Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach

Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio

Abstract The Big Data, originating from the digital breadcrumbs of human activities, sensed as a by-product of the technologies that we use for our daily activities, let us to observe the individual and collective behavior of people at an unprecedented detail. Many dimensions of our social life have big data "proxies", as the mobile calls data for mobility. In this paper we investigate to what extent such "big data", in integration with administrative ones, could be a support in producing reliable and timely estimates of inter-city mobility. The study has been jointly developed by Istat, CNR, University of Pisa in the range of interest of the "Commssione di studio avente il compito di orientare le scelte dellIstat sul tema dei Big Data ". In an ongoing project at ISTAT, called "Persons and Places" – based on an integration of administrative data sources, it has been produced a first release of Origin Destination matrix – at municipality level – assuming that the places of residence and that of work (or study) be the terminal points of usual individual mobility for work or study. The coincidence between the city of residence and that of work (or study) – is considered as a proxy of the absence of intercity mobility for a person (we define him a static resident). The opposite case is considered as a proxy of presence of mobility (the person is a dynamic resident: commuter or embedded). As administrative data do not contain information on frequency of the mobility, the idea is to specify an estimate method, using calling data as support, to define for each municipality the stock of standing residents, embedded city users and daily city users (commuters).

Key words: Big data, urban population, inter-city mobility, data mining

Dino Pedreschi
University of Pisa, Pisa, Italy e-mail: pedre@di.unipi.it

Giuseppe Garofalo, Roberta Vivio
ISTAT, Roma, Italy e-mail: surname@istat.it

Barbara Furletti, Lorenzo Gabrielli, Fosca Giannotti, Letizia Milli, Mirco Nanni
KDDLAB ISTI CNR, Pisa, Italy e-mail: name.surname@isti.cnr.it

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 performative 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 [1], where it has been highlighted the existence of correlation between the similarity of individuals movements and their proximity in the social network; the inference of origin-destination tables for feeding transportation models [2]; 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 [3]. 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 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 so-called *Sociometer*, a data mining tool for classifying users by means of their calls habits. 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 [5]. 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 [6].

In this paper we extend the basic method to work on a larger territory and to include the flows of people between different territorial units (in this study, the municipalities), whereas the Sociometer is focused on the presence over a single area. The objective is to produce statistics that are comparable with those obtained in an ongoing project at ISTAT, called ”Persons and Places”, where residences and flows of people are studied using administrative data sources. Achieving success along this direction would mean to be able to safely integrate existing population and flow statistics with the continuously up-to-date estimates obtained from GSM data, thus a first step towards exploiting *big data* in official statistics.

2 Objectives and experimental setting

The purpose of this work is to deploy the massive and constantly updated information carried by mobile phone call data records (CDRs) for estimating population statistics related to residence and mobility. In this section, we will first describe what information CDRs contain and 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.

2.1 Call Detail Records (CDRs)

GSM (Global System for Mobile Communications) Network is a mobile 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 telcom operator stores for billing purposes. The format of the CDR used in this work is the following: $< Timestamp, Caller\_id, d, Cell\_1, Cell\_2 >$, where $Caller\_id$ 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.

The dataset used in this work consists of 7.8 million CDRs collected from Jan 9^{th} to Feb 8^{th}, 2012. The dataset contains calls corresponding to about 232,200 users with a national mobile phone contract (no roaming users are included).

We notice 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 Sec. 3, where we introduce a sophisticated methodology for handling the data and partially overcome the incompleteness issue.

2.2 User categories and O/D Matrix

The spatial granularity considered in this work is the municipality level. In particular, our study focused on the 39 municipalities in the province of Pisa, Tuscany. Municipalities host a largely variable number of residents, ranging from less than one thousand for the smaller ones, up to around 86000 for the central municipality of Pisa, with an average of ca. 10000. Each municipality is spatially covered by an average of 3-4 GSM antennas.

