Use of Mobile Phone Data to Estimate Visitors Mobility Flows

Lorenzo Gabrielli, Barbara Furletti, Fosca Giannotti,
Mirco Nanni(✉), and Salvatore Rinzivillo

KDDLAB, ISTI CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
{lorenzo.gabrielli,barbara.furletti,fosca.giannotti,
mirco.nanni,salvatore.rinzivillo}@isti.cnr.it

Abstract. Big Data originating from the digital breadcrumbs of human activities, sensed as by-product of the technologies that we use for our daily activities, allows us to observe the individual and collective behavior of people at an unprecedented detail. Many dimensions of our social life have big data "proxies", such as the mobile calls data for mobility. In this paper we investigate to what extent data coming from mobile operators could be a support in producing reliable and timely estimates of intra-city mobility flows. The idea is to define an estimation method based on calling data to characterize the mobility habits of visitors at the level of a single municipality.

Keywords: Big data · Urban population · Inter-city mobility · Data mining

1 Introduction

Mobile phones today represent an important source of information for studying people behaviors, for environmental monitoring, transportation, social networks and business. The interest in the use of the data generated by mobile phones is growing quite fast, also thanks to the development and the spread of phones with sophisticated capabilities.

The availability of these data stimulated the research for increasingly sophisticated data mining algorithms customized for studying people habits, mobility patterns, for environmental monitoring and to identify or predict events. Some examples include the discovery of social relations studied in [16], where it has been highlighted the existence of correlations between the similarity of individuals movements and their proximity in the social network; the inference of origin-destination tables for feeding transportation models [10]; and, based on roaming GSM data (users arriving from other countries), the study of how visitors of a large touristic area use the territory, with particular emphasis on visits to attractions [11]. For data mining purposes, GSM data proved to be significant in terms of size and representativeness of the sample. In general, having information about the localization or the behavior of human or moving entities permits

C. Canal and A. Idani (Eds.): SEFM 2014 Workshops, LNCS 8938, pp. 214–226, 2015.
DOI: 10.1007/978-3-319-15201-1_14

to build support tools for applications in several domains such as healthcare, coordination of social groups, transportation and tourism.

In this work we propose and experiment an analysis process built on top of the *Sociometer*, a data mining tool for classifying users by means of their calling habits. The calling activities are used to infer the presence of the user and to construct an aggregated and compact call profile. The first prototype of the Sociometer has been developed during the project "Tourism Fluxes Observatory - Pisa", having the aim of producing a presence indicator of different categories of people in the city of Pisa [7]. The project, carried out in cooperation with the Municipality of Pisa, aimed at studying the fluxes of tourists visiting the town in order to evaluate the overall quality of the reception system on the territory, and to install a permanent monitor system. The Sociometer has been tested with positive results on real case studies both in Pisa and Cosenza [8].

In this paper we apply the Sociometer to classify the users moving in the city of Pisa. In particular, we concentrate in the urban area of the city and focus only on the sub-population of visitors – which complements previous analyses performed, mainly focused on residents and commuters, i.e. classes of users visiting the territory on a regular basis. Our objective is to produce statistics that are capable of estimating the probability of observing visitors moving across the urban area rather than arriving and staying in a limited zone. Indeed, such larger-scale visitors represent the group of people that might benefit most from an improved information about city attractions, navigation assistance and public transportation services. Therefore, it is crucial to better understand what kind of mobility (strictly localized vs. over all the city) visitors tend to follow, and in which measure.

The advantage of having defined the call profiles is that the analysis is no more based on the original GSM raw (big and privacy sensitive) data, but on an aggregated privacy-preserving summary of the original data. This allows the Telco operators to disclose only information that satisfy the required level of privacy, respecting the laws and preserving their customers. At the same way the analysts can work with data that are still meaningful. To this aim we also developed a method to measure and handle the privacy risks involved in the distribution of individual habits.

