Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era"

Pierdomenico Fiadino
Eurecat - Technology Centre of Catalonia
Barcelona, Spain
pierdomenico.fiadino@eurecat.org

Victor Ponce-Lopez
Eurecat - Technology Centre of Catalonia
Barcelona, Spain
victor.ponce@eurecat.org

Juan Antonio Torrero-Gonzalez
Orange
Madrid, Spain
juan.antoniotorrero@orange.com

Marc Torrent-Moreno
Eurecat - Technology Centre of Catalonia
Barcelona, Spain
marc.torrent@eurecat.org

Alessandro D'Alconzo
Austrian Institute of Technology
Vienna, Austria
alessandro.dalconzo@ait.ac.at

ABSTRACT

The exploitation of cellular network data for studying human mobility has been a popular research topic in the last decade. Indeed, mobile terminals could be considered ubiquitous sensors that allow the observation of human movements on large scale without the need of relying on non-scalable techniques, such as surveys, or dedicated and expensive monitoring infrastructures. In particular, Call Detail Records (CDRs), collected by operators for billing purposes, have been extensively employed due to their rather large availability, compared to other types of cellular data (e.g., signaling). Despite the interest aroused around this topic, the research community has generally agreed about the scarcity of information provided by CDRs: the position of mobile terminals is logged when some kind of activity (calls, SMS, data connections) occurs, which translates in a picture of mobility somehow biased by the activity degree of users. By studying two datasets collected by a Nation-wide operator in 2014 and 2016, we show that the situation has drastically changed in terms of data volume and quality. The increase of flat data plans and the higher penetration of "*always connected*" terminals have driven up the number of recorded CDRs, providing higher temporal accuracy for users' locations.

CCS CONCEPTS

• **Networks → Network mobility**; **Mobile networks**;

KEYWORDS

mobile networks, call detail records, human mobility

ACM Reference format:
Pierdomenico Fiadino, Victor Ponce-Lopez, Juan Antonio Torrero-Gonzalez, Marc Torrent-Moreno, and Alessandro D'Alconzo. 2017. Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era". In *Proceedings of Big-DAMA '17, Los Angeles, CA, USA, August 21, 2017,* 6 pages.
https://doi.org/10.1145/3098593.3098601

1 INTRODUCTION

According to a projection by GSMA, 70% of people worldwide will be mobile subscribers by the end of 2017, reaching almost 85% in developed countries [10]. Another figure [2] states that half of mobile customers own Internet-capable mobile devices and the largest majority of people access the Internet through cellular networks [1]. The increasing penetration of mobile terminals is not new: the rising trend started two decades ago. With these numbers, it is not a surprise that both network operators and the research community look at mobile technologies as an unprecedented information source. Every terminal produces an enormous amount of meta-data that can be exploited to study aggregated behaviors and trends.

Current systems for observing human mobility gather information from dedicated infrastructures (i.e., sensors deployed in key areas), which would be too expensive to deploy at scale. Another typical source consists in surveys, which suffers of lack of scalability and error proneness. Mobile devices and cellular networks can help overcoming these limitations as the high number of terminals can be opportunistically exploited as existing sensors, without the need of dedicated hardware.

In this paper, we discuss about the quality of mobility information provided by a type of cellular meta-data known as Call Detail Record (CDR). A CDR is a summary ticket of a telephone transaction, including the type of activity (voice call, SMS, 2G/3G/4G data connection), the user(s) involved, a time-stamp, technical details such as routing information, and the identifier of the cell offering connectivity to the hand-terminal during the transaction. The latter is especially interesting as it allows to localize the associated action within the boundaries of the cell's coverage area.

CDRs might not be as rich as other cellular data (e.g., signaling data including hand-over information), but they have been a popular study subject since the 2000s. The reason behind this interest should be sought in their rather high availability, as network operators are collecting these logs for billing which makes them a

Big-DAMA '17, August 21, 2017, Los Angeles, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5054-9/17/08…$15.00
https://doi.org/10.1145/3098593.3098601

ready-made data source. Indeed, there is a fairly rich literature on the their exploitation for a plethora of applications, ranging from road traffic congestion detection and public transport optimization to demographic studies. Despite the initial interest, the research community gradually abandoned this data source as it has become clear that the location details conveyed by CDRs were biased by the users' activity degree. In other words, we can observe the position of a user in conjunction with an activity, which translates in the impossibility of locating a user with fine temporal granularity.

