information

MDPI

Article

Discovering and Understanding City Events with Big Data: The Case of Rome

Barbara Furletti, Roberto Trasarti *, Paolo Cintia and Lorenzo Gabrielli

Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy (ISTI-CNR), 56100 Pisa, Italy; barbara.furletti@isti.cnr.it (B.F.); paolo.cintia@isti.cnr.it (P.C.); lorenzo.gabrielli@isti.cnr.it (L.G.)

* Correspondence: roberto.trasarti@isti.cnr.it; Tel.: +39-50-621-3006

Received: 31 March 2017; Accepted: 25 June 2017; Published: 27 June 2017

Abstract: The increasing availability of large amounts of data and digital footprints has given rise to ambitious research challenges in many fields, which spans from medical research, financial and commercial world, to people and environmental monitoring. Whereas traditional data sources and census fail in capturing actual and up-to-date behaviors, Big Data integrate the missing knowledge providing useful and hidden information to analysts and decision makers. With this paper, we focus on the identification of city events by analyzing mobile phone data (Call Detail Record), and we study and evaluate the impact of these events over the typical city dynamics. We present an analytical process able to discover, understand and characterize city events from Call Detail Record, designing a distributed computation to implement Sociometer, that is a profiling tool to categorize phone users. The methodology provides an useful tool for city mobility manager to manage the events and taking future decisions on specific classes of users, i.e., residents, commuters and tourists.

Keywords: city events detection; big data analytics; distributed systems; sociometer; mobile phone data; case of study; Rome

1. Introduction and Context

Mobile devices are nowadays one of the main means by which people disseminate digital tracks of their everyday activities: trips, purchase transactions, preferences, opinions, and so on. In particular, mobile phones and the data they produce revealed to be a high-quality proxy for studying people mobility in different domains, such as environmental monitoring [1,2], transportation planning [3], smart cities and social relationship analysis [4,5].

For data mining analytics and applications, these data are very significant in terms of size and representativeness. In [6] the authors demonstrate how the number of residents observed through mobile phone data is highly correlated with the number of residents identified by official estimates. Currently, a hot topic in the modernization of official statistics is precisely how to use big data in combination with traditional data sources, in order to improve quality, timeliness and spatio-temporal granularity of statistical information [7]. Despite their limits in spatial precision compared to other location data such as GPS tracks, mobile phone data are of uttermost interest due to their global availability for any countries, and the ability to portray mobility independently from the transportation means. This is documented by many contributions realized within the D4D challenge (http://d4d.orange.com/), and for example in (i) [8], where Call Detail Records (CDRs) from an African city were used to reconstruct an Origin and Destination (OD) matrix describing typical traffic flows; or in (ii) [9] where the authors, created temporal profiles of the call activities in order to identify the different location types (residential, business, mixed area) in a city.

In the exploration of city dynamics, authors in [10] illustrated several urban phenomena of the city of Graz by using different type of mobile phone records: CDRs for traffic intensity, handovers for

traffic migration, and volunteers' tracks for reconstructing individual movements. In [11], authors reviewed several techniques for extracting knowledge from mobile phone data in order to perform a global sensing of the cities and reason about the mechanisms driving the city life. Pollution at urban level has been also studied the support of mobile phone data: in [12], for example, the authors used the CDR to estimate the population density and draw a distribution map over the city to evaluate the actual population exposure to pollutants.

In the domain of city and human dynamics, with this paper we have a twofold objective: on the one hand we want to detect and characterize important and unusual events at urban level; to the other hand, we want to estimate the composition of the population who attended to them. The goal is to identify and isolate significant and unusual peaks of presences among all the mobile signals of presences registered during the days, thus quantifying the events and their impact on the city dynamics and population composition. With respect to state of the art, this work exploit a more precise identification and characterization of the events of the different city users through the categorization of phone users. Such an estimation is performed through the Sociometer, a tool for urban demographics which processes mobile phone records to characterize city users into behavioral categories [13]. Users categorization represents a further semantic layer that enrich the information provided by each single CDR, thus introducing a novel approach w.r.t. to [14,15]. The former proposes a method to detect unusual events relying on users' mobility profile, considering each antenna the user connected to as a location. Then, they identify unusual crowds detecting users whom are aggregating in areas unusual for them, according to their corresponding mobility profiles; the latter proposes a supervised approach to learn the pattern of an event, that differs from our method since it is based on the availability of a list of known events.

The result we obtain is particularly useful to support the understanding of phenomena and provide new knowledge to decision makers and urban planners. For example, it was possible to eliminate noise signals introduced by the so called "people in transit". Moreover, this distinction allowed us to focus on the impact of the events on the city users identifying what kind of people are interested in different kind of events, and try to answer to typical questions like: does the event attract people from outside the city? Or, is it rather attended especially from residents? How does the city users composition change during these events, w.r.t. the normal days? Our aim is to provide updated knowledge in order to improve demographic and urban investigations and support decision makers in planning purposes.

The paper is organized as follows: Section 2 presents the motivations of this work and the definition of the problem. Section 3 presents the basics of the sequential version of the Sociometer, while Section 3.1 describes the new distributed version discussing advantages and computational issues; Section 4 shows an interesting case study over the city of Rome, while Section 5.2 presents an evaluation and validation of the system w.r.t. the case study. Finally, Section 6 contains the conclusions and the plans for future works.

2. Motivations and Problem Definition

Given a big dataset of mobile phone records (Call Detail Records—CDRs) covering six months of observations over one of the Italian biggest and touristic cities, we focus on the problem of identifying and isolating "important" events. An event is a significant, unusual, and relatively bounded spatio-temporal happening which attracts many people toward a point of interest (POI). An event can be political, religious, a popular demonstration, a music festival, a sport event, and so on. A POI can be a place, a park, a stadium, a street, a church, a theater, and so on.

The objective is to estimate the impact of these "unusual" events on the "usual" city dynamics and over the city users. This means trying to isolate the presences of people attending the event from people who typically (every day or periodically) are found there, and characterize them in mobility categories among residents, dynamic residents, commuters and visitors (see details in Section 3). The distinction between types allows us to define the typical composition of the city users and the detection of hidden

anomalies among the complex scenario of a city. It is worth noticing that, for this experiment, a real time data collection service was not available. Hence, we propose an offline method, relying on processing data with a four weeks moving window.

Our methodology opens new scenarios for real-time demography. We are currently collaborating with the Italian Statistical Bureau, which is studying how to integrate Big Data analytics for official statistics [16].

3. The Analytical Process

In this section, we describe the general process to detect and characterize events in a particular area, providing for each one a description of the involved methods. The process of detecting events starts from the analysis of a CDR dataset. This dataset is pre-processed in order to select the records according to a four-weeks sliding temporal window, also assigning each record to its relative region of interest. The Sociometer [13] is applied to classify users who make calls within the spatio-temporal selected window, into six different categories (see Figure 1):

- a *Resident* is an individual who lives and works in A, and therefore his/her presence is significant across all days and all time slots in A;
- a *Dynamic Resident* is an individual who lives A, but works/studies in a different area B. The presence in A is expected to be always significant, excepted during working/studying days and working/studying hours;
- a *Commuter* is an individual who lives in a different area B but works/studies in A. The presence in A is expected to be almost exclusively concentrated during working/studying days and working/studying hours;
- a *Visitor* is an individual who lives and works/studies outside, and visits A only once or occasionally. It is a Visitor if the user perform at least two calls during the stay.
- a *Passing by* is similar to a Visitor category but it contains only users performing a *single call* in A during the entire period of observation. This category is useful to identify, and filter when necessary, people in transit, or characterize particular kinds of visits.

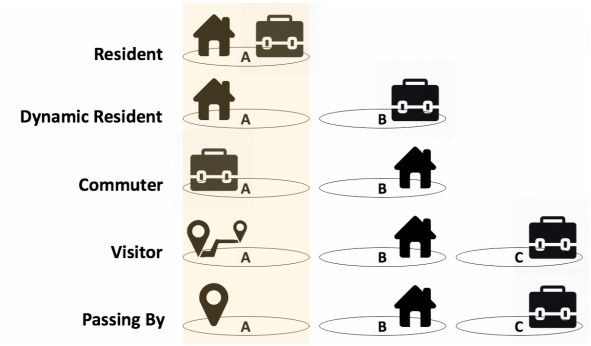

Figure 1. The behavioral categories used to classify users performing calls in the area A in a specific time window.

