Identifying users profiles from mobile calls habits

Barbara Furletti
KDDLAB - ISTI CNR
Pisa, Italy
barbara.furletti@isti.cnr.it

Lorenzo Gabrielli
KDDLAB- ISTI CNR
Pisa, Italy
lorenzo.gabrielli@isti.cnr.it

Salvatore Rinzivillo
KDDLAB - ISTI CNR
Pisa, Italy
salvatore.rinzivillo@isti.cnr.it

Chiara Renso
KDDLAB - ISTI CNR
Pisa, Italy
chiara.renso@isti.cnr.it

ABSTRACT

The huge quantity of positioning data registered by our mobile phones stimulates several research questions, mainly originating from the combination of this huge quantity of data with the extreme heterogeneity of the tracked user and the low granularity of the data. We propose a methodology to partition the users tracked by GSM phone calls into profiles like resident, commuters, in transit and tourists. The methodology analyses the phone calls with a combination of top-down and bottom up techniques where the top-down phase is based on a sequence of queries that identify some behaviors. The bottom-up is a machine learning phase to find groups of similar call behavior, thus refining the previous step. The integration of the two steps results in the partitioning of mobile traces into these four user categories that can be deeper analyzed, for example to understand the tourist movements in city or the traffic effects of commuters. An experiment on the identification of user profiles on a real dataset collecting call records from one month in the city of Pisa illustrates the methodology.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms

Keywords

GSM Data, User profiles, SOM

1. INTRODUCTION

4.4 billions or users worldwide, 838 GSM networks spread in 234 countries, 1.44M new GSM subscribers every day are only a few of the impressive numbers that witnesses the enormous diffusion of the GSM phenomena since its first network launched at the beginning of '90 [7]. This massive quantity of mobile phones are moving everyday with their human companions, leaving tracks of theirs movements. These tracks represents the mobility of millions of people in the Earth surface and the opportunities to use these data for analysing and understanding human mobility are tremendous. Research literature has seen a growing interest in techniques for analysing mobility of users based on GSM position data. This research has also been driven by an increasing number of applications that has found in mobile phone data a good partner for discovering interesting results on people behavior. The advantages of relying on these kind of data, compared to standard survey based data collection, is that they offer a wide coverage of the people presence in an area, they are heterogeneous from the point of view of the tracked person and they tend to be up to date and easily upgradeable with new automatic data collection. However, these huge quantity of humans location data comes with a price. Due to privacy reasons, the telecommunication provider must anonymize the data. Thus the analysis that can be done on such data does not distinguish the different user profiles. Therefore the heterogeneity of these data, besides being a strong point, is also a weak point. The mobility analysis that can be performed may suffer from biases due to the wide difference in mobility behavior of tracked users. How to determine, among all the positions collected in a city, which ones correspond to specific categories of users such as residents or visitors? Is it possible to distinguish them looking at they mobile usage?

In this paper we face this problem proposing a methodology to partition a population of users tracked by GSM mobile phones into four predefined user profiles: residents, commuters, in transit and tourists/visitors. Several applications may benefit from the analysis of a partitioned set of users based on this mobility characteristics. For example, being able to distinguish between residents and commuters may help in traffic management to better understand how traffic is affected by the residents mobility compared to the commuters. Having identified the tourists/visitors, it is essential to study how the city is receiving people from outside and how their movements are affecting the city. Again, being able to combine the mobility of resident population with the temporary population (like commuters, visitors or people in transit) may give a measure of the sustainability of the incoming population with respect to resident one. The population on a territory consumes resources like water,

UrbComp'12, August 12, 2012. Beijing, China

air and produces negative effects on the surroundings, like garbage, pollution, noise. In the cases where these resources are limited, the incoming tourist population may break the sustainable equilibrium of such resources. Thus, the ratio between residents and incoming people should be monitored in order to prevent critical situations.

The methodology introduced here aims at inferring the population profile among the GSM call positioning data, namely GSM Call Records (CDR) identifying, with a certain degree of approximation, which calls may correspond to predefined users categories among residents, commuters, in transit and tourists. The basic idea of the methodology is to perform two steps: the top-down step and the bottom-up step, thus combining the deductive power of queries on the call records to the inductive power of a machine learning step based on the SOM [5] technology. More in particular, the top down step tends to identify classes of users based on a predefined call behavior that may approximate a given typology of users. Just to give an example, users that generally call any time during the day or night for a long period may be considered resident, while users that only calls in a restricted period may be considered in transit or tourists. Since the border between these definitions may be not so crisp when data are sparse (e.g. how clearly to distinguish between tourists and in transit, or with a person that makes few calls?) a bottom-up step is performed to compute sets of users with similar calling behavior integrating the results of the top-down step.

