RC25533 (IRE1504-011) April 24, 2015
Computer Science

IBM Research Report

Comparing Urban Sensing Applications Using Event and Network-Driven Mobile Phone Location Data

Fabio Pinelli, Giusy Di Lorenzo, Francesco Calabrese
IBM Research
Smarter Cities Technology Centre
Mulhuddart
Dublin 15, Ireland

Research Division
Almaden – Austin – Beijing – Cambridge – Dublin - Haifa – India – Melbourne - T.J. Watson – Tokyo - Zurich

Comparing urban sensing applications using event and network-driven mobile phone location data

Fabio Pinelli, Giusy Di Lorenzo, Francesco Calabrese
IBM Research – Ireland
Email: {fabiopin, giusydil, fcalabre}@ie.ibm.com

***Abstract*—In this paper we address the use of mobile phone location data to build urban sensing applications. In the past decade, several research works have proposed the use of different types of location data from the telecommunication network to characterise people mobility in the city. Thus, several applications to infer urban dynamics where proposed. However, different papers have used different types of mobile phone location data, making it is difficult to understand whether a particular dataset provided by a telecom operator is indeed effective for a specific urban sensing application. In this paper we address this issue by comparing the quality of the insights extracted from different types of mobile phone location data, with specific reference to two urban sensing applications: people count by location, and people flow between locations. Experiments executed on a real dataset provided by a telecom operator in Belgium show the advantages of using network-driven mobile phone location data (collected regardless on whether people are using their phone) compared to the widely used Call Detail Records.**

I. Introduction

Mobile phone location data from telecom operators in the form of Call Detail Record (CDR) has been widely studied, especially to extract insights into urban dynamics [5]. Such massive data can be useful to extract patterns of human mobility at an incredible scale. This data allows sampling the location of a mobile device every time the device is actively interacting with the network, e.g. at call time, while sending an SMS, or while connecting to the Internet with Smartphones. The disadvantage of such data collection method is that the spatio-temporal sampling of each individual mobile phone user trajectory over time might be very uneven, and perhaps biased to specific locations (e.g. home locations) or times (e.g. during the evenings or during working hours). Moreover, different users might interact more or less with the network, resulting in more or less mobility information recorded from them. This could result in under-sampling the population, or more problematically, biasing the extracted insights.

We then ask the question whether insights extracted from actively collected-mobile phone location data are a good proxy for human mobility. To answer this question, we compare such results, with results extracted by both actively and passively sampling user location, which constitute a richer set of location information.

We used a real dataset collected from a telecom operator in Belgium, which had a system which allowed to collect both CDR, records of Internet connections (which we call IPDR), and passively generated data (which we call Signaling). Since each location event was tagged with the specific type of event generating it, we were able to decompose the dataset in three different ones: only CDR, CDR + IPDR, and all data. We specifically take as reference a set of urban sensing applications which have been proposed in the past, and compare the patterns extracted from both datasets, to evaluate the limitations of active-only user sampling.

The paper is structured as follows: Section II reports related work in the area of urban sensing using mobile phone location data. Section III describes the process under which mobile phone location data is generated in telecommunication network. Section IV describes the data used in this paper for the evaluation. Section V describes the results of an application-independent comparison of the datasets. Section VI evaluates the insights extracted from the different datasets, considering specific urban sensing applications. Section VII concludes the paper with a discussion on the provided comparison, and draws conclusions.