Using this information, we are able to compute the users presences in the area considering only a particular category obtaining six time series (e.g., the distribution in time of presences of people belonging to the category of *Residents*). Then, each time series is separately processed to detect relevant peaks of presences. This task is accomplished by comparing the density of population against the typical one computed as the average density. The residual density is divided by the standard deviation in order to consider the typical fluctuation of the values (i.e., the standard score normalization). The resulting time series is processed to extract the data points with a value higher than the 95th representing a peak of presences which may represent an important event. It is worth

remarking that the time series we are analyzing contains a high value in terms of knowledge, since we know the users category we are comparing. This differs from other works in literature [14,15], that put more efforts on analyzing the peaks of call/presences without a preliminary semantic enrichment of the dataset. To characterize the events we study the composition of the population in that particular day comparing it to the *typical* one and then we determine their *origin* exploiting the classification of his activity in the long term. In details this second characterization is realized computing an OD Matrix focused on the users participating to an event where their origins are determined by the multi-classification in the long period (and not only during the event): we consider the origin of a user as the area where he is classified as *resident* or *dynamic resident* in the majority of the time windows analyzed, in case the user is not classified in any area with those labels, we consider him as an *outsider*. The results will give all the information for answering to questions about the provenance of the people attending to a specific event.

3.1. Sociometer

As introduced before, one of the key methods of the analytical process is the Sociometer. We briefly reports here only the main steps of analysis useful for understanding how we used it and how the new parallel version is implemented. Figure 2 shows these main steps.

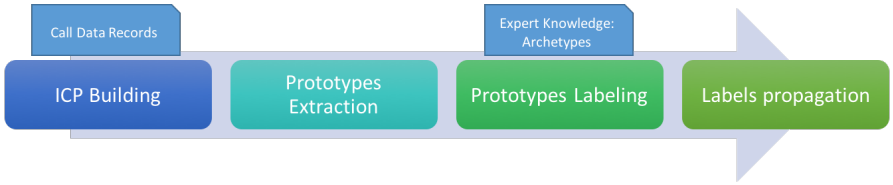

Figure 2. The analytical process of the Sociometer decomposed into its the main sub-tasks.

Starting from a set of CDRs, in the *ICP Building* step, we compute the Individual Call Profiles (ICPs). An ICP (see Figure 3) is a compact representation of the presence matrix of an individual during the period of observation and in three different daily time windows (early morning, central day and night). An individual is present over a matrix cell if she/he made at least one call. During this transformation the values in the matrix is divided by the number of days in each column (i.e., 5 for weekdays and 2 for weekend). This means that each cell contains a value in the interval $[0, 1]$ representing the percentage of days in which there is at least a call. To guarantee the comparability of the ICPs for all the users, the values of presence per week (i.e., the couple columns representing the $<$ weekdays-weekend $>$ pair of each week) are reordered in order to have the higher values of presences in the first positions. This transformation does not affect the classification.

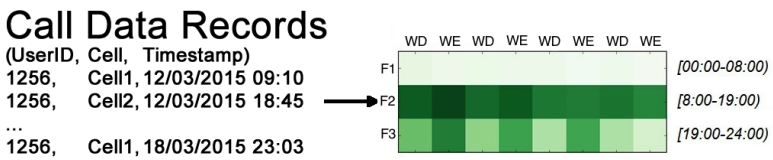

Figure 3. Example of Individual Call Profile: a compact representation of user's phone activities extracted from his CDRs.

In *Prototypes Extraction*, by using the K-Means algorithm, a set of clusters are extracted from the ICPs. The corresponding k centroids, called *stereotypes*, are the set of representative behaviors of the population. It is worth noticing that the value of k for the prototype extraction is selected empirically considering the *Sum of Squares Error* (SSE) and the fact that: (i) a small number lead to the mixture of behaviors, (ii) a large number to the splitting them into subcategories maybe loosing in generality. Studying the variation of the prototypes extracted we selected a trade-off value equal to 100.

A *Prototype Labeling* step is performed in order to label the *stereotypes* in the behavioral category introduced in Section 3. This process is done automatically by means of the labeling of the *stereotypes* w.r.t. some reference profiles called *archetypes* which synthesize the typical behavior of an individual who belong to a city users category. In [13] is widely described how the semi-automatic labeling process obtains the same results of a manual labeling process of the *stereotypes*. It is worth saying that, in this way, we optimize the trade-off between efficiency and precision in the learning phase, also facilitating the reproducibility.

Finally, in the *Label Propagation* sub-task we apply the classification model for classifying new instances. This centralized version takes days (4-nodes cluster, each one with 6-cores Intel XEON@2.93Ghz, 24 GB Ram and 2 TB storage capacity) to analyze a city scale dataset for a single temporal window of four weeks (200 M calls for 7 M users). In the event detection analysis we are presenting, the temporal span is one order of magnitude greater, and thus this version of the Sociometer is not usable. A re-engineering was necessary to make the system efficient for a large-scale data processing.

3.2. Scaling up to Big Data

In this section, we describe the distributed version of the Sociometer able to analyze big data in a scenario which evolves in time. This version depicted in Figure 4, is based on the *Spark* (http://spark.apache.org/) technology, and it re-implements and optimizes the analytical process of Figure 2.

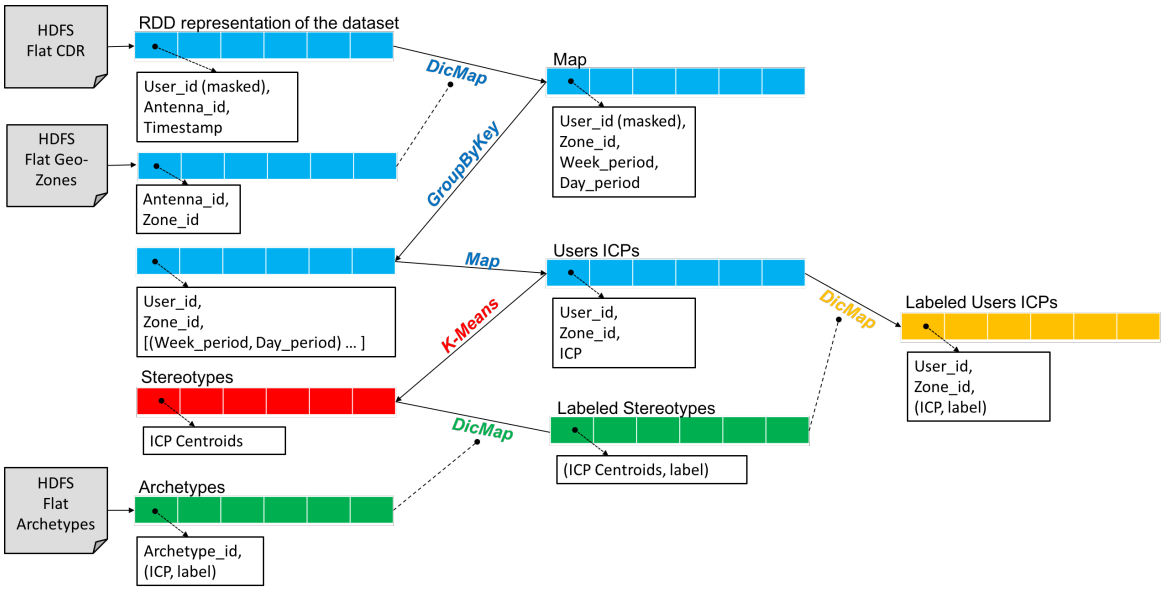

Figure 4. The Spark process: (i) ICP Building process (blue); (ii) K-Means application (red); (iii) Stereotypes Labeling (green); and (iv) Label Propagation (orange).

Starting from the anonymized Call Detail Records (CDRs) provided by the Telecommunication company (The CDRs provided by the Telecommunication Company contain anonimyzed user ids, which are changed at the source before the data is transferred), the system creates a Resilient Distributed Dataset (RDD) composed of all the data entries loaded from the Hadoop Distributed File System (HDFS): each position contains a single call of a single user with the information about the antenna serving the user's call and the timestamp of the call event. It also loads the information about the geo-localization of the antennas and the corresponding signal coverage. The zones are application dependent i.e., they may be a city, a district or single tower (each tower has more antennas mounted on it). The mapping between antennas and zones is usually given by the experts considering their

objectives, for example in this paper we will use the mobility agency partitioning. By using this information the system applies the custom function *DicMap*, which uses a dictionary to perform a *Map* transformation. The resulting RDD contains: (i) the calls where the antenna is transformed in zone, and the timestamp is transformed into two values indicating the *Week period* i.e., whether the call is performed during the weekdays or weekend; (ii) the *Day period* i.e., a value indicating in which period of the day the call started (Early Morning, Daylight, and Night). The next step is realized by a *GroupByKey* function, where the key is represented by the user and the zone, obtaining a smaller RDD containing for each position all the calls performed by a user in a specific zone. After that, a *Map* builds the *ICP* by aggregating those calls for each position and computing the frequencies of calls during the weekdays and weekend in the three possible day periods. The result is a compact representation of the individual profile as exemplified in Figure 3. By using the K-Means implementation provided by Spark, the set of centroids are computed obtaining the stereotypes which have the same representation of the ICPs. The system then loads the archetypes and uses them for labeling the stereotypes based on a proximity criterion using an n-dimensional Euclidean distance. The same procedure is followed for classifying the "new" (and unlabeled) ICPs as exemplified in Figure 5. In the picture, a_1 and a_2 are the archetypes, while s_1 and s_2 the stereotypes; δ' and δ'' are the distances of the ICP u from a_1 and a_2 respectively, while σ' and σ'' are the distances of u from s_1 and s_2 respectively. If u is compared directly with a_1 and a_2, then it will be assigned to a_2 because $\delta'' < \delta'$. If we label first s_1 and s_2 with the closest archetypes, u will be assigned to s_1 (inheriting the label of a_1) because $\sigma' < \sigma''$.

