Journal of Visual Languages and Computing 42 (2017) 31–45

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Wi-Fi based city users' behaviour analysis for smart city

Pierfrancesco Bellini, Daniele Cenni, Paolo Nesi[1,*], Irene Paoli

Distributed Systems and Internet Technologies Lab, Department of Information Engineering, University of Florence, Florence, Italy

ARTICLE INFO

Article history:
Received 10 January 2017
Revised 11 August 2017
Accepted 13 August 2017
Available online 18 August 2017

Keywords:
People flows
Smart city
Wi-Fi access point location
GPS
Sensor positioning

ABSTRACT

Monitoring, understanding and predicting city user behaviour (hottest places, trajectories, flows, etc.) is one the major topics in the context of Smart City management. People flow surveillance provides valuable information about city conditions, useful not only for monitoring and controlling the environmental conditions, but also to optimize the deliverying of city services (security, clean, transport,..). In this context, it is mandatory to develop methods and tools for assessing people behaviour in the city. This paper presents a methodology to instrument the city via the placement of Wi-Fi Access Points, AP, and to use them as sensors to capture and understand city user behaviour with a significant precision rate (the understanding of city user behaviour is concretized with the computing of heat-maps, origin destination matrices and predicting user density). The first issue is the positioning of Wi-Fi AP in the city, thus a comparative analyses have been conducted with respect to the real data (i.e., cab traces) of the city of San Francisco. Several different positioning methodologies of APs have been proposed and compared, to minimize the cost of AP installation with the aim of producing the best origin destination matrices. In a second phase, the methodology was adopted to select suitable AP in the city of Florence (Italy), with the aim of observing city users behaviour. The obtained instrumented Firenze Wi-Fi network collected data for 6 months. The data has been analysed with data mining techniques to infer similarity patterns in AP area and related time series. The resulting model has been validated and used for predicting the number of AP accesses that is also related to number of city users. The research work described in this paper has been conducted in the scope of the EC funded Horizon 2020 project Resolute (http://www.resolute-eu.org), for early warning and city resilience.

1. Introduction

The understanding of city users' behaviour is one of the most challenging activities in a Smart City context: how the tourists (short, medium and long term) are moving and using the city, how the commuters are arriving and leaving the city, etc. City services are mainly related to mobility, government, energy, culture, events, commercial activities, environment, etc. Among these services, mobility is considered as a commodity; thus, transportation and mobility analyses are valuable aspects always considered for an effective definition of Smart City. According to Giffinger et al. [24], Smart Mobility is among the key factors of a modern Smart City, including local and international accessibility, availability of ICT infrastructures, sustainable, innovative and safe transport systems. Caragliu et al. [13] include traditional transport communication infrastructures among the essential requirements for Smart Cities.

In the context of mobility and transport, traffic/flow analysis is a major prerequisite for planning traffic routing. Hence, it is a central part of the so called Intelligent Transportation Systems (ITS) for public transportation. Traffic flow analysis is commonly used to ease the transportation management, for regulating the access control to the cities, for Smart Parking, for traffic surveillance providing information about road conditions and travel, or for monitoring and controlling the environmental conditions, such as harmful emissions (e.g., CO_2, PM10, ozone). The European Commission indicates, among the main topics that should be considered with special attention in the framework of the CARS 2020 process, the implementation and promotion of ITS, including Smart Mobility [CARS 2020] [14].

Some of the techniques adopted for traffic monitoring and management can be declined for people flow analysis and thus to support understanding the city user behaviour. For the city municipality it is very important to know the movements of city users within a certain precision, and detecting where and how they are crossing the city and exploiting services by using different kinds of

* Corresponding author.
E-mail addresses: pierfrancesco.bellini@unifi.it (P. Bellini), daniele.cenni@unifi.it (D. Cenni), paolo.nesi@unifi.it (P. Nesi), irene.paoli@unifi.it (I. Paoli).
[1] http://www.disit.dinfo.unifi.it, http://www.dinfo.unifi.it, http://www.unifi.it.

http://dx.doi.org/10.1016/j.jvlc.2017.08.005
1045-926X/

transportation solutions: car, bike, walking, taxi, car sharing, buses, tram, etc., targeting services into the city [[4],[5],[11]].

