FP7 Project ASAP
Adaptable Scalable Analytics Platform

ASAP D9.4

Results of the performance and scalability study

WP 9 – Applications: Telecommunication Data Analytics

Nature: Report

Dissemination: Public

Version History

Version	Date	Author	Comments
0.1	2/2/2017	PC, VR, RT	Draft
0.2	16/2/2017	RB, PC, VR, RT	Draft
1.0	21/2/2017	RB, PC, VR, MRS, RT	Draft
1.1	27/2/2017	RT, RB	FORTH Review
Final	8/3/2017	RB, PC, VR, MRS, RT	Final Version

RB - Roberto Bertoldi - WIND
PC - Paolo Cintia - WIND
RT - Roberto Trasarti - WIND
MRS - Maria Rita Spada - WIND
VR - Vittorio Romano – WIND

Acknowledgement This project has received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement number 619706.

Executive Summary

This Deliverable D9.4 *"Results of the performance and scalability study"* provides a report and on the results of the evaluation study on the prototype and of the final version of the application. The performance has been evaluated in terms of query execution time, number of parallel queries supported and the scalability in terms of computation nodes used while achieving speedup. The target performance for the application has been to batch-analyze data collected over days in hours, and data from hours in minutes. This deliverable will also reports on the compliance with data privacy regulations. This deliverable includes also analytical results obtained during the evaluation study.

Contents

List of Figures

1 WP9 Introduction

The main objective of this Work Package (WP) is the design and development of an analytics application on Wind Telecommunications customer data (TDA), targeted towards tourism and mobility scenarios.

1.1 TDA Goals and Requirements

The ASAP Telecommunication Data Analytics (TDA) application will show how a number of analytical services describing the mobility of people can be created on the basis of the data collected by Wind's mobile network during routine operation. In particular three applications will be targeted:

- Event Detection
- Tourism Observation and Analysis
- Ride Sharing

The first two are directed to a manager who can take decision considering a set of indicators, and the third which represent an application for a recommendation system for the telecommunication company final user.

The envisaged use cases will be integrated into the ASAP framework and will be evaluated using several measurement methods. The application will include the development of the proposed analytics tools. Starting from the envisaged use cases some of the subtasks are developed and integrated in the ASAP platform.

In order to fulfil the goals the foreseen telecommunication data analytics (TDA) application will have the following features: an engine for the statistical analysis of the data developed in Spark, a visualization API used to query the dataset and a display, on a graphical dashboard, of the desired information required by the data-scientist/mobility-manager (e.g. a spatial representation of the traffic patterns mined, and an extraction instrument that extracts mined data into different formats for further analysis). In particular, the implementation and outcome of the process will be evaluated in terms of:

- Performance
- Manageability
- Dashboard Usability

In particular the event detection application is designed to let the mobility manager/marketing analyse different features of an event: spatio-temporal characteristics, social aspects and statistical properties. To do this what is needed is a dashboard with the following features:

- a geographic map on which the user can select the area of interest and that can be used to:
    - Visualize peaks in an interactive chart
    - Display/highlight of correlation patterns through color coded markers that reflect social diversity (refer to *Sociometer*)
    - Tooltips to show additional metadata
    - Distribution of user classification

- advanced search dialog and date range selector used to:
    - Visualize/select specific events on filtered CRM data, e.g:
      "visualize all the events for the period 1-15 January 2015 related to people in the age group 18-25 years old".

The deliverable will present the process behind the construction of such dashboard which will be described in details in Deliverable D6.4 *"ASAP Dashboard"* [1].

1.2 Task 9.4 - Evaluation

Following the former activities on Task 9.1 where the datasets used by the ASAP partners and based on Wind's CRM and CDR data have been defined, Task 9.2 where the requirements for the TDA use case have been defined, Task 9.3 where the analytics application that realizes the use case of task T9.2 has been developed, the conclusive Task 9.4 that produces this Deliverable D9.4 has the objective to design and conduct evaluation studies with the aim of measuring the performance and scalability of the developed method, as well as quality of the mined patterns. In particular expert in the fields have been involved in the process to understand the usefulness of the methods, the quality and the validity of the results.

2 TDA Use Cases Development

In the context of the ASAP project, a new application has been designed and implemented to take better advantage of the new big data approach for mobile applications. The ASAP telecommunications application (TDA) will show how a number of analytical services describing the mobility of people can be created on the basis of the data collected by Wind's mobile network during routine operation.

From this initial design of the application described in detail in Deliverable D9.2 *"Use Case Requirements"* [2], we selected a subset of modules to be developed: the *user profiling*, the *Sociometer* and the *peak detection*. Moreover some additional modules are defined and developed in order to build complete workflows going from the data to the publication of the results as shown in Figure 1.

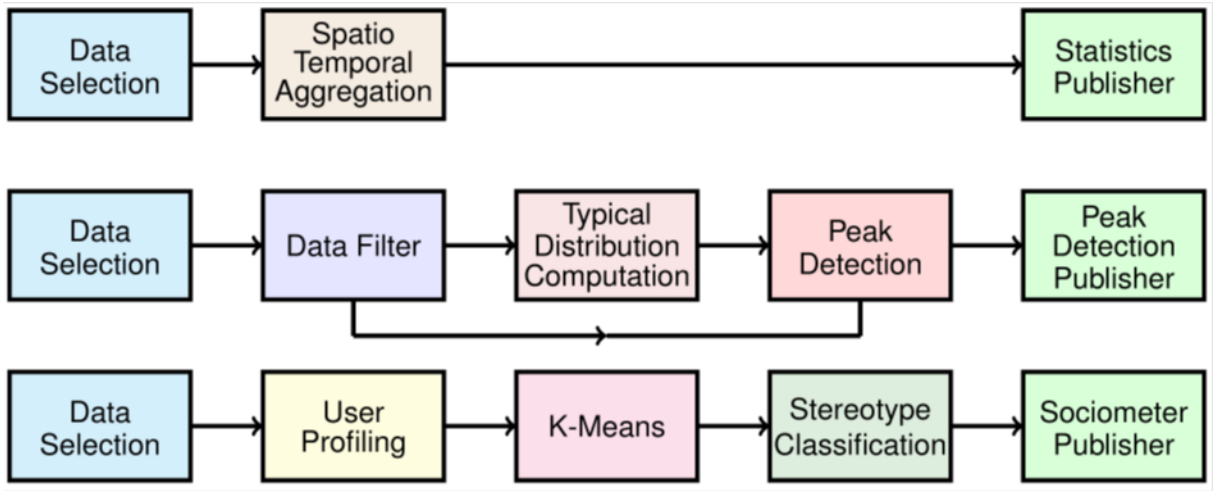

Figure 1: The workflows of the Event Detection and Tourism applications.

It is important to notice that all the modules start assuming that the data is available. This means that a phase of extract, transform, load (ETL) and data processing has already been executed in the workflow and is not the focus of this document.

2.1 Data Management

In this section we will briefly describe the kind of data used by the TDA, the processes used to extract and transfer the data stored in the Wind servers to the platform as well as the techniques used to guarantee the privacy of the users.

2.1.1 Call Data Records (CDR)

The cellular phones are probably the most popular devices we carry everywhere nowadays. Since mobile phones functionality is based on the communication to an antenna covering a local area, the active connection (e.g. a call or SMS) to a certain antenna represents a spatio-temporal position information of the user. This information, as collected by Wind, provides a spatio-temporal fingerprint of the users moving in an area covered by mobile telecommunication services. An example of call data record (CDR) is reported in Figure 2. The call started in *cell1* and ended in *cell2* will result in a single row in the log as follows (as reported in the Deliverable D9.3 *"Specification and early prototype"* [3]):

<id32876, 10/10/2015 10:00:00, 5:00, Cell_1, Cell_2, VOICE>

containing the id of the user, the type of the event, when the event started, the initial cell, the end cell and the duration.

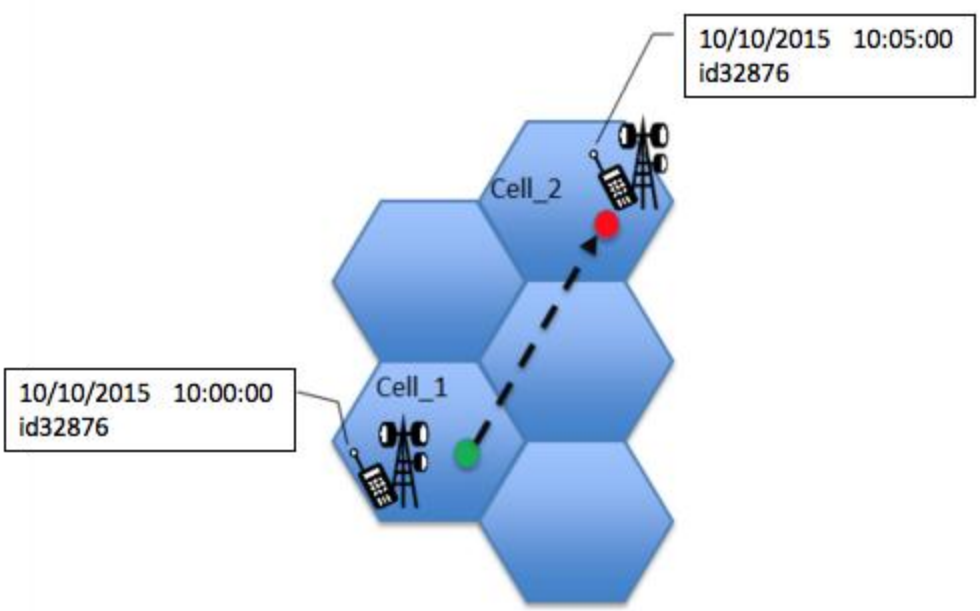

Figure 2: Example of CDR log produced by user activity.

This type of data raises three considerations related to privacy and scalability aspects: the first one is that mobile telecommunications traffic data embeds a great deal of information about the user habits and behavior; the second one is that mobile telecommunications traffic data, when collected from the provider side, comes in large volume thus posing new challenges, not only from the collection and storage point of view, but also from the analysis and mining side; thirdly, the analysis of this large volume of personal data poses several privacy issues.

2.1.2 ETL Procedure

A continuous flow of data from the users is stored in Wind's data warehouse (DWH) which comprises multiple databases containing different types of user's data (CDR, CRM). The datasets being used in the ASAP project are a result of the combination of various datasets (CDR, CRM) which are extracted following the ETL process depicted in Figure 3. This ETL process takes also into account privacy aspects of the data being extracted and the resultant datasets comply with current privacy rules and regulations. The dataset contains all the call data records registered in the region of the City of Rome for several months. Each month of traffic activity (50 GB of data in the average) corresponds to about 5.6 million lines of records per day. The amount of data being extracted monthly implies that by the end of the ASAP project the size of the accumulated datasets will be about 1 TB.

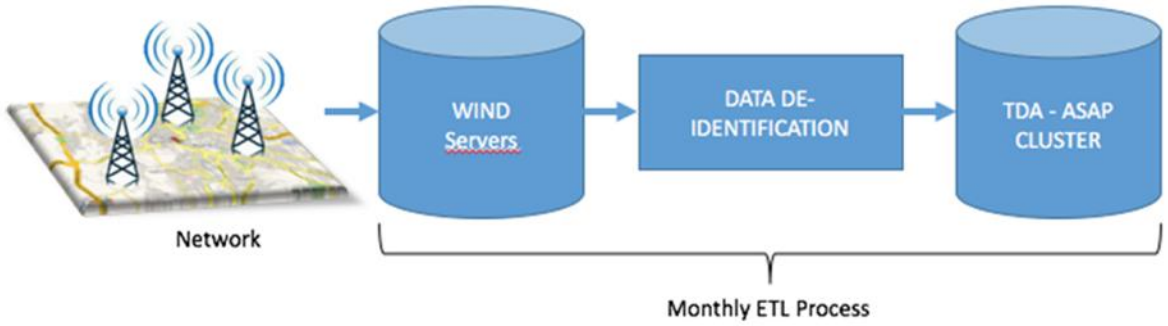

Figure 3: The implemented ETL process to extract the data from Wind servers.