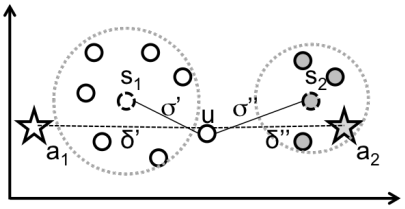

Figure 5. A simplified example (in a 2-dimensional space) of the automatic labeling.

The redesign of the Sociometer with the distributed paradigm offered by Spark results in a great reduction of complexity and execution time. Considering m nodes, the first and second steps is highly parallelized: (i) in the ICP Building, the process is decomposed in u sub-tasks, one for each user, so the computation complexity results in $O(u/m)$ (ii) an efficient distributed version of K-Means is provided by the framework MLLib-Spark (http://spark.apache.org/docs/latest/mllib-guide.html) reducing the complexity to $O(uidk/m)$ where i is the maximum iterations and d is the number of dimension of the elements; (iii) the Stereotypes Labeling becomes $O(ka/m))$ where a is the number of the archetypes (generally very small), and finally (iv) the Label Propagation is $O(ku/m)$. The computing time is reduced allowing us to run many more experiments in less time, and as we will show in the experimental section, we are able to manage very big datasets in terms of users and time window. In particular, in Section 4 we will analyze seven months of data using a mobile time window of four weeks. The result is a complete overview of the composition of the population in an areas over time in a very detailed way that, for our current knowledge, no other tools or survey can reach.

4. A Real Case Study: City of Rome

In this section, we exploit the Sociometer for studying the dynamics of a big Italian city. After a description of the dataset used for the experiments, we present the case study and discuss the validation.

4.1. Data Acquisition

Thanks to a collaboration of an Italian telecommunication company within a European project called ASAP (https://www.asap-fp7.eu/), we collected a massive CDR dataset and we had the possibility of testing our analytical process. The focus of our study is on the big Italian city of Rome. The data was provided under an obligation of confidentiality and for this reason it can not be published or disclosed. The dataset, composed of CDRs, covers a period of seven months between 1 January and 31 July 2016. Each day contains data for 1.2 GB, for a total size of more than 3 billions of CDRs in the whole period. The distinct phone users, with an Italian phone contract (no roaming data of foreign people are included), are about 14 millions.

Spatially, the dataset covers the extended area of Rome as in Figure 6, i.e., it contains CDRs of the users who had been served by the cellular antennas inside this area. The case study focuses on five Administrative areas of Rome provided by the Mobility Office of the city (https://romamobilita.it/). As we will see in the next, we discovered and analyzed some events which took place in four *Points Of Interest* (POIs) within these areas. These POIs are shown in the zoom of Figure 6.

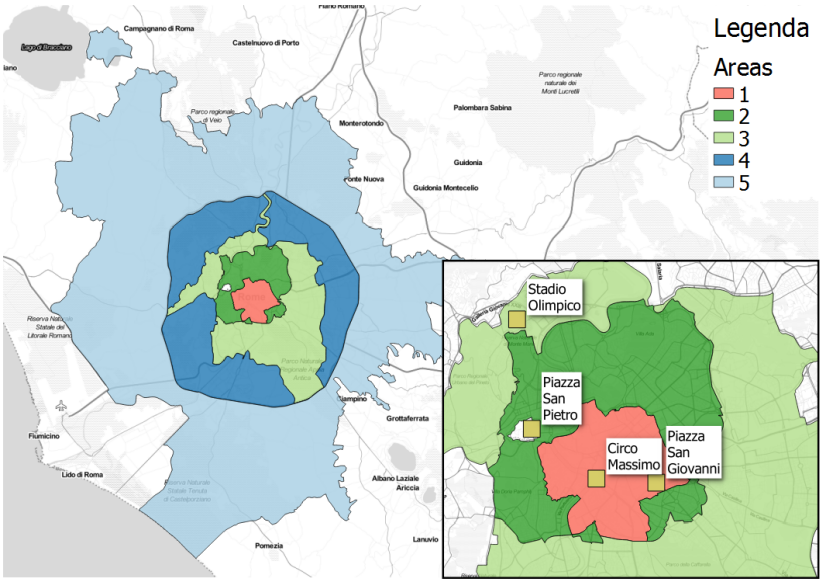

Figure 6. The area of Rome and the five administrative areas provided by Municipality of Rome. Area 1 is the center of Rome where the main touristic and historical attractions are located. Area 2 is mainly a residential zone where are also located many tertiary and business activities, as well as a sport center and a cultural area. Area 3 is residential but contains many social buildings and parks; Area 4 as a diversified settlement system, where the urban areas alternating natural reserves and agricultural areas. This area is bounded by a highway. Finally, the extended Area 5, which contains the rest of the municipality of Rome, includes the suburb of Rome and it is mainly agrarian.

The aim of the whole process we exploited is to provide a tool to perform a deep mobility analysis for a given territory. CDRs are the most valuable data source in terms of diffusion across population. Mobile phone activity is part of the every-day activities of, more or less, the whole population. Even considering the market share of a single telecommunication company, the available data still represents a very high percentage of the total population w.r.t. to other data sources commonly used in the literature. The distinction of users categories provides us with fundamental insights on the usage of a particular area and is a way to add a semantic layer to the total daily number of calls/presences in a certain area. In the next sections we describe how such additional semantic layer allows us to identify occurring events in a given area: e.g., an increasing of visitors, indeed, is useful to spot something unusual that is happening. Such an increase, tough, could not have a big impact on the total number of presences. Moreover, users classification is also the basis to compute OD matrices, useful for both

further event analysis and, more generally, for assessing the relationships between the different areas of Rome in terms of human mobility. Due to the sensitive nature of the data that has been used and that is contained in the various datasets, we have taken into account the privacy issues during the entire process of analysis customizing and applying the privacy risk analysis method presented in [17]. This methodology implements and satisfies the constraints issued by the European Union for data protection (Article 6.1(b) and (c) of Directive 95/46/EC and Article 4.1(b) and (c) of Regulation EC 45/2001. European Union for Protection of personal data) and follows the principle therein given.

4.2. The Analysis

The availability of the huge amount of CDRs described in Section 4.1, allowed us to carry out an extensive experimentation over the city of Rome, investigating how people use and live one of the biggest Italian cities. As we will show in the next, the efficient tools of analysis and algorithms we developed have been indispensable for the identification of interesting and hidden behaviors which would not otherwise emerge. Furthermore, the period under analysis is very interesting because Rome was the place of many religious, cultural and recreational events which attracted people from both the surroundings and more distant locations.

One of the first analyses that one typically can perform over CDRs is the calls distribution over a period of time. This analysis produces a first indicator of the phone traffic and (indirectly) of people presences. Nevertheless, as shown in Figure 7, the extracted information is not sufficient to highlight significant patterns able to tell something about the city dynamics. Statistical approaches on this kind of data are not really effective due to the complexity and multitude of dynamics of the city. Following the classical statistical analysis and extracting the *number of calls per day* (time series of Figure 7 (Left)), we ended up with quite trivial results. Even decomposing the calls distribution by using the administrative areas as shown in Figure 7 (Right), we did not obtain much more information except for a decrease in the call trend in the summer months. Nevertheless, no particular irregularities to be used as stimulus for further investigations, rise from these analysis. To further investigate, we decompose the time series into different components (one for each city users' category), similar to the wave decomposition in signal processing context [18], in order to analyze them separately with their hidden sub-patterns. In particular, we used this strategy to spot anomalies in the time series and to highlight events occurred in the area. To better appreciate the quality of results we show this analysis applied on four well known POIs of Rome: Piazza San Pietro (St. Peter's Square), Stadio Olimpico (Olympic Stadium), Circo Massimo (Circus Maximus), and Piazza San Giovanni in Laterano (St. John in Lateran's Square) (see Figure 6).

