# Towards an Adaptive Completion of Sparse Call Detail Records for Mobility Analysis

Guangshuo Chen, Aline Carneiro Viana, Carlos Sarraute

# Towards an Adaptive Completion of Sparse Call Detail Records for Mobility Analysis

(Highly focused technical solution)

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

Carlos Sarraute
Grandata Labs
Bartolome Cruz 1818 Vicente Lopez
Buenos Aires, Argentina
charles@grandata.com

***Abstract*—Call Detail Records (CDRs) are a primary source of whereabouts in the study of multiple mobility-related aspects. However, the spatiotemporal sparsity of CDRs often limits their utility in terms of the dependability of results. In this paper, driven by real-world data across a large population, we propose two approaches for completing CDRs adaptively, to reduce the sparsity and mitigate the problems the latter raises. Owing to high-precision sampling, the comparative evaluation shows that our approaches outperform the legacy solution in the literature in terms of the combination of accuracy and temporal coverage. Also, we reveal those important factors for completing sparse CDR data, which sheds lights on the design of similar approaches.**

***Keywords*—Call detail records, user mobility, human trajectories, location boundaries.**

## I. Introduction

In the past decades, the proliferation of personal mobile devices makes Call Detail Records (CDRs) a very promising source of location information [1]. Collected by mobile network operators for billing purposes, CDRs document the details about *when*, *where* and *how* mobile phone subscribers generate voice calls or text messages, usually across remarkably large populations. The rich information from CDRs has led to a dramatic increase in mobility-related studies, such as identifying important locations [2], optimizing paging in cellular networks [3], and understanding dynamics of human mobility [4].

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, and thus may not record a user's whereabouts with a stable and consistent frequency. The incomplete mobility information from CDRs causes possible biases on characterizing mobility-related features [5], [6], [7]. To deal with the sparsity, sometimes heavy filters have to be applied on CDRs to select users having enough mobility information [8].

Data completion aims at filling spatiotemporal gaps in CDRs as much and accurate as possible. It is to locate users continuously in time by leveraging the information of users' instantaneous whereabouts. Though it does not fully conquer the sparsity, as locations logged by CDRs are usually incomplete [7], data completion can relieve the temporal sparsity of CDRs and problems the latter raises. The legacy solution for completing sparse CDR data is to hypothesize that a cell tower location documented in a CDR is available and representative for a period (typically one hour) rather than only at a time instant when an activity happens, as used in [8], [9]. This solution is actually a reflection of human nature, *i.e.*, one tends to stay in the vicinity of her voice call places most of the time [10].

A major drawback with the legacy solution is that it always expands all CDRs by the same period. One this point, previous findings have shown that using a fixed period at all time is inadequate. In the scenario of determining whether and when a mobile subscriber stays at home during the nighttime, Hoteit *et al.* [6] found that estimating the home period adaptively by historical CDRs outperformed the legacy solution with using a fixed period $(10pm, 7am)$ in terms of accuracy. In the general scenario of completing CDRs during the daytime, we found that the legacy solution aforementioned might lead to a significant spatial error in our previous work [7], which reported significantly that the spatial error was positively correlated with the cell size.

The studies above reveal the importance of having an adaptive approach for data completion for better accuracy. So far, however, there has been little discussion about this aspect. Although [6] proposed an adaptive solution for the scenario of identifying user's home, it is not a universal design and does not consider the environmental information like the cell size.

In this paper, we keep on focusing on data completion for CDRs. We explore *(i)* what can be extracted from CDRs as features for their completion, and *(ii)* which features are critical to the design of a universal adaptive approach. Our results contribute to the effort on reliable CDR data completion in the following ways.

- Our investigation is based on two real-world datasets. Compared with previous GPS datasets, the dataset which we leverage as ground-truth still features high temporal resolution but covers movements of a larger number of users. Details are provided in Sec. II.
- We propose two adaptive approaches for completing sparse CDR data and assess their quality on hundreds of thousands of CDRs leveraging the ground truth information. They outperform the legacy solution: keep-

ing a low spatial error and shortening uncompleted periods. Also, we shed light on the main features which are related to completing CDRs through learning real-world data. Details are provided in Sec. III.