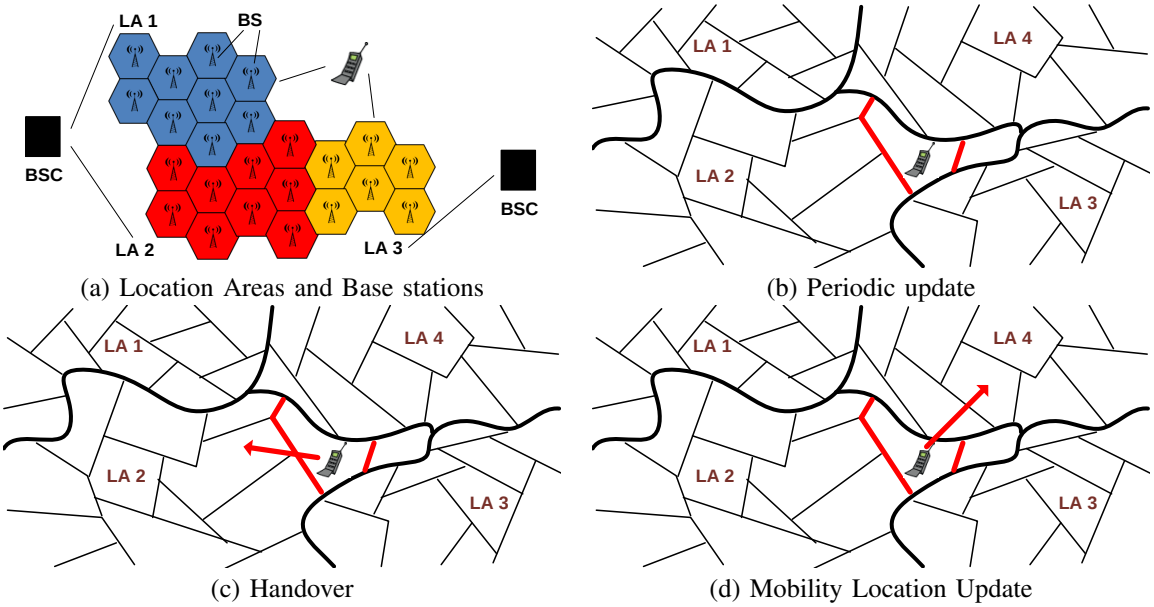
II. Related Work

In the past decade, there has been a rising interest in using mobile phone location data to infer user trajectories [1], [8], and to study human mobility and their patterns [7]. Different types of data have been used in these studies. CDR data were used in [7],[2],[6]. CDR information, enriched with records from Internet access was exploited in [4]. Data from idle phones were also used in [9] to estimate the road traffic. However, to the best of our knowledge, no work so far has specifically compared the different types of datasets that a telecom operator can collect. Moreover, no work so far has analysed the limitations of using a specific dataset for a given urban sensing application.

III. Mobile Phone Network Data Generation

In this section we describe different types of location information that can be collected by a telecom operator related to interaction of the mobile phones and the telecommunication network. When a mobile phone is switched on, it regularly notifies its position reporting the actual cell where it is currently located. The notification of the mobile phone position can be triggered by *events* (call, sms, or Internet usage) or by updates of the *network* (for a more detailed description of the technologies and standards used to derive the position of mobile phones see [11]).

Event-Driven Mobile Phone Network Data. Today, there are two primary sources of these data: communication and Internet usage. Most telephone networks generate Call Detail Record (CDR): records produced by a telephone exchange documenting the details of a phone call or sms passed through



(a) Location Areas and Base stations (b) Periodic update

(c) Handover (d) Mobility Location Update

Fig. 1. (a) Location area and base stations; (b) Periodic update; (c) Handover; (d) Mobility Location Update.

the device. A CDR is composed of data fields that describe the telecommunication transaction such as the user id of the subscriber originating the transaction, the user id receiving the transaction, the transaction duration (for calls), the transaction type (voice or sms), etc. Each telecommunication operator decides which information is emitted and how it is formatted. As an example, there could be the timestamp of the end of the call instead of the duration.

The second source of data is Internet usage. In telecommunications, an IP Detail Record (IPDR) provides information about Internet Protocol (IP)-based service usage and other activities. The content of the IPDR is determined by the service provider, the Network/Service Element vendor, or any other community of users with authority for specifying the particulars of IP-based services in a given context. Examples of IPDR data fields are: user id, type of the website, time of event, number of bytes transmitted, etc. It is important to note that the margin of error in this case varies widely according to whether the device to which the IP address is attached is mobile, and to the density and topology of the underlying IP network.

Both communication and Internet usage can be associated to the cell phone towers used during the interaction.

Network-Driven Mobile Phone Network Data. A cellular network is a radio network of individual cells, known as base stations. Each base station covers a small geographical area which is part of a uniquely identified location area. By integrating the coverage of each of these base stations, a cellular network provides a radio coverage over a much wider area. A group of base stations is named a Location Area (LA), or a routing area. A LA is a set of base stations that are grouped together to optimise signaling (see Figure 1(a)).

Typically, tens or even hundreds of base stations share a single Base Station Controller (BSC). The BSC handles allocation of radio channels, receives measurements from the mobile phones, controls handovers from base station to base station.

In such a context, different types of location update can happen:

1) **Periodic Update**, which is generated on a periodic base and provides information on which cell tower the phone is connected to (see Figure 1 (b)).
2) **Handover**, which is generated when a phone involved in a call moves between two cell areas (see Figure 1 (c)).
3) **Mobility location update**, which is generated when the phone moves between two Location Areas (see Figure 1 (d)).

Location updates also happen when the phone changes type of connectivity it uses to access the telecommunication infrastructure (e.g., from 2G to 3G). Finally, operators might install systems to monitor the signaling messages from the links between the cellular Radio Access Network and Core Network (specifically on the A, Gb, IuPS and IuCS interfaces). The frequency of these updates strongly depend on how the operator has deployed the different connectivity technologies.

Another important aspect is how the user's location can be detected. Location information can be extracted as part of the interaction data between the mobile phone and the telecommunication infrastructure. In most cases it is represented by the cell tower position or the cell sector to which the mobile phone is connected.

An operator might decide to record some of the above information for further uses. This might involve installing additional hardware and software to be able to connect to the data streams, and of course specific storage capacity. Usually data is only stored for a limited time. Thus, it is possible that in a real setting only one type of information can be made available for use for urban sensing applications. In this paper, we have been able to get access to all above described location data, and thus we can compare the quality of the insights

extracted by any combination of the datasets. This study can serve as a basis for choosing which investments and effort an operator has to put in place to collect specific data used to provide effective urban sensing applications.