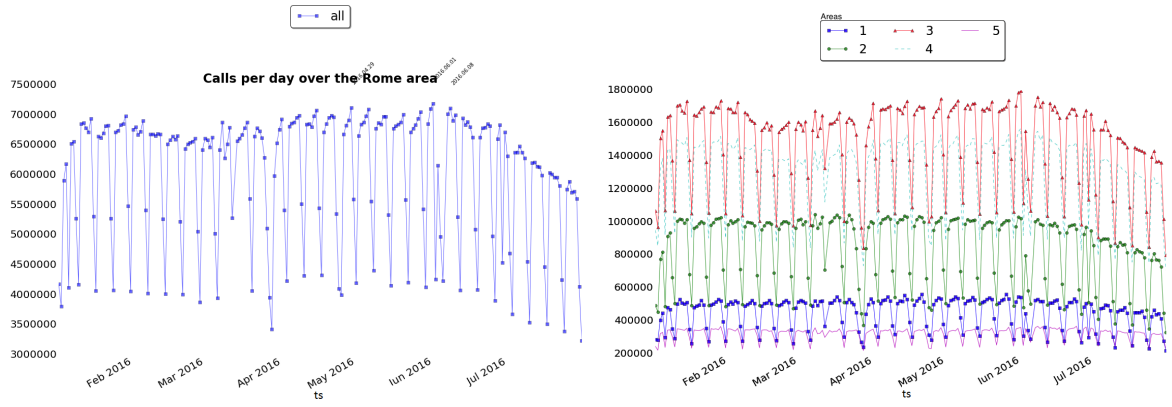

Figure 7. (**Left**) Distribution of the calls along the entire time window all over Rome, and (**Right**) for the five administrative areas, separately.

Piazza San Pietro is one of the most famous squares located in front of the St. Peter's Basilica in the Vatican City. It is the location where both weekly and special religious events take place all the time during the years and in particular this year dedicated to the Jubilee of Mercy. By running the Sociometer on these spatial contexts, e.g., Piazza San Pietro, we obtained the time series shown in Figure 8 (Top). Thanks to this first step, we can see how the residents dominate in terms of volumes, and how they hide the other categories which seems to have less effect on the area. To overcome this issue, we rescaled the distribution by normalizing it w.r.t. the typical distributions, i.e., by "subtracting" from the distributions the daily and typical presence patterns. The normalization procedure foresees the computation of the typical distribution of a week for each time series obtaining two values on a weekly basis for each day: avg^n, std^n. avg is the average number of distinct users for the n-th day of the week (0 = Monday, 6 = Sunday) and and std is the standard deviation of the same day. Using those values we rescaled the time series as follows:

$$v^d_{normalized} = \frac{v^d - avg^n}{std^n}$$

where n is the relative day of the week of the absolute day d. Such a normalization technique is motivated by the number of hourly presences over an area, that follows a Gaussian distribution. Hence, to identify interesting events we need to rely on mean and standard deviation, in order to discover unusual values of users presences for the areas we are analyzing.

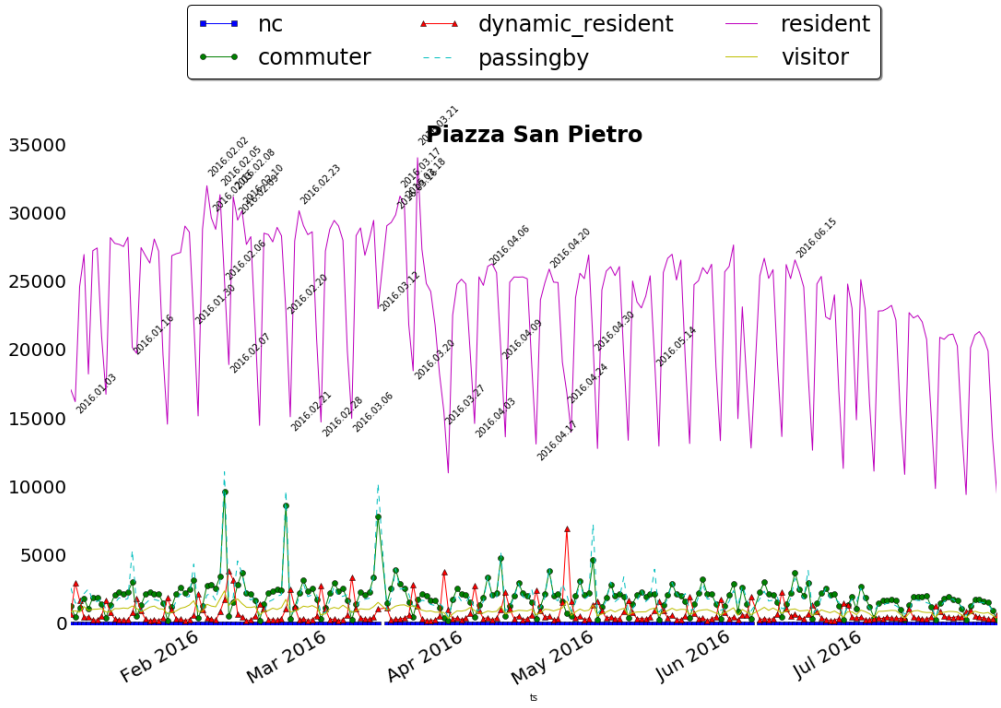

Figure 8. *Cont.*

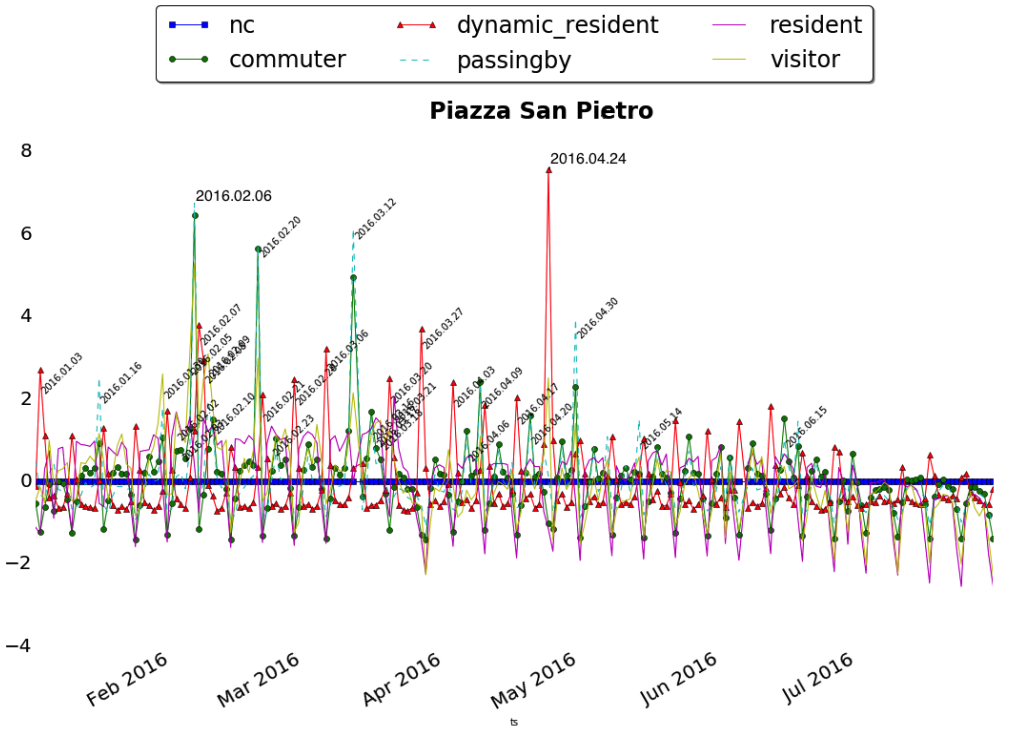

Figure 8. (**Top**) Distribution of people's presences by categories around Piazza San Pietro; (**Bottom**) the corresponding rescaled normalized distribution of people's presences by categories.

We also applied a post processing step on the class *Visitors* in order to distinguish the short term visitors and people in-transit, from the others. We called them *Passing by*, i.e., users who made a single call in all the period, and thus we registered their presence only for a single day. Clearly, this category is the result of a heuristic, and due to the nature of the data, we are not actually able to track the real presence but only register an "appearance" whenever a user performs a call. As stated this bias is related to the nature of the data and thus affects all the methods which aim at estimate presences of people starting from their call activities.

Globally, processing the dataset in we have annotated:

- 3M residents
- 3M dynamic residents
- 1M commuters
- 9M passing by
- 1.5M visitors

Figure 8 (Bottom), shows the normalized distribution of the presences. Now the deviations emerge and become clearer since the *typical* behavior, less interesting for our purpose, has been eliminated. It is important to notice that, if the normalization is applied on the original data without the Sociometer analysis, the average and the standard deviation values of the residents would eliminate the peaks discovered for the other classes.

Going further with the analysis, we can study the *peaks* which represent, in fact, unusual aggregations of people who share the same place at the same time for different purposes as for example for attending an event. By investigating the peaks we actually discovered that, in correspondence to those dates, very important religious events took place. In particular, considering Piazza San Pietro for example, we can study its peaks comparing the distribution of the classes of the typical day against the actual distribution during the event to verify if something changes.