The profiling methodology we are proposing is accompanied by an experiment studying the behavior of user profiles analysing mobile phone positioning data, namely GSM Call Records (CDR). Each dataset collects the (anonymized) records storing the location and duration of calls of mobile phone users in the GSM network. The dataset of CDR has been collected by an Italian telephone company in the area of Pisa, in Tuscany, Italy. The city counts about 90,000 inhabitants and it is the location of an ancient and prestigious University that attracts more than 10,000 students from everywhere in Italy. Pisa is well known all over the word as being a tourist attraction for its leading tower. There are estimations that every year one million tourists visit Pisa. Therefore this area is suitable to perform an experiment in trying to infer the different users profiles moving in the city.

The structure of the paper follows. After a selection of related works presented in Section 2, Section 3 introduces some basic concepts used though the paper like the Call Records definition and the user profiles. The methodology is introduced in the following Section 4 where the two steps top-down and bottom-up are presented. The experiments are illustrated in Section 5, while Section 6 draws the conclusions.

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.

Famous experiments on analysing GSM data for studying people movement have been run on Rome [3] and Graz [11]. They use GSM data to realize a real-time urban monitoring systems. They 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. The final objective is to realize a wide range of services for the city such as traffic monitoring and tourists movement analysis.

A different approach comes from Schlaich et al. [12] 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 [1], 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. [9], 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 their proposed 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. [10] 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 crawled 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 [2] 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. Altough 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 is to characterize the user profile based on the call habits of the tracked users.

There is a broad research area that focuses on the inference of significant places or activities from mobility data represented as GPS trajectories [8]. Examples of these works are in [15, 6]. In paper [6] the authors infer activities carried out by moving people (e.g. AtHome, Shopping) and the transportation mode. The inference of activities is based

[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.

on temporal patterns (since different activities have different temporal duration), the location where people stopped, transition relations, since one activity may or may not be followed by another, and a number of common sense constraints. Authors of the paper [15] infer the similarity between users based on their GPS trajectories. They associate to a user trajectory the semantic location history - the sequence of Points of Interests visited - that is used to measure the similarity between different users. However, the GPS data offers a better spatio-temporal granularity level compared to GSM data so many of the techniques available for GPS cannot be used for GSM call record and new methods have to be invented.

3. BASIC CONCEPTS

The objective of this work is to propose a methodology for user profiling in GSM data. We propose a method to infer a possible segmentation of a population of GSM users into different behavior categories. This is an essential step for better understand and study people mobility from unsupervised mobility data.

Indeed, being able to differentiate population ranges enables a number of new applications where the mobility behavior is relative to a specific user segmentation. However, when the mobility data is not directly annotated with the user profile, the association of an anonymous trajectory to a given segment is far to be trivial.

The strategy we propose here is based on the identification of *residents*, *commuters*, *people in transit* and *visitors/tourists* from GSM call records.

GSM network and Call Data Records.

GSM (Global System of Mobile communications) network allows the mobile phone communications based on a system of antennas that transmit the signal to a spatial area that is called Local Area Network. All mobile phones inside that area may receive the signal and therefore are connected to the network. When they are connected they are enabled to make calls or send SMS (Short Text Messages) to another GSM phone connected to the network. When a call is engaged, the telephone company registers data about the location and duration of the call for billing purposes. These data are called Call Data Record (CDR) [13] and we can simplify the standard format as follows:
$< Caller\_ID, ID\_Cell\_Start, Start\_Time, ID\_Cell\_End, Duration >$
where: $Caller\_ID$ is the anonymous identifier of the caller, ID_Cell_Start and ID_Cell_End are the identifiers of the cell where the call starts and ends respectively, $Start\_Time$ is the date and time when the call starts, and $Duration$ is the call duration.

Mobile users profiles.

We are interested in inferring the profile of users moving in a city. For the sake of the current study, stated A the spatial area under analysis, the categories we are interested in, are the following:

Resident. A person is resident in an area A when his/her home is inside the A. Therefore the mobility tends to be from and towards his/her home.