IV. Available Data

In this paper we used anonymised mobile phone location data from a telecom operator in Belgium, for users in the area around the city of Mons, Belgium. To safeguard personal privacy, individual phone numbers were anonymised by the operator before leaving storage facilities. In particular, each data item is of the form: $< id, timestamp, cellid, type >$, where id is a user identifier, and the $type$ field allows to specify the reason for the location data. In particular, such reason can be:

- Event-driven signaling due to:
  - Callsetup, generated at the beginning of a call (either originated or terminated);
  - SMS, generated at time of SMS message being sent;
  - Internet data packets (IPDR), generated for internet traffic;
- Network-driven signaling due to:
  - location updates;
  - radio access network;
  - data sessions.

The data covers users connected to 150 distinct cell towers in the city area. For each cell tower, we were given the coordinate and the azimuth of each cell sector. Thus we were able to derive a voronoi tessellation of the space, following the approach presented in [3]. Some cells cover the same area (i.e. 2G and 3G antennas installed on the same tower), resulting in being able to discriminate among 58 distinct locations in the city.

The available data covers one week in October 2014. We use the available data to simulate 3 different scenarios:

- availability of only CDR information, in which we only use the CALL and SMS data items, and are representing cases in which only CDR information is provided.
- availability of CDR and IPDR, in which we use the above data, together with IPDR, to represent cases where all Event-driven signaling information is provided.
- availability of all signaling information (CRD+IPDR+Signaling), which will be our reference for comparison.

Figure 2 depicts an illustrative example of temporal sequence of events for a user in the dataset. We also depict the trajectory that we are able to detect, given the tree different scenarios. The example clearly shows that for this user, the availability of all information allows detecting 3 different visited locations, and an estimated stop time for locations 2 and 3. If no Signaling information is available, only two visited places could be detected, and the estimated stop time would also be reduced, with lowest accuracy in the case of only CDR information available.

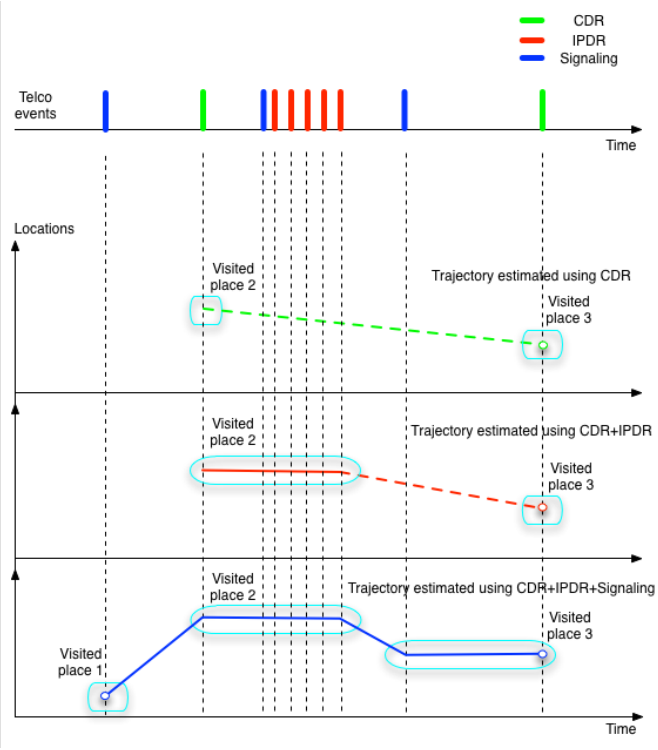


Fig. 2. Example of temporal sequence of events (top), and estimated trajectories using three different datasets (bottom). To simplify the reading, the locations have been drawn in one dimension (as opposed to the two dimensions (latitude and longitude).

At the general level, the advantages of using Network-driven location data (in addition to event-driven) include: i) sampling more users (people who are not making calls/SMS/Internet connections); and ii) having more samples of user locations, particularly at times where users are not too active, e.g. at night). Motivated by this example, in the following sections we quantitatively and qualitatively analyse the difference of the different datasets from the point of view of extracting accurate trajectories. This is firstly done by extracting application-independent characteristics. Then, we selected frequently used urban sensing applications designed to make use of mobile phone location data, and compared the accuracy of the extracted insights among the different datasets, highlighting in which cases one dataset is preferable compared to the others.

V. Application-independent comparison

We can compare the three datasets along different dimensions: number of sampled users, number and timing of events, spatial dispersal, and finally we can try to classify users based on the available data.

A. Sets of users

Let us define as $\#CDR_i$, $\#IPDR_i$ and $\#Sign_i$, the number of CDR, IPDR and Signalling events for user i. A first comparison between the three datasets is in terms of the

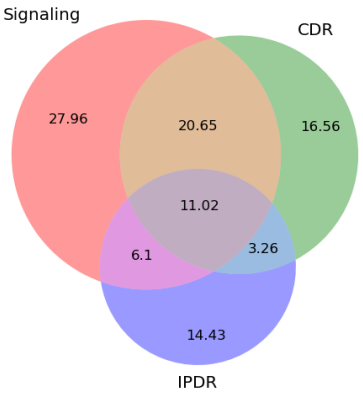


Fig. 3. Percentage of users per data type and relative intersections.