Conclusions are finally discussed in Sec. IV.

## II. Datasets

We leverage two datasets collected from a major cellular operator in Mexico: the target dataset which is composed of CDRs, and the flow dataset to build ground truth information of user movements.

The target dataset contains CDRs of $36,735$ users recorded from April 1st to August 31st, 2015. On each of these days, CDRs are collected during $[10am, 6pm]$, prevailing working hours. Each CDR provides the detailed information of a user's activity (*i.e.*, a phone call or a text message), consisting of the involved devices (*i.e.*, caller/callee or message sender/receiver, as anonymized identifiers), the activity time, the routing cell tower location, and the activity duration.

The flow dataset is composed of flows collected during $[10am, 6pm]$ across the same population as the target dataset[1]. Each flow describes the lifecycle of a TCP or UDP session, and consists of the device identifier, the session time, and, particularly, the cell tower location where a session ends. Therefore each user has a fine-grained discrete trajectory of locations by her flows. Due to the operator's limits of privacy, we can only have three days of flows (July 19th, 20th and August 9th, 2015). On these days, we use the flows to construct continuous mobility information as ground truth. For that we complement each discrete trajectory by expanding each flow from a time instant to a continuous period, as illustrated in Fig. 2(1).

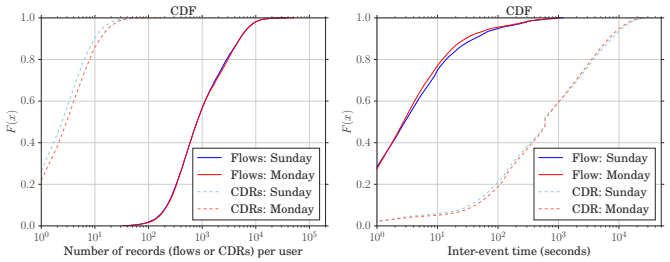
We plot the cumulative distribution function (CDF) of the number of records and the inter-event time in Fig. 1(a) and Fig. 1(b), respectively. The figures reveal that: *(i)* each user has far more flows than CDRs ($95\%$ of users have less than $10$ calls but more than $200$ flows); *(ii)* these flows spread the 8-hour observing period with a dense temporal coverage (in $95\%$ of cases, a user has two consecutive flows within $100$ seconds, but only in $20\%$ of cases, has two consecutive CDRs).

Overall, the flow dataset contains fine-grained mobility information. Its high temporal granularity ensures flows capture all handovers of cell towers in the observing period of each day, and thus supports the use of trajectories in this flow dataset as ground truth in our analysis.

## III. CDR Data Completion

In this section, we propose two adaptive approaches for completing CDR data. The approaches are driven by real data and aim at filling temporal gaps of unknown locations between consecutive activities. We evaluate their performance

[1]The following data pre-processing steps are carried out prior to our analysis, in order to guarantee that every user's movement satisfies an appropriate temporal granularity in the flow dataset. We first apply the *recursive look-ahead filter* on each user's flows to tackle the undesirable effects of *cell-tower oscillation* [11]. We then filter out those who have two consecutive flows within higher than 20 minutes. We refer the reader to [7] for all details.




(a) Number of Records (b) Inter-event Time

Fig. 1. (a) CDF of the number of CDRs (as dashed lines) and flows (as solid lines) per user. (b) CDF of the inter-event time between two consecutive CDRs (as dashed lines) and flows (as solid lines).