2.2 TDA Workflow

In this section a more detailed description of the sub-tasks and the modules is presented.

Spatio-Temporal Aggregation: The CDRs are aggregated in space and time in order to obtain a time series of the number of calls or SMS. This basic statistic is very useful to have an overview of the data feeding the platform without disclosing the raw data.

Peak Detection: With this analysis we want to detect relevant peaks representing an event. Comparing the density of population within a region in a given moment against the expected density for that area at that hour of the day can do this. This process is part of the methodology presented in [4] In detail it is realized by means of two modules: typical distribution computation and peak detection.

User Profiling: The spatio-temporal profile is an aggregated representation of the presence of a user in a certain area of interest during different pre-defined timeslots. This profile is constructed starting from the CDR and with reference to a particular spatial representation. The CDR spatial coverage describes the distribution of the mobile antennas on the territory, which can be used to estimate the corresponding coverage. A spatio-temporal profile codes the presence of a user in the area of interest in a particular time (or timeslot) identified by the information in the CDR.

Sociometer: Exploiting the methodology called *Sociometer* it is possible to classify the users using the presence of the users in the network cell. Once the profiles have been created, the *Sociometer* classifies them implementing a set of domain rules that describes the mobility behavior categories: *resident*, *dynamic resident*, *commuter*, *visitor* or *passingby.*

Publisher Modules: as can be seen in Figure 1 each workflow ends with a publisher module which is able to produce a report of the result obtained. In particular they translate the results in a JSON[1] or GeoJSON[2] format readable by the webLyzard visualization tool.

[1] www.json.org
[2] www.geojson.org

A more detailed description of the modules is given in Deliverable D9.3 *"Specification and early prototype"* [3].

2.3 Data Anonymization Procedure

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 [5] and already tested in the work presented at CPDP in 2013 [6]. This methodology implements and satisfies the constraints issued by the European Union for data protection in [7] and follows the principle therein given.
In summary the risk analysis follows the idea that, given a dataset and a specific application, it is possible to define the set of possible attacks with regard to different levels of knowledge.
A more detailed description of the process is given in Deliverable D9.3 *"Specification and early prototype"* [3].

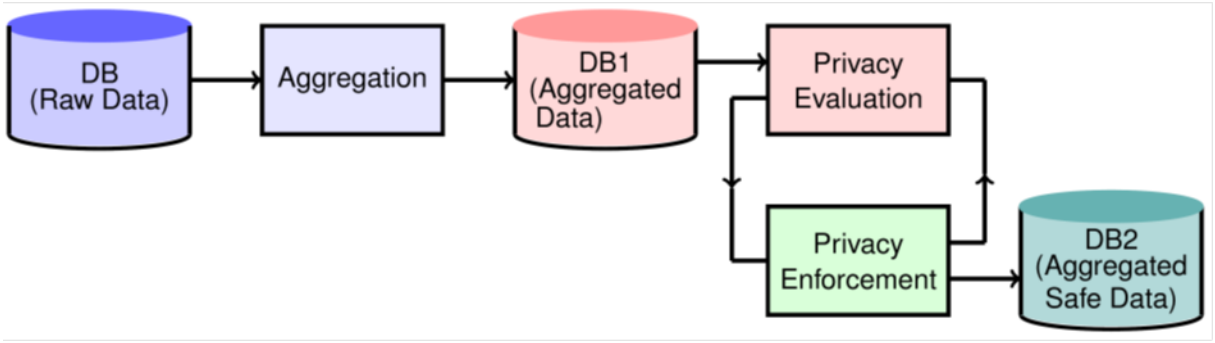

Figure 4: The Privacy Risk framework.

As shown briefly in Figure 4 the data is firstly aggregated to fit the requirements of a specific application, then the privacy is evaluated in order to detect the portion of data having a high risk and finally that data is deleted or transformed to generate the safe data.

The implementation of the methodology presented in [8] allows us to check the privacy risk over the Wind dataset and the results are reported in Figure 5. The results show that, considering an acceptable risk of 10% (a user is not distinguishable in a group of 10 users) the 80% of the profiles are safe. The other 20% of user profiles lead to a real risk on the privacy of the users, and therefore a privacy enforcement method must be applied on them.

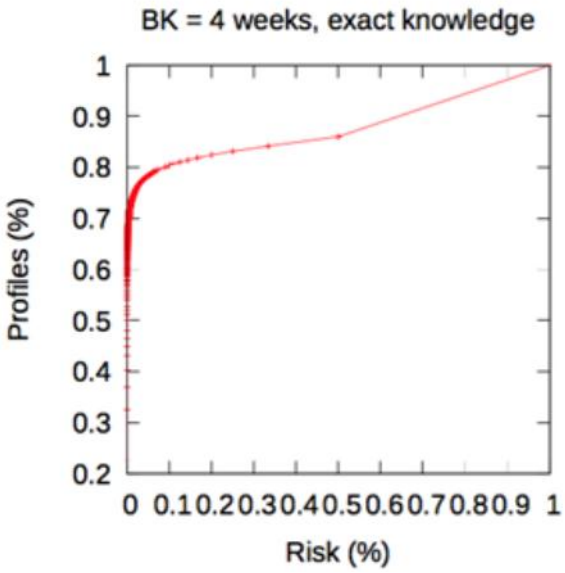

Figure 5: Cumulative curve of the privacy risk in disclosing users profiles computed from Wind Call Data Records (City of Rome, November 2015).

The most radical solution is to delete them. The real effects of this deletion on the results of the *Sociometer* are evaluated empirically with good results, but a formal definition of them is still under study as well of other enforcing methods in substitution of the simple deletion.

2.4 Integration with ASAP Platform

All the information about the integration with the ASAP platform and IReS are reported in the Deliverable D7.3 *"ASAP System Prototype"* [9] as well as the integration with the visualization portal.

3 Analytical Results

In this section we present the latest status of the core algorithm and last analytical results, moreover an evaluation in terms of quality of the results and computational cost is provided.

3.1 Scaling up to Big Data

The limitations of the sequential version of the *Sociometer* described so far can be overcome by re-engineering the system and adopting engines for large-scale data processing. In this section we describe the distributed version of the *Sociometer* able to analyze Big Data in a scenario which evolves in time.

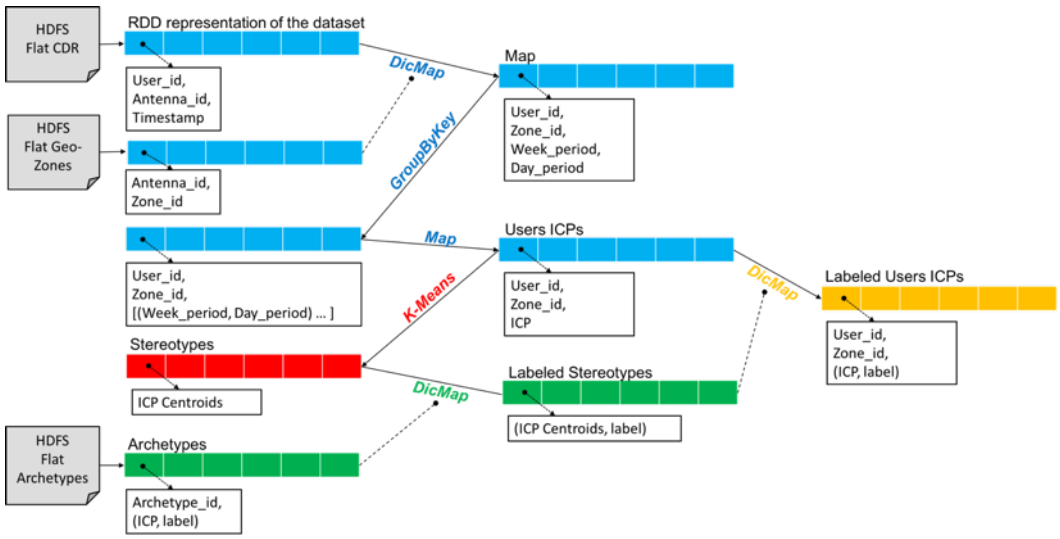

Figure 6: The Spark process: (i) ICP Building process (blue), (ii) K-Means application (red), (iii) Prototypes Labelling (green), and (iv) Label Propagation (orange).

Figure 6 shows the workflow of the process realized using Spark primitives. At the beginning, 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 loads also 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). 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 if the call is performed during the weekdays or weekend (Week period); (ii) a value indicating in which period of the day the call started (Day period between 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. By using the K-Means implementation provided by Spark, the set of centroids are computed obtaining the Prototypes which have the same representation of the ICPs. After that, the system loads the archetypes and uses them for labeling the Prototypes based on the proximity criterion. The same procedure is followed for classifying the "new" (and unlabeled) ICPs.

The redesigning 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-Spark2 reducing the complexity to

O(uidk/m), (iii) the Prototypes Labeling becomes *O(ka/m)*), and finally (iv) the Label Propagation is *O(ku/m)*. This new implementation guarantees that all the tasks complexity is linearly reduced with the number of available nodes.

3.2 Rome Area Case Study

In this section we will describe the methodologies and the analyses performed on the Rome area.

3.2.1 Event Detection and Tourism Analysis

The dataset, composed of CDRs, covers a period of seven months between January 1st and July 31st, 2016. Each day contains data for 1.2 GB, for a total size of 350 GB of call records in the whole period. The distinct phone users, with an Italian phone contract (no roaming data of foreign people are included), are about 14 million. Spatially, the dataset covers the extended area of Rome as in Figure 7, i.e. contains CDRs of the users who had been served by the cellular antennas inside this area, during making calls.. The case study focus on five Administrative areas of Rome provided by the domain expert of the Municipality of Rome. These areas are commonly used by the Local Administration as "upper" districts. Area 1 - Mura Aureliane (Aurelian Walls) is the center of Rome where the main touristic and historical attractions are located. Area 2 - Anello Ferroviario (Rail Ring) 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 - Secondo Sistema Anulare (Second Ring System) is residential but contains many social building and parks; Area 4 - Grande Raccordo Anulare (GRA) (is a big highway around the city) has a diversified settlement system, where the urban areas alternating natural reserves and agricultural areas. This area is bounded by a big ring road. 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 experiments, besides proving the ability of the new *Sociometer* in handling big data, aim at showing its usefulness in the identification of different types of city users, and how they use and live the city. In particular, we will show how the *Sociometer* can help in the hard task of recognizing events through the analysis of the different categories of people. The analytical process we followed is inspired to the one used in [10] for the analysis of the city of Paris, but in this case the availability of such a big data allows us to improve our knowledge of a big city and to exploit the *Sociometer* in its analysis.

The availability of the huge amount of CDRs, 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.

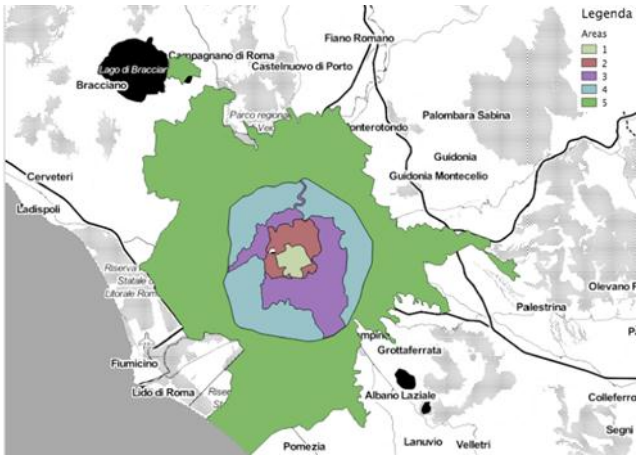

Figure 7: The area of the City of Rome and the five Administrative Areas.