2 Related Works

The use of GSM traces for studying the mobility of users is a growing research area. An increasing number of approaches propose to use GSM data for extracting presence and/or movement patterns and users behavior. We already cited in Sect. 1 the Sociometer [6] as a method for identifying mobility behavior categories starting from call profiles, and the possibility to perform number of analysis about presences and flows of peoples in various cities [7,8]. Among the literature we can recall a famous experiments on analysing GSM data for studying people movement have been run on Rome [3] and Graz [4]. GSM data are used to realize a real-time urban monitoring systems with the aim of realizing a wide range of services for the city such as traffic monitoring and tourists movement analysis. The authors get detailed real time data by installing additional hardware on top of the existing antennas to get an improved location of the users in the networks.

A different approach comes from Schlaich et al. [14] where the authors exploit the GSM handover data - the aggregated number of users flowing between cells - to perform the reconstruction of vehicles trajectories. The objective is to study the route-choice-behavior or car drivers in order to determine the impact of traffic state.

Another use of GSM data is the identification of interesting users places as in [2], where the authors propose a method for the identification of meaningful places relative to mobile telephone users, such as home and work points. They use GSM data (both calls and handovers) collected by the phone operator. The localization precision is the cell which is the same accuracy level of the identified interesting points. They distinguish between personal anchor points like home, work and other person-related places as the locations each user visits regularly, as for example a gym.

In Pereira et al. [12], the authors exploit cellular phone signaling data[1], focusing on the prediction of travel demand for special events. Similar to the previous approach, their analysis identifies the home location: here is defined as starting point of people's trips. However, they observed that mobility data are dependent on mobile phone usage, and this may bias the results. Therefore they propose to integrate the GSM dataset with external data (e.g. ticketing statistics or taxi trips) with the aim of increasing the quantity and the quality of the data, in particular in term of spatial resolution.

Quercia et al. [13] uses GSM data for recommending social events to city dwellers. They combine the locations estimated by mobile phone data of users in the Greater Boston area and the list of social events in the same area. After extracting the trajectories and stops from GSM calls, they crawl the events from the web. Then, they divide the area of Boston in cells and locate each events and each stop in the corresponding cell. Therefore, by crossing the events and the stops, they identify a set of potential users participating to events.

Mobile phone records are analysed also in [1] where the authors propose a visual analytics framework to explore spatio-temporal data by means of SOM (Self-Organizing Map) analysis. They propose a method to cluster the dataset by either of the two dimension and evaluate the resulting aggregation on the other one. Although they show the potentialities of using SOM for analysing mobile phone records, they do not focus on identifying user profiles.

All these approaches, as well others that can be found on the literature, offer different perspectives on how GSM data can be exploited to study the human mobility and the huge potentialities of these kinds of data. Differently from these approaches, the aspect we want to study in this paper are the flows across a city of the a particular category of people: the visitors.

[1] These data consist of location estimations which are generated each time when a mobile device is connected to the cellular network for calls, messages and Internet connections.

3 Objectives and Experimental Setting

The purpose of this work is to demonstrate how the massive and constantly updated information carried by mobile phone call data records (CDRs) can be exploited to estimate visitors movements within an urban area and their flows across the observed territory.

In this section, we will first describe what information CDRs contain and we will provide details about the dataset used in the experiments. Then, we will introduce the user categories and the mobility measures we aim at inferring from CDRs.

3.1 Call Detail Records (CDRs)

GSM is a network that enables the communications between mobile devices. The GSM protocol is based on a so called *cellular network architecture*, where a geographical area is covered by a number of antennas emitting a signal to be received by mobile devices. Each antenna covers an area called cell. In this way, the covered area is partitioned into a number of, possibly overlapping, cells, uniquely identified by the antenna. Cell horizontal radius varies depending on antenna height, antenna gain, population density and propagation conditions from a couple of hundred meters to several tens of kilometers.

A Call Detail Record (CDR) is a log data documenting each phone communication that the TelCo operator stores for billing purposes. The format of the CDR used in this work contains a subset of information as follows:

$$< Timestamp, Caller\_id, d, Cell\_1, Cell\_2 >$$

$Caller\_id$ is the anonymous identifier of the user that called, $Timestamp$ is the starting time of the call, d is its duration, $Cell\_1$ and $Cell\_2$ are the identifiers of the cells where the call started and ended (See Fig. 1). Only voice communications are included in the dataset.