Moreover, we compared also the actual day after/before the event (where no peaks are detected) with the day of the event. Figure 9 shows people's composition during the case corresponding to

the peak of 6 February, that we discovered to be the day of the arrival of Padre Pio's body at the Basilica. This day does not register a big increment of the residents (as one could image from Figure 8), but rather of new visitors arrived on purpose (visitors and passing by). We can also notice that the day after the event the composition of the population becomes, again, similar to the typical Sunday highlighting how the people's composition in the city returns to the normality after the big event.

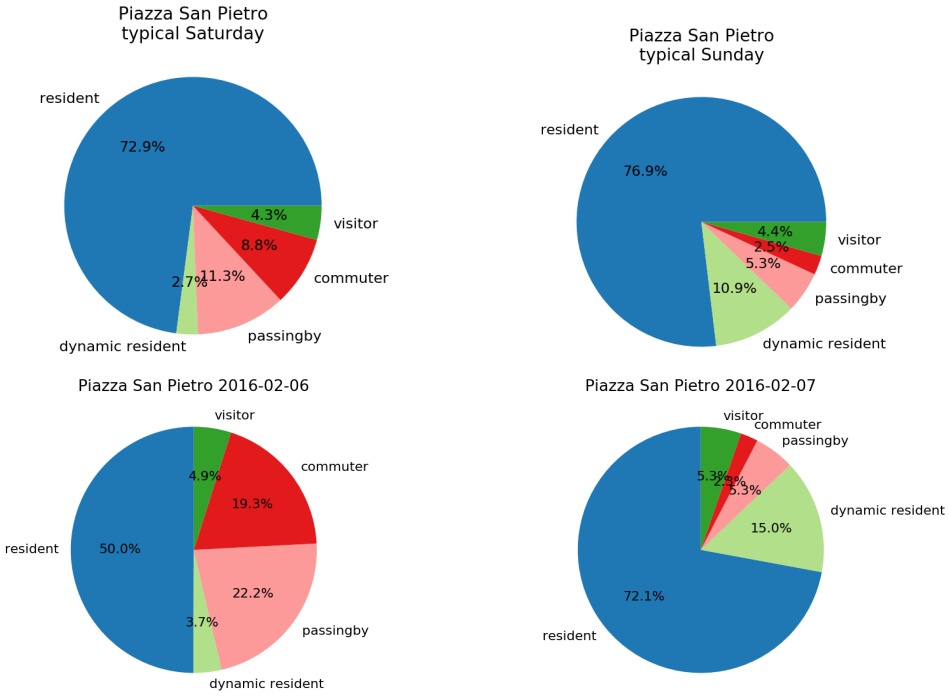

Figure 9. The comparison between the typical city people's composition (during a typical Saturday and Sunday), and the one on 6 February (where the peak appeared) and 7 February, at Piazza San Pietro.

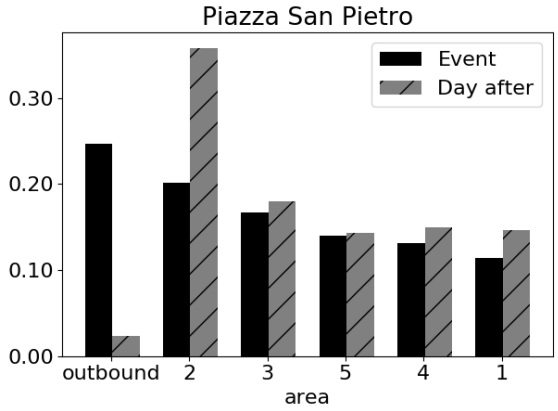

Figure 10. The size of flows coming to the event (i.e., 6 February: the arrival of Padre Pio's at St. Peter's Basilica) compared to a normal day (i.e., the day after the event).

To complete the analysis, we investigated the provenances of the attending people by reconstructing the OD matrix. We assigned as origin, one of the five administrative areas of Rome where each individual has been classified by the Sociometer as *resident*, and a further origin, *outbound*, has been used for the individuals coming from other locations (i.e., users which are not classified as resident in any of the five areas). From the OD matrix of Figure 10, we see that the day of the event

the majority of people come from outside Rome with an high reduction of the percentage of people coming from administrative area 2, while the day after the external visitors are very few (Figure 10) and the normal flow from administrative area 2 is restored.

Different is the case of 24 April, the day of the *Celebration of the Holy Mass with Pope Francis in Saint Peter's Basilica* during the *Jubilee for Boys and Girls*, which seems to be mainly a local event. In fact, the category of majority is dynamic resident registering a relative presence of 24.5% (Figure 11 (Right)) w.r.t. the typical daily presence which usually ranges from 2.7% to 10.9% (typical Sunday in Figure 9). Furthermore, the attendees come mainly from administrative area 2 and outbound flow is quite small as shown in Figure 12.

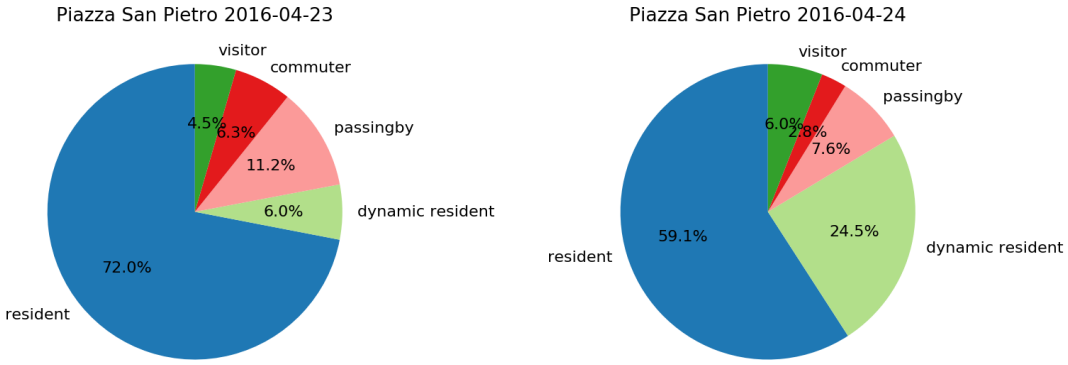

Figure 11. City people's composition (**Left**) the day before the Jubilee for Boys and Girls, and (**Right**) the day of the event at Piazza San Pietro.

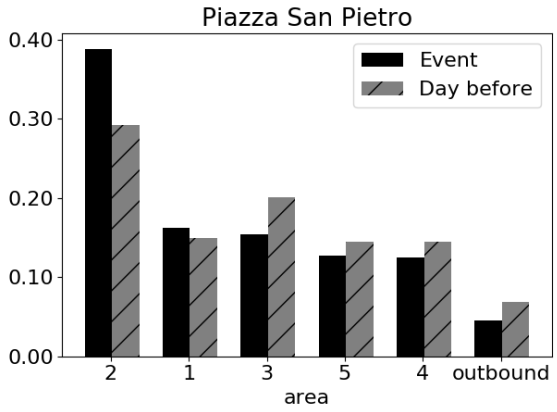

Figure 12. The size of flows coming to the event (i.e., 24 April: Jubilee for Boys and Girls) compared to a normal day (i.e., the day before the event).

In the area of the Stadio Olimpico, it is interesting to see how the month of May is characterized by a sequence of peaks as shown in Figure 13. Nevertheless, thanks to the Sociometer the fact that different classes of people are involved in different days and different events becomes clear. In particular, we focused on three of them: the *Tennis with Stars* on 9 May, a charity exhibition involving champions in tennis and football, the *International BNL tennis championship* on 10 May, an international event valid for the world tennis rating, and the *Italian Soccer Cup* on 21 May between Milan and Juventus Italian soccer teams. While the first tennis event affected especially local citizens (residents and dynamic residents), the tennis championship, as well as the soccer match, attracted fans mainly from the outside (passing by and visitors).