Usually telecom operators do not provide detailed information about city user behaviour: they may provide the number of people connected to each cluster of cellular antennas at a given time slot during the day, but not how the people move actually in the city, passing from one cluster/cell to another. Moreover, the telecom operator collect the cellular traffic from all the city users including residences that are stably at home and thus are not walking and using the city services on the road. Telecom operators are also constrained by the national contract from operator to the citizen in term of privacy and data use. At this regard, specific tracking services for mobile devices are needed and, when applied, the citizens have to be informed via an informed consent (e.g., terms of use, privacy policy).

The typical descriptor of people flow analysis in the city is the so called OD matrix (Origin Destination Matrix). The OD matrix presents on both axes the city zones, while the single element (at the intersection) contains the number of people (or the probability) of passing from the zone of origin to the zone of destination, in a given time window, for a given kind of users, for a given day of the week. Therefore, the OD matrix estimation is the main target results to understand the city usage, and thus it is a very relevant data source for traffic/people flow prediction and management. In particular, OD matrices are can be used as default descriptors of the traffic conditions and are used for (i) planning optimized routes predicting shortest and viable paths exploited by routing and path algorithms; (ii) providing infotraffic services on desktop or mobile devices, via the so called Advanced Traffic Management Systems (ATMS), ITS managing busses and vehicles (intelligent Transportation Systems), and UTS (urban traffic systems) managing semaphore networks; (iii) planning evacuations.

OD matrices are typically time dependent, and thus their dynamic real-time estimation may be needed, or at least the estimation of their values every 15 min, and distinguishing from the different days of the week (working days, festive and pre-festive days). Their values are of primary interest if they represent the maximum or at least sustainable traffic values, disregarding when the traffic infrastructure cannot sustain the traffic flow. In the context of traffic flow, some methods for computing OD matrices use parametric estimation techniques (e.g., Maximum Likelihood, Generalized Least Squares, Bayesian inference). Maximum Likelihood methods minimize the likelihood of computing the OD matrix and the guessing traffic. Other methods based on traffic counts include Combined Distribution and Assignment (CDA) [15], Bi-level Programming [19],[30], Heuristic Bi-level Programming [32], Path Flow Estimation (PFE) [35], or Neural Networks [25]. For example, Ashok and Ben-Akiva [3] used a Kalman filtering technique to update the OD matrix. Time dependent offline estimation deals with time-series of traffic counts.

Typically, building an OD matrix for mobility requires installing devices to count every single vehicle, and eventually recording the speed of each vehicle on the road. A traffic counter is a device that records vehicular data (i.e., speed, type, weight). At this regard, the US Federal Highway Administration defines three main traffic counting methods: human observation (manual), portable traffic recording devices and permanent automatic traffic recorders (ATR). Thus, at the level of traffic flow observation several different techniques are used: video cameras, pneumatic road tubes, piezo-electric sensors embedded in the roadway as inductive loop detectors, magnetic sensors and detectors, microwave radar sensors, Doppler sensors, passive infrared sensors, passive acoustic array sensors, ultrasonic sensors, laser radar sensors. Most of these sensors use intrusive technologies and require pavement cut; in some cases, lane closure is required, the devices are sensitive to environmental conditions and require an expensive periodic maintenance.

Several solutions have been proposed to solve the problem of an effective sensor placement for traffic counting. For example, in [16] Contreras et al. present a novel approach for studying the observability problem on highway segments, using linearized traffic dynamics about steady-state flows. They analyze the observability problem (sensor placement) and propose a method that compares scenarios with different sensor placements. In [6] Ban et al. present a modelling framework and a polynomial solution algorithm to determinate optimal locations of point detectors, for computing freeway travel times. They use an objective function to minimize the deviation of estimated and actual travel times; the problem is discretized in both time and space, using a dynamic programming model, solved via a shortest path search in an acyclic graph. In [31] the performance of the sensors is measured in terms of estimation error covariance of the Best Linear Unbiased Estimator of cumulative flows in the network. Sensors are placed to minimize the sum of the error covariance and of a cost penalizing the number of sensors, using the concept of Virtual Variance. Ivanchev et al. [27] defines a measure of importance for a node in a traffic network and use it to solve the sensor placement problem, by maximising the information gain (i.e., users' routing choices). It presents a method for finding the optimal number of sensors to be placed, modelling, and maximising the utility stemming from the trade-off between cost, performance, robustness and reliability of the sensor placement. Bao et al. [9] describes some spatial distributions of traffic information credibility and proposes different sensor information credibility functions, to describe the spatial distribution properties. The authors propose a maximum benefit model and its simplified model to solve the traffic sensor location problem. In [7] Ban et al. propose a modelling framework to capture a sequential decision-making process for traffic sensor placement. Optimal sensor deployment for a single application is determined by a staged process or dynamic programming method; sensor locations for new applications can be optimally solved by the DP method considering existing sensors.