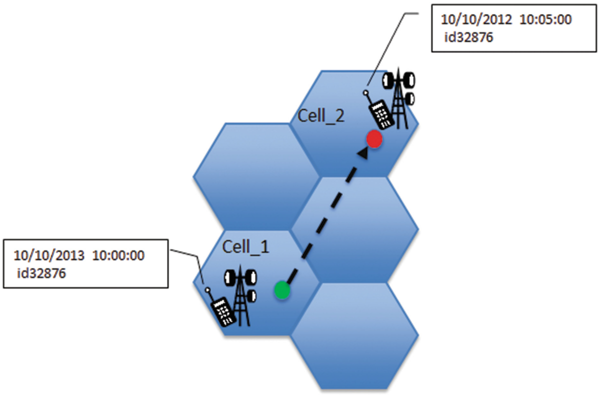


Fig. 1. Exemplification of the cellular network and communication.

The dataset used in this work consists of around 7.8 million CDRs collected from Oct 9th to Nov 9th, 2012. The dataset contains calls corresponding to about 232,200 customers of the Italian TelCo operator *Wind SpA*, with a mobile phone contract (no roaming users are included).

It is important to point out that a major limitation of CDRs is the fact that the localization of individuals occurs only during phone calls, that can lead to an incomplete view of their mobility. We discuss this point in Sect. 4, where we introduce a methodology to partially overcome the incompleteness issue.

3.2 Spatial Granularity

The spatial granularity considered in this work takes into account the spatial resolution of the cells covering the area of study. In the urban area of Pisa, the coverage of each cell is relatively large, therefore it often does not allow a precise relationship between a Point of Interest (POI) and the cell itself. This means

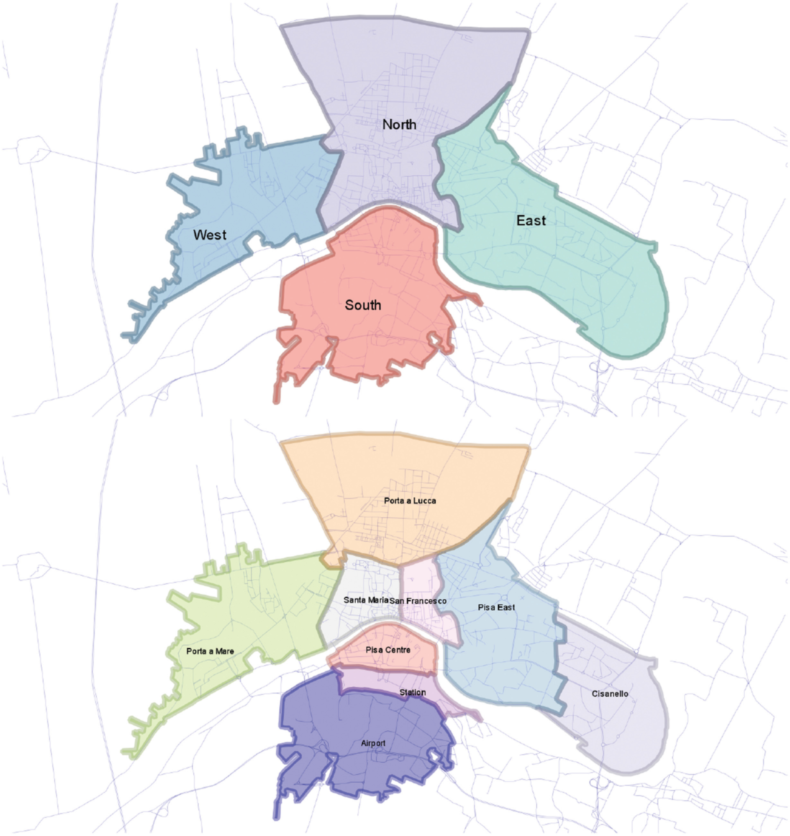


Fig. 2. City partitions adopted for the study: (Top) "Cardinal points" - Four zones (North, South, East, West); (Bottom) "City districts" - Nine areas.