The goal of this work is to study where we are at with CDRs. We argue that the spreading of flat rates for voice and data traffic encourages users to generate more actions. In particular, competitive data plans are nowadays characterized by very high data volume limits, letting customers keep their terminals *always connected*. The popularity of instant messaging and social network applications with background sync behavior completes the picture of today's typical smartphone usage. Having said this, we observed a drastic change in CDR datasets over the last years. In support of this argument, we analyze two anonymized datasets collected in 2014 and 2016 from a major operator. Our goal is to show that a significant share of users are characterized by a high frequency of actions, accountable, mostly, to data connections. Our first dataset (2014) has been successfully employed in commercial mobility studies (briefly mentioned in this paper). In the course of our work, we have witnessed a remarked rise of action rate, which gives good perspectives for the future. We believe that the lessons learned from our operational experience could be of interest for operators and data practitioners.

Our contributions are manifold: (i) we conduct a longitudinal study showing how CDR datasets *improved* over a 2-year period; (ii) we define a set of indicators apt to measure the quality of CDR datasets in terms of number, frequency and uniformity of distribution of actions over time; (iii) we give initial indications on how to select *high quality* user samples.

The remainder of this paper is organized as follows. In Section 2 we present the state of the art. In Section 3 we introduce the two studied datasets. In Section 4 we propose a set of metrics to estimate the activity degree of users. Section 5 touches upon some applications and provides guidelines for the use of CDRs in the context of human mobility. Section 6 gives the final conclusions.

2 STATE OF THE ART

The use of mobile terminals and cellular networks as source of mobility information has drawn the research community's attention in the last decade. In general, one can distinguish between terminal-based and network-based mobility monitoring approaches. The former consists in smartphone applications that report GPS (Global Positioning System) positions, and optionally other details from the radio interface, to a central server [5, 6, 13, 18]. In [18], for example, GPS is used as part of a cloud-based system to gradually learn user's driving behavior and hence predict the future traffic conditions. The same approach is adopted by consumer applications, such as Waze [1], that rely on a large user base to collect road statistics and offer traffic-aware navigation.

	ds2014	ds2016
Length	31 days (Q3-2014)	31 days (Q2-2016)
# of records/day	350M	1.1B
# of users/day	9M	11M (inc. roamers)
Data volume/day	50GB	120GB

Table 1: Summary of the two CDR datasets.

GPS data is extremely valuable to study mobility with high accuracy in positioning and speed estimation. However, it requires the deployment of dedicated monitoring applications, resulting in smaller datasets than those collected with network-based approaches. Network-based monitoring approaches, conversely, are based on the collection of data from the network itself, i.e., they do not require modifications to the terminals.

Most of previous studies based on network-based approaches make use of Call Detail Records (CDRs) [3, 9, 17]. CDRs are tickets to support operators' billing procedure, summarizing a transaction, such as a call, text message or data connection. They include meta-data such as the time-stamp and cell identifier (useful for inferring the terminal's location in the context of the radio access network's topology). Since CDR datasets only log the position of users when an action occurs, their exploitation for the characterization of human mobility has been criticized [14]. In [17], authors highlight some concerns in this direction and also show that users are in-active most of the time. The main limitation lies in the fact that the mobility perceived from the observation of CDRs is highly biased, as it strictly depends on the specific terminals' action patterns. In other words, users are *visible* during few punctual instants over the entire day, making us miss most of their movements.

We argue that the situation is drastically changed and nowadays CDRs are much richer than the past, due to the increasing usage of data connections and background applications that has consequently increased the number of records. The number and frequency of tickets is now larger and allow a finer-grained tracking of users' positions. Some recent mobility studies, such as [12], are successfully based on CDRs. We expect a resurgence of research work tackling the study of CDRs in the next years. Indeed, the research community is already targeting some open issues, such as the lack of CDR standardization across operators [16].