One of the first analysis 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 6, the extracted information are 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 the complexity and multitude of dynamics of the city. As we will show in the next the classical statistical analysis and extracting the number of calls per day [time series of Figure 8 (left)] ended up with quite trivial results. Even decomposing the call distribution by using the administrative areas, see Figure 8 (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 investigate further, we used the *Sociometer* to decompose the time series into different components, similar to the wave decomposition in signal processing context [11] in order to analyze them separately with their hidden sub-patterns. In particular, we used the *Sociometer* to spot anomalies in the time series and to highlight events occurred in the area. To better appreciate the quality of results we applied this analysis to four well known Points of Interest (POIs) of Rome: San Pietro square (St. Peter's Square), Olympic Stadium, Circo Massimo (Circus Maximus), and San Giovanni square (St. John in Lateran's Square). San Pietro Square is one of the most famous square located in front of 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".

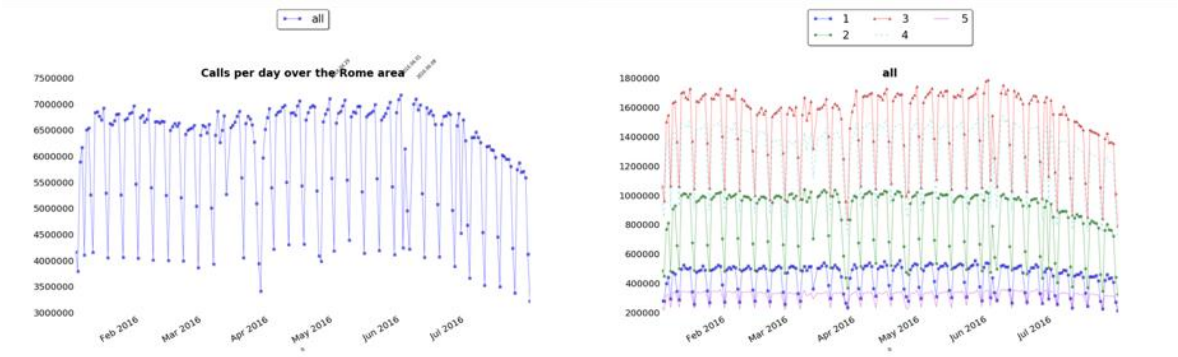

Figure 8: (left) Distribution of the calls along the entire time window all over Rome, and (right) for the five Administrative Areas, separately.

By running the *Sociometer* on these spatial contexts, e.g. San Pietro square, we obtained the time series shown in Figure 9 (top). Thanks to this first step, we can see how the residents dominate in terms of volumes, and how they hides the other categories which seems to have less effect on the area. To overcome this issue, we rescaled the distribution by normalizing it with respect to 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, for each day, two values on a weekly basis: (avg^n ,std^n); avg^n is the average number of distinct users for the n-th day of the week (0 = Monday, 6=Sunday) and std^n 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*.

We also applied a post processing step (after the *Sociometer*) on the class Visitors in order to distinguish the short term visitors or people in-transit from the others. We called them Passingby, i.e., users who made a single call in all the period, and thus we register their presence only for a single day. Clearly, this category is the result of an 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. The Figure 9 (bottom), shows the normalized distribution of the presences.

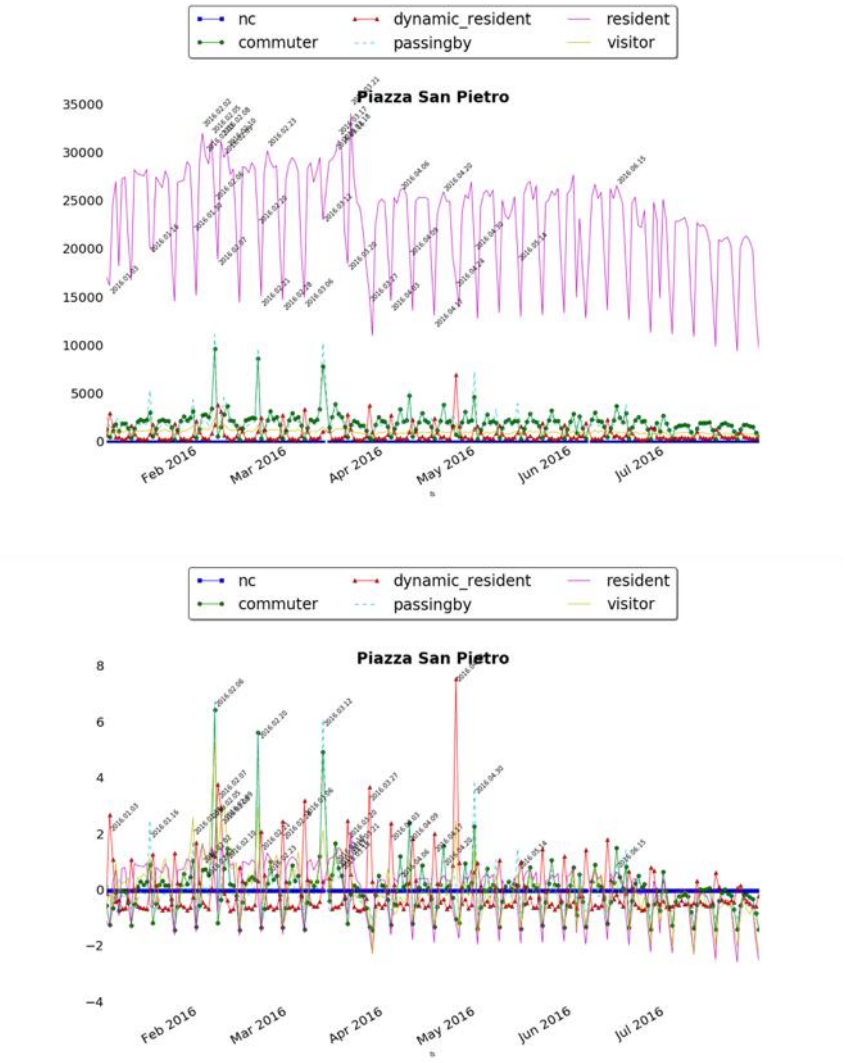

Figure 9: (top) Distribution of people's presence by categories around San Pietro Square; (bottom) the corresponding rescaled normalized distribution of people's presence by categories.

Now the deviations emerge and become clearer since the typical behavior, less interesting for our purpose, have 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 in the Figure 9 which represent,

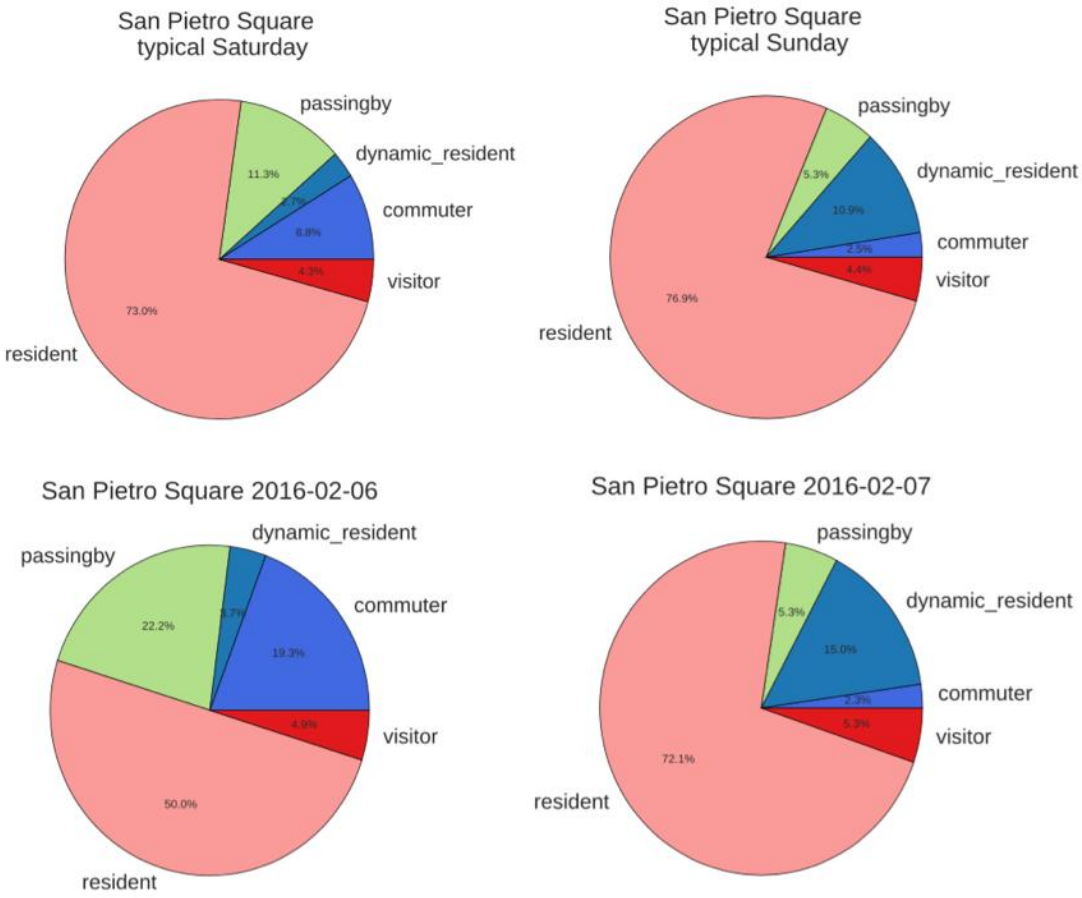

Figure 10: The comparison between the typical city users' composition (during a typical Saturday and Sunday), and the one on February 6th (where the peak appeared) and February 7th, at San Pietro Square.

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 San Pietro square 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 10 shows people composition during the case corresponding to the peak of February 6th, 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 see in Figure 9), but rather of new visitors arrived on purpose (visitors and passingby). 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 composition in the city returns to the normality after the big event. To complete the analysis, we investigated the provenances of the attending

people by reconstructing the Origin-Destination (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 11 (left), we see that the day of the event the majority of people come from outside Rome and from the Administrative Area 2, while the day after the external visitor are very few, see Figure 11 (right).