Commuter. A person is a commuter between an area B and an area A if his/her home is in B while the workplace is in A. Therefore the daily mobility of this person is mainly between B and A.

In Transit. An individual is "in transit" over an area A if his/her home and work places are outside area A, and his/her presence inside area A is limited by a temporal threshold T_{tr} representing the time necessary to transit through A. In other words, the user does not perform any main activity inside A. Depending on the application this temporal threshold T_{tr} may vary from few minutes to few hours.

Tourist or Visitor. The definition given by The World Tourism Organization defines tourists 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" [14]. We can rephrase and formalize this definition as: a person is a tourist in an area A if his/her home and work places 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 A. In particular here the presence has to be concentrated in a finite temporal interval inside the time window. 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 activities related to work, visiting relatives, health reasons, etc.

4. METHODOLOGY

The proposed analytical process is based on a step-wise approach: first, domain knowledge is used to label each user according to a set of rules that define each profile; second, the profiles that do not fit in any of the hypothesis templates are analyzed by means of a machine learning approach to determine relevant groups of users according to their calling behavior. Therefore when an individual makes (at least) a phone call inside a network cell we say this individual is *present* into the cell area. The presence pattern is then defined by temporal constraints on the detected presence. However, these definitions combined with the characteristics of the GSM call data may give misleading classifications. For example, a resident user who rarely calls may be misclassified as a tourist or a person in transit, while a resident that only use phone at work may be classified as a commuter. Again, defining a good threshold to identify tourists may be difficult and certainly depends on the application. Although the "in transit" profile is well defined once a temporal threshold is fixed, the other profiles, especially the "tourist" population, is characterized by a "fuzzy" and non clear characterization.

To face this problem the user profile methodology we introduce here proposes the combination of a deductive and an inductive technique that we name *top-down* and *bottom-up*. In the top-down approach a set of spatio-temporal constraints are used to describe the individual categories following the definitions given by the domain experts and that we introduced in Section 3. The constraints are then implemented in the mobility data management and mining system M-Atlas [4], through the use of the provided query language. In the bottom-up step the assignment of users to categories is refined using a clustering algorithm, namely Self Orga-

nizing Map (SOM) [5]. Clearly, since the top-down step is based on a set of rules provided by the domain experts, they may fail at classifying behaviors on the borders of the definitions. Therefore, all those individual that have few phone calls or whose phone calls behavior does not clearly fall into the well-defined categories remain unassigned.
The bottom-up approach aims at integrating the results of the first step by using a data-driven approach to identify relevant group of users that present similar behaviors that can be classified as one of the available profiles.

The advantages of the described technique lies on the fact that GSM data - due to the widespread use of mobile phone and the heterogeneous classes of their users - allows to analyze the mobility behavior of a huge amount of people and a broad range of users categories. Furthermore, the use of an inductive step allows a refinement of the preliminary results obtained with the top-down approach.

Top-down Approach.

During this phase the residents, commuters and in transit categories are retrieved from the CDR dataset with a proposed set of spatio-temporal constraints that depend on the time window of the data collection and reflect the indications given by the domain experts.

Resident users are those whose CDR data show a continuous presence in the monitored period during the late afternoon and night (since we assume that during this period individuals stay at home) and the weekly minimal presence to be reasonable to establish as home.

Users that tend to have a sparse presence of calls during the period but concentrated only during the weekdays in the conventional working/studying times, are classified as commuters. The assumption is that the commuters spend nights at the home place (an area outside our interest) and weekdays at the work/study place in the area of interest, and never appear during the weekends.

People in transit are directly identified by a simple constraint that limits the presence to a fixed time range depending on the dimension of the area under analysis. The constraint tries to encode the average time needed to cross the area without stops for activities. The idea is to capture people passing on motorways and freeways near a city or crossing the city by using urban roads. This gap can vary from less then 1 hour, in case of small towns, up to several hours (two, three or more), in the case of big cities.

For example, a simple query in SQL Like to identify people in transit is the following:

```
SELECT a.Caller_ID
FROM Cdr_Calls_Table a, Coverage_Table b
WHERE (a.Last_Call - a.First_Call) < Time_Limit AND
b.Location = 'Pisa'
```

where `Cdr_Calls_Table` is the table containing the CDR data, `Coverage_Table` is the table containing the spatial coverage of the network cells, `Last_Call` and `First_Call` are the timestamps of the last and first call respectively, and `Time_Limit` is the estimated time needed to cross the area of interest.