In addition to CDRs, there are other sophisticated monitoring approaches that rely on the capture of the signaling between terminals and the network (including hand-overs and location area updates) in a passive fashion [4]. In our previous work published in 2012 [8], we have studied the differences between CDRs and signaling data in terms of number of actions per user. Depending on the monitored interfaces, these approaches greatly vary in terms of cost and data quality [15]. Although the analysis of signaling is promising, there is a general lack of studies based on actual operational signaling data (some exceptions are [8, 11]), as complex dedicated monitoring infrastructures for the extraction and immense data storage systems are required.

3 DATASETS DESCRIPTION

The ultimate goal of this work it to define a set of quality indicators for CDR datasets and evaluate the applicability of this data type for

[1] Waze homepage - http://www.waze.com

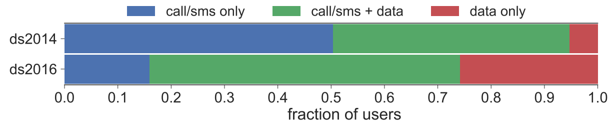


Figure 1: Fraction of users per type of actions. Data connection customer share grows from 50% to 85%.

	ds2014	ds2016	Δ
SMS (snt.+rcv.) / day	0.5	0.1	−0.4
calls (inst.+rcv.) / day	1.8	2.5	+0.7
data conn. records / day	10.9	50.1	+36.9
total actions / day	13.2	52.7	+39.5

Table 2: Averarage number of daily actions per user by action type. There is a remarked shift increase of data connections. SMS records have become marginal.

the inference of human mobility patterns. To this end, we use two anonymized CDR datasets collected in a Nation-wide network with the collaboration of a major cellular operator. The two logs cover the activities of all operator’s millions of customers in Spain, over two different periods, in the 3rd quarter of 2014 and 2nd quarter of 2016 (named ds2014 and ds2016 from now on). The time-span covered by the two datasets allows the observation of the mid-term evolution, in particular for what concerns the number of actions recorded per user per unit of time.

Table 1 summarizes the characteristics of the datasets. Given the considerable amount of data, it has been necessary to adopt suitable technologies for the storage and the analysis. In particular, our set-up consisted in a cluster of 6 nodes (5 workers and 1 master node), with a total HDFS (Hadoop Distributed File System) storage of 10TB. We adopted a mix of different big data tools, such as Apache HIVE for the phase of data cleaning and preprocessing and Spark for the analytics. Our main objective consisted in inferring urban mobility in the city of Barcelona, Spain, producing origin-destination matrices, geographical clustering of densely visited areas and studies of urban trajectory at district level. Despite the limited geographical area of interest, we considered the whole datasets (i.e., whole Spain), in order to observe the origin area of visitors and the main city’s entry points. We present few sample results in Section 5.

Before studying the datasets, it is worth discussing about the type of actions recorded in the CDRs. Each time a user executes an action – e.g., start/answer a call, send/receive a text message (SMS), start a data connection, etc. – the meta-data associated with the telephone transactions are recorded by the operator for billing purposes, independently from the user’s tariff plan. The records are accompanied by technical details, such as the routing information, not considered in this work. Finally, CDRs include the identifier of the cell to whom the mobile terminal was attached during the transaction. This field is particularly relevant for studying the mobility of users, because it reveals an estimation of the geographical position of the hand-held.