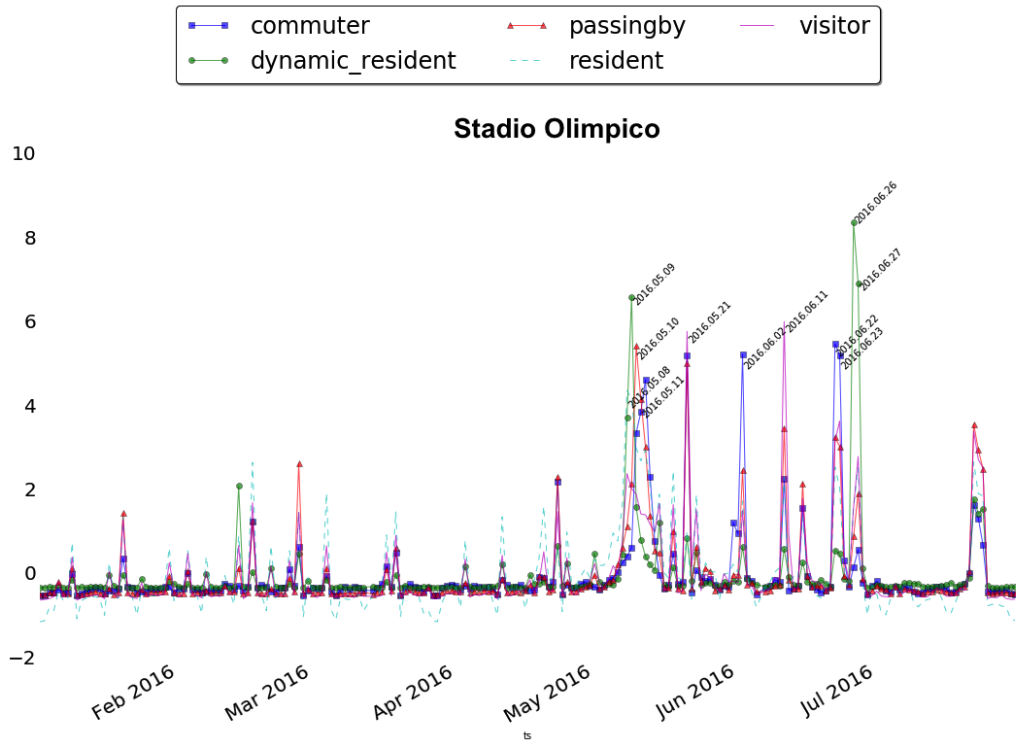

Figure 13. Time series for Stadio Olimpico (separated and normalized).

Circo Massimo is instead an ancient Roman chariot racing stadium, now become public park and used for big concerts and events. Figure 14 shows the anomalies registered in this area. The event on 30 January, *The family day* had impact especially on people never seen before (i.e., passing by), while other events like the *Good Deeds Day* (it is an event which unites people from over 75 countries to do good deeds for the benefit of others and the planet), the 10th of May, and *Race for the Cure* (an initiative dedicated to health, sports, welfare and solidarity), the 15th of May, had impact especially on local citizens. Interesting is the case of 16 July which registered an unusual peak of presences of all the categories. This corresponds to the Bruce Springsteen's Concert. As last case, we show in Figure 15, the anomalies registered around Piazza San Giovanni in Laterano. This square is situated just in front of the homonym Archbasilica and often location for social and political events. In the figure two main anomalies emerge. The first big one corresponds to 1 May, in Italy the *Labour day*. This event attracts especially local citizens but also many people from other parts of Italy who usually participate with organized tours. This is clearly stated by the fact that all the classes are involved but while the dynamic residents and residents are almost the same, representing the locals, the passing by are half of them. The second one on 7 May is another socio-political event organized by several Unions, Foundations and Organizations to stop the Trade Liberalization Treaty. This event, with evidently less local relevance, had impact especially of external visitors and passing by. The two peaks of 1 May and 7 May represent actually the two main events in the period of observation. Even if the normalization is applied (as explained in Section 4.1), the effect of a massive presence of people in these two days is predominant w.r.t the other minor public gathering.

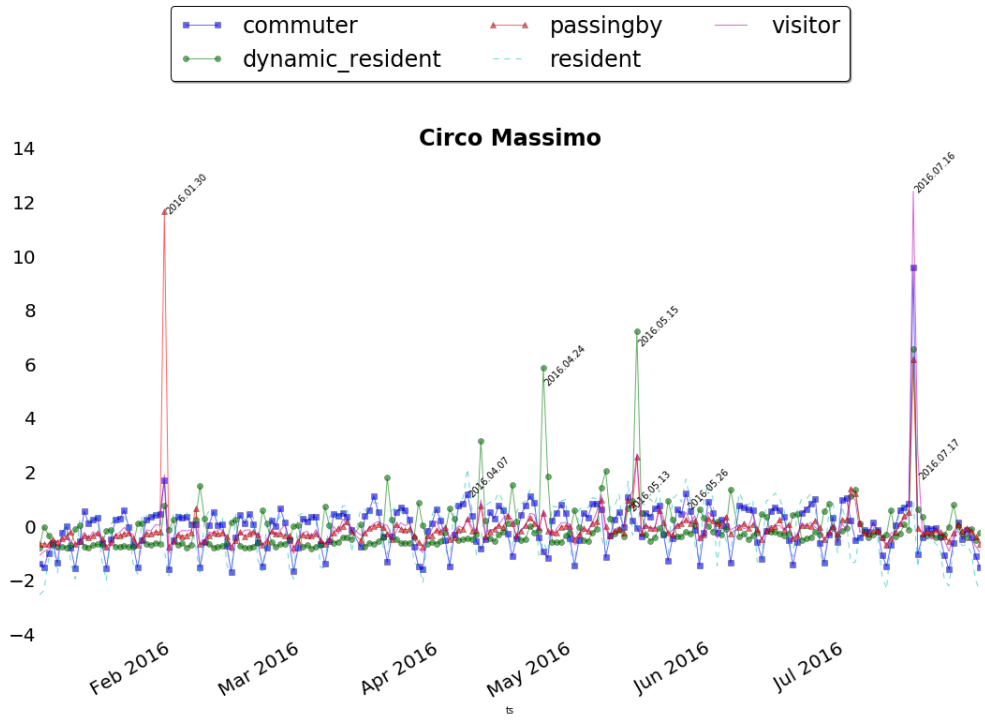

Figure 14. The time series for Circo Massimo (separated and normalized).

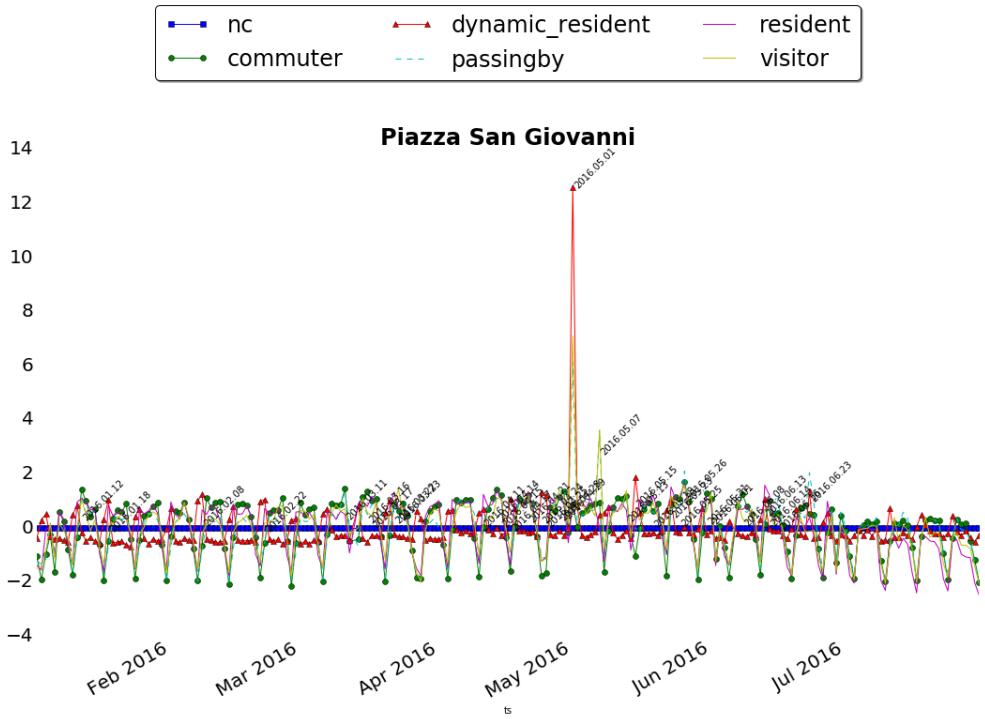

Figure 15. Rescaled distribution of people's presence by categories around Piazza San Giovanni in Laterano.

With these examples we wanted to highlight the fact that, thanks to the Sociometer, it is possible to discover events and characterize them in detail, as well as to understand their influence on people's composition in the city. Although in this case we focused on famous POIs and events in order to assess

the validity of the results, it has to be clear that the same process may be used to spot unknown events and it can be extensively applied on different locations or POIs of a city (e.g., train stations, universities, parks, and other touristic areas).

5. Validation and Empirical Evaluation of the Results

In this section we study the results obtained under two different points of view, the first considering the robustness of the stereotypes extracted by the Sociometer, and in a second moment we analyze the results against some ground truth. This kind of check over a long period of time is possible thanks to the efficient solution we implemented of the Sociometer and the availability of almost one year of data.

5.1. Stereotypes and Archetypes

Our objective is to prove the importance of using the mixed approach of extracting behavior from the data (bottom-up approach) and match it with the knowledge given by the experts (top-down approach). In particular we will show how the real behavior of the users coming from the data differ from the one the experts have in mind, in fact the first shows the complexity of the people habits represented by the calling behavior and the second is too perfect to be used directly. In other words the process shifts the *perfect* concepts of the the experts (archetypes) to the real ones coming from the data (prototypes). To this aim it could be useful to recall that an archetype is the abstract representation of a typical ICP provided by the domain expert, while a stereotype is a representative behavior extracted from the data after the clustering step.