that a cell may contain more than one POI and a POI, if it is large, may belong to more than one cell. Thus, in this study we use a higher level of granularity, and we define two types of partitions of the urban area: "cardinal points" (Fig. 2 top) and "city districts" (Fig. 2 bottom). In the former case the city is divided in four areas according to the cardinal points; while in the latter case the city is divided according to the major districts. Both partitions follow the natural division provided by the Arno River. To better compare the flows measured over the two partitions, each area of the first partition is defined as an aggregation of zones of the second partition.

4 Methodology

The basic idea of the methodology and at the basis of the Sociometer is that the behavior category of an individual within a specific municipality can be inferred by the temporal distribution of his/her presence in the area. For example, people commuting to a municipality for work will usually appear there only during working hours and only during working days – obviously with some exceptions, which however, are expected to be occasional. In this work we are interested in the movements of visitors, a class of users characterized by a sporadic presence on the territory, usually appearing only for a short time period (a few days). As explained in [6], a formal definition of visitors is given by The World Tourism Organization that identifies them as "people traveling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes" [15]. In other words, a person is a visitor in an area A if his/her home and work place are outside A and the presence inside the area is limited to a certain period of time T_{to} that can allow him/her to spend some activities in the city. In particular the presence has to be concentrated in a finite temporal interval inside the time window. It should also be occasional therefore, he/she does not appear anymore during the observation period. It is also important to point out the distinction that this definition includes not only the classical *tourism* as visiting cultural and natural attractions, but also the activity related to work, visiting relatives, health reasons, etc.

We already mentioned that CDR may describe the movements of users only partially, since the localization is available only when a user performs a call. For frequent callers, thus, there is a strict correspondence among movements and calls. For users that make low use of their phone, instead, sensing their movements may be underestimated. When analyzing visitors movements, it is crucial to take into account the previous observations. On one hand, the classification of a user u as visitor is based on a narrow period τ where he/she is observed performing a call. Thus, the narrower is the period τ the larger is our confidence that u is a visitor. Obviously, there is still some probability that u may be a local users that uses his/her mobile phone just very seldom, and therefore his/her calling footprint is wrongly classified as that of a visitor.

On the other hand, once we have identified the sub-population of visitors, we want to make inferences about their movements within the city. Since the period of activity of user u within the territory is limited, he/she may be able to perform very few calls, resulting in an underestimation of his/her movements.

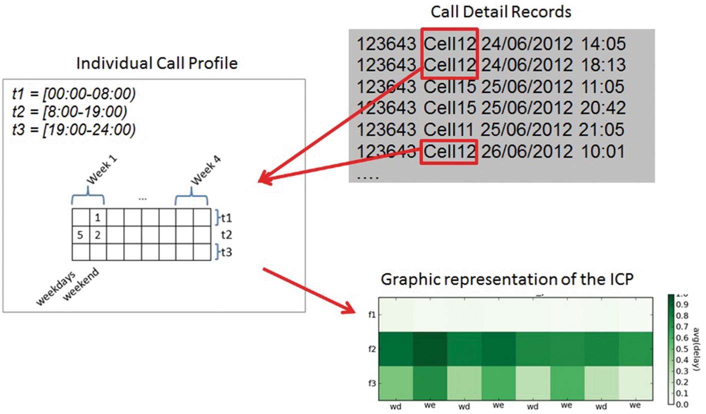
In summary, a dependable inference on visitor movements is based on the dualism between these two dimensions: the period of permanence within the area and the number of calls performed during that period. In the next sections we will show how to reason upon these two dimensions to determine the confidence about our predictions.