Of course a number of users whose call behavior does not precisely falls in these three definitions remains unclassified after this first step, the bottom-up step is thus necessary to analyse the unclassified set of users trying to assign a category basing on a temporal profile.

Bottom-Up Approach.

The bottom-up approach has the twofold purpose of both identifying tourists and refining the results obtained by the top-down phase for residents and commuters.

The behavior of each user is modeled by means of the concepts of space (where a call is started and where it is terminated) and time. We exploit these two dimensions to define a temporal profile for each user.
Given a user u, a *Temporal Profile* TP_u is a vector of call statistics according to a given temporal discretization. For example, using a time discretization by day and a measure of frequency of calls, each entry of TP_u would contain the number of calls performed by the user in the corresponding day. Since we are interested in a specific area, we define a *Space constrained Temporal Profile* TP_u^A as a Temporal Profile where only the calls performed in the cells contained within the area A are considered.
This spatial projection is crucial when studying commuters in order to distinguish the call behavior at work and at home. To explore different time patterns, we define also two time transformations of a (Space constrained) Temporal Profile: (i) *time projection* by a cycle period, where the time intervals of the vector are referred to relative position in a time cycle like week, month, and so on; (ii) *time shifting*, where the time intervals of the vector are shifted in order to have the first entry corresponding to the first activity of the user. Clearly, the available statistics can be chosen according to the specific analytical scenario. For example, it can be considered the number of calls, the total duration of the calls, or a boolean operator that yields true if at least a call has been performed in a specific time period. Figure 1 shows two examples of extraction of temporal profiles from the call behaviors of two users, using call frequency as measure for each cell.

The temporal profiles defined above can be analyzed according to their relative similarities by means of a Self Organizing Map [5]. A SOM is a type of neural network based on unsupervised learning. It produces a one/two-dimensional representation of the input space using a neighborhood function to preserve the topological properties of the input space. As most neural nets, a SOM constructs a map in a training phases using input examples and uses the map for classifying a new input vector. The procedure for placing a vector from data space onto the map is to first find the node with the closest weight vector to the vector taken from data space. Once the closest node is located, it is assigned the values from the vector taken from the data space. SOM forms a sort of semantic map where similar samples are mapped close together and dissimilar apart. In our case the weighted vectors used for the analysis are the Temporal Profiles extracted for each user. For example, using a discretization of one hour, a daily temporal profile for a user consists of a vector with 24 entries. The dimensions of the vector may change accordingly with different aspects of the analysis like, for example, the temporal profile in a week, in a single day, or in the whole period by hour (see Figure 1). The SOM algorithm produces a set of nodes, where each node represents a group of users with similar temporal profile. By analyzing the profile that describes each group, it is possible to

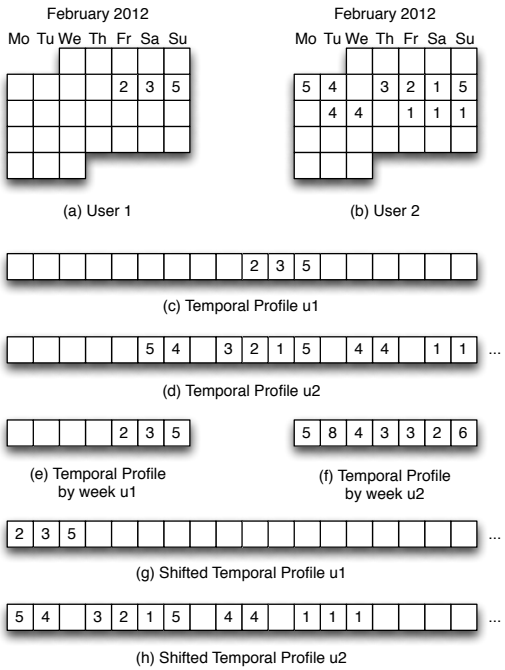

Figure 1: Example of the extraction of temporal profiles from the call activities of two users. Each square represent a day and the number within a cell is the number of calls of the corresponding user in that day.

assign a class to the group itself, basing on the definitions of resident, commuter and tourist given above. In particular, we label as "tourists" the nodes that correspond to a temporal profile localized in a short and consecutive time period (as for example few days or a week). As stated before the temporal profile of a tourist can not be defined a priori because it has a wide variability depending on the season, the location and other unpredictable events. This method can thus help to discover the touristic degree of an area without using particular apriori knowledge. This phase is useful also to re-calibrate the top-down results. In fact, the analysis of the profiles emerging from the groups may give information about local habits and may suggests how to set the temporal constraints to adapt the model to the local habits of the area of study.