To get a visual representation of the ICPs, due to the high number of dimensions of them, we used a multi-dimensional scaling technique on the set of prototypes generated every two weeks and the archetypes defined by the experts. Figure 16 (left), (right) shows this distribution in 2-dimensional space. It is possible to notice that the actual behavior in the data is very close to each other making the problem of classification hard, and that some archetypes are very far from them. Moreover, in Figure 16 (right) we only highlight the portion of space where prototypes are present, showing how *visitors* and *dynamic residents* are very close, while *commuters* and *residents* are more distinct. To deepen the analysis, we also performed a density based clustering technique on the different classes of prototypes comparing them with the original archetypes. In Figure 17 the archetypes are shown as vectors: starting from the same definition of the ICP, each six positions of the vector represent a week, where the first four represent the weekdays and the last two the weekend. In this way we can see how the archetypes are very regular and the variability of the same class archetypes is given only by small variations of the same general patterns (usually a constant in the weight).

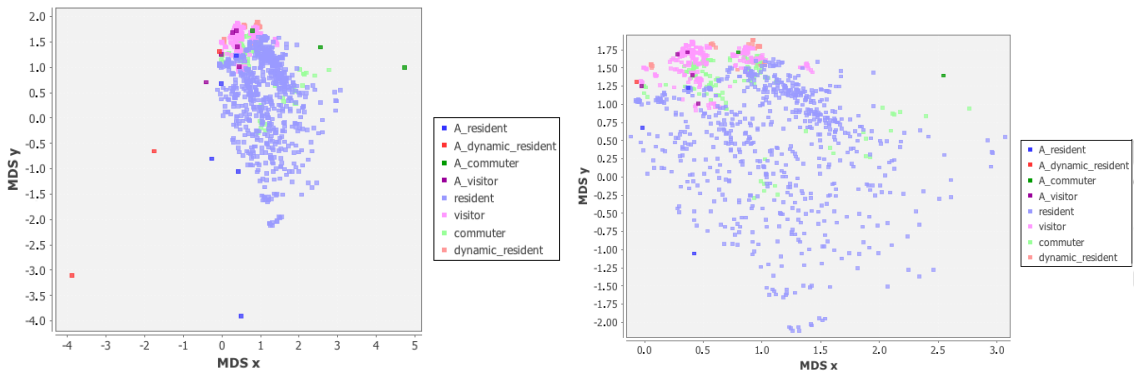

Figure 16. The distribution of archetypes and prototypes computed every 2 weeks in a 2-dimensional space.

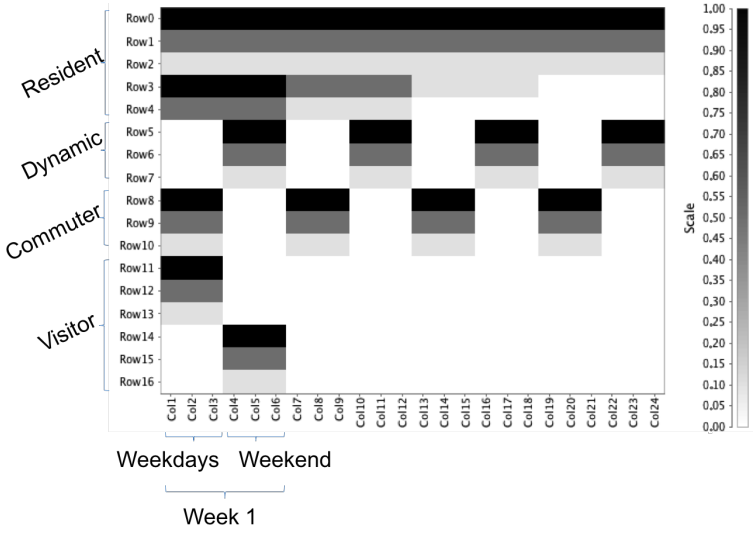

Figure 17. The vector representation of the archetypes defined by the domain experts.

We can say that if a prototype represents the behavior of a group of users in a specific time window, the prototypes clusters represent a more abstract global behavior. In practice, it is something similar to what the experts want to represent with the archetypes they provides. At a first look the results in Figure 18 show clearly how the reality if more complex than the conceptual view of the experts; in particular, by looking at the *resident* we can notice the variability of behaviors due to the fact that, using a specific area in a more intensive way, the variability of how the calls are performed varies significantly. More interesting is the case of the *commuters*: although both of the clusters contain the pattern represented in the archetypes, there is a marked variation in the first one where the users use the area during one of the weekend. We can explain this, with the fact that several types of shops have employees with a weekend shift of work each month. For the *visitor* class the results are closer to the archetypes, suggesting that visitors tend to stay longer than the weekend. Finally the *dynamic resident* class shows a more complicated situation than the one described by the experts, the results show that the general pattern followed by the users of this class tends to be very close to the *visitor* class sometimes (clusters 3 and 5) which was also highlighted by Figure 16. This kind of analysis will also be used as feedback to the experts in order to show the emerging reality from the data and maybe considering to redefine the archetypes.

5.2. Empirical Evaluation

This evaluation aims at comparing the information extracted from the CDRs in the area of Rome, with official data provided by the the Public Administration of Rome. The latter are essentially surveys data and traffic data collected by using sensors and processed to produce reports about the population and its mobility. To remain consistent with the previous analysis, we use here the same spatial partition presented in Section 4.1.

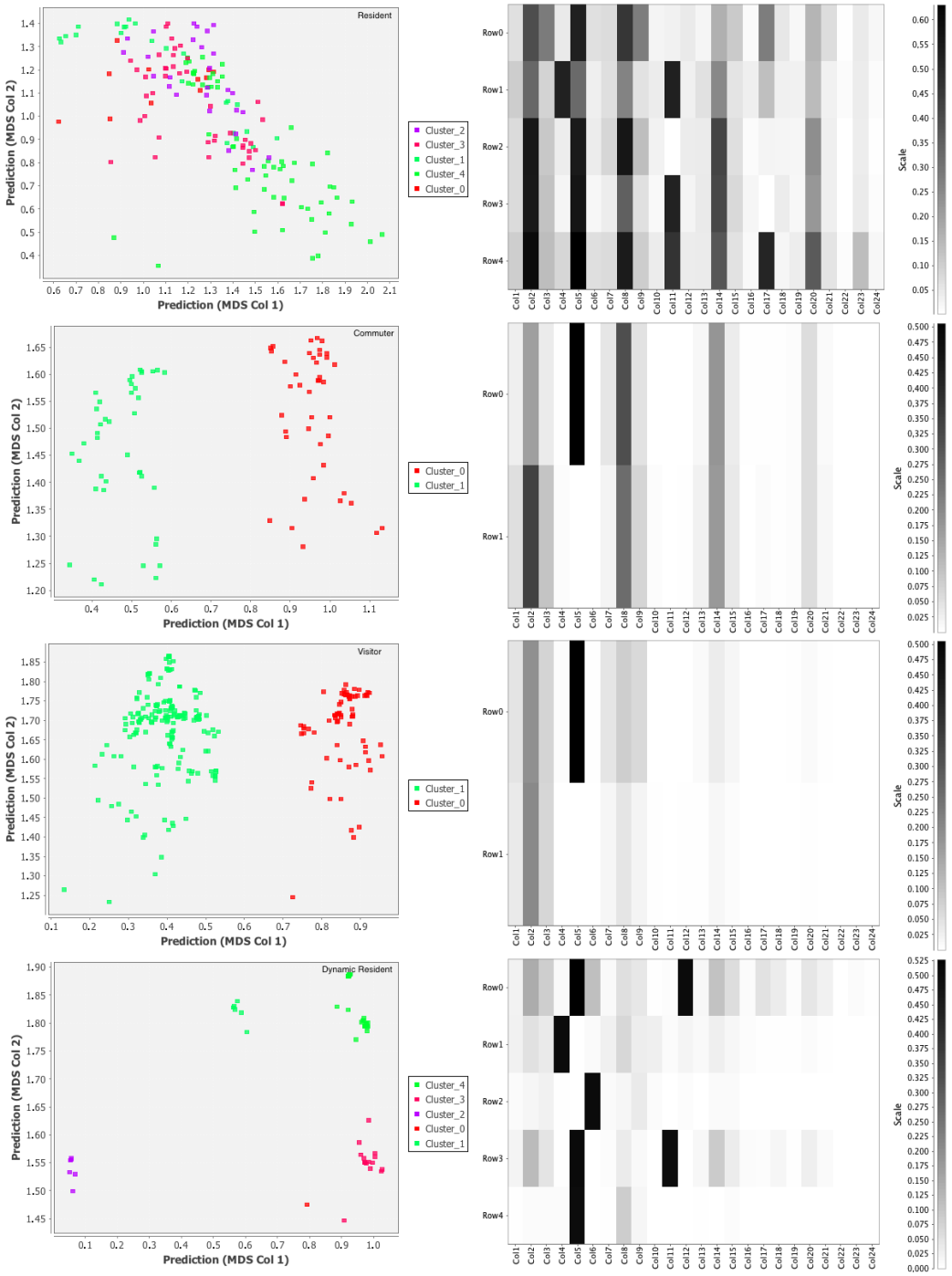

Figure 18. Prototypes clusters and their global behaviors.