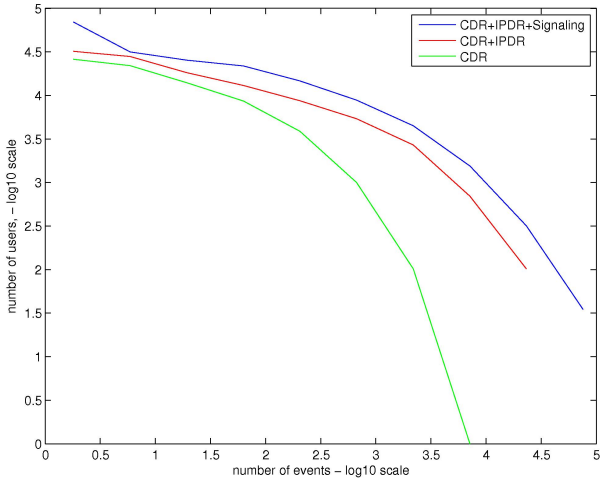


Fig. 4. Distribution of user by number of events

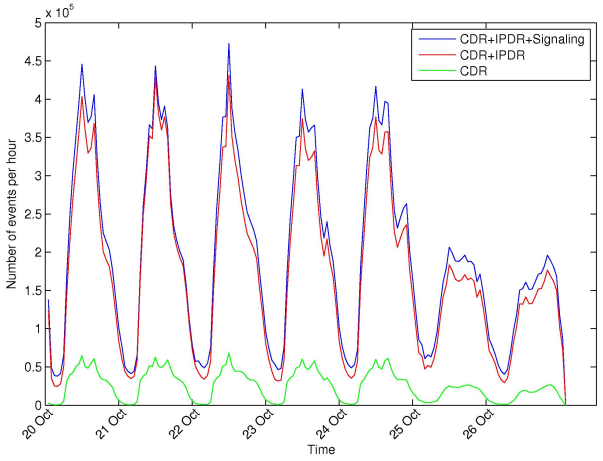


Fig. 5. Number of events per hour

set of sampled users. We counted, for each user, the number of the three different types of events (CDR, IPDR and Signaling). Figure 3 shows a Venn diagram of the unique users by data type. Only for about 11% of the users we can see all three different types of events. This is due to several reasons:

- Not all users have smartphones for which IPDR can be generated;
- Some users are only seen very temporary in the dataset (users only traversing the city), and so only Signaling information is available;
- We can also have users for which only CDR information is generated (without any Signaling information). This can be explained by people spending only a limited time in the area under analysis, and for which no location updates happened in that period.

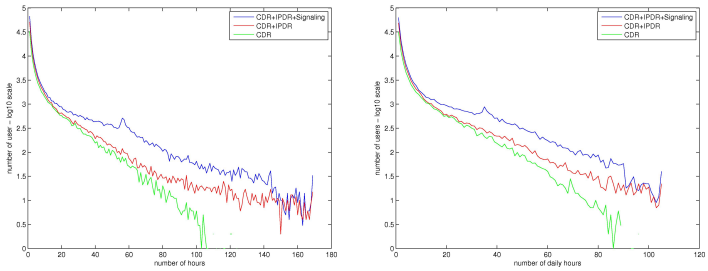
B. Number and timing of events

Not all users generate the same number of events. The number of events by user follows a long tail distribution, as shown in Figure 4. Clearly considering only CDR or CDR+IPDR events, the average number of events per user is smaller. This is even more clear if we look at the total number of events per hour, see Figure 5. CDR events represent around 10% of all events, signaling about 2%. Clearly, majority of events are due to IPDR, given the number of packages being downloaded or uploaded for each internet connection.

We then ask the question whether this decrease in the number of events is concentrated in particular hours of the day, or is equally spread over time. Figure 6(a) show the distribution of users by number of distinct hours for which there is at least one event. This is computed, user by user, by counting the number of distinct hourly intervals (from 00 to 59) in which we have at least one record for that user. Curves CDR and CDR+IPDR look very close, to show that the large amount of IPDR events are on average concentrated in the same number of hours as the CDR events. Moreover, majority of Internet usage tends to be bursty and associated to people' use of mobile apps for a limited amount of time. On the contrary, Signaling events are able to sample the user location over many more hours. However, if we only consider daily hours (from 6 to 22), as reported in Figure 6(b) we notice that the difference in terms of number of monitored hours decreases. This behaviour indicates that the probability to be able to locate a user in space during daily hours is relatively similar for the 3 different datasets. This is very important for many urban sensing applications, as we will see i the following section.

C. Spatial dispersal

We analysed the distribution of users by number of distinct visited locations. Figure 7 shows that majority of users visits less then 6 locations in the week. There is no much change between the different datasets, showing the all datasets are able to detect the most visited locations for each user, which usually are also the ones that characterise most of the user's mobility (e.g. home and work locations).