We perform this analysis from two perspectives dependent on the chosen temporal profile. In the first case we extract the temporal profiles for each user and we transform each profile by a left shifting operator in order to make the data more dense (see Figure 1 (g) and (h)), since the users who has visited the area in different time periods can be compared also by the length of their staying. Such operation, on one hand, loses an absolute temporal reference, thus is not directly possible to associate each entry to a specific time period. On the other hand, the SOM algorithm can easily identify group of users with compatible periods of visit to the area. In this way, for example, it is possible to identify the typical duration of presence of the users and, hence, assign the tourist/visitor class to the nodes with compatible temporal profiles. In the second case we extract the original temporal profile according to the absolute time alignment. In this case the resulting SOM tends to highlight similar and compatible presence profile of longer stay people, allowing to separate commuters and residents, by exploiting the calling habits of these users in particular during the weekends.

With respect to other clustering techniques, the SOM allows an easier and clearer visualization of the results. This technique seemed to us very useful for precessing the profiles input extracted from the GSM data, and visualizing the complex results.

5. EXPERIMENTS

We tested the proposed approach on a case study in the city of Pisa. We used a large dataset of GSM data collected in the province of Pisa by one of the Italian mobile operators. The data consist of around 7.8 million CDR records collected from January 9^{th} to February 8^{th} 2012. The data contains calls corresponding to about 232.200 users with a national mobile phone contract (no roaming users are included in the dataset). Our approach is based on a set of temporal constraints over the users' temporal profiles. As a preliminary validation analysis of the method, we analyze the temporal presence of users in the province of Pisa. Figure 2 shows the cumulative distribution of the duration of stay of users in the province: a point (x, y) on the chart represents the number of distinct users y that were observed in the area at most for x days. From the chart we can roughly partition the population on the basis of the domain knowledge: people staying less than four days are candidate visitors or in transit, while the others can be considered as resident/commuters. This is a very naive segmentation, however allow us to estimate how effective this approach is. In particular, using the four-days threshold the candidates resident commuters are around 107k. This number is compatible with the customer statistics provided by the telecom operator in the area, thus informally validating the proposed approach.

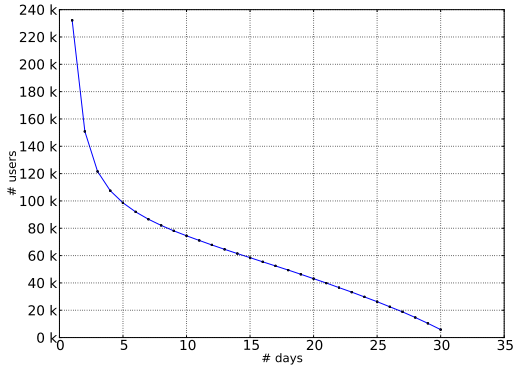

Figure 2: Cumulative distribution of the number of users per length of staying.

Since our aim is to study the mobility of residents and visitors in the area of Pisa, from the whole network we first selected the cells overlapping the urban area of the city. The urban center of the city is crossed by the river Arno and its corresponding cells are highlighted in pink in Figure 3. The larger gray area corresponds to the administrative territory

of the city that includes also the seaside and the large park called *San Rossore*. Once we have determined the area of interest, we filtered the calls by considering only those calls performed in the selected cells.

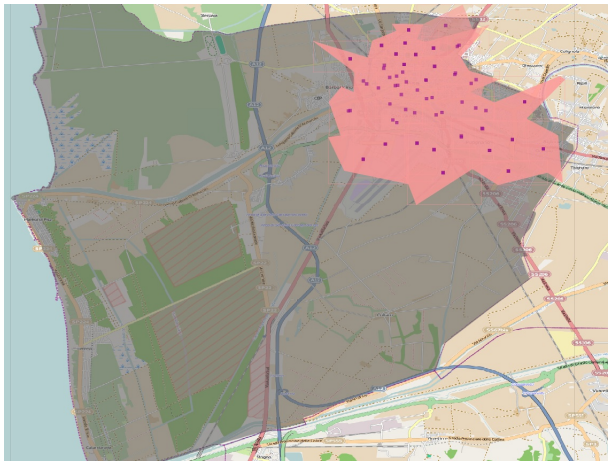

Figure 3: Coverage of the urban area of Pisa.