Figure 19 shows the residents in the five areas according to the official data (first column), compared with (i) the number of users labeled as resident or dynamic residents by the Sociometer and (ii) the number of presence computed using only the calls. It is evident that there are differences between the official data and the other two statistics, but this is due to the fact that only a portion of the population is captured by a single telecommunication company. Anyway, using only the calls as indicator for the residents is not sufficient, in fact there is an overestimation in each area which vary between 1.88 and 8.23 times the official data. Since the Sociometer provides separate statistics for the different categories, we are able to clean the data removing the commuters and visitors and leading to a more realistic indicator and maintain a better ratio w.r.t. the official data (i.e., a value between 0.58

and 1.46). Nevertheless, the administrative area 1 remains overestimated in both cases. Showing the results to the experts (Interview to Giuseppe Sindoni: responsible of population census at the Italian Statistical Bureau (ISTAT)—https://goo.gl/zgA47C), they explained that estimating people's presence in administrative area 1 is very difficult because of a high percentage of people who are not *visible* by standard official ways and therefore underestimated by official statistics. In the light of this statement, we are aware that mobile phone data can be a good proxy for the estimation of the actual population. Unfortunately, at the moment, no additional data are available to make a complete analysis and to support this thesis further.

Area	Official population	Sociometer population	User Presence	Sociometer/Official	Presence/Official
1	107,247	156,318	883,243	1.46	8.23
2	346,215	281,031	1,131,858	0.81	3.26
3	971,467	431,170	1,325,726	0.44	1.36
4	658,308	392,618	1,268,063	0.59	1.92
5	634,446	373,009	1,197,277	0.58	1.88

Figure 19. Comparison between the official census from 2011, number of residents and dynamic residents labeled by the Sociometer and the number presence (users performing a call).

6. Conclusions

In this paper, we tackled the problem of identifying events at city level through the observation of a big mobile phone dataset for a long period of time. By applying data mining and statistical methods we isolated the important happenings, and we studied their impact on people composition and city dynamics. The availability of a large amount of CDRs and the more and more urgent request to analyze these data stimulated us to implement efficient algorithms and analytical processes able to cope with such Big Data. In particular we re-implemented a new version of the Sociometer based on Spark technology allowing us to make it usable in a big city context.

By means of an extensive case study on the city of Rome, we proved the ability of the Sociometer in managing these large dataset maintaining the classification accuracy. Furthermore we demonstrated its usefulness for identifying anomalous presences and distinguish between different events at urban level. We focused on four points of interest in the city of Rome, and by separating the presences of different people categories and bringing out real anomalies, we showed how occasional events can emerge from the mass of the systematic ones.

From this study we understood that some kind of events attract especially local citizens (as for example the socio-political events), but others tend to mainly attract people from outside (as the football matches or the big singer's concerts). It is not a trivial result that we identified, among all the events involving the considered POIs, which were the most significant in terms of impact of presences providing an interesting quantitative and qualitative (i.e., demographic) analysis. For some POIs we found that only few events were actually very impactful in terms of attendance, as the case of Piazza San Giovanni. Other POIs, like Piazza San Pietro and Circo Massimo, even though they are highly-frequented places all over the year, hosted instead many important events which can be isolated from the minor ones, and properly studied. This ability to select and distinguish among less or more interesting events represents a valuable contribution of the approach. For the Public Administration the estimation of people attending a public event is often very hard when no form of ticketing is provided, and this method gives a valuable support in this context.

A validation of the results coherence is performed considering the stereotypes extracted in the various time-windows together with an empirical evaluation with public available statistics. The feedback from the Local Administration of Rome and the Italian Statistical Bureau was decisive for the refinement of the process, and their interest in this kind of analysis made us more and more aware

of the necessity of putting the big data analysis besides the traditional statistical tools and survey to support the decision making process.

Acknowledgments: This work has been partially funded by the EU under the FP7-ICT Program: Project ASAP (https://www.asap-fp7.eu/), and under H2020 Program: Project SoBigData (http://www.sobigdata.eu/).

Author Contributions: Barbara Furletti and Roberto Trasarti conceived and designed the experiments and analyzed the results; Paolo Cintia and Lorenzo Gabrielli implemented the algorithms and performed the experiments; Everybody wrote the paper according to their contributions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lane, N.D.; Miluzzo, E.; Hong, L.; Peebles, D.; Choudhury, T.; Campbell, A.T. A survey of mobile phone sensing. *IEEE Commun. Mag.* **2010**, *48*, doi:10.1109/MCOM.2010.5560598.
2. Quercia, D.; Lathia, N.; Calabrese, F.; Lorenzo, G.D.; Crowcroft, J. Recommending Social Events from Mobile Phone Location Data. In Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010; pp. 971–976.
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.* **2011**, *12*, 141–151.
4. Eagle, N.; Pentland, A.; Lazer, D. Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15274–15278.
5. 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, San Diego, CA, USA, 21–24 August 2011.
6. Furletti, B.; Gabrielli, L.; Garofalo, G.; Giannotti, F.; Milli, L.; Nanni, M.; Pedreschi, D.; Vivio, R. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. In Proceedings of the 47th Meeting of the Italian Statistical Society (SIS 2014), Cagliari, Italy, 11–13 June 2014.
7. Barcaroli, G.; De Francisci, S.; Scannapieco, M.; Summa, D. Dealing with Big data for Official Statistics: IT Issues. In Proceedings of the Meeting on the Management of Statistical Information Systems (MSIS 2014), Dublin, Ireland, 14–16 April 2014.
8. 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 Proceedings of the D4D Challenge—3rd Conference on the Analysis of Mobile Phone datasets (NetMob 2013), Cambridge, MA, USA, 1–3 May 2013.
9. Andrienko, G.; Andrienko, N.; Fuchs, G. Multi-perspective analysis of D4D fine resolution data. In Proceedings of the Orange D4D Challange 2013, Cambridge, MA, USA, 1–3 May 2013; pp. 383–396.
10. Ratti, C.; Sevtsuk, A.; Huang, S.; Pailer, R. *Mobile Landscapes: Graz in Real Time*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 433–444.
11. Calabrese, F.; Ferrari, L.; Blondel, V.D. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Comput. Surv.* **2014**, *47*, 25.
12. Gariazzo, C.; Pelliccioni, A.; Bolignano, A. A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic. *Atmos. Environ.* **2016**, *131*, 289–300.
13. Furletti, B.; Gabrielli, L.; Trasarti, R.; Giannotti, F.; Pedreschi, D. City users' classification with mobile phone data. In Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 1007–1012.
14. Gundogdu, D.; Incel, O.D.; Salah, A.A.; Lepri, B. Countrywide arrhythmia: Emergency event detection using mobile phone data. *EPJ Data Sci.* **2016**, *5*, 25, doi:10.1140/epjds/s13688-016-0086-0.
15. Dong, Y.; Pinelli, F.; Gkoufas, Y.; Nabi, Z.; Calabrese, F.; Chawla, N.V. Inferring Unusual Crowd Events from Mobile Phone Call Detail Records. In *Machine Learning and Knowledge Discovery in Databases, Proceedings of the European Conference (ECML PKDD 2015), Porto, Portugal, 7–11 September 2015*; Appice, A., Rodrigues, P.P., Santos Costa, V., Gama, J., Jorge, A., Soares, C., Eds.; Springer: Cham, Switzerland, 2015; pp. 474–492.

16. Giulio Barcaroli, P.R. Big Data in Official Statistics: Ongoing Work at Statistics Italy. Available online: https://ec.europa.eu/jrc/sites/jrcsh/files/j3-bigdata-03-Barcaroli_Righi.pdf (accessed on 1 January 2017).
17. Basu, A.; Monreale, A.; Corena, J.C.; Giannotti, F.; Pedreschi, D.; Kiyomoto, S.; Miyake, Y.; Yanagihara, T.; Trasarti, R. A Privacy Risk Model for Trajectory Data. In Proceedings of the Trust Management VIII—8th IFIP WG 11.11 International Conference (IFIPTM 2014), Singapore, 7–10 July 2014; pp. 125–140.
18. Soltani, S. On the use of the wavelet decomposition for time series prediction. *Neurocomputing* **2002**, *48*, 267–277.