Some of the above-mentioned techniques can be used to produce vehicle classification (e.g., rural cars, business day trucks, through trucks, urban cars). Recently, other techniques have been adopted as RFID, Bluetooth, Real Time Location System (RTLS) and Wi-Fi access points [17],[36]. In some cases, the position of the vehicle can be monitored from the GPS position of mobile devices installed on the vehicle itself, or simply by using smartphone navigators (e.g., Google Maps, TomTom, Waze), that provide positions and velocities of the vehicles. In these two cases, vehicle's tracking is authorized by the users (through an informed consent) that install the device or run the mobile application on the smartphone or navigator. RFID is quite unsuitable to detect devices because of the small range of action. Bluetooth can be more suitable but it is expensive, since specific stations to collect the passages are needed. Wi-Fi access points are less reliable in detecting the presence of high speed people as in motorized sources with respect to physical devices, and GPS-related methods. In [43] the Wi-Fi analysis has been used to assess the passages of pedestrians in buildings. In [38], the quality and feasibility of using multiple solutions based on Wi-Fi and Bluetooth for people tracking has been has been presented providing the evidence that Wi-Fi count may be more reliable. In [23], an early experiment in tracking people flow by exploiting Wi-Fi data has been reported exploiting direct MAC address tracking. In [2], a small scale experiment has been performed for tracking a limited number of people (8000) in well-known and restricted area with 20 AP. The effective precision and assessment was not provided. In [41], a similar experience has been analysed, with the aim of extracting trajectories.

1.1. Aims and structure

On the other hand, we addressed the usage of Wi-Fi Access Points (APs) as devices to have indication of people flow and density in the city. Wi-Fi solution is viable given the high distribution of mobile devices, the low cost of a Wi-Fi AP, and the fact that a huge number of APs is already installed in the cities. This solution is quite cheap and easy to implement, also considering that many municipalities offer free Wi-Fi connectivity, and the needed coverage can be easily obtained with a small effort adding a few more APs, or just reconfiguring those already present, and thus the proposed approach is used only for selecting those to be reconfigured. Therefore, the paper presents mainly two major results:

1) Methodology for identification of the best placement for Wi-Fi Access Points, as detectors for collecting data for user behaviour understanding, maximising the precision of OD matrix computation as one of the most attended results. The methodology aims at limiting the costs to obtain reasonable data for massive and systematic measuring of the whole city flows by humans. The study and solution has been validated by using the data set introduced in [37] which covers cab mobility traces, collected in May 2008 in San Francisco (USA). This result has been used to identify the most suitable AP in Florence for reconstructing OD and flows.
2) The data collected from the Firenze Wi-Fi network (instrumented for people flow tracking) have been analysed to derive a number of information and knowledge: (i) most frequent places and hottest areas in the city, represented as heat map; (ii) daily user behaviour patterns around AP in the city to understand how the city is used; (iii) OD matrix to extract people movements; and a (iv) predictive model for guessing number of Wi-Fi connections for each time slot and AP (which are directly related to people presences, behaviour and flows). This result poses the basis for exploiting the produced model and instrument for early warning. That means as a tool for detecting dysfunctions or un-expected patterns in the city user movements at their early inception.

The proposed AP positioning strategy, combined with the data analysis of Wi-Fi data, constitute an innovative methodology to understand user behaviour at low costs in urban areas. Such services include: enrichment of traffic sensor data (i.e., physical road sensors, cellular data), notification tools for alerts or events with huge crowds (e.g., people's flood detection, emergencies, manifestations), development of traffic/people routing and optimization algorithms, resilience management and realtime monitoring tools, building of green areas or recreation activities in zone at high density of pedestrians, control of air pollution, and city cleaning, increasing city security.