In the following we summarize the user classification process, at the basis of the quantitative mobility analysis proposed in this paper. The process, introduced in [6], performs a form of active transductive learning, i.e., a process that selects a sample of data to be labeled by the analyst, and exploits that sample to classify the whole dataset. After introducing the individual call profiles (ICFs) (Sect. 4.1), we will describe a semi-automatic methodology for classifying call profiles (Sect. 4.2). In this process, a human expert is asked to manually label a small number of representative call profiles, which are then used to automatically label all other call profiles. After the classification step, we associate each ICP of visitors to the corresponding sequence of CDRs, in order to reconstruct their movements. From the sequence of CDRs we determine an individual indicator stating if a user as crossed one or more city areas.

4.1 Individual Call Profiles (ICPs)

ICPs are the set of aggregated spatio-temporal profiles of an analyst computed by applying spatial and temporal rules on the raw CDRs in order to identify the presences. The resulting structure is a matrix of the type shown in Fig. 3. The temporal aggregation is by week, where each day of a given week is grouped in weekdays and weekend. Given for example a temporal window of 28 days (4 weeks), the resulting matrix has 8 columns (2 columns for each week, one for the weekdays and one for the weekend). A further temporal partitioning is applied to the daily hours. A day is divided in several timeslots, representing interesting times of the day. This partitioning adds to the matrix new rows. In the example we have 3 timeslots (t1, t2, t3) so the matrix has 3 rows. Numbers in the matrix represent the number of events (in this case the presence of the user) performed by the user in a particular period within a particular timeslot. For instance, the number 5 in Fig. 3 means that the individual was present in the area of interest for 5 distinct weekdays during Week1 in timeslot t2 only.

Figure 3 exemplifies the whole process of constructing the ICP from the raw data: starting from the dataset of the calls, the spatio-temporal aggregation rules are applied and the corresponding presences are inferred. The matrix is filled with the number of presences in each time slot. Coloring the slots based on the presence density, we get a simple representation of the profiles that give an immediate idea of the category a user belongs to. In the example the profile is of a resident because the presence is uniform in the whole windows of observation both in the weekdays and in the weekends.



Call Detail Records

```
123643 Cell12 24/06/2012 14:05
123643 Cell12 24/06/2012 18:13
123643 Cell15 25/06/2012 11:05
123643 Cell15 25/06/2012 20:42
123643 Cell11 25/06/2012 21:05
123643 Cell12 26/06/2012 10:01
....
```

Fig. 3. Example of Individual Call Profile: from the calls, the individual presence is derived

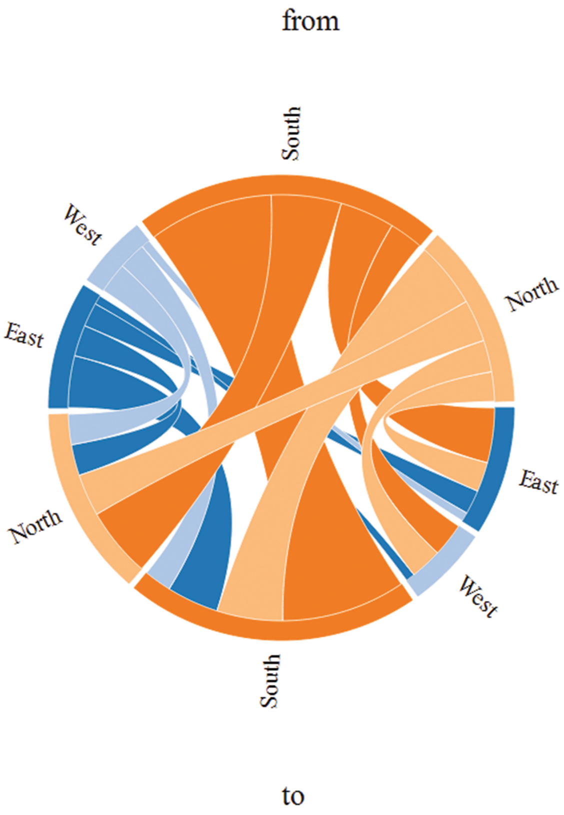


Fig. 4. Visitor flow transitions among cardinal areas according to Fig. 2 (Top).