(a) all hours (b) daily hours

Fig. 6. Distribution of user by number of hours for which at least one event

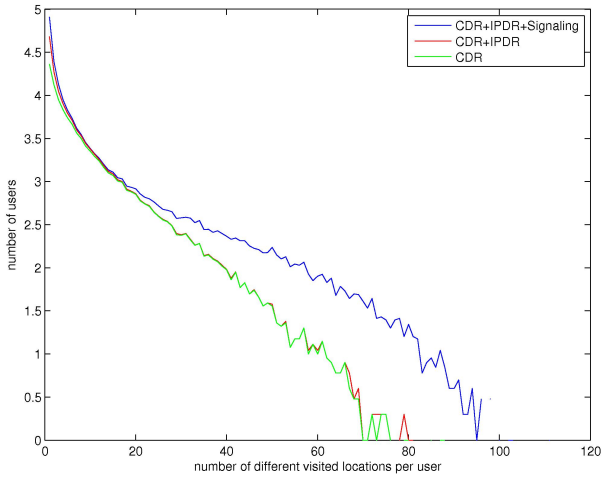


Fig. 7. Distribution of users by number of distinct visited cells

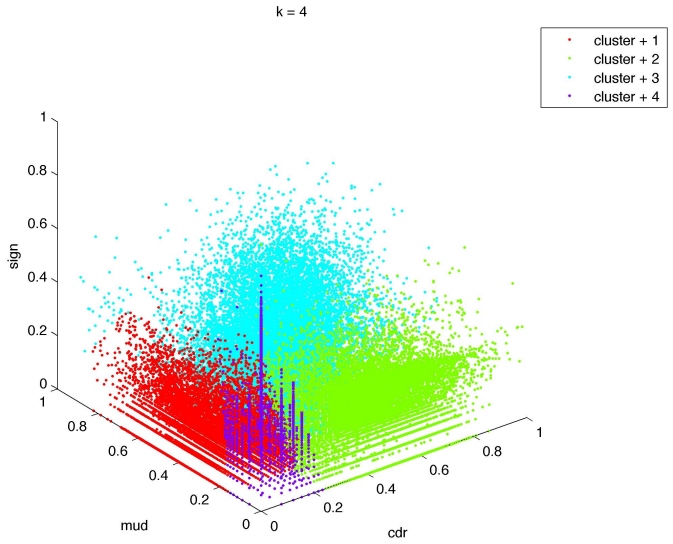
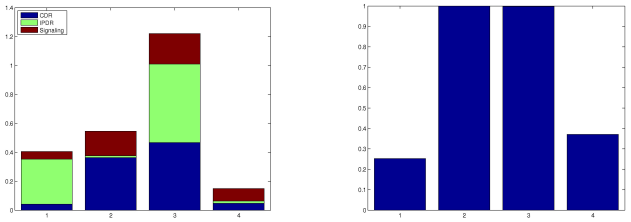


Fig. 8. Cluster of users based on relative number of events



(a) Cluster centroids (b) Percentage of CDR-only users

Fig. 9. Clustering results

D. Classes of users

We cluster users by the number of different types of events. Figure 8 shows 4 clusters extracted running K-means on the set of users[1], given the feature vector for each user i:

$$\left(\frac{\#CDR_i}{\max_i \#CDR_i}, \frac{\#IPDR_i}{\max_i \#IPDR_i)}, \frac{\#Sign_i}{\max_i \#Sign_i}\right)$$

The representatives of each cluster are shown in Figure 9(a). Cluster 1 corresponds to very low interacting users. Cluster 2 corresponds to users mainly using their phones for calls and sms. Cluster 3 corresponds to users mainly using their phones for Internet access. Cluster 4 corresponds to highly interacting users, for which both the volume of call, sms and Internet connection is high. In Figure 9(b), we reported the percentage of users having only CDR in the 4 clusters. We can see that Clusters 2 and 4 are well represented in the CDR dataset, while the other two classes of users are only partially represented. This shows that using only CDR information could result in partially biasing the results toward some specific classes of users.

VI. Application-dependent comparison

In this section, we have taken frequently used examples of urban sensing applications using mobile phone location data: i) the count estimation over time, such as the number of people being in a certain location in a given time interval; ii) the Origin Destination (OD) flow estimation, such as the number of people travelling from a certain origin to a destination in a given time interval. We extracted this two kinds of information from the 3 datasets and we compared the obtained results.

We provide an extensive series of results considering all the users, and a summary of the results considering only the 11% of users having all three CDR, IPDR and Signaling information recorded.