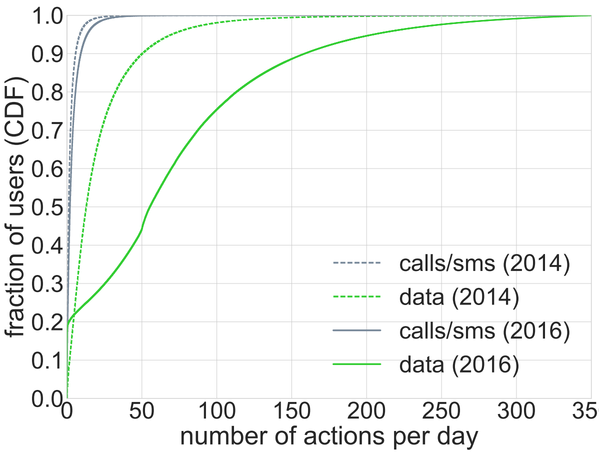
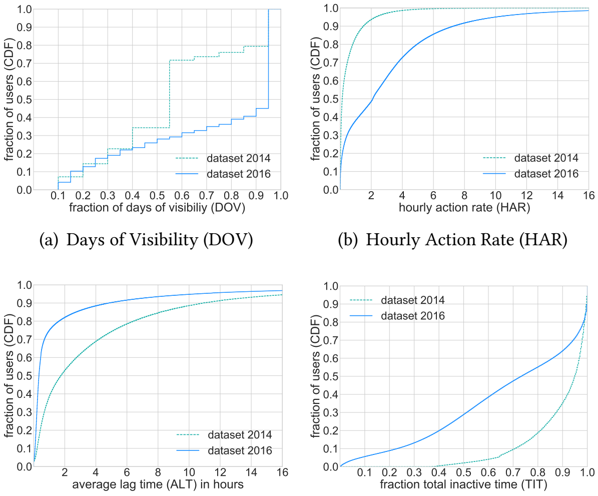


Figure 2: Distribution of daily action count per user in ds2014 and ds2016, separating per type of action.

As a matter of fact, CDRs do not generally include signal strength or other radio details that let triangulate the terminal with high precision, therefore the estimated position corresponds to the antenna location. The position accuracy depends on the network planning and the coverage area of the cell (ranging from 50 meters for pico-cells, up to several kilometers for macro-cells).

Let us now characterize the datasets. Apart from the difference in the absolute number of customers, the largest substantial change from ds2014 to ds2016 is the action rate per user (cfr. Table 2). This is mainly due to a much higher fraction of data connection users. The penetration of smartphones, and in general Internet access through mobile networks, has increased 2 years, as showed in Figure 1. In 2014, only 50% of customers where Internet users. This percentage has grown up to almost 85% in 2016. As a consequence, the action type share has also changed. Table 2 reports the average number of daily actions per user in ds2014 and ds2016. Over two years, users have reduced the number of sent and received text messages (SMS), going from an average of 0.5 messages per day to a negligible 0.1 (an effect of the traditional SMS market being replaced by online messaging systems). For what concerns calls, we measured an average of 1.8 and 2.5 records respectively (the increase is probably caused by the higher number of flat rates). The main change is related to data connections: the number of daily data actions has dramatically increased in ds2016, surging from 10.9 to 50.1. Overall, the daily action count has grown four times.

The increased share of data connection users, together with the increment of data-related logged actions, has driven a remarkable shift of the overall distribution of daily actions by users. Figure 2 depicts the cumulative distribution function (CDF) of average daily actions by user in ds2014 and ds2016, discriminating by action type (voice calls and text messages have been merged for the sake of clarity). The distribution confirms the previous considerations and suggests that the newer dataset provides more tickets (and therefore geographical information) per user. 50% of users in 2016 executed at least 50 daily actions related to data connections. Only 10% did the same in 2014.



(a) Days of Visibility (DOV)

(b) Hourly Action Rate (HAR)

(c) Average Lag Time (ALT)

(d) Total Inactive Time (TIT)

Figure 3: Distribution of activity degree indicators among users in ds2014 and ds2016. All indicators favour ds2016 in terms of users' activity degree.

4 ACTIVITY DEGREE INDICATORS

In the previous Section, we have seen that the average number of daily *action tickets* has dramatically increased from 2014 to 2016, mostly due to the higher number of flat data plans. This information, however, is not enough for evaluating the quality of CDRs for the study of users' mobility. As showed in [17], users tend to concentrate the interactions with their terminals in short time spans (e.g., while commuting), which generates bursts of records. Through CDRs, we have no information about users' activities (and hence their position) while they are idle, which prevents the observation of their movements (or the absence thereof).