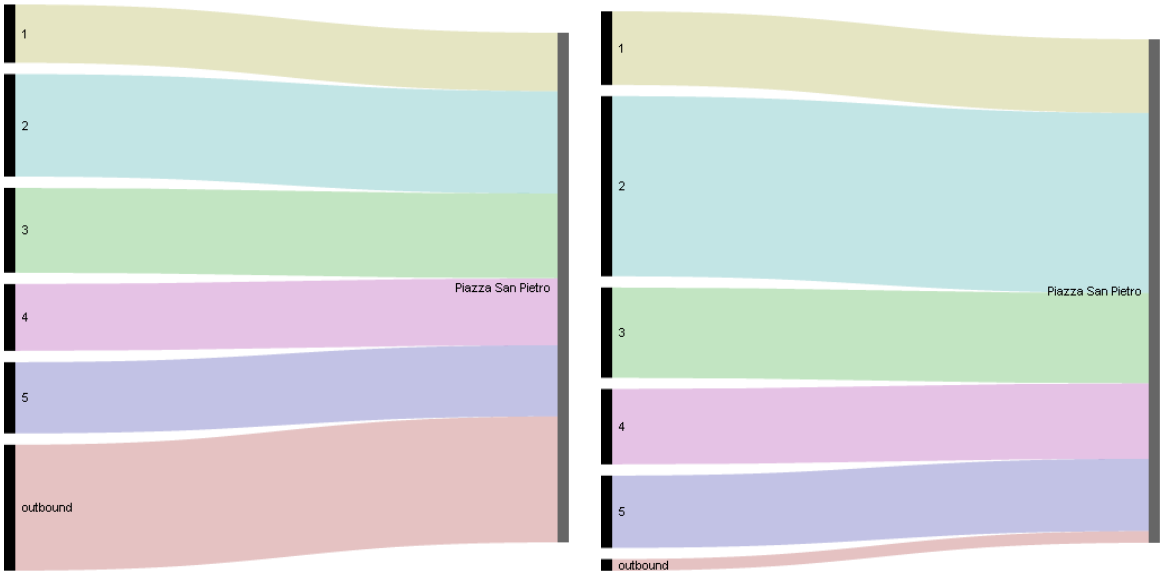

Figure 11: (left) OD matrix on February 6th, the day of Padre Pio arrival at St. Peter's Basilica, and (right) OD matrix the day after.

Different is the case of April 24th 2016, the day of 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)) with regard to the typical daily presence which usually ranges from 2.7% to 10.9% (typical Sunday in Figure 13). Furthermore, the attendees come mainly from Administrative Area 2 and outbound flow is quite small as shown in Figure 12.

In the area of the Olympic stadium is interesting to see how the month of May is characterized by a sequence of peaks as shown in Figure 14. Nevertheless, thanks to the *Sociometer* becomes clear that different classes of people are involved in different days and different events. In particular, we focused on three of them: the "Tennis with Stars" on May 9th, a charity exhibition involving champions in tennis and football, the "International BNL Tennis Championship" on May 10th, an international event valid for the world tennis rating, and the "Italian Soccer Cup" on May 21st between Milan and Juventus Italian soccer teams.

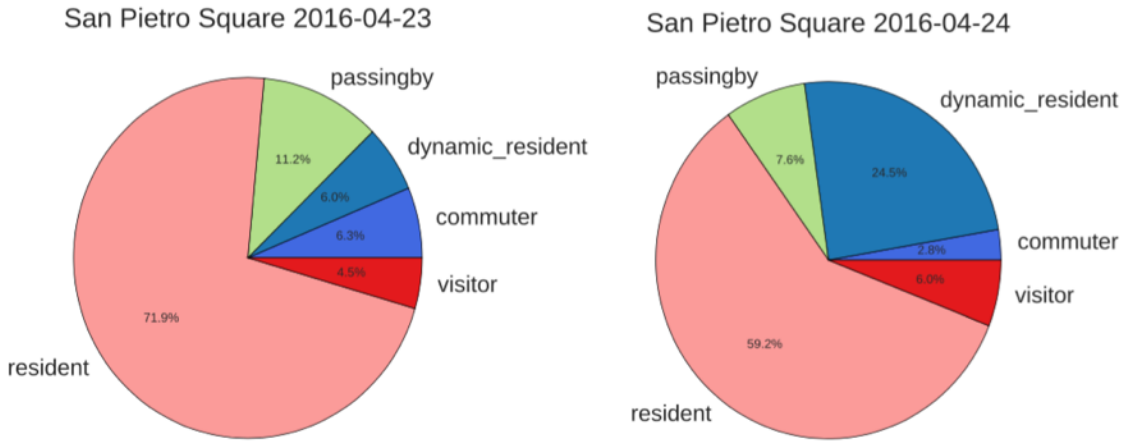

Figure 12: City users composition (left) the day before the Jubilee of boys and girls, and (right) the day of the event at San Pietro Square.

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 (passingby and visitors).

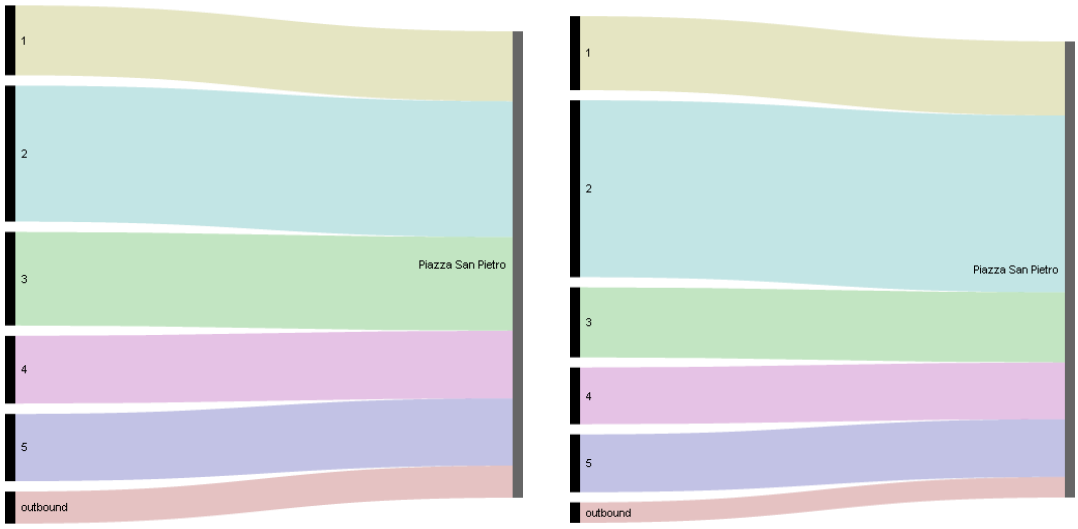

Figure 13: (left) OD matrix the day before the "Jubilee of Boys and Girls", and (right) OD matrix the day of the event at San Pietro Square.

The Circus Maximus is instead an ancient Roman chariot racing stadium, now has become a public park and used for big concerts and events. Figure 15 shows the anomalies registered in this area. The event on January 30th 2016, The "Family Day" had impact especially on people never seen before (i.e. passingby), while other events like the "Good Deeds Day" (May 10th) and "Race for the Cure" (May 15th) had impact especially on local citizens. Interesting is the case of

July 16th which registered an unusual peak of presences of all the categories. This corresponds to the “Bruce Springsteen’s Concert”.

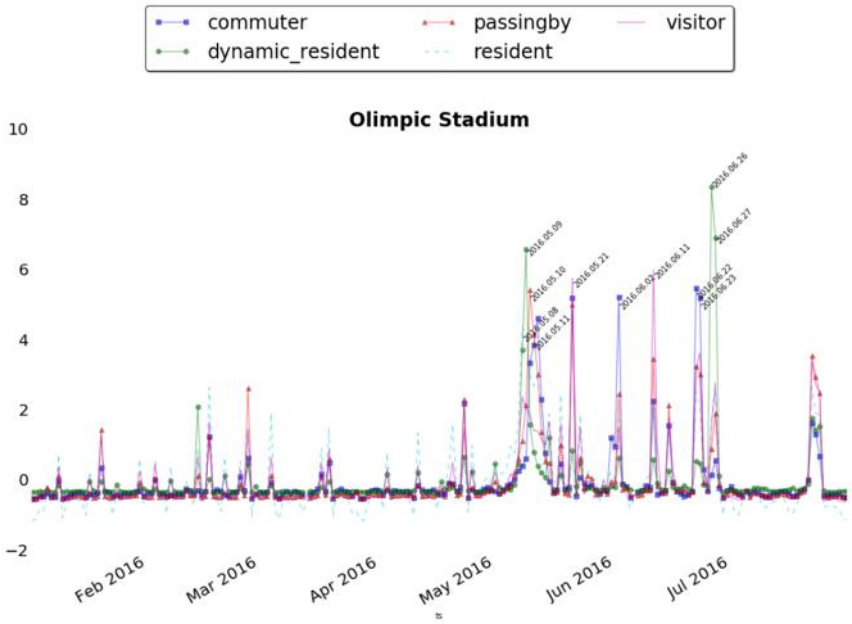

Figure 14: Time series for the Olympic Stadium (separated and normalized).

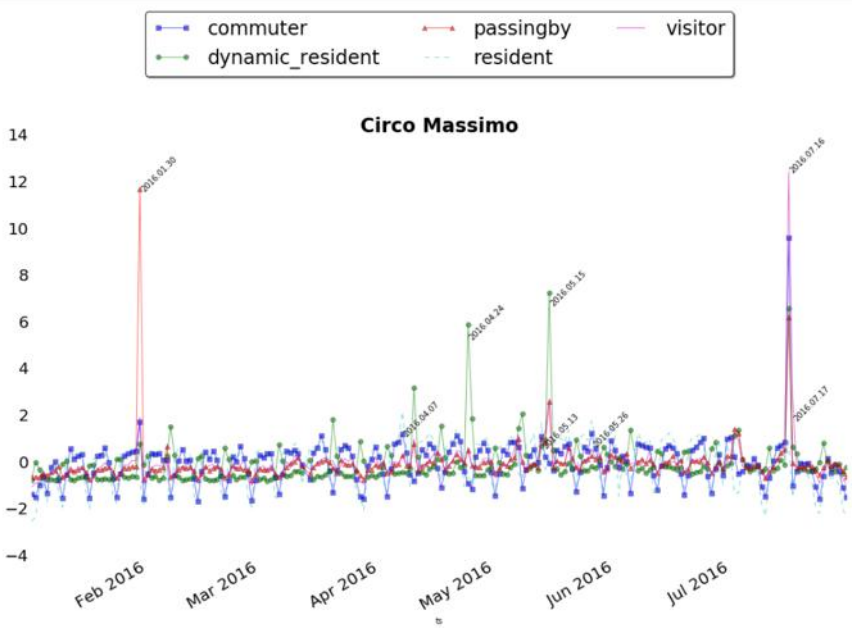

Figure 15: The time series for the Circus Maximus (separated and normalized).

As last case, we show in Figure 16, the anomalies registered around St. John in Lateran's Square. This square is situated just in front of the homonym Archbasilica and often location for social and political events. In the Figure 15 two main anomalies emerge. The first big one corresponds to May 1st, in Italy the Labor day. This event attracts mostly 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 May 7th 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 passingby.

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 a 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 a different locations or POIs (Points of Interest) of a city (e.g. train stations, universities, parks, and other touristic areas).

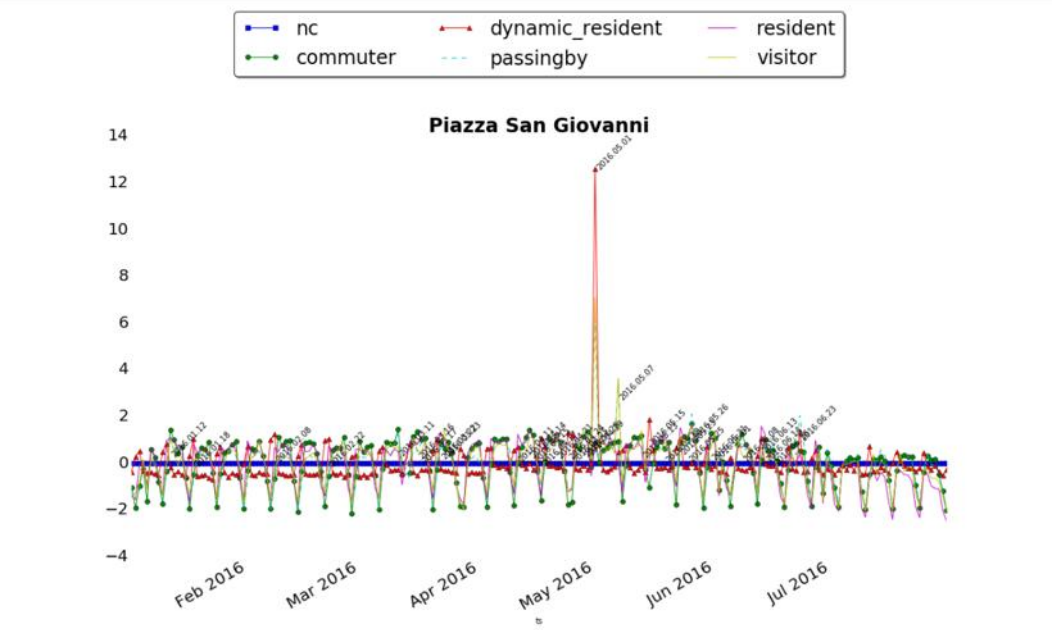

Figure 16: Rescaled distribution of people's presence by categories around St. John in Lateran's Square.