As mentioned in Section 4, the top-down approach consists in a set of spatio-temporal queries expressed in the query language of M-Atlas [4]. The spatial constraints, which are the same in all the cases, define the area of interest i.e., the urban area of Pisa and the corresponding GSM coverage. This coverage is expressed as a geometric intersection of the base station positions with the urban census surface. The temporal constraints are different for all the categories of users we want to identify as detailed in the following:

Resident
C1 - Temporal range: at least 1 call in [19:00 - 6:59] during the weekdays.
C2.1 - Daily presence: at least 2 distinct weekdays per week, that satisfy C1.
C2.2 - Daily presence: at least 1 day in the weekend without temporal range.
C3 - Weekly presence: at least 3 weeks, in which C1, C2.1 and C2.2 are satisfied.

Commuter
C1.1 - Temporal range: at least 1 call in [9:00 - 18:59] during the weekdays.
C1.2 - Temporal range: no calls in [19:00 - 8:59] during the weekdays.
C2.1 - Daily presence: at least 2 distinct weekdays per week, that satisfy C1.1 and C1.2.
C2.2 - Daily presence: never during the weekends.
C3 - Weekly presence: at least 3 weeks, in which C1.1, C1.2, C2.1, C2.2 and C3 are satisfied.

People in Transit
C1 - Temporal range: calls during at most 1 hour.
C2 - Daily presence: at most 1 day in which C1 is satisfied.
C3 - Weekly presence: at most 1 week, in which C1 and C2 are satisfied.

The result of the top-down approach is shown in Fig. 4. This method is able to capture only a low number of commuters and residents because the temporal constraints are very strict and selective. On the contrary, people in transit are well identified. The high percentage of this kind of users is justified by the presence of an highway and a freeway close to the town.

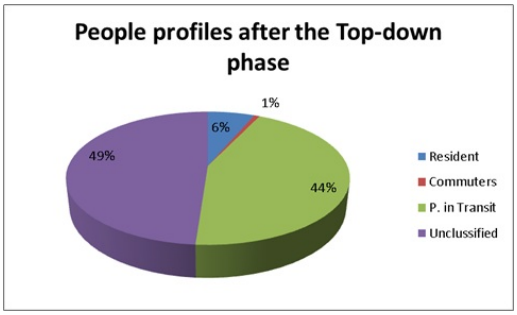

Figure 4: People profiles after the top-down phase.

Thus, starting from the unclassified users, we can apply the SOM method to identify temporal profiles with similar characteristics, i.e. we can group together people who have the same calling patterns. In particular, we are interested in two aspects of the temporal profile: the duration of the stay in the city and the typical temporal location of a user call. To address the first problem, we perfomr a transformation of the temporal profiles by applying a temporal shift. The objective it to align all the user activities at the beginning of the time window. The results provided by SOM is shown in Figure 5. The resulting map shows a set of nodes, where each node contains a set of user profiles. For an immediate readability of the results, each node shows the cardinality of its the population, the circle is proportional to the population and the time chart shows the temporal distribution of the user activities in the specific time interval. In the map of Figure 5*(Left)*, the shifted temporal profiles consist of vectors of 31 entries, one for each day of the time window. Since we are dealing with rotated profiles, the extension of the temporal distribution in each node provides an immediate estimation of the duration of the stay of the corresponding users. From the map it is evident how the temporal profiles are grouped: on the bottom left corner of the figure there are the temporal profiles corresponding to short visits of the city; the upper right side of the figure shows the profiles that span for the whole period and it is possible to identify even nodes that present a clear commuter-like pattern with high frequency during the workdays and a smaller activity during weekends. It is important to point out the presence of three larger nodes corresponding to short visits ranging from one day (node with 5750 profiles) to three days (nodes in the upper left corner). The shifting transformation can be inverted to observe the actual temporal distribution of the activities during the period of study. Figure 5*(Right)* shows, for each node in the *(Left)* map, the corresponding absolute temporal distribution of the activities. The cardinality statistics are left as a reference between the two figures. It can be noted how the short-ranged temporal profiles are uniformly distributed across the whole period. For instance, the larger node containing temporal profiles of a single day presents a quite uniform presence of users across the month considered in the study. On the other hand, as it could be expected, the profiles with a larger extent do not vary too much, since