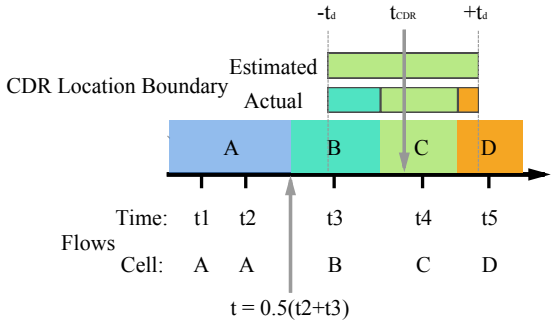



Fig. 2. A demo of (1) constructing the ground truth and (2) estimating a location boundary by the `static` approach: (1) Suppose five consecutive flows recorded at time $t_1, \ldots, t_5$ and at cell locations $A, B, C, D$. The two flows at $t_1$ and $t_2$ are merged as they are observed consecutively at the same cell $A$. Each transition between two cells is assumed to occur at the mid-time of corresponding consecutive flows. In this way, a continuous trajectory is built. (2) The `static` approach estimates a fixed-period location boundary $(-t_d, +t_d)$ attached with a CDR at time $t_{\text{CDR}}$ at the cell $C$, so as to assume the user remains at the cell $C$ during this period, while actually she moves from the cell $B$ to $D$, indicating that only a sub-period is accurate in this location boundary.

by comparing with the legacy solution introduced in Sec. I which we refer as the `static` approach in the following.

The `static` approach is to hypothesize that a user always stays in the corresponding cell during a period centered at the time of each CDR, as illustrated in Fig. 2(2). By the `static` approach, we present the idea of *location boundary*. Each location boundary contains a period corresponding to the completion of a CDR representing that a user's whereabout during this period. In the `static` approach, all CDRs are completed by symmetric location boundaries of the same size $(-t_d, +t_d)$.

The period in a location boundary should be estimated according to the very situation of the activity. For instance, a location boundary deserves a large $t_d$ if the user is walking when making a call, but fits a small $t_d$ if she is on a high-speed train. The arbitrary determination of $t_d$ in the `static` approach leads to a significant spatial error in practice due to the complexity of a user's realistic behavior [7].

With regards to this, we introduce two novel adaptive approaches, where $t_d$ (or $t_d^{(s)}, t_d^{(e)}$) in a location boundary is adaptively determined by its own CDR, unlike the `static` approach which uses a unified threshold. They are *(i)* the `sym-adaptive` approach makes a CDR into a symmetric location boundary as $(-t_d, +t_d)$ and *(ii)* the `asym-adaptive`

approach makes a CDR into an asymmetric location boundary as $(-t_d^{(s)}, +t_d^{(e)})$.

In the following, we introduce how a location boundary is estimated in the two adaptive approaches in Sec. III-A. After that, we compare all the three approaches from two perspectives: *accuracy* and *coverage*, discussed in Sec. III-B.

### A. Determining adaptive location boundaries

**Observable factors from CDRs:** Leveraging a large number of a user's CDRs collected during the long-term observing period, we can learn her behavior for identifying a location boundary though the following factors in three categories:

*1) Event-related factors:* These are the metadata of a CDR, including the activity's `time`, `type` (call/message) and `duration`[2].

*2) Long-term behavior factors:* The radius of gyration (`URg`) of a user, the number of a user's locations (`ULoc`) appearing in the observing period, and the number of a user's active days (`UDAY`). These factors characterize a user by giving senses of *(i)* her long-term mobility and *(ii)* her habit on generating calls and text messages, computed by her CDRs produced during the 5-month observing period.

*3) Location-related factors:* The first factor in this category is related to the cell size[3], *i.e.*, the average call radius (`CR`). Since we have no knowledge of the actual cell coverage, we assume a homogeneous propagation environment and an isotropic radiation of power in all directions at each cell tower, so that we are able to roughly estimate each cell's `CR` using a composition of Voronoi cells extracted from CDRs which covers the area, as in [7].

The rest of the factors describe the location where the activity happens regarding its importance to the user. For that, we learn from the algorithm presented by Isaacman *et al.* [2], which is designed to determine prominent locations which the user usually spends a large amount of time and/or visits frequently. Their algorithm firstly clusters all locations which appear in a user's trajectory of CDRs, and then identify for each cluster whether a cluster is important by measuring these observable factors derived from each cluster in the following, as used in our work: *(i)* the number of days on which any cell tower in the cluster was contacted (`CDay`); *(ii)* the number of days which 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 the cluster (`CTower`); *(v)* the distance from the activity location to the centroid of the cluster (`CDist`).