3.2.2 Ride Sharing

The objective of this application is to understand if it is possible to create a service of recommendation for a ride sharing system using the CDR data.

Many works are related with analysis to understand carpoolers and the ride sharing phenomena. In [12] the authors describe the characteristics of carpoolers, distinguishing among different types of carpooler, and identifying the key differences between carpoolers. In [13] it is introduced the methodology for extracting the mobility profiles used also in this work, and the criteria to match common routes. Something similar is illustrated in [14]. The authors extract home and work locations, and the social ties among the users for matching the users according to similar mobility pattern. [15] studies how to overtake the psychological barriers associated with riding with strangers and exploit it to find compatible matches for traditional groups of users and to find rides in alternative groups. An approach widely followed for analyzing carpooling is the agent based model (ABM). In [16] an ABM is designed to optimize transports by the ride sharing of people who usually cover the same route. The information obtained from this simulator are used to study the functioning of the clearing services and the business models. In [17] the authors use a multi-ABM to investigate opportunities among simulated commuters and by providing an online matching for those living and working in close areas. [18] present a conceptual design of an ABM for the carpooling application to simulate the interactions of autonomous agents and to analyze the effects of changes in factors related to the infrastructure, behavior and cost.

The ride sharing application differs from the previous in terms of objectives and final user, in fact the car sharing is a service which is directed to final user in order to create a recommendation system able to match compatible users in terms or systematic movements. To develop this application we need to detect the systematic movements of each users, in particular we consider two separated time frames: a morning time frame, and an afternoon time frame, in which the users usually move, respectively, from home to work and from work to home.

The first step is to identify the movements performed by individuals from *L1* to *L2* (*L1*→*L2*) and from *L2* to *L1* (*L2*→*L1*). It is important to notice that we are looking for movement between these two special areas even if they are not contiguous, i.e. other areas were traversed between them, as shown in Figure 17 (*A* is distinct from *L1* and *L2*).

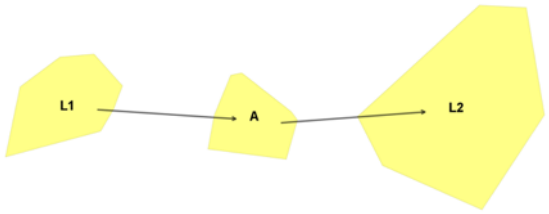

Figure 17: Example of flow between *L1* and *L2*.

The second step consists in selecting only the systematic movements, which is done by applying two different constraints: (i) request a minimum number of movements between the pair; and (ii) request a minimum value for the lift measure of the pattern *L1 L2*, which we define as:

$$LIFT(L1 \rightarrow L2) = {P(L1 \cap L2)}/{(P(L1) * P(L2))}$$

LIFT measures the correlation between *L1* and *L2*, resulting high if they appear together often with regard to the frequency of *L1* and *L2* taken separately. The main purpose is to normalize the frequency of *L1 L2* with regard to the frequency of calls of the user, since otherwise the candidate movements of frequent callers would be excessively favoured in the selection. The constraint on the number of movements is usually adopted in literature to exclude extreme cases where the LIFT (or other correlation or relative frequency measures) is not significant. More important in fact, is the LIFT measure to select. In our case, after a preliminary exploration we chose to select only pairs that appeared at least 3 times. Performing this analysis over all the user we extract their *mobility profiles* that will be used for the recommendation system. Notice that not all the users have 2 *routines*: (*L1*→*L2*, time period 1) and (*L2*→*L1*, time period 2), but there are cases in which a pair is not frequent enough or other cases in which there are more than 2 because the mobility of the users is different in different days (but frequent enough). Starting from the *routines* which constitute the user *mobility profiles*, our first objective is to test whether a routine is compatible with another one. If a routine *r1* is compatible with a routine *r2* then the user that systematically follows *r1* could leave his/her car at home and travel with the user that systematically follows *r2*.

The concept of compatibility is the following: given a routine *r1 = (l1, l2, time)*, it is considered compatible to the routine *r2* if *distance(r1.l1, r2.l1)+distance(r1.l2+r2.l2) < max_space_dist* and *r1.time = r2.time*, in other words, if the location where the two users start and the locations where the users end their routines are close (less than a threshold) and the time period is the same.

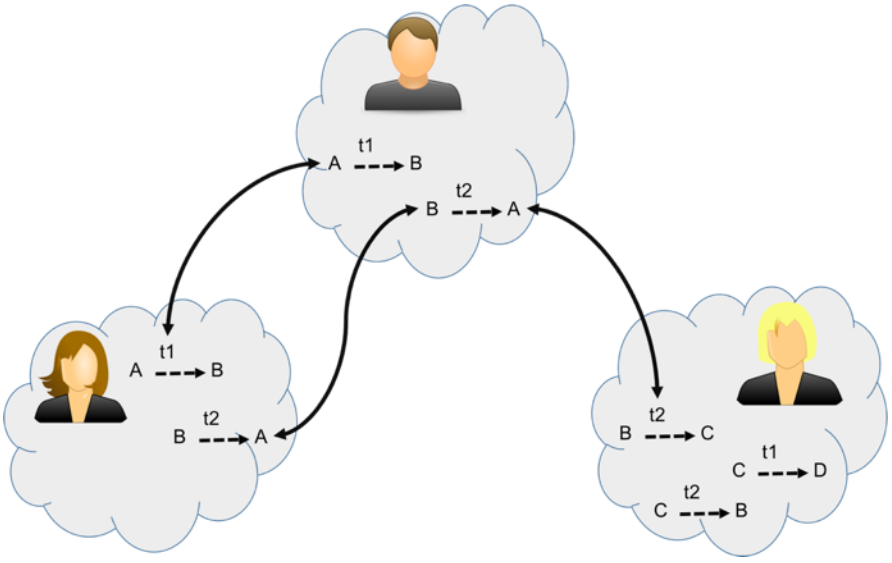

Figure 18: Example of ride sharing candidate network.

A ride sharing network represents all the links induced by the routine compatibility relation. Figure 18 shows a simple example of ride sharing network with three users considering the *distance(A,C)<max_space_dist*. The network is multi-dimensional, since two nodes can be connected by several different routines.

This network represents the base of the recommendation system which is able to detect automatically the possible candidate and can pro-actively suggest comparing their habits to see if they can share the car. This application is made to directly push notification to the users (registered to the service), as said before the application has different kind of final users and is not directly connected to the webLyzard Dashboard. In the following Figure 19 some example of communities that can be found in the network are shown:

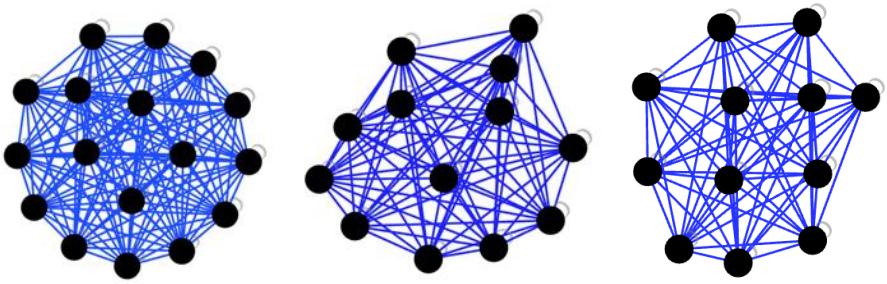

Figure 19: Example candidates of ride sharing communities.

A ride sharing application may use this information to suggest those group of people to share a ride because they may follow the same routines.

3.3 Integration with webLyzard Portal

Thanks to the integration of the analysis with the webLyzard portal all the results shown in the previous analyses can be visually represented and navigated in a dynamic Dashboard. The integration is realized using publishing modules integrated in the ASAP platform which send automatically the result to the front-end server.

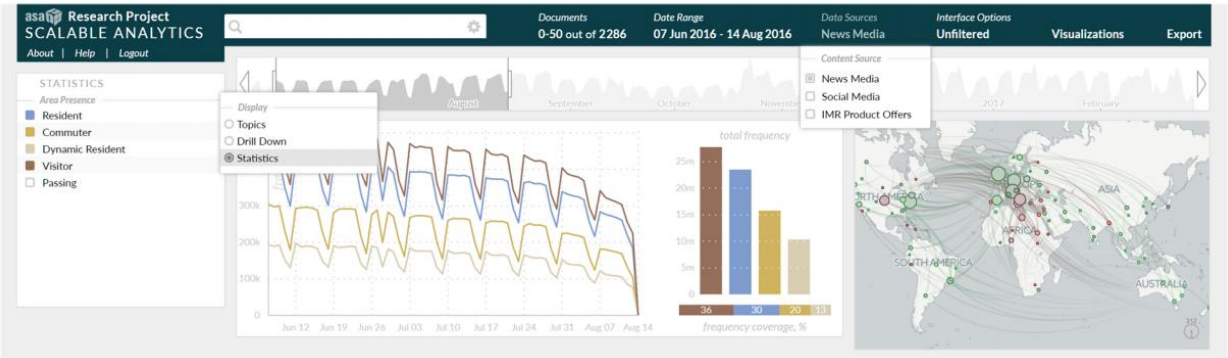

Figure 20: An example of visualization of results in the webLyzard portal.

In Figure 20 an example is shown, additional details about the technology used and dashboard functionalities and usability are reported in Deliverables D6.4 *"ASAP Dashboard"* [1], D6.5 *"Usability Report"* [19] and D10.7 *"Showcase"* [20].

In order to extract and visualize the statistical information produced by the various algorithms provided by Wind as a result of the activity in WP9 and the development of the TDA

(Telecommunication Data Analytics) application and then integrated into the ASAP workflow – (e.g., *Sociometer*, Origin/Destination (O/D) Matrix, etc.), the *Statistical Data AP*[3] was developed in WP6 and first reported in D6.3. Statistical data is hard to visualize on-the-fly as the only thing that various datasets have in common is the temporal dimension, all the other dimensions (location, values, relations, etc.) being typically dependent on custom dataset schemas. The Wind datasets themselves do not look similar as they contain data about *classification of users* who visited a certain region (e.g., *Sociometer*), the relation between a user's location and its destination on area level (e.g., O/D Matrix for Rome) or at a *Point of Interest* (POI) (e.g., O/D Matrix for Points of Interests in Rome), or even simply about the number of calls at cell tower level across the whole city. While the significance (the interpretation) of the data and the number of data points might differ, in order to align all these datasets into a common visualization framework, the Statistical Data API was built around ideas from the *RDF Data Cube Vocabulary* (QB)[4], but designed to support the JSON format to enable rapid visualization of large datasets. Following the philosophy of the QB vocabulary, each time series constitutes a statistical indicator, whereas each data point is an observation. The slices of a particular dataset (e.g., users classified as residents that visit a certain area) do not need to be defined directly in the dataset, and they can be specified at runtime via the dashboard's slices editor, once the data is integrated via the API.