To demonstrate how today's high penetration of smartphones has drastically changed the situation, we propose a number of activity degree indicators aimed at measuring not only the amount of records, but also how uniformly the activities are spread over time. By applying these indicators, our goal is to count how many users respect certain "*quality standards*", i.e., their activity degree is sufficiently high and stable in order to allow the observation of movements with fine granularity, hence reducing the risk of drawing partial or misleading trajectories. In the remainder of this Section, we briefly describe the set of indicators and we show how the two datasets under study compare. These indicators are applied to each anonymous user in the datasets.

Days of Visibility (DOV). The first indicator consists in counting the number of distinct days in which users generate activities and therefore *reveal* their position. This indicator does not add much more information than simply counting of actions per day, however it could provide useful insights in case of macro-mobility studies (e.g., at Nation/Regional level). Figure 3(a) shows the distribution of the DOV indicator in terms of fraction of active days in ds2014 (dashed line) and ds2016 (solid line). The curves show a much higher chance of observing a user for more days in the new dataset (50% the users are visible every day in ds2016), hence providing better coarse information of the visited regions). An important remark should be made about the collection period: ds2016 was collected over summer, notoriously a vacation period. This has two consequences: (1) people are generally less active (in particular business phones are idle for longer periods) and/or on holiday abroad, preventing us from observing their logs; (2) more international tourists (roaming on operator's network) could be active during their visit (e.g., one week). These considerations reinforce the conclusion that ds2016 largely dominates in terms of *quality*.

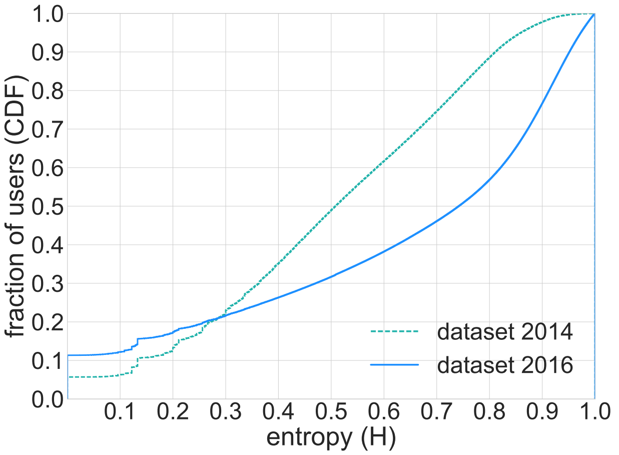


Figure 4: Distribution of entropies in ds2014 and ds2016. The second one is much more skewed towards the right end.

Hourly Action Rate (HAR). For each anonymous user, we calculate the average number of records per hour, giving more precise time-related information wrt. the previous indicator. As showed in Figure 3(b), the distribution of users' HAR greatly favors ds2016 again. In fact, 50% of users generate an average of 2 actions per hour or more. The same percentage of the most active users in ds2014 were executing a number of hourly actions an order of magnitude smaller. Note that good values for this indicator are still not a guarantee of uniformity.

Average Lag Time (ALT). We define as *lag time* the interval between a pair of consecutive actions. It follows from the definition that we seek short average lag times, i.e., users with an action rate that spans the observed period, independently from the hourly rate. This indicator is much more powerful than the previous ones, because it tends to be less biased by burstly activity patterns. As for the other indicators, the two datasets largely diverge for what concerns ALT (cfr. Figure 3(c)). In particular, more than 80% of users in ds2016 are characterized by an ALT smaller than 2 hours, while only 50% respect the same requirement in ds2014.