4.2 Profile Classification

The classification method we propose is composed of two parts. First, we extract representative call profiles, i.e. a relatively small set of synthetic call profiles, each summarizing an homogeneous set of (real) ICPs. This step reduces the set of samples to be classified, which can then be handled manually by a human expert, based on the class definitions given above and his/her own experience and judgment. Finally, the labels assigned to the representative profiles are propagated to the full set of ICPs.

In the first step the standard K-means algorithm is used, which aims to partition n ICPs into k homogeneous clusters, and the mean values of the ICPs belonging to each cluster serves as prototype/representative of the cluster. The algorithm follows an iterative procedure. Initially it creates k random partitions, then, it calculates the centroid of each group, and it constructs a new partition by associating each object (ICP) to the cluster whose centroid is closest to it. Finally the centroids are recalculated for the new cluster, reiterating the procedure until the algorithm reaches a stable configuration (convergence). The similarity between two ICPs, which is the key operation of K-means, is computed through a simple Euclidean distance, i.e. comparing each pair of corresponding time slots in the two ICPs compared. Also, the centroid of a cluster is simply obtained by computing, for each time slot, the average of the corresponding values in the ICPs of the cluster. The choice of the parameter K is made by performing a wide range of experiments, trying to minimize the intra-cluster distance and maximizing the inter-cluster distance. The value chosen as most suitable was K = 100. Once extracted the representatives (RCPs), we asked the domain experts to label them. The second step, i.e. the propagation of the labels manually assigned to the RCPs, followed a standard 1-Nearest-Neighbor (1-NN) classification step. That corresponds to assign to each ICP the label of the closest RCP. Extensions of the solution can be easily achieved by adopting a K-NN classification, with $K > 1$, where the majority label is chosen.

4.3 Mobility Indicator

Our basic objective is to determine whether a user has moved across the city during the period of observation, or not. Since we are dealing with movement patterns of visitors, we associate each visitor to his/her *landing cell*, i.e. the cell where he/she initiated his/her calling activities. This cell might be the airport when the visitor arrives via plane, or the bus parking at the north of the city if he/she arrives by bus, etc.

Given the base cells of a user, we define the corresponding *Mobility Indicator* as the number of distinct areas visited by the user. Starting from the landing cell, we can also estimate the Origin-Destination Matrix of visitors within the city, since the consecutive visit of two areas imply a movement between them – though the incompleteness issue mentioned in previous sections might lead to introduce some errors, since some intermediate visits to other areas might be missed. Figure 4 shows the flows of the visitors among the cardinal areas of our partition obtained with the dataset which spans over a period of one month, as described in Sect. 3.1.

We can appreciate how the incidence of self-loops, i.e. people staying still in a region, is greater in the southern area, which contains the main transportation facilities of the city (airport, train and central bus stations) to arrive to the city. From East and West we cannot appreciate any self loop, suggesting that those routes are mainly used to cross the city.

If we consider the partition in districts (Fig. 5), it is easier to observe a transition among two adjacent districts (e.g., airport (Aeroporto) and train station (Stazione)). It is however difficult to measure large flows across distant districts.

5 Evaluation

In this section we summarize the experimental results obtained by computing some population and flow statistics over the city of Pisa. After the classification step, we identify around 90k users classified as visitors. Since our objective is to