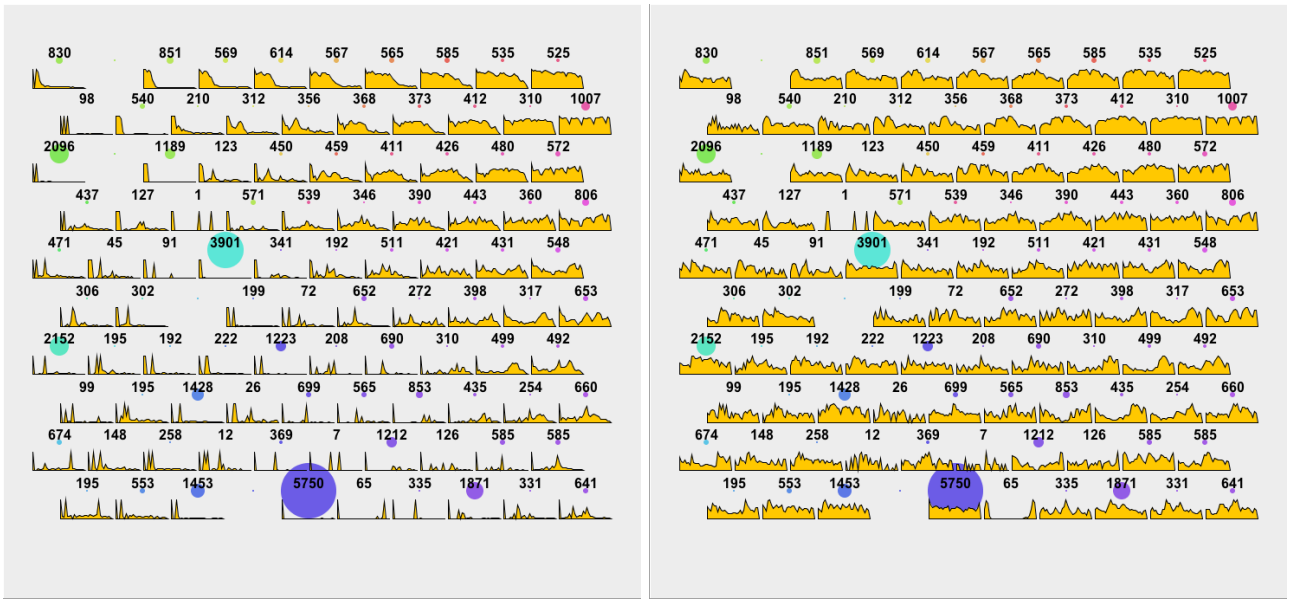

Figure 5: SOM clusters with rotated Temporal Profiles. *(Left)* Each node shows the distribution for shifted temporal profiles. *(Right)* Each node shows the actual temporal distribution for the corresponding set of users.

their width limits the shifting transformation.

To better understand the temporal distribution of user activities in the period, we apply the SOM method to the unshifted temporal profiles. The resulting SOM map is showed in Figure 6. In this map, we can notice how the commuter-

Figure 6: SOM clusters with (un-shifted) Temporal Profiles.

like patterns are even more evident in the bottom left corner. The population corresponding to these nodes is even larger than the nodes present in the shifted version. Actually, these nodes are contributed by users with different habits: the frequent callers have a regular temporal profile that remains unchanged even after the temporal shift; the infrequent callers, on the contrary, do not present such distribution by themselves but their aggregated distribution reconstructs this temporal pattern. The influence of personal calling activity may be crucial in some analysis, since a too specific distribution may be biased by an incomplete vision of the real phenomena, in particular when people do not use the mobile phone during their movements or activities. The case above of commuters distribution is just an example of this kind. This aspect is even more evident when we consider some of the larger node of the map in Figure 6. According to the node with 2059 entries (on the center right side of the figure), a lot of people were present in Pisa for a single day, specifically on January 26. Actually, that day, around 16 in the afternoon, an earthquake happened in northern Italy and it was perceived by the population in Pisa. That event triggered the need for a lot of users to communicate and call their relatives producing the peak we observed in the node. This particular example is emblematic in showing how a peak in the phone traffic not necessarily implies an increase in population density. In this case, the peak is due to infrequent callers that were forced to call by an external event.