The integration of the Wind data into the ASAP dashboard required several steps:

1. Conversion of the Wind data into the JSON Statistical Data format of webLyzard;
2. Creation of dedicated publishers for the various datasets to automatically upload data to the webLyzard repositories;
3. Additional aggregation where it was needed (e.g., averages);
4. Integration in the ASAP dashboard and visualization.

The dashboard provides actionable knowledge with respect to the ingested statistical data, for example to understand daily or weekly patterns of user movement or number of calls at a cell tower, or to visualize the interplay between the chosen indicators (e.g., relations between slices from same indicator or different related indicators, but also relations with other types of data like news media or social media data). The integration into the ASAP dashboard of Wind data provides new insights through the advanced slicing and visualization mechanisms:

- Data can be visualized in correlation with news media and social media data available through the other windows (e.g., Documents, Tag Cloud or Keyword Graph);
- The *Statistics* menu offers the possibility to present data slices based on the indicator structure and quickly displays new graphics as soon as the needed slices are selected. For example, the slices for the *Area Presence* indicator are based on user types (visitor, resident, dynamic resident, commuter or passingby), whereas the slices for the *O/D Matrix* indicator are based on both origin and destination to support fine-grained exploration of user movement based on origin. For the *O/D Matrix* indicator it is

[3] https://api.weblyzard.com/doc/ui/#/Statistical_Data_API

[4] https://www.w3.org/TR/vocab-data-cube/

therefore possible to both analyze the visiting behavior of users from a certain region, as well as the visitors coming to a certain region from all the other regions.

- The *Line Chart* surfaces weekly or monthly patterns and is ideal for understanding user behavior, while the *Bar Chart* is a frequency-based aggregated representation.
- The *Stacked Bar Chart* component combines data from multiple sources (e.g., social media, statistical indicators).
- The *geographic map* showcases data that contains geolocation coding, therefore being ideal for identifying the various hot spots in town during various events (e.g., Champions League final, Pope's announcement, etc.).

Figure 21 shows a screenshot of the ASAP dashboard with area presence data from December 2016. The indicators to display are selected through the *Statistics* sidebar. The slices displayed on the sidebar reflect the structure of the data (e.g., classification by user type, by location, by both) and can be set by registered portal users. The main content area shows the strong correlation between users from multiple locations in the *Line Chart*. The *Document View* will remain unchanged, enabling a comparison of statistical data and geotagged documents. In the upper right corner of Figure 21, for example, the *geographic map* depicts the local distribution of cell towers in Rome (blue markers), and overlaying this information with sentiment annotated Twitter postings (green = positive sentiment; red = negative sentiment).

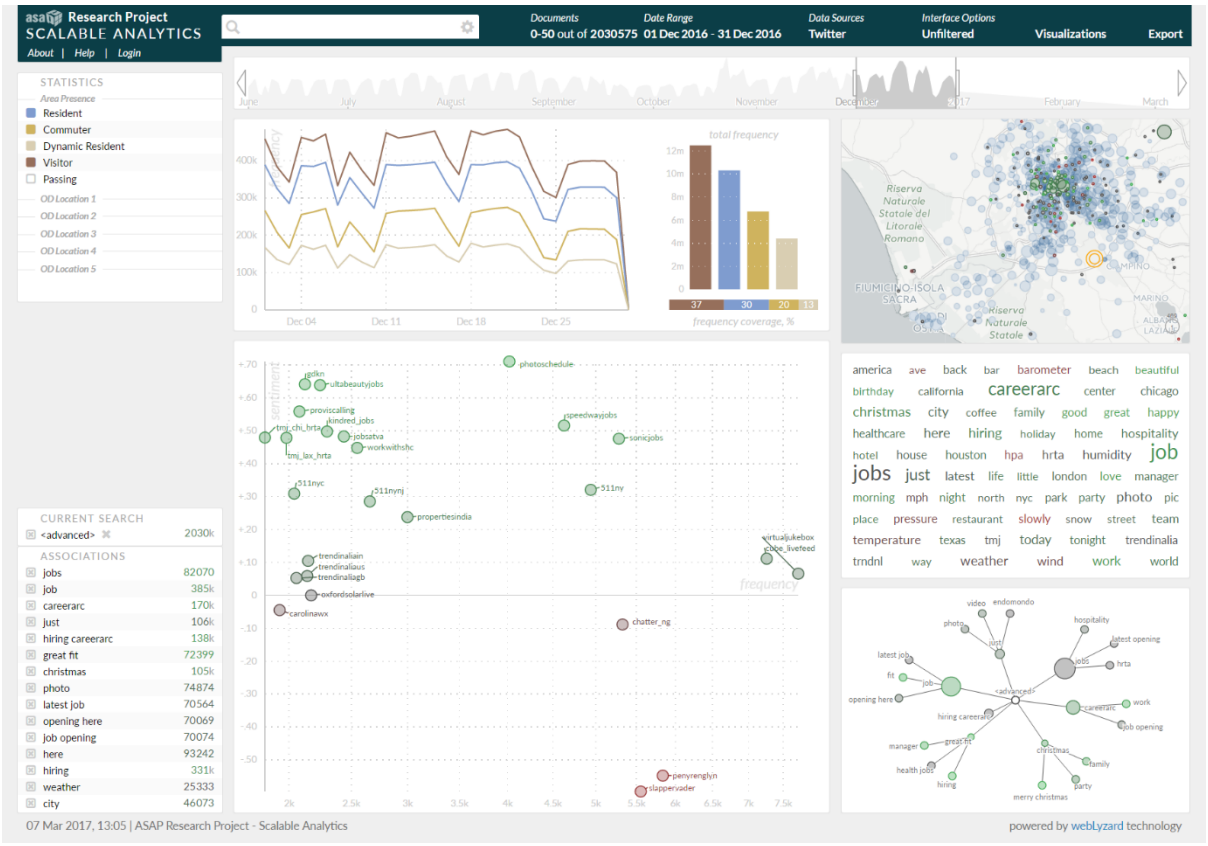

Figure 21: Screenshot of the ASAP dashboard with Wind Area Presence indicators, sliced by User Type in the trend chart and projected onto a geographic map in the upper right corner, together with geotagged Twitter postings.

Figure 22 shows the maximized version of the geographic map including the tooltip for on-the-fly query refinements, which can zoom to street level to visualize cell tower data. Tested with queries that delivered up to 100 million documents, the geographic map has been specifically extended for the Wind use case to allow for the display of statistical data and the color coding of data attributes (e.g., each data point has a different size, color, opacity that reflects its metadata). It was also designed to support the display of data regardless of its provenance, therefore it can allow us to do overlays with social media data on top of the call data, making it easier for the dashboard users to identify hotspots of communication during various events.

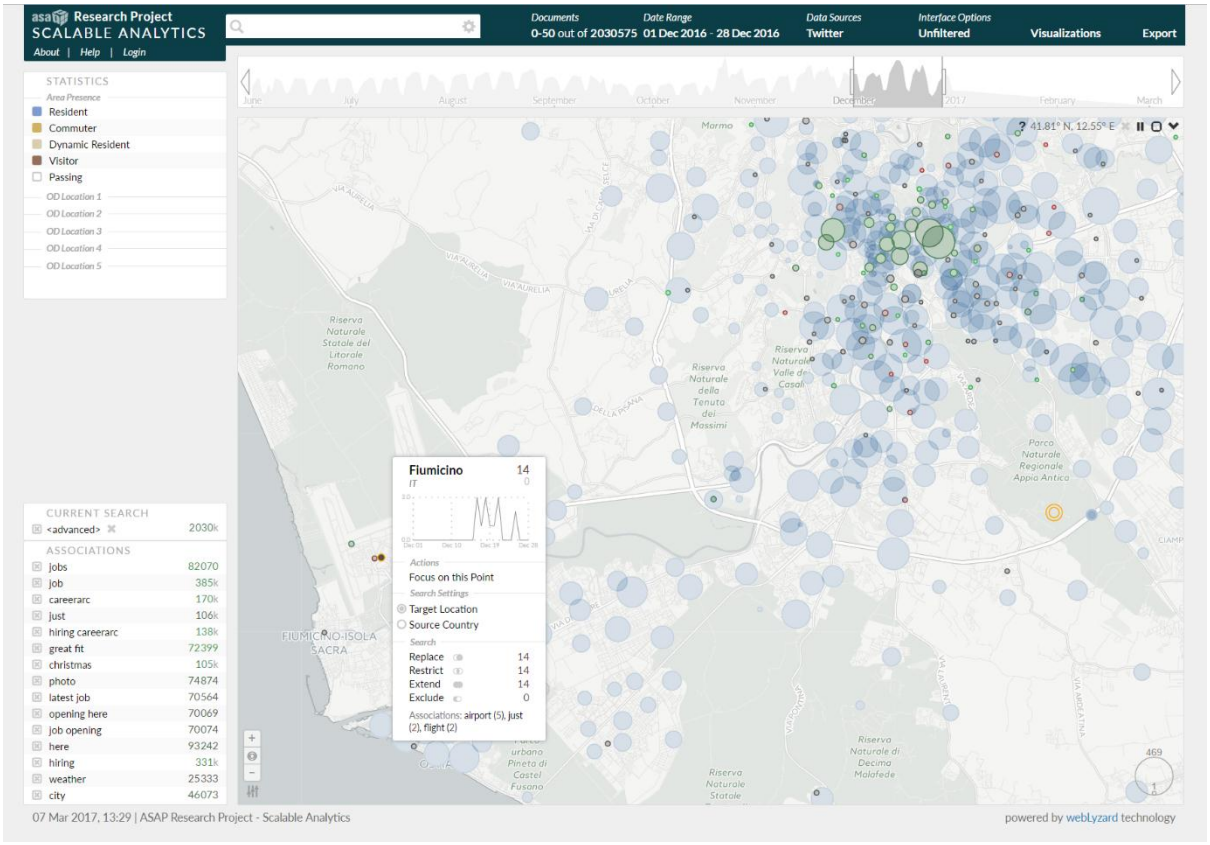

Figure 22: Geomap for Area Presence at Cell Tower level (with blue) with an overlay of social media data from Twitter.

The *Stacked Bar Chart* of Figure 23 combines statistical indicators produced by Wind (blue area) with sentiment extracted from news media and social media (green, grey and red areas), complementing the geographic hotspot analysis with a longitudinal dimension, for example to identify peak days or cyclic patterns.

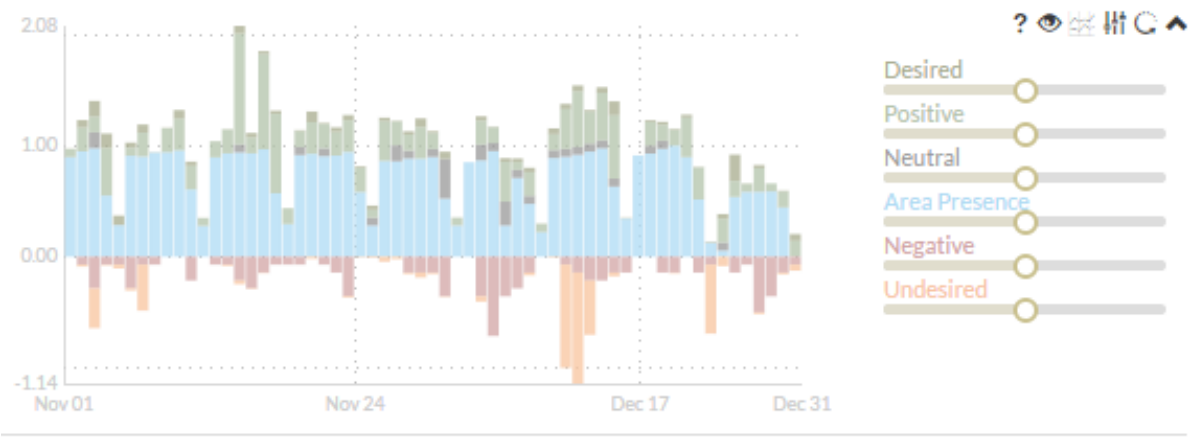

Figure 23: WYSDOM visualization combines sentiment data from news media and social media with statistical indicators produced by Wind.