**Training and estimating**: We develop our approaches by firstly training our model with a set of $65,791$ CDRs collected on August 9th. The other two days, i.e., July 19th and 20th, on which we have the ground-truth information are then utilized into the testing phase. In this way, the model is trained having as input the above described factors and the location boundaries extracted from the ground-truth information. More specially, given a CDR, its user and location, we derive a vector $(x_1, \cdots, x_n)$ from the above factors, such as: *(i)* the categorial factor `type` is converted to two binary features by one-hot encoding; *(ii)* the `time` is converted to two separate time differences (in seconds) from the activity time to $10am$ and to $6pm$; *(iii)* other factors are used as values they are.

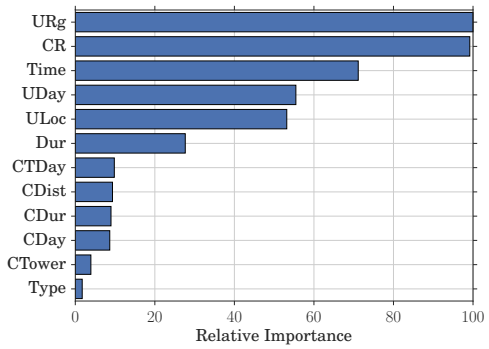



Fig. 3. Relative Importance of features in determining accurate location boundaries.

In the `sym-adaptive` approach, this vector and the symmetric location boundary $(-t_d, +t_d)$ of each activity extracted from the ground-truth information are provided as input to a model trained by Gradient Boosting Regression Trees (GBRT) [12], under the squared loss function. Once the training phase is completed, the testing phase consists in estimating the interval $(-t_d, +t_d)$ by feeding the vector derived from $136,562$ CDRs collected on the other days (July 19th and 20th) into the regression formula.

In the `asym-adaptive` approach, the factors vector and the two limits of an asymmetric location boundary, i.e., $t_d^{(s)}$ and $t_d^{(d)}$, of each activity extracted from the ground-truth information are separately provided as input resulting in two trained GBRT models. The training phase is the same as in the `sym-adaptive` approach but works independently on the two models.

Fig. 3 shows the relative importance of factors with respect to the estimation of a location boundary after the training phase of the adaptive approaches. This figure allows us drawing the following main conclusions, valid for both approaches.

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

### B. Coverage and accuracy

To validate the location boundaries given by the three previous discussed completion approaches, we again use the CDRs collected on July 19th and 20th, on which we have the ground-truth information. Note that our approaches are applicable to CDRs on other days, though there is no ground-truth information. Hereafter, we use the days having ground-truth information only for comparison and validation reasons. For

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

[3]We exclude the antenna-related information such as the transmit power or RF propagation environment, which a mobile operator rarely provides.

the `static` approach, $t_d$ is set to $5/15/30/45/60/90/120$ minutes for location boundaries, respectively.

Additionally, since the goal of data completion is to fill the temporal gaps, we compare the three approaches in terms of *accuracy* and *coverage*. For that, we use the following measures to quantify the completed CDRs:

- *Spatial error*: The average cumulative error in the distance over time during a location boundary, quantifies accuracy in terms of activity.
- *Coverage*: The filling rate, defined as the ratio of the covered duration (*i.e.*, the sum of the periods of a user's location boundaries) to the observing period, represents to what degree a user's CDRs are completed.
- *Accuracy*: The ratio of the accurate duration (*i.e.*, the sum of all accurate sub-periods) to the covered duration (*i.e.*, the sum of all periods) on a user's location boundaries, quantifies accuracy in terms of the user.

Intuitively, an ideal data completion approach should cover the observing period as much and precise as possible, *i.e.*, satisfying high accuracy and coverage simultaneously.