Total Inactive Time (TIT). This metric is rather intuitive: it tells how long a user is inactive over the entire observation period. It follows that we seek datasets in which users have low values for TIT. Figure 3(d) shows how ds2014 and ds2016 compare with respect to users' amount of inactive periods, in terms of fraction of 1-hour bins over the observation period. The distribution of the older dataset is concentrated towards high percentages of TIT, while the one for ds2016 is more uniform. This indicator is highly

indicator	threshold
DOV – Days of Visibility	$DOV >= 75\%$
HAR – Hourly Action Rate	$HAR >= 1$
ALT –Average Lag Time	$ALT <= 30$ minutes
TIT – Total Inactive Time	$TIT <= 75\%$
H – Entropy	$0.7 <= H <= 0.9$

Table 3: Thresholds used for urban mobility studies.

indicator	avg. ds2014	avg. ds2016
DOV – Days of Visibility	96.5%	97.7%
HAR – Hourly Action Rate	2.25	4.98
ALT –Average Lag Time	888.9	823.3
TIT – Total Inactive Time	62.7%	43.2%
H – Entropy	0.75	0.84

Table 4: Average indicator values for the HAU samples.

influenced by the portion of time of the day under study: if we consider the entire day (including nights, as we have done for all the Figs. 3), the TIT is higher. Given that more and more users tend to keep their smart-phones turned on over night, it is more likely that the users are still visible more often (cfr. applications with background data activities), which explains the substantial change in 2016. It should be noted that, depending on the use-case, we could still be interested in observing actions in inactive periods such as nights to assess the lack of movements.

Entropy (H). The final indicator is the most complex one and summarizes the others. The entropy is a measure of the uniformity of an empirical distribution. Be X a distribution, the entropy is defined as $H(X) = -\sum_{i=1}^{n} p(x_i)log(p(x_i))$, where $x_1, \ldots, x_n$ is the range of values for X, and $p(x_i)$ is the probability that X takes the value x_i. The entropy is normalized to a scaling factor $log(n_0)$, where n_0 is the number of distinct x_i. For our experiments, we consider X the distribution of actions of a user over time and $p(x_i)$ the fraction of actions in the 1-hour time-bin x_i. It follows that the extreme case of a user whose all activity is concentrated in a single bin will give 0 entropy. The opposite case of an extremely uniform user (all bins with the same action count) will produce entropy 1. In the light of this, high entropy users (closer to 1) are to be preferred for mobility studies as their activity pattern is more uniform. Figure 4 shows that the entropy distribution among users in ds2014 is normal-like, centered at 0.5, while the distribution for ds2016 is skewed towards higher values, which confirms quality improvement.

5 LESSON LEARNED AND GUIDELINES

The combination of the aforementioned indicators allow the characterization of users activity degree and, most importantly, the stability of their activity patterns (note that the indicators do not *capture* the quality requirements if considered individually).

We now focus on the application of these metrics for the extraction of samples of *high quality users*, i.e., users whose activity patterns make them suitable for mobility studies with restrained bias. To this end we have defined a set of thresholds for the indicator values, summarized in Table 3. We have applied these conditions in order to select such a sample, named HAU (Highly Active Users) from now on. The sizes of the two obtained samples correspond to 7.5% and 12.5% of users, respectively. The resulting average values of the indicators for the HAUs are reported in Table 4: beside being larger, the 2016 HAU sample presents better characteristics (HAR, TIT and H in particular).

One could argue that the sample size is small, however it should be considered that we have included the entire day for the calculation of the indicators. As said, including night hours negatively impacts all indicators (especially TIT) and it particularly affects ds2014, in which the fraction of *always-on* terminal is smaller. Restricting the analysis to diurnal hours (e.g., 08:00am-10:00pm, the most congested period and of interest for the administrations and businesses), we drive the sample size up to 25% and 38% respectively.

The reported thresholds could serve as guidelines derived from our operational experience in the study of urban mobility. The parameters have been empirically tuned in collaboration with the local authorities taking into account historical statistics of the area, such as travel times with public transportation, area geometry, etc. In particular, we considered the validity period of a public transport ticket (90 minutes) and the average displacement time (roughly 60 minutes). As a side node, we have excluded from our analysis the users with extremely high entropy ($H <= 0.9$). The rationale lies in the fact that a purely uniform behavior is a typical trait of M2M (machine-to-machine) communications, such as IoT (Internet of Things) devices, which shall not be considered for mobility, as, in general, they are not representative of human behavior.