A. Count estimation

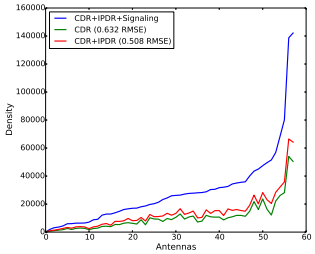
This application involves the estimation of number of users by location, as a time series. This information is highly relevant for many sectors, such as Retail, Property, Leisure and Media, since it allows to compare locations in terms of expected crowd. Clearly, an accurate estimation of the time series of number of users by location is crucial to provide trustable insights. Given a user locations dataset, we apply the following method to estimate the number of users by location:

- select a time interval (e.g. 1 hour);
- find, for each user, the location at which it has been seen for most of the time;
- assign to each location and time interval, the count of users based on the step above;

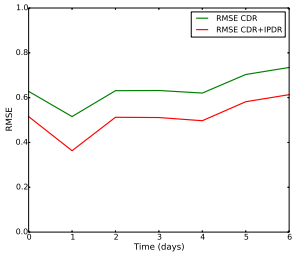
[1] k=4 was chosen based on maximising the average silhouette of the clusters

	CDR+IPDR+Signaling		CDR		CDR+IPDR	
Antenna ID	Rank	Density	Rank	Density	Rank	Density
7934	1	20742	1	10867	1	13546
7932	2	20636	2	9488	2	12413
8736	3	10838	3	5185	3	6754
8000	4	9412	4	4725	**5**	5950
7933	5	8555	5	4538	**4**	5934
19976	6	7744	**7**	3610	6	5561
8256	7	7183	**16**	2104	**10**	3570
8001	8	6663	**9**	3146	**9**	4083
8032	9	5979	**11**	2488	**12**	3260
8034	10	5954	**6**	3808	**7**	4642

TABLE I. RANK AND COUNT ESTIMATION OF TOP 10 ANTENNAS IN SIGNALING ON THE THREE DATASETS.



(a) by antenna



(b) error by day of week

Fig. 10. Density estimation

We computed user count time series for each location in the city covered by a cell tower, starting from the three different datasets. Cumulative count estimation by location is show in Figure 10(a), ranked by increasing value of count (based on the reference dataset). If results using the different datasets were similar, we would expect to see a non decreasing curve for the estimated count using either CDR or CDR+IPDR. In order to compare the different count estimations, we computed two measures of error:

- Root mean square error, calculated as

$$\sqrt{\sum_{i=1}^{n_{loc}} \sum_{t=1}^{n_{times}} \left(\frac{count_R(t,i) - count_c(t,i)}{count_R(t,i)} \right)^2}$$

  where $count_R$ is the reference count estimated using all data (CDR+IPDR+Signaling), while $count_c$ is the count computed using either CDR or CDR+IPDR. The error ranges from 0 to 1, and low values correspond to low error.

- Normalised discounted cumulative gain (nDCG) [10] which is used in recommender systems to measure rank quality. We have chosen this measure to evaluate whether the estimated ranking of crowded locations is kept the same by using CDR or CDR+IPDR only information. The error ranges from 0 to 1, and high values correspond to low error. This measure is different from the RMSE, since it does not take into account the absolute estimated count for each location or time, but just the relative ordering of such counts by location. This measure is directly useful in application scenarios such as choosing the most crowded place between a set of locations.

An example of the results of the estimated counts and the ranking for a set of antennas is reported on Table I. We selected the top 10 antennas in terms of count estimation for the Signaling dataset and reported the corresponding values and ranks obtained using the other 2 datasets. As we can see, there are big differences among the datasets if we consider the absolute values and this phenomenon is measured through the RMSE. Instead, the differences in terms of rank are smaller, and this is measured with the nDCG computation.

Table II shows the average errors for the 2 datasets. As expectable, the error is higher if we only consider CDR information. Moreover error measured in terms of nDCG is much lower (since the value is very close to 1), and there is no much difference in using CDR or CDR+IPDR data.

	CDR		CDR+IPDR	
	RMSE	nDCG	RMSE	nDCG
Count estimation	0.51	0.994	0.34	0.995
Out flow estimation by antenna	0.42	0.993	0.28	0.995
O/D flow estimation by pair	0.50	0.997	0.39	0.998

TABLE II. RMSE AND NDCG OBTAINED ON DIFFERENT ANALYTICS WITH CDR AND CDR+IPDR DATASETS.

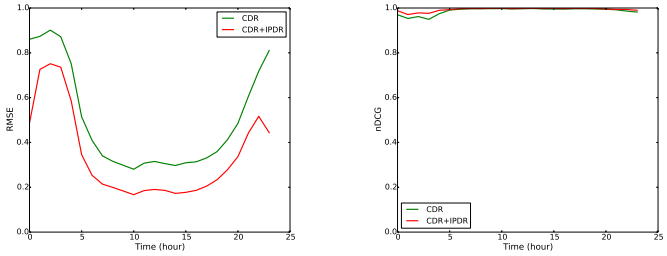
This shows, that for the purpose of comparing user counts by locations, CDR information is on average a good source of information. Figure 10(b) shows the average RMSE computed on the 7 distinct days. As it can be seen, the error is higher over the weekend (last 2 days). Moreover, Figure 11 shows the errors as function of the hour of day (averaged over all days). It's interesting to see that error is large over the night hours, and instead quite low during the daylight hours.