The examples provided here are intended as a proof of concept that semantic technologies in conjunction with visual tools can automatically transform statistical data into valuable repositories of actionable knowledge.[5] For a telecommunications company, the temporal context and the patterns of user movement are of particular importance. The *Statistics* sidebar developed in WP6 enables the rendering of various slices. The *Line Chart* presents the relations between the weekly behavioral patterns of various user groups, while the *Bar Chart* provided an aggregated comparison between groups. The *geographic map* reflects the major increase in scalability achieved in Year 3 of the ASAP project, showing the entire result set of a query instead of just 50 top-ranked documents. The adaptive tooltip shown in Figure 22 supports on-the-fly query refinements, either to *Replace* the search query with a new term, or using the Boolean operators AND *(Restrict)*, OR *(Extend)*, and NOT *(Exclude)* to refine the search.

4 Evaluation

The results shown in the previous sections will be analyzed considering three different points of view:

- Functional: the robustness of the process and if the results are sound compared to the limited ground truth available.
- Technical: considering the runtime costs and how the process behaves varying the amount of data analysed.
- Marketing: taking into consideration the usability of the final results and the potentiality of the information provided.

[5] Brasoveanu, A.M.P., Sabou, M., Scharl, A., Hubmann-Haidvogel, A. and Fischl, D. (2017). Visualizing Statistical Linked Knowledge for Decision Support, *Semantic Web Journal*, 8(1): 113-137.

4.1 Functional Evaluation

In this section we study the interaction between the prototypes and archetypes during the experiment in the city of Rome. 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

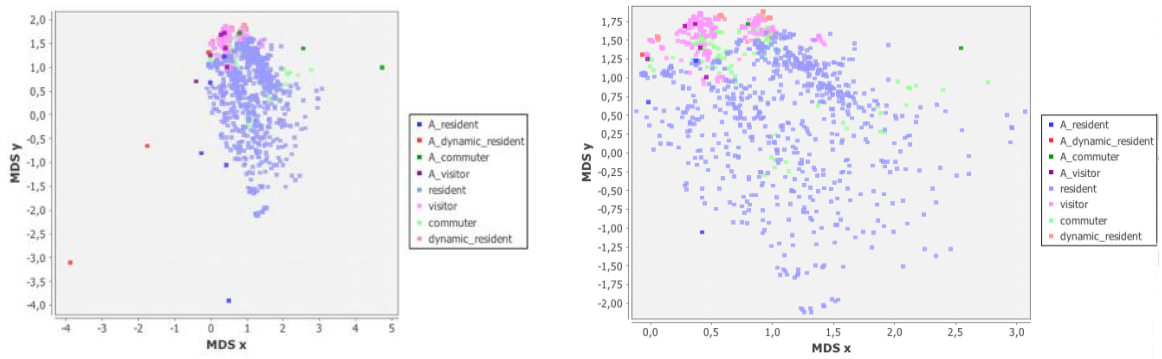

Figure 24: The distribution of archetypes and the prototypes computed every 2 weeks in a 2-Dimensional space.

we will show how the real behaviors of the users coming from the data differ from the one the experts have in mind, in fact the first show the complexity of the people habits represented by the calling behavior and the second is too perfect to be used directly. In other word the process shifts the perfect concepts of the expert (archetypes) to the real ones coming from the data (prototypes). To get a visual representation of the ICPs, due the high number of dimensions of them, we used a multi-dimensional scaling techniques on the set of prototypes generated each 2 weeks and the archetypes defined by the experts. Figure 24 (left) and (right) shows this distribution in 2-dimensional space. It is possible to notice that the actual behavior in the data are very close to each other making the problem of classification hard, and that some archetypes are very far from them. Moreover, In Figure 24 (right) only the portion of space where prototypes are present is highlighted showing how the visitors and the dynamic residents are very close while the commuters and the residents are less mixed and then easier to distinguish. To deeper the analysis, we also performed a density based clustering technique on the different classes of prototypes comparing them with the original archetypes. In Figure 25 the archetypes are shown as vector: starting from the same definition of the ICP, each six position of the vector represents a week where the first three represent the week-days (morning, day, night) and the last three 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).

We can say that if a prototype represents the behavior of a group of users in a specific time window, the prototype clusters represents a more abstract global behavior. In practice, it is something similar to what the expert want to represent with the archetypes he/she provides. At a first look to the results shown in Figure 26 it is clear how the reality if more complex than the conceptual view of the expert; in particular, by looking to the resident we can notice the

variability of behaviors due by the fact that, using a specific area in a more intensive way, the variability of how the perform calls varies significantly.

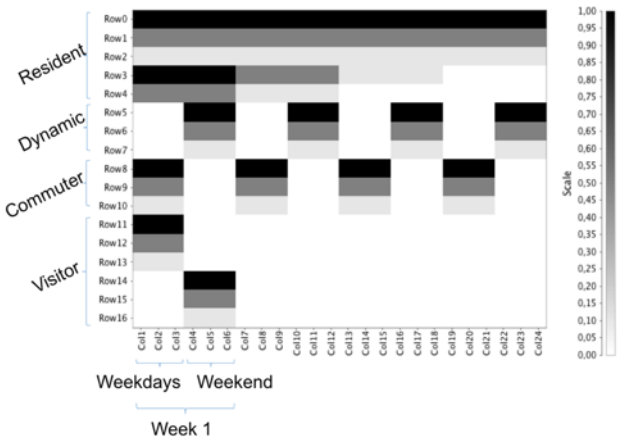

Figure 25: The vector representation of the Archetypes defined by the domain expert.

More interesting is the case of the commuters: although both of the clusters contains 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, but only the weekend visitors defined by the experts is not very clear in the data 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 expert, the results show that the general pattern followed but the users of this class tends to be very close to the visitor class sometimes (cluster 3 and 5) which was also highlighted by Figure 24. This kind of analysis will also be used as feedback to the expert in order to show the emerging reality from the data and maybe considering to redefine the archetypes. Finally an empirical evaluation is performed with the aims to compare the information extracted by using the *Sociometer* over the CDRs in the area of Rome, with official data provided by the Public Administration of Rome. These lasts are essentially surveys data and traffic data collected by using sensors and processed to produce reports about the population and their mobility. To remain consistent to previous analysis, we use here the same spatial partition presented in section 4.1. Figure 27 shows the residents in the five areas according to the official data (first column), compared with (i) the number of users labelled 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 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 with regard to the official data (i.e., a value between 0.58 and 1.46).

Nevertheless, Area 1 remains overestimated in both the cases. Showing the results to the experts, they explained that estimating people presence in Area 1 is very hard 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 further support this thesis.

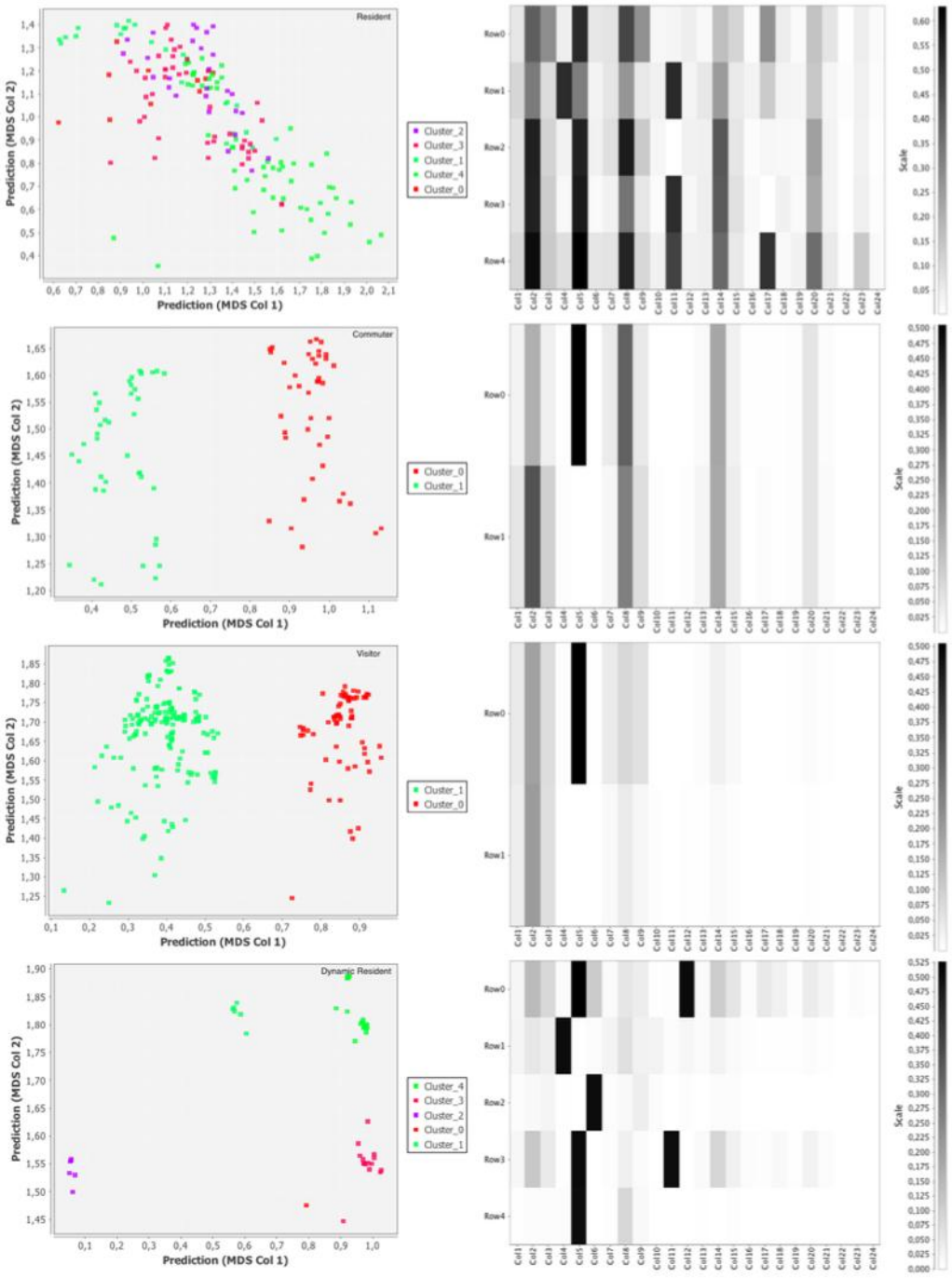

Figure 26: Prototypes clusters and their global behaviors.

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 27: 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).

4.2 Technical Evaluation

In this section we evaluate the computational cost of the analysis in terms of running time considering different amount of data. As described in previous section the process is executed for every window of four weeks for the entire period, the runtime of each execution is reported in Figure 28.