Clearly, the tuning of these values strictly depends on the use case and on the required level of granularity. In case of studies tackling long-distance trajectories (e.g., highways at Nation-scale), the thresholds can be regulated in a less strict fashion, as a coarser position sampling rate would not negatively impact the results. We reckon that this work still lacks general guidelines for rigorously tuning these indicators, in particular for the entropy, which is the less intuitive one. We leave this, together with an entropy-based classifier for IoT devices useful for mobility (e.g., vehicle black-boxes), as future work.

In our studies, we selected HAUs focusing on the city of Barcelona, Spain. In the following paragraph, we shortly report two sample studies (cfr. Figure 5). Figure 5(a) shows a map of clusters where Barcelona visitors concentrate their activities. The clustering has been achieved applying Dbscan (Density-Based Spatial Clustering of Applications with Noise) algorithm [7], using as items the cells weighted by visitor number. The top-clusters (corresponding to touristic attractions or important way points such as stations) are highlighted with colors (less important clusters are in black). Gray cells correspond to *noise*, i.e., cells flagged as no part of any cluster because of the smaller visitor density. Figure 5(b) contains a chord diagram showing the transitions between pairs of districts. In both cases it is important to base the analysis on HAU samples, in order to reduce the mobility bias. Note that a high-frequency position sampling is important for observing also static behaviors.

A detailed description of such results is out of scope and we leave it for a separate publication. However, the examples above provide a high level concept of the potentiality of CDRs. As a final remark, it is worth mentioning that we have accurately verified the statistical relevance of the HAU samples by comparing demographic information (i.e., age, gender, residence, etc.) with the entire customer population. We found negligible deviations that suggest the absence of demographic-related biases. In other words, there are no specific customer segments (e.g., younger users) affected by the increase of logged actions.

6 CONCLUSIONS AND FUTURE WORK

We have studied two large-scale CDR datasets collected over a period of two years. For the sake of characterizing the quality of the mobility information they provide, we have proposed a set of indicators aimed at quantifying the amount and the uniformity over time of action tickets, corresponding to calls, text messages and data connections. The datasets highly differ in terms of position sampling rate and therefore achievable quality of movement observations. Indeed, our investigation revealed that the user habits have drastically changed over time, making newer datasets a finer-grained source of mobility data compared to the past. We foresee that this trend of improvements will continue in the future, while studies and commercial products based on CDRs are gaining relevance.

As for future work, we plan to continue our longitudinal study by characterizing additional datasets collected over 2017. We expect to observe further improvements impacting all our quality indicators., Finally, we are interested in characterizing roamers: CDRs for this user class are an interesting mean to infer the mobility of international visitors (which would provide valuable insights for studies targeting tourism). Roamers, however, do not currently follow the same activity patterns of common users and therefore require further assessments (at least until the forthcoming European regulations for roaming charges are enforced, which is expected to happen in 2017).

REFERENCES

[1] 2015. Spring Survey. http://www.pewglobal.org/2015/06/23/spring-2015-survey. (2015). Accessed: 2017-03-17.

[2] 2017. Mobile phone internet user penetration worldwide 2014-2019. www.statista.com/statistics/284202/mobile-phone-internet-user-penetration-worldwide. (2017). Accessed: 2017-03-17.

[3] H. Bar-Gera. 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies* 15, 6 (2007), 380 – 391. https://doi.org/10.1016/j.trc.2007.06.003

[4] N. Caceres, J. P. Wideberg, and F. G. Benitez. 2007. Deriving origin destination data from a mobile phone network. *IET Intelligent Transport Systems* 1, 1 (March 2007), 15–26. https://doi.org/10.1049/iet-its:20060020

[5] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. 2011. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems* 12, 1 (March 2011), 141–151. https://doi.org/10.1109/TITS.2010.2074196