The first objective of this work is to correctly estimate, for each municipality, the population that belongs to each of the following categories, already calculated by ISTAT inthe ongoing project “Persons and Places” using administrative data:

- **Standing residents in A**: residents who have formal residence and place of work (study) in the same municipality A, or who do not work (study).
- **Embedded city users in A**: people that spend long periods for working (studying) in a municipality A (e.g., most days of the week), while being formally resident in another municipality, different from A.
- **Daily city users in A**: people who commute to municipality A, having formal residence in another municipality, different from A.

Each category represents a different way of living the territory and, correspondingly, a different usage of its resources.

In ISTAT project it has been produced a first release of Origin Destination matrix – at municipality level – assuming that the places of residence and that of work (or study) be the terminal points of usual individual mobility for work or study. The coincidence between the city of residence and that of work (or study) is considered as a proxy of the absence of intercity mobility for a person (we define him/her a static resident). The opposite case is considered as a proxy of presence of mobility (the person is a dynamic resident: commuter or embedded). In particular, commuter and embedded are not distinguished.

As we will discuss later, the analysis process we developed on GSM data allows to infer slightly different user categories. In particular, Standing residents and Embedded city users are not distinguished, yet, and therefore the statistics we will produce aggregate them together.

3 Methodology

The basic idea followed in this paper is that the user category of an individual within a specific municipality can be inferred by the temporal distribution of his/her presence in the area. People commuting to a municipality for work, for instance, will usually appear there only during working hours and only during working days – obviously with some exceptions, which however are expected to be occasional. On the other hand, as already mentioned, CDRs can describe the locations of users only partially, therefore the distributions of presence we can extract from them are usually an underestimation of the real ones. For this reason, after introducing the individual call profiles (Section 3.1) representing such incomplete distributions of presence, and the expected typical distributions of presence for our key user categories (Section 3.2), we will describe a semi-automatic methodology for classifying call profiles (Section 3.3). 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.

3.1 Individual Call Profiles (ICPs)

ICPs are the set of aggregated spatio-temporal profiles of a user computed by applying spatial and temporal rules on the raw CDRs. The structure is a matrix of the type shown in Fig. 1. 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. 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.1 means that the individual was present in the area of interest for 5 distinct weekdays during Week1 in timeslot t2 only.

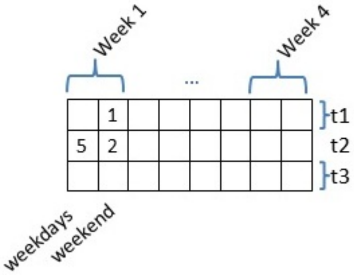


Fig. 1 Individual call profile

3.2 Profile Classes and Calling Activity Templates (CATs)

The Individual Call Profiles computed above provide a synthesis of the users' presence that makes it relatively easy to characterize some specific classes. In particular, in this study we will consider four classes, described as follows:

- **Residents** (or **Static Residents**): are those individuals that live and work in the same area, and therefore their presence is significant across all days and all time slots for a specific municipality.

- **Dynamic Residents**: people who reside in some municipality A, but work in a different one (B). The presence in A expected to be significant always, excepted during working days and working hours (time slot t2).
- **Commuters**: people that reside (Dynamic Resident) in some municipality B and whose work or study place is in A. The presence in A is expected to be almost exclusively concentrated during working days and working hours (time slot t2).
- **Visitors**: people that visit a municipality only once or a few times.

In particular, if compared to the classes introduced in Section 2.2 adopted in the ISTAT project “Persons and Places”, the lack of administrative information about the GSM users does not allow to distinguish between Standing Residents and Embedded city users, since in practice their physical presence on the residence/embedded area tends to be identical. On the other hand, the physical presence of users allows to easily distinguish (at least in principle) Dynamic vs. Static residents, since the former usually are not present in the residential municipality during working hours. This small mismatch between the two classifications will be considered in later sections, when we compare the population estimates obtained with the two methods (the one based on official data vs. the one based on GSM data).