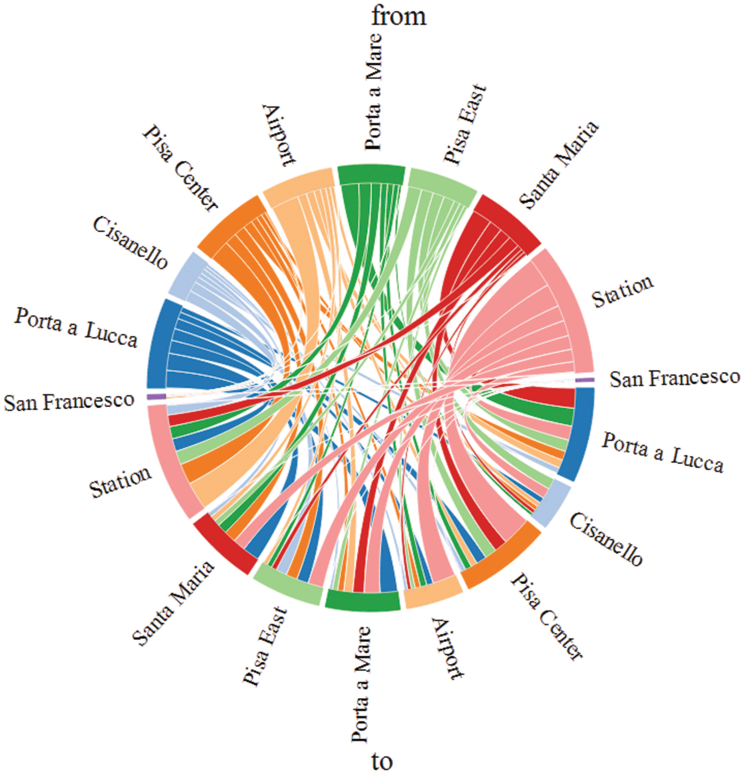


Fig. 5. Visitor flow transitions among districts according to Fig. 2 (Bottom).

determine the percentage of visitors who cross the city to visit different areas, we want to establish the percentage of users with a positive Mobility Indicator. To determine such percentage, however, we have to take into account the limitations about the dualism of precision of the classification and coverage of movement sensing. Not having the support of external evidences to determine a dependable threshold for the two dimensions, we derive a *Mobility Indicator Curve*, connecting the percentage of mobility to a *minimum support threshold* for the observed number of calls for each user.

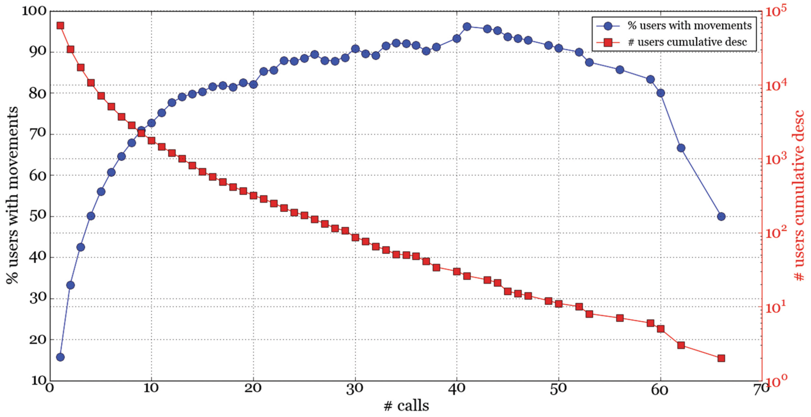


Fig. 6. Mobility Indicator Curve: relationship between the number of calls performed by each user and the probability of visiting more than one zone.

Figure 6 shows the resulting Mobility Indicator Curve for the cardinal area partition (analogous results may be observed for the other partition). If we consider all the visitors with at least one call, the percentage of mobility is very low. This is mainly due to the low duration of each call, thus preventing a user to cross too many cells. Even choosing a very permissive threshold of at least two calls, we can observe that around one third of the population moves across the areas. This percentage increases when selecting a sub-population within an higher minimum call threshold. The curve reaches relatively stable values at around 15 calls – i.e., the sensitivity of the mobility index with smaller thresholds appears to be too high, suggesting to require at least 15 calls. At the same time, the red curve shows an exponential decrement of the number of users for each threshold, thus adopting a large minimum support threshold would result in selecting just a tiny and statistically poor sample. Reasonable trade-offs, aimed at keeping at least some hundreds of users in the sample, should then not exceed 30 calls. Within this range of choices – between 15 and 30 calls as minimum threshold – we can see that the mobility index ranges between 80 % and 90 %, thus indicating that the mobility is apparently quite high.