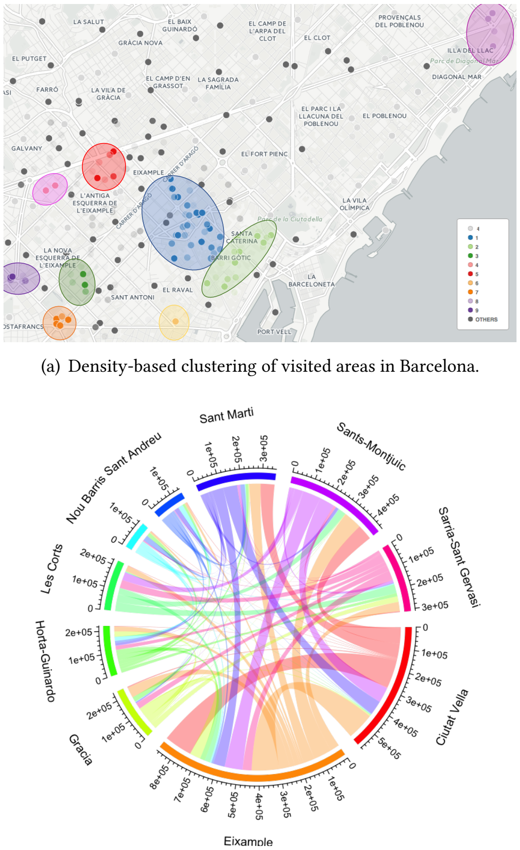
[6] C. de Fabritiis, R. Ragona, and G. Valenti. 2008. Traffic Estimation And Prediction Based On Real Time Floating Car Data. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*. https://doi.org/10.1109/ITSC.2008.4732534

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu. 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise *(KDD'96)*. AAAI Press. http://dl.acm.org/citation.cfm?id=3001460.3001507

[8] P. Fiadino, D. Valerio, F. Ricciato, and K. A. Hummel. 2012. *Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data*. Springer Berlin Heidelberg, Berlin, Heidelberg, 66–80.

[9] M. C. González, C. Hidalgo, and A. Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453 (June 2008), 779–782. https://doi.org/10.1038/nature06958 arXiv:physics.soc-ph/0806.1256

[10] Matthew Iji. 2017. GSMA Intelligence - Unique mobile subscribers to surpass 5 billion this year. https://www.gsmaintelligence.com/research/2017/02/unique-mobile-subscribers-to-surpass-5-billion-this-year/613. (2017). Accessed: 2017-05-24.



(a) Density-based clustering of visited areas in Barcelona.

(b) Transitions among Barcelona districts.

Figure 5: Results of urban mobility studies based on a sample of Highly Active Users (HAU).

[11] A. Janecek, D. Valerio, K. Hummel, F. Ricciato, and H. Hlavacs. 2015. The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. *IEEE Transactions on Intelligent Transportation Systems* (2015). https://doi.org/10.1109/TITS.2015.2413215

[12] S. Jiang, J. Ferreira, and M. C. Gonzalez. 2017. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Trans. on BigData* (2017). https://doi.org/10.1109/TBDATA.2016.2631141

[13] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira, Jr., E. Frazzoli, and M. C. González. 2013. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities *(UrbComp '13)*. ACM, New York, USA. https://doi.org/10.1145/2505821.2505828

[14] G. Ranjan, H. Zang, Z. Zhang, and J. Bolot. 2012. Are Call Detail Records Biased for Sampling Human Mobility? *SIGMOBILE Mob. Comput. Commun. Rev.* 16, 3 (Dec. 2012). https://doi.org/10.1145/2412096.2412101

[15] F. Ricciato. 2006. Traffic monitoring and analysis for the optimization of a 3G network. *IEEE Wireless Communications* 13, 6 (Dec 2006), 42–49. https://doi.org/10.1109/MWC.2006.275197

[16] F. Ricciato, P. Widhalm, F. Pantisano, and M. Craglia. 2017. Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing* 35 (2017), 65 – 82. https://doi.org/10.1016/j.pmcj.2016.04.009

[17] C. Song, Z. Qu, N. Blumm, and A. L. Barabási. 2010. Limits of Predictability in Human Mobility. *Science* (2010). https://doi.org/10.1126/science.1177170 arXiv:http://science.sciencemag.org/content/327/5968/1018.full.pdf

[18] J. Yuan, Y. Zheng, X. Xie, and G. Sun. 2011. Driving with Knowledge from the Physical World. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 316–324. https://doi.org/10.1145/2020408.2020462