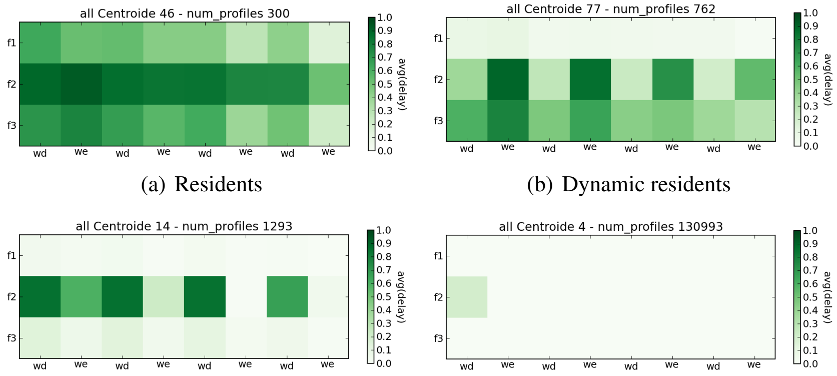
The four classes defined here can be easily translated to classification rules that, given an individual call profile (ICP), automatically assign it to the proper class. Unfortunately, the ICPs are affected by two sources of noise: (i) the expected distribution of presences for a given class does not always perfectly match real presences, although the deviations should be relatively small; (ii) ICPs provide only a sample of actual presences, since only those corresponding to (at least) one phone call are detected. The latter issue leads to significant perturbations in the distribution of presences, which cannot be easily removed by any filtering or rescaling transformation. Empirical evaluation, indeed, proved that trying to fit the ICPs to the ideal presence distributions, the result is very poor. In order to deal with the variability present in the ICPs, we developed a semi-automatic procedure, described in the next section.

3.3 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 judgement. 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 was 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 here 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 was 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), they have been labeled by domain experts in agreement with those reported in Sec. 3.2. Figure 2 reports some examples of RCPs obtained on real data, one per user category. Darker (resp. lighter) colors represent higher (resp. lower) frequencies.



(a) Residents

(b) Dynamic residents

(c) Commuters

(d) Visitors

Fig. 2 Examples of (labelled) RCPs

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 Evaluation

In this section we summarize the experimental results obtained by computing some population and flow statistics over the province of Pisa, Italy, and by comparing them with analogous estimates obtained through official data within the “Persons and Places” project at ISTAT.

The GSM data available for our study comes from a single telecom operator, therefore not covering all the phone users on the territory. For this reason, to make the GSM dataset and the ISTAT dataset comparable, in all statistics we rescaled our results taking into account the market share of the operator in each single municipality considered.

In the following subsections we evaluate, for each of the 39 municipalities of the province of Pisa, the number of Residents, Dynamic Residents and Systematic flows (i.e., those generated by commuters), and test the correlation with administrative data. The Visitors category was not considered here, since it was not possible to obtain corresponding statistics from official data, and therefore at the moment it is not possible to perform comparison on this side.

4.1 Residents

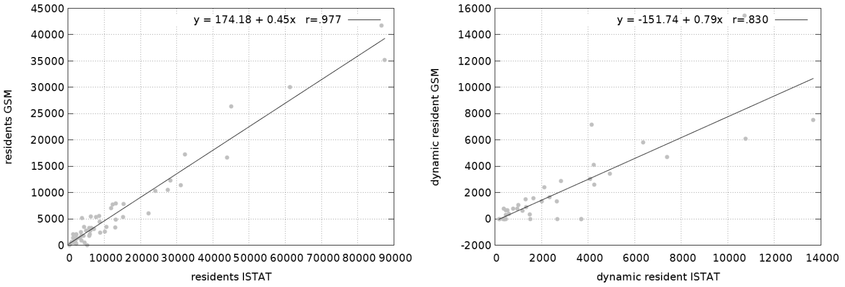
The GSM residents – Static residents – are represented by individuals who reside or live for study or work in the given municipality during the period of observation. We compare those data to persons who have a registered residence in a municipality (ISTAT).

Figure 3(a) plots, for each municipality, the corresponding values obtained through ISTAT data (horizontal axis) and through GSM data (vertical axis). The plot visually shows a clear correlation between the two variable, which is confirmed by an high R value – $R = 0.977$ ($R^2 = 0.955$). This confirms that our method provides good estimates the number of Residents in the area.