The publication of the final results cannot put at risk the individual privacy because this information is a simple aggregation that does not contain any sensitive information about the single users. This means that an attacker, by accessing this kind of data, cannot infer any information about a user. The ICP reconstruction instead, may be more problematic for the individual privacy because requires to access the CDR data that contain all information about the user calls. However, since the only information that the analyst needs for performing the analysis is the set of ICPs, we propose an ”protocol” where, the computation of the ICPs is delegated to the TelCo operator that sends them to the analyst for the computation of the other steps. As described in [5], we supply to the TelCo operator a tool for evaluating the risk of privacy in disclosing ICPs so that it can decide supply only a subset of data that are compliant to the required level of privacy.

6 Conclusions

In this work we developed an analytical process to determine the probability of observing a population of visitors moving across an urban area. The method is based on a classification step capable of determining the class of mobile phone users by analyzing their call habits. The population of users tagged as visitors is further analyzed by reconstructing their respective movements. To overcome the limitation of partial observation for movements due to individual call habits, we introduce a methodology to relate the observations available for each user and the confidence of the prediction of observing a movement. The experimental results show that visitors have a high tendency of moving across the city, even for coarser spatial granularities.

Acknowledgments. This work has been partially funded by the European Union under the FP7-ICT Program: Project DataSim n. FP7-ICT-270833, and Project Petra n. 609042; and by the MIUR and MISE under the Industria 2015 program: Project MOTUS grating degree n.0000089 - application code MS01_00015.

References

1. Andrienko, G., Andrienko, N., Bak, P., Bremm, S., Keim, D., von Landesberger, T., Poelitz, C., Schreck, T.: A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. J. Locat. Based Serv. **4**, 3–4 (2010)
2. Ahas, R., Silm, S., Järv, S., Saluveer, E.: Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. **17**, 1 (2010)
3. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-time urban monitoring using cell phones: a case study in rome. IEEE Trans. Intell. Transp. Syst. **12**, 141–151 (2011)
4. Ratti, C., Sevtsuk, A., Huang, S., Pailer, R.: Mobile Landscapes: Graz in Real Time. MIT Senseable City Lab, Massachusetts (2005)

5. Furletti, B., Gabrielli, L., Monreale, A., Nanni, M., Pratesi, F., Rinzivillo, S., Giannotti, F., Pedreschi, D.: Assessing the privacy risk in the process of building call habit models that underlie the sociometer. Technical report. http://puma.isti.cnr.it/dfdownload.php?ident=/cnr.isti/2014-TR-011&langver=it&scelta=Metadata
6. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Identifying users profiles from mobile calls habits. In: The Proceedings of UrbComp (2012)
7. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Turism fluxes observatory: deriving mobility indicators from GSM calls habits. In: The Book of Abstracts of NetMob (2013)
8. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Analysis of GSM calls data for understanding user mobility behavior. In: The Proceedings of Big Data (2013)
9. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. **20**, 695–719 (2011)
10. Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Mede, P.V.D., Bruijn, J.D., Romph, E.D., Bruil, G.: MP4-A project: mobility planning for Africa. In: D4D Challenge @ 3rd Conference on the Analysis of Mobile Phone datasets (NetMob 2013)
11. Oltenau, A.-M., Trasarti, R., Couronne, T., Giannotti, F., Nanni, M., Smoreda, Z., Ziemlicki, C.: GSM data analysis for tourism application. In: Proceedings of 7th International Symposium on Spatial Data Quality (ISSDQ) (2011)
12. Pereira, F.C., Liu, L., Calabrese, F.: Profiling transport demand for planned special events: prediction of public home distributions (2010). www.scienceDirect.com
13. Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., Crowcroft, J.: Recommending social events from mobile phone location data. In: International Conference on Data Mining, ICDM (2010)
14. Schlaich, J., Otterstätter, T., Friedrich, M.: Generating trajectories from mobile phone data. In: The Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies (2010)
15. Wikipedia. Tourism. http://en.wikipedia.org/wiki/Tourism
16. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.-L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 11. ACM, New York (2011)