Fig. 4(a)(b) plots the distribution of the spatial error over all location boundaries. It confirms that the spatial error increases as $t_d$ becomes larger when using the `static` approach, which is revealed in our earlier work [7]. More importantly, the performance of the two adaptive approaches is nearly as good as the `static` approach with $t_d = 30$ minutes in terms of the spatial error.

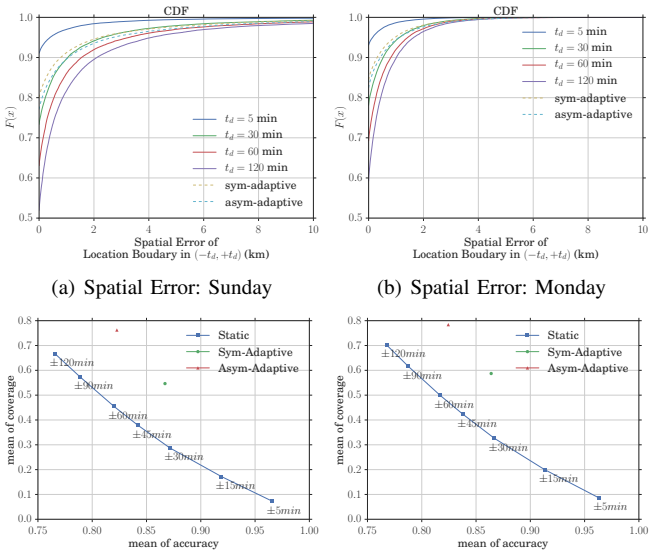
To further compare the approaches in terms of the combination of accuracy and coverage, we plot in Fig. 4(c)(d) the mean of accuracy versus mean of coverage over the observing users. We notice that using a large $t_d$ contributes to enhancing coverage with reducing accuracy as a price, indicating that the `static` approach cannot achieve high accuracy and high coverage together.

However, the two adaptive approaches show splendid performance. The sym-adaptive approach reaches the same level of accuracy as the `static` with $t_d = 30$ minutes. Yet it has a better temporal coverage (approximately as good as the `static` with $t_d = 90$ minutes). This indicates that the sym-adaptive approach can complete more time while it can still ensure accuracy. As to the asym-adaptive approach, it performs better in terms of coverage with losing a small degree of accuracy, compared with the sym-adaptive approach, and still outperforms the `static`.

Overall, we see a clear advantage of the `sym-adaptive` approach over the others in Fig. 4, as it achieves the best combination of fair accuracy and high temporal coverage.

## IV. Conclusion

In this paper, we focused on data completion and proposed two novel data-driven approaches which utilized multiple factors of a user's behavior and determined location boundaries adaptively for completing CDRs. The comparative evaluation showed that the proposed approaches outperformed the legacy




(a) Spatial Error: Sunday (b) Spatial Error: Monday

(c) Accuracy/Coverage: Sunday (d) Accuracy/Coverage: Monday

Fig. 4. CDF of the spatial error of location boundaries computed on (a) Sunday and (b) Monday; mean of accuracy versus mean of coverage per user on (c) Sunday and (d) Monday, across the `static`, `sym-adaptive` and `asym-adaptive` approaches.

solution in the literature. Further research should be done to investigate a heuristic approach which does not rely on the contextual information.

## 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

[1] 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, 2015.

[2] 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 *Pervasive computing*, 2011, pp. 133–151.

[3] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *ACM MobiCom 2007*, 2007, pp. 123–134.

[4] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[5] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, pp. 33–44, 2012.

[6] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis," in *ACM CHANTS 2016*, 2016, pp. 45–50.

[7] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Relevance of Context for the Temporal Completion of Call Detail Record," INRIA Saclay, Research Report RT-482, Nov. 2016. [Online]. Available: https://hal.inria.fr/hal-01393364

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[9] 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.

[10] M. Ficek and L. Kencl, "Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model." *IEEE INFOCOM 2012*, pp. 469–477, 2012.

[11] M. A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 435–454, 2010.

[12] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.