Fig.3(b) shows similar estimates for Dynamic Residents. These values are compared to those who are not active for work or study in the same municipality of residence. As we saw, instead, our GSM-based method directly associates the Dynamic Resident status to users. The plot and the R statistics for this case ($R = 0.830$, $R^2 = .689$) show that results are still relatively good, yet less accurate than the previous ones. Among the possible causes for that, we identified the actual lack of data for some small municipalities, which could be filled in future experiments. Another reason is a not complete comparability of the two classifications for the city users drafted on administrative data and on GSM data.

4.2 Daily mobility flows

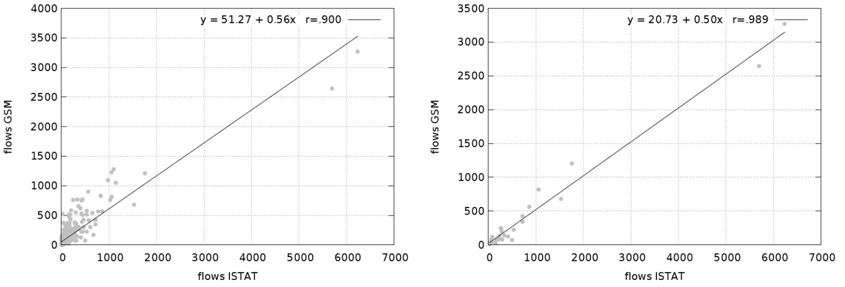
Through the processing of census data, ISTAT requires respondents the municipality of residence and work. In our case, we calculate the flow from home to work by selecting for each individual the pair of the municipalities in which it is classified, respectively, resident and commuter. Fig. 4 (a) shows that, in spite of a relatively good R statistics ($R = 0.900$, $R^2 = 0.810$) there is a significant number of pairs for which the flow estimates are not very accurate. Most likely, this is due (again) to



(a) Residents (b) Dynamic residents

Fig. 3 Correlation between GSM and ISTAT Resident and Dinamic resident

the fact that for many small municipalities no entry is represented in our dataset. However, if we consider only the flows towards Pisa, shown in Fig. 4 (b), we can see that the estimates improve considerably (now, $R = 0.989$ and $R^2 = 0.978$). The explanation is that the Pisa municipality is a strong attractor for the area, and therefore the set of corresponding flow samples is much larger. In general, we note that the flow estimates with our method are more accurate with larger towns, since they tend to attract larger systematic flows.



(a) All flows (b) Pisa

Fig. 4 Correlation between sistematic flows measured by ISTAT and Sociometer

5 Conclusions

In this work we developed a population and flow estimation based on mobile phone *big data*, used here as proxy of the presence and mobility of individuals. The results obtained are generally encouraging and, for some specific statistics, very accurate in comparison to analogous statistics obtained with official data.

The experience summarized here is part of an ongoing project, and several lines of improvements are planned for the future, including the following: (i) adopt more efficient and effective clustering methods for the extraction of RCPs; (ii) perform user-centric classifications, instead of ICP-centric ones, i.e. classify all the ICPs of a user together, exploiting the relations and dependencies that exists among them, e.g. each user should have exactly one residential area (clear exceptions apart); (iii) extend the experimentation to larger areas, in order to both increase the sample of population covered and avoid *border effects* due to flows coming from/directed to outside our area of study.

References

1. Wang, D., Pedreschi, D., Song, C., Giannotti, F., and 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, NY. 2011.
2. Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Mede, P. V. D., Bruijn, J. D., Romph, E. D., and Bruil, G. MP4-A project: Mobility planning for Africa. In D4D Challenge @ 3rd Conf. on the Analysis of Mobile Phone datasets (NetMob 2013). 2013.
3. Oltenau, A.-M., Trasarti, R., Couronne, T., Giannotti, F., Nanni, M., Smoreda, Z., and Ziemlicki, C. GSM data analysis for tourism application In Proceedings of 7th International Symposium on Spatial Data Quality (ISSDQ). 2011.
4. F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal, 2011
5. B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo Turism fluxes observatory: deriving mobility indicators from GSM calls habits In the Book of Abstracts of NetMob 2013
6. B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behaviour In the Proceedings of Big Data 2013