In conclusions, we can observe that using CDR or CDR+IPDR as proxy for user count per location works relatively well in application scenarios when preserving the ranking is important, such as choosing the most crowded place between a set of locations, especially if the focus is on daylight hours over weekdays.

B. Flow estimation

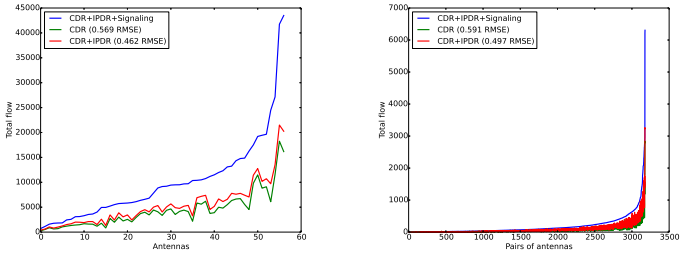
Origin Destination (OD) matrices are a widely used information for urban planning and development, specifically in the transportation community. Generally, this information is estimated using census information and/or travel surveys. However, recent work has used Mobile phone location data to estimate such information [2, 4, 6]. In this section we want to evaluate whether OD matrices extracted from Event-driven information are a good proxy of OD matrices. For the comparison, we used the method presented in [4] to compute the OD matrices starting from a given set of location data. Even if starting from the same set of users, we can already see that the total OD flow volume in the city is different. Figure 12 shows such volumes by time of day. While the total flow counts by hours are different, the temporal trends seem to be maintained. To perform a better comparison, we evaluated the accuracy of estimated OD matrices at different spatio-temporal granularity.

1) Antenna-based: In some application scenarios (e.g. estimating the number of visitors of a certain location), we are interested in the cumulative incoming or outgoing flow from a given location. In order to measure the error in this case, we ranked antennas by increasing flow volume (based on the reference dataset) and we plotted the corresponding volumes



(a) RMSE (b) nDCG

Fig. 11. Density estimation error by hour of day



(a) by antenna (b) by pair

Fig. 13. OD flow, ranked by increasing value (based on ALL data)

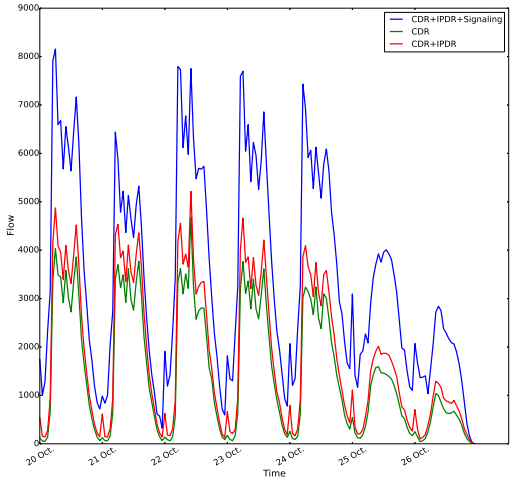
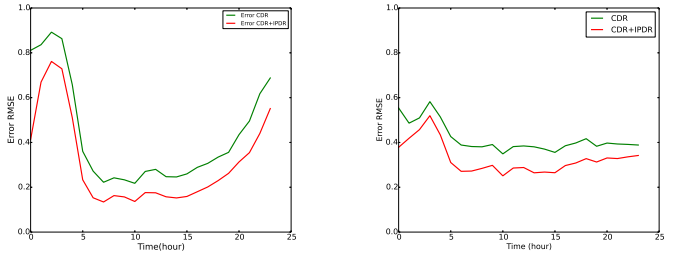


Fig. 12. OD flow over time, by data type. Time goes from 00:00 of October 20th, to 23:59 of October 26th.



(a) by antenna (b) by pair

Fig. 14. Error of OD flow by hour of day

computed using the other data, see Figure 13(a). If results using the different datasets were similar, we would expect to see a non decreasing curve for the estimated OD. However, we see some fluctuations, which are more explicit for the OD extracted solely on CDR data.

Figures 14(a) and 15(a) shows the relative error measurement, for different hours of day, and days of the week respectively. As also seen in the count estimation, the error is lower over weekday daylight hours.

2) Pair-based: In application scenarios in which we are interested in specific OD pair flows (e.g. in transportation planning where we would like to estimate the number of travellers between two locations in order to design a new road or transit service), we are interested in flows from two specific locations. In order to measure the error in this case, we ranked the OD flow by pair, based on order using the reference dataset, see Figure 13(b). Unlike Figure 13(a), the fluctuations are much more evident, showing a much higher error on a pair-by-pair basis. Figures 14(b) and 15(b) shows the relative error measurement, for different hours of day, and days of the week respectively. As opposite to what seen in the antenna based error estimation, the error is lower over weekends and nights. This is due to the fact that over these periods, majority of OD pairs experience no flow (using any of the datasets) and this brings down the average error. However, if we consider weekday daylight hours, the error is quite high, showing that OD flow extracted from CDR or CDR+IPDR are not a good proxy for OD flow extracted using the reference dataset.

