps in CDRs is hence an interesting approach to mitigate the sparsity in CDRs. In this paper, we leverage an original dataset of GPS logs as ground truth, and we mimic sparse CDRs by subsampling such ground-truth data. Then, we first assess the capability of sparse CDRs of modeling important features of individual trajectories: our results confirm previous findings in the literature. Secondly, 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. In addition, we propose original CDR data completion solutions, and show that they outperform previous proposals in reducing the spatial error in the completed data and in shortening the time periods where no location information is available.

2 Datasets

Our study is based on two datasets represented as follows:



(a) Home locations (b) Work locations (c) Radii of gyration

(d) All samples (e) CDR Completion Accuracy (f) CDR Completion Coverage

Figure 1: CDF of the spatial error of (a) home locations, (b) work locations, (c) radii of gyration and (d) all samples from the ground-truth and CDR data. CDF of (e) the spatial error between the ground truth and the completed CDR data and (f) the temporal coverage of completed CDR data.

- The **fine-grained data** that is used as our ground truth is obtained through an Android mobile phone application, MACACOApp[†]. The application collects data related to a user's digital activities (*e.g.*, mobile services, uplink/downlink traffic volumes, network connectivity, etc.), and user's locations as GPS logs[‡] at every 5 minutes. The data collection spans 18 months and covers 84 users who live in 6 different countries and travel worldwide. In this paper, we focus mainly on collecting GPS logs of users without taking into account the type of traffic generated.
- The **coarse-grained data** is a downsampled version of the MACACOApp data used to mimic sparse CDR data. The downsampling 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 sensible manner, we leverage real-world CDRs, collected by a cellular operator in Mexico. Specifically, we downsample the MACACOApp data according to the real distribution of the inter-event time between each user's consecutive CDRs on each hour, which allows taking into account the differences emerging across hours. We refer the reader to [HCVF16] for more details about our downsampling technique.

3 Biases in Call Detail Records

We compare the ground-truth and the CDR data in terms of the results they yield in human mobility analysis, and perform the following three tests:

1. **Identification of home and work locations**: The identification of significant places where people live and work is an important step in characterizing human mobility. Using our two datasets, we determine for each user *(i)* the home and *(ii)* the work location as the most frequent location during the

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

[‡] MACACOApp collects only GPS locations in longitude and latitude coordinates but it does not provide any information about the phone calls and the sent or received SMS by the users

nighttime $t^H = (22h, 7h)$ and the common working period $t^W = (9h, 17h)$, respectively. The CDF of the geographical distance (in km) between the real home/work locations (i.e., those obtained through ground truth data) and the estimated ones (i.e., those obtained through CDR data) are presented in Fig. 1a and Fig. 1b. We can notice that the errors related to home locations are fairly small (below $1km$ for all users, and within $100m$ for 94%), which indicates that 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. However, the errors associated with work locations are sensibly higher (within $300m$ for 75% of users but larger than 10 km for 12% of them). The large errors are due to the presence of users who do not have a stable work location, and might be working in different places depending on the day of the week.

2. **Span of movement**: To study if CDRs can be used to determine the geographical span of the movement of individual users, we employ the radius of gyration r_u as a relevant metric [GHB08]. It is defined by: $r_u = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ell_u^i - \ell_u^*)}$ where ℓ_u^* is the center of mass of all locations recorded in a user's spatiotemporal trajectory $\{\ell_u^1, \cdots, \ell_u^n\}$. The CDF of the errors between the radius of gyration computed using both datasets is shown in Fig. 1c. We notice that in only for 30% of the users, the difference between real and estimated radiuses of gyration is lower than $100m$). However, for 26% of the users, the difference is larger than $1km$.

3. **Complete trajectories**: We compare the ground truth and CDR trajectories by computing the geographical distance between each ground-truth sample and the corresponding CDR sample that is the closest in time. As shown in Fig. 1d, for 83% of points, the error in the equivalent CDR is minimal (*i.e.*, $100m$ 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 the same location [FK12]. However, for around 11% of samples, the information in CDRs is highly erroneous (with spatial displacements of at least $1km$). The errors of these samples 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.

In summary, these results confirm previous findings [RZZB12], and further prove that CDRs are not very suited to the analysis of transient movement patterns of individuals. In addition, the fact that our CDRs yield similar performances (i.e., the same statistical distributions) to those observed in the literature using real-world large-scale CDR datasets corroborates the validity of our downsampling approach.

4 Call Detail Record completion

Data completion aims at filling the spatiotemporal gaps in CDR data. Hereafter, we mention some of the most important techniques from the literature.

- The `static` solution adopted in [KGEYF16] in which the user remains static at the same location where he is last seen in the CDR data.
- The `continuous` solution adopted in [HSS+14] in which the user continuously moves between consecutive samples in CDR data without stopping at all. It uses the linear and cubic interpolations to reconstruct a user's trajectory from discrete samples according to his radius of gyration.
- The `stop-by` solution proposed in [JKKK12] assumes 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.

In addition to those above, we propose three new techniques, which represent refinements of the `stop-by` solution by leveraging the fact that CDR data allow identifying the home location of individuals with high accuracy. They extend the `stop-by` solution by adding home boundaries.

- The `stop-by-home` technique assumes that 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. Instead of considering t^H as the fixed boundaries for all users, each user $u \in \mathcal{U}$ has a relaxed and flexible home boundary computed as 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.

In the following, we compare the different completion techniques in terms of two metrics: the accuracy and the temporal coverage.

1. **Completed data accuracy**: we compute the geographical distance between each ground-truth sample and its time equivalent CDR sample. The results are shown in Fig. 1e. We notice that the `static` and `continuous` techniques provide very poor accuracy; both techniques yield an error of 3*km* or more for around 10% of spatiotemporal samples. On the other hand, the `stop-by-home` and `stop-by-flexhome` techniques largely improve the data precision, with an error that is lower than $100m$ in $90-92\%$ of cases. However, they introduce some very large errors, above $50km$, mainly due to situations where the user is traveling and is very far from his actual home location overnight. 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 $100m$ of the ground-truth position, and only 1% that yield an error larger than $3km$.
2. **Completed data coverage**: we evaluate the temporal coverage (i.e., time per day in hours during which user's position cannot be identified) for the `stop-by` and derived solutions in Fig. 1f. We notice that 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` technique.

Overall, the combination of the results in Fig. 1e and Fig. 1f 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).

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.

References

[FK12] 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*, pages 469–477, March 2012.

[GHB08] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[HCVF16] Sahar Hoteit, Guangshuo Chen, Aline Viana, and Marco 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, pages 45–50, New York, NY, USA, 2016. ACM.

[HSS+14] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64:296 – 307, 2014.

[IBC+] 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 *Proc. of Pervasive*.

[JKKK12] Hang-Hyun Jo, Márton Karsai, Juuso Karikoski, and Kimmo Kaski. Spatiotemporal correlations of handset-based service usages. *EPJ Data Science*, 1:1–18, 2012.

[KGEYF16] Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El-Yacoubi, and Marco Fiore. Population estimation from mobile network traffic metadata. In *IEEE WoWMoM*, 2016.

[NFRS16] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica. Large-scale mobile traffic analysis: A survey. *IEEE Communications Surveys Tutorials*, 18(1):124–161, 2016.

[RZZB12] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *SIGMOBILE Mob. Comput. Commun. Rev.*, 16(3):33–44, December 2012.