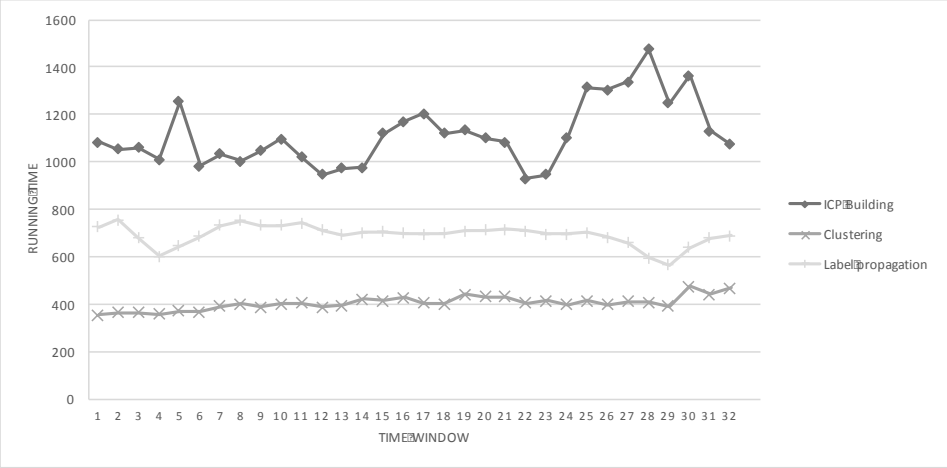

Figure 28: Execution time of the three main steps of the process in the different time windows.

The average running time for the ICP Building is 1100 seconds with a standard deviation of 157, for the Clustering is 406 seconds with a standard deviation of 26 and finally the label propagation takes 678 with a standard deviation of 65. This represent a good improvement if compared to the original runtime of the *Sociometer* taking hours for a computation of a single time window.

Additional details about the technical evaluation are reported in Deliverable D3.3 *"IReS Platform v.2"* [21].

4.3 Marketing Evaluation

According to the European Data Economy and global trends Wind believes that as an innovative industry it will lead the process using Big Data and Data Science in different sectors; the experience gained in the ASAP project is fundamental because it gives many elements useful to select a particular sector in order to start translating data into a business advantage.

In the ASAP project we learned that mobile phones and the data they produce can serve as a high quality proxy for studying people mobility in different domains, such as environmental monitoring, transportation planning smart cities and social relationship analysis. The Big Data are of high value: thanks to their capacity to generate knowledge on the customer base, contributing to the definition of new services. A number of analytical services describing the mobility of people or other properties (e.g. the Internet of Things, IoT) can be created on the basis of the data collected during routine operations to create a very large services portfolio always taking into account the Privacy aspects.

Different analyses enable different business areas:

- Socioeconomic: these enable planners to make economic activity predictions based on telecommunications activity and consumption data to create more customized services. The consumer behavior will offer new bases to provide next-generation services in different arenas as the automotive or the assurances.
- Operations: mobile data can drive efficiency improvements and enable agencies to optimize their traffic, network flows based on aggregated location data.
- Smart communities: boosting socioeconomic development and improving the well-being of the population. Focused on health improvement, city and transportation planning and other public community issues. Apps use mobile data to predict availability for on-street parking, without the need for sensors or information on user interactions.
- Marketing and sales: applications include the extraction of marketing insights regarding consumer retail behavior from consolidated location, demographic, and communications data. Operators could also enable targeted marketing opportunities based on personal location, segmentation, and search information.

According to all the elements described in the previous paragraph, Wind focuses the attention on the TDA application because its use cases are very useful for many contexts: information will be employed and combined in order to infer a model and predict patterns of tourist flows through different localities and regions, relating to the occurrence of popular events and the load of the tourist season. The ASAP Telecommunication Data Analytics (TDA) application shows how a number of analytical services describing the mobility of people can be created on the basis of the data collected by Wind's mobile network during routine operation. In particular three applications have been targeted using the available basic elements to create many different services:

- Event Detection: analyses the different features of an event, including its spatio-temporal characteristics, social aspects, and statistical properties. By controlling input parameters such as the time interval, the spatial area and additional CRM attributes, analysts gain a detailed understanding of evolving events.
- Ride Sharing: provides functions for mobility managers and individual drivers alike, for example, the visualization of routine trips in a specific area, together with an optimized car sharing solution for managing such trips. A driver can use this application as a recommender system to identify specific ride sharing opportunities.
- Tourism Observation: the analysis of dynamic tourist flows allows mobility managers to identify common movement patterns of visitors, using a map-based dashboard and with the option to provide spatio-temporal constraints as input.

As indicated, a new application has been designed and implemented to take better advantage of the new big data approach for mobile applications. The ASAP telecommunications application (TDA) will show how a number of analytical services describing the mobility of people can be created. According to the TDA features and targets, the following aspects will be investigated with the feedback from the Marketing people involved in the TDA evaluation process.

Additional details about the technical evaluation are reported in the Deliverable D6.5 "Usability Report" [19].

Taking into account the three main fields of application targeted by the TDA, Event Detection, Ride Sharing and Tourism Observation, the following parameters have been considered:

- Marketing parameters
- Applicability of the TDA (typical fields of application)
- Market target
- User target
- Types of applications
    - The fields of application being evaluated are:
        - Events analysis (Events prediction)
        - Ride Sharing (Transportation)
        - Tourism Observation (Services optimization)
- Visualization aspects that are most relevant to the Marketing and envisaged fields of application *(see the Dashboard evaluation Questionnaire Deliverable D6.5 "Usability Report"* [19]
- Evaluation from the point of view of final end users as outlined in the non-functional requirements analysis

5 Concluding Remarks and Next Steps

In this concluding section some possible actions that could be useful to consolidate and to address the evolution in the Big Data area are suggested and briefly discussed.

In the ASAP project we investigated the opportunity to use Big Data technologies to cover traditional analytics and new streaming analytics scenarios, in particular Tourism Services. In the ASAP Project we consolidated a framework that will allow us to run a complete process spanning from the stage of data injection to the data availability analysis and that will be useful to various future applications.

In the future, the collection, processing and analysis of data from the CDR could be useful, to improve the Data Service Experience Modeling using the output results of the TDA application. In fact the ASAP capabilities will be useful to obtain a Network Site Ranking by clustering network sites based on the traffic or other network parameters. It could be possible then to better define the network development strategies (e.g. network expansion plan, roll-out or site maintenance) to improve also the usability experience of the Customers Base.

The experience gained from the ASAP Project will be useful also to address the development of Network Customer Experience metrics based on elements evaluated using the ASAP platform.

We would also like to highlight the importance of new visualization tools developed in the project which are highly useful to improve the meaning and quality of the output produced by the requested analysis, taking into account the different needs coming from the selected context, as explained in section 3.3.

We also underline the possibility to extend the datasets by including other kinds of information: this will require particular attention to the privacy aspects of the data being extracted and the resulting datasets so that they can comply with current privacy rules and regulations to cope with aggregated safe data. In addition, the impacts on the data cluster size and time of processing have been analysed.

The analysis and evaluation performed on the TDA will be fundamental to identify the requirements for the optimal release of new services.

Recommendations, support for modelling, knowledge management of requirements patterns and dependency form from data extraction and management, everything will be useful to process high update workloads such as the ones processed by Telcos or other service operators with a fast analytical approach.

The ASAP Project stimulates the creative use of multiple data sources to perform researches and to launch new integration activities of Big Data in order to create new business opportunities . The use of mobile phone data to estimate the embedded population on an territory at municipality level and to evaluate the tourism demand was the first step. An analysis process will be built on top of the ASAP tools (e.g. *Sociometer*) to create new opportunities in new Market contexts like Industry 4.0, Smart Communities, etc.

Since the beginning of the ASAP Project, Wind cooperated tightly with other departments in the Company group (Commercial, Regulatory, Antitrust, Privacy and Wholesale Affairs) to be able to cope with the aspects useful for new data-based applications, e.g. risk assessment and privacy preserving, to create a framework to be used for Cloud, Sentiment Analysis and other contexts. Improvements in different areas will be addressed in cooperation with different departments in the Company group.

To further extend the application of the results obtained with the ASAP Project we could also define a school staging (scholarship) for the study of new test cases and their application with an innovative concept in markets like the Automotive/Assurance, in the PA (Public Administration) or in other business sectors.

Bibliography

[1] ASAP Project. *Deliverable D6.4 "ASAP Dashboard".* FP7 Programme.

[2] ASAP Project. *Deliverable D9.2 "Use Case Requirements"*. FP7 Programme.

[3] ASAP Project. *Deliverable D9.3 "Specification and early prototype"* . FP7 Programme.

[4] Roberto Trasarti, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, Cezary Ziemlicki. *Discovering urban and country dynamics from mobile phone data with spatial correlation patterns.* Telecommunications Policy Journal, Volume 39, Issues 3–4, 2013.

[5] Furletti B., Gabrielli L., Monreale A., Nanni M., Pratesi F., Rinzivillo S., Giannotti F., Pedreschi D**.** *Assessing the Privacy Risk in the Process of Building Call Habit Models that Underlie the Sociometer.* Technical report CNR - ISTI, Italy, 2014.

[6] Sergio Mascetti, Anna Monreale, Annarita Ricci, Andrea Gerino. *Anonymity: A Comparison Between the Legal and Computer Science Perspectives.* European Data Protection: Coming of Age 2013: 85-115.

[7] European Union for Protection of personal data**.** *Article 6.1(b) and (c) of Directive 95/46/EC and Article 4.1(b) and (c) of Regulation EC (No) 45/2001.*

[8] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti and D. Pedreschi. *Privacy-by-design in big data analytics and social mining.* EPJ Data Science, 3:10, 2014.

[9] ASAP Project. *Deliverable D7.3 "ASAP System Prototype".* FP7 Programme.

[10] Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, Salvatore Rinzivillo. *Analysis of GSM calls data for understanding user mobility behavior*. Big Data, 2013.

[11] Lorenzo Gabrielli, Barbara Furletti, , Roberto Trasarti, Fosca Giannotti, Dino Pedreschi. *City users' classification with mobile phone data.* Big Data, 2015.

[12] Roger F Teal. *Carpooling: who, how and why.* Transportation Research Part A: General, 21(3):203–214, 1987.

[13] Roberto Trasarti, Fabio Pinelli, Mirco Nanni, and Fosca Giannotti**.** *Mining mobility user profiles for carpooling*. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1190–1198. ACM, 2011.

[14] Blerim Cici, Athina Markopoulou, Enrique Frias-Martinez, and Nikolaos Laoutaris. *Assessing the potential of ride-sharing using mobile and social data: a tale of four cities.* Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 201–211. ACM, 2014.

[15] Gonçalo Correia and José Manuel Viegas**.** *Carpooling and carpool clubs: Clarifying concepts and assessing value enhancement possibilities through a stated preference web survey in Lisbon, Portugal.* Transportation Research Part A: Policy and Practice, 45(2):81–90, 2011.

[16] Marcelo Armendáriz, J Burguillo, A Peleteiro, Gérald Arnould, and Djamel Khadraoui. *Carpooling: A multi-agent simulation in netlogo.* Proc. ECMS, 2010.

[17] Tom Bellemans, Sebastian Bothe, Sungjin Cho, Fosca Giannotti, Davy Janssens, Luk Knapen, Christine Körner, Michael May, Mirco Nanni, Dino Pedreschi, et al. *An agent-based model to evaluate carpooling at large manufacturing plants*. Procedia Computer Science, 10:1221–1227, 2012.

[18] Sungjin Cho, Ansar-Ul-Haque Yasar, Luk Knapen, Tom Bellemans, Davy Janssens, and Geert Wets. *A conceptual design of an agent-based interaction model for the carpooling application.* Procedia Computer Science, 10:801–807, 2012.

[19] ASAP Project. *Deliverable D6.5 "Usability Report".* FP7 Programme.

[20] ASAP Project. *Deliverable D10.7 "Showcase"*. FP7 Programme.

[21] ASAP Project. *Deliverable D3.3 "IReS Platform v.2"*. FP7 Programme.