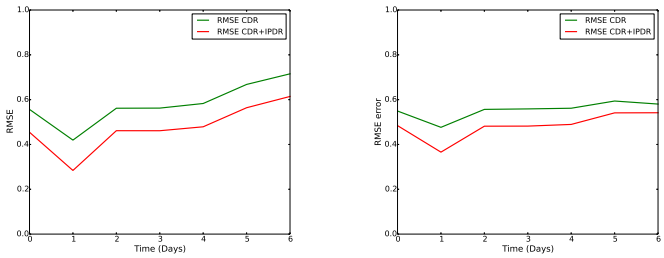
We perform the same comparisons among the 3 datasets only for the users that present at least one entry on all of them. Therefore, we select only the 11% of users being the intersection presented in Figure 3. For the sake of readability, we report in Table III only the overall results on the different analytics. As expected the RMSE decreases since the differences between the datasets, in terms of number of observed users, are smaller, as it is possible to observe in Figure 16 for the case of OD flow. Instead the nDCG presents very similar results.

	CDR		CDR+IPDR	
	RMSE	nDCG	RMSE	nDCG
Count estimation	0.63	0.993	0.50	0.995
Out flow estimation by antenna	0.57	0.992	0.46	0.994
O/D flow estimation by pair	0.59	0.997	0.49	0.998

TABLE III. RMSE AND NDCG OBTAINED ON DIFFERENT ANALYTICS WITH CDR AND CDR+IPDR DATASETS CONSIDERING ONLY USERS WITH AT LEAST ONE ENTRY FOR EACH DATASET.

VII. CONCLUSION AND FUTURE WORKS

In this paper we have compared different types of mobile phone location data, with respect to different urban sensing applications. The goal was to evaluate whether CDR information alone, which is collected based on user-generated events, would be sufficient for specific urban sensing applications like user count estimation and flow estimation. The results of the comparison on real mobile phone location data show the opportunities and limitations of using event-driven location



(a) by antenna (b) by pair

Fig. 15. Error of OD flow by day of week (from Monday to Sunday)

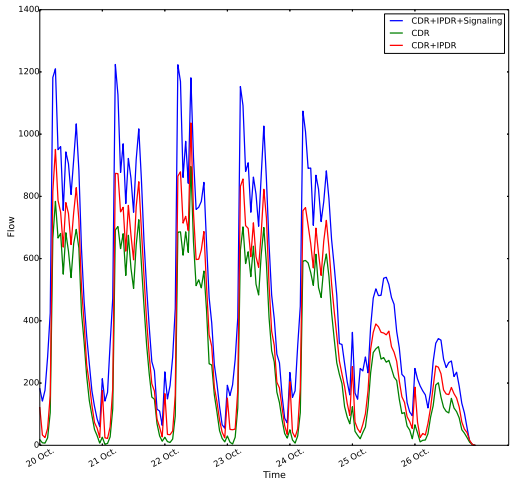


Fig. 16. OD flow over time, by data type. Time goes from 00:00 of October 20th, to 23:59 of October 26th.

information as opposed to network-driven and more frequently updated location information.

While some of the reported results might depend on the specific mobile phone usage of a particular country or region, and on the configuration of the monitoring system that the telecom operator used, the methodology we presented to evaluate the limitation of each dataset are general, and can be applied for other telecom operators to measure the effectiveness of using each individual dataset for urban sensing applications. Moreover, we are planning to extend the analysis to others, and more complex, urban sensing applications, such as event detection, trajectory pattern extraction and trajectory clustering.

ACKNOWLEDGMENT

The authors would like to thank Mobistar for providing access to the anonymized data.

REFERENCES

[1] T. Bao, H. Cao, Q. Yang, E. Chen, and J. Tian. Mining significant places from cell id trajectories: A geo-grid based approach. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 288–293, July 2012.

[2] N. Caceres, J. Wideberg, and F. Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15 –26, 2007.

[3] R. Caceres, J. Rowland, C. Small, and S. Urbanek. Exploring the use of urban greenspace through cellular network activity. In *Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)*, 2012.

[4] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE*, 10(4):36 –44, april 2011.

[5] F. Calabrese, L. Ferrari, and V. D. Blondel. Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20, Nov. 2014.

[6] G. di lorenzo, M. L. Sbodio, F. Calabrese, M. Berlingerio, R. Nair, and F. Pinelli. Allaboard: Visual exploration of cellphone mobility data to optimise public transport. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 335–340, New York, NY, USA, 2014. ACM.

[7] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[8] S. Hoteit, S. Secci, S. Sobolevsky, G. Pujolle, and C. Ratti. Estimating real human trajectories through mobile phone data. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 2, pages 148–153, June 2013.

[9] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs. Cellular data meet vehicular traffic theory: Location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 361–370, New York, NY, USA, 2012. ACM.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[11] S. Wang, J. Min, and B. Yi. Location based services for mobiles: Technologies and standards. In *IEEE ICC*, Beijing, 2008.