6. CONCLUSIONS AND FUTURE WORKS

The wide coverage of GSM networks enables numerous applications for the understanding of people mobility behavior. However, the data anonymization combined with the heterogeneity of people that carry a GSM phone limits the analysis that can be performed due to the lack of a user profile. To face this problem in this paper we propose an approach for inferring user profiles from GSM data. In particular, we concentrate our efforts in identifying mobility profiles based on the presence of user in an area, that is in turn based on the call habits of that user. The methodology

aims at identifying four categories of users: residents, commuters, in transit and tourists/visitors. The process in based on two phases: a top-down step where GSM Call Records data are queried, combined with a bottom-up step based on a machine learning algorithm to find homogeneous groups of users. We show - through an experiment run on a real dataset - how the process identifies users profiles. Finding the user profiles may enable wide spectrum of applications from traffic management - for example relating the commuter mobility with the resident one - or tourism - where the touristic flows in a city may be analyzed. The identification of user profiles is based on the analysis of a large dataset of call logs that carry almost no semantic information about the users' metadata. In fact, such data is usually subject to several restrictions (e.g. privacy) that does not allow the diffusion of demographic and personal information of a single individual. In this scenario, we try to compensate the deficiency of semantic information by relying on a unsupervised method to separate relevant groups with similar temporal profiles. Current work include the investigation of the accuracy problem. Indeed, the outcome of the method should be evaluated at least on two levels of accuracy: a quality measure of the resulting segmentation (based for example on classical measures of cluster quality like separateness and cohesion) and a validation assessment of the population of each group with a reference ground truth. The first task is straightforward using classical clustering methods, but does not guarantee the real adherence of the result to the reality. On the contrary, the latter is more challenging since it may assess the results against some form of ground truth and it is particularly useful when the input data do not come with a rich semantic information, like in the present case. We plan to investigate this latter issue from several points of view. For example, a possibility consists in the comparison of the profiles of resident users with the resident population measured by national statistical bureau. However, this ground truth describes only partially the phenomenon since we miss the dynamic aspect that are difficult to measure with precision with classical statistical measures like, for example, the continuously changing ratio of residents with commuters and tourists. Our proposal consists in the comparison of each of these profiles with a series of observations coming from different datasets. In particular, we are currently working on a project with the Municipality of Pisa to collect, integrate and analyze a set of statistical indicators about touristic presence and mobility. In this context, we plan to compare the temporal distribution of the extracted profiles with the data coming from several information sources related to the touristic presence in the city. Examples are the aggregated number of hotels reservations, issued museum tickets, social photo services (i.e. Flickr and Panoramio), microblogging services (i.e Twitter), trajectories of private vehicles equipped with GPS devices, touristic buses presence, etc. The method we propose is not a direct comparison of two variables, i.e. GSM profiles and the estimation of tourists in the city, but we aim at proposing a more complex methodology where different datasets are used to provide a specific vision on the same phenomenon, i.e. the touristic presence in the city, by means of the combination of the different points of view giving a complementary vision of the whole scenario. This is a very ambitious task that involves the management of different data sources. It is worth pointing out that this work would be relevant also in other in different application scenarios where a classical statistical validation is not possible.

7. REFERENCES

[1] R. Ahas, S. Silm, S. Järv, and E. Saluveer. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1), 2010.

[2] G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C Poelitz, and T. Schreck. A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. *Journal of Location Based Services*, 4(3–4), 2010.

[3] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12:141–151, 2011.

[4] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB J.*, 20(5):695–719, 2011.

[5] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer, 2001.

[6] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational markov networks. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence, IJCAI'05*, 2005.

[7] Nokia. Nokia siemens networks. http://www.slideshare.net/NokiaSiemensNetworks/20-years-of-gsm-past-present-future-8512655.

[8] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. A. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *Accepted at ACM Computing Surveys*, 2012.

[9] F. C. Pereira, L. Liu, and F. Calabrese. Profiling transport demand for planned special events: Prediction of public home distributions, 2010. Available online at www.scienceDirect.com.

[10] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *International Conference on Data Mining, ICDM*, pages 971–976, 2010.

[11] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer. Mobile landscapes: Graz in real time, 2005. MIT Senseable City Lab.

[12] J. Schlaich, T. Otterstätter, and M. Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies*, 2010.

[13] Wikipedia. Call data record. http://en.wikipedia.org/wiki/Call_detail_record.

[14] Wikipedia. Tourism. http://en.wikipedia.org/wiki/Tourism.

[15] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. Finding similar users using category-based location history. In *GIS*, pages 442–445, 2010.