These features are becoming always more requested since according to the last European Directives for large events, the limited number of people is mandatory and also tools for monitoring and to react to critical events. The research work described in this paper has been conducted in the context of the Resolute Horizon 2020 project (http://www.resolute-eu.org) which has been founded by the European Commission. RESOLUTE is focussed on city resilience assessment and management and thus the solution proposed is a component of the risk assessment model and tool on the basis of the distribution of population in real time, on predicting them, and on for early detection of critical situations in the city [10].

This paper is structured as follows. In Section 2, the definition of the user behaviour representation analysis and tools including OD matrix are discussed, together with the possible data for their estimation and study (data collected in San Francisco). Section 3 presents a number of models for AP positioning and comparatively evaluate them by using various AP scenarios and their capability of computing OD matrix with respect to the real data used. In the second part of the article, we describe the results obtained for city user behaviour understanding derived from the application of the proposed methodology for AP positioning in instrumenting the city of Florence (Italy) by reconfiguring the Wi-Fi network. Thus, as a second result, in Section 4, the data collected from the Firenze Wi-Fi network has been analysed in different manners by data mining to extract the most frequented places, the typical city users' behaviour (presenting the proposed OD Spider Flow tool), the clustering of city users' behaviour inferring patterns in data about city usage. Finally, Section 5 presents the predictive model which is estimated in real time for each single AP in the city, experimental results of a forecasting model for predicting number of connections. Conclusions are drawn in Section 6.

2. User behaviour analysis vs data set

User behaviour is urban area is represented by a set of tools: (i) trajectories, (ii) hottest places also represented as heat maps, (iii) Origin Destination Matrices, (iv) analysis of regency and frequencies. These results and model can be mathematically obtained with some specific algorithms processing singles GPS traces of the movements.

In more details, the OD matrix representing flows among the zones of the city (considered for example as zip codes z or smaller areas) is defined as

$$OD_{n,n} = \begin{pmatrix} z_{1,1} & \cdots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{n,1} & \cdots & z_{n,n} \end{pmatrix} \quad (1)$$

where: $z_{i,j}$ represents the total number of traffic counts from z_i to z_j (i.e., in our context how many cabs moved from z_i to z_j) defined as

$$z_{i,j} = \sum_{t \in T} n_t(i, j), \quad (2)$$

and, T is the set of unique cab traces, $n_t(i, j)$ is the number of traffic counts from z_i to z_j for the trace t.

This means that, if the aim consists in identifying the best position for the sensors (may be Wi-Fi AP as in this case), one should have the data representing the whole set of people movements in the city, that is unrealistic, no one has those data neither the telecom operators. On the other hand, Fei et al. [21] present a nonlinear two-stage stochastic model to compute sensor location (classical traffic flow detector as spires) maximizing the quality of origin-destination matrix (OD) by starting from the traffic flow data. In this case, the authors presented an iterative heuristic solution algorithm, Hybrid Greedy Randomized Adaptive Search Procedure (HGRASP), to find the near-optimal locations. This approach is feasible when the flows are known. The validation of any AP positioning methodology for the people/flow count is not a trivial task. In principle, one should install the APs in certain positions and demonstrate, making measures on the real context, that they produce strongly correlated data with the real people flows, among the different areas of the city. Since this approach is very expensive and unfeasible for a number of configurations, we adopted an indirect method described in the following.

2.1. Reference data from San Francisco vs AP

Due to the above described difficulties for our analysis, we exploited the data set introduced in [37], that includes cab traces in San Francisco, collected in May 2008. The dataset reports all the cab traces by providing precise GPS positions for each of them. In

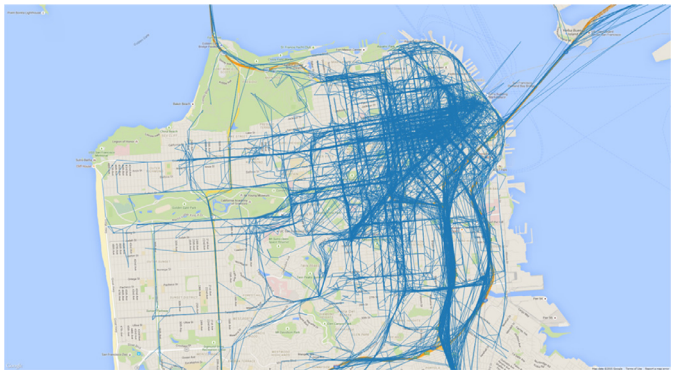

Fig. 1. Trace flows in San Francisco on a working day of May, 8:00 a.m.–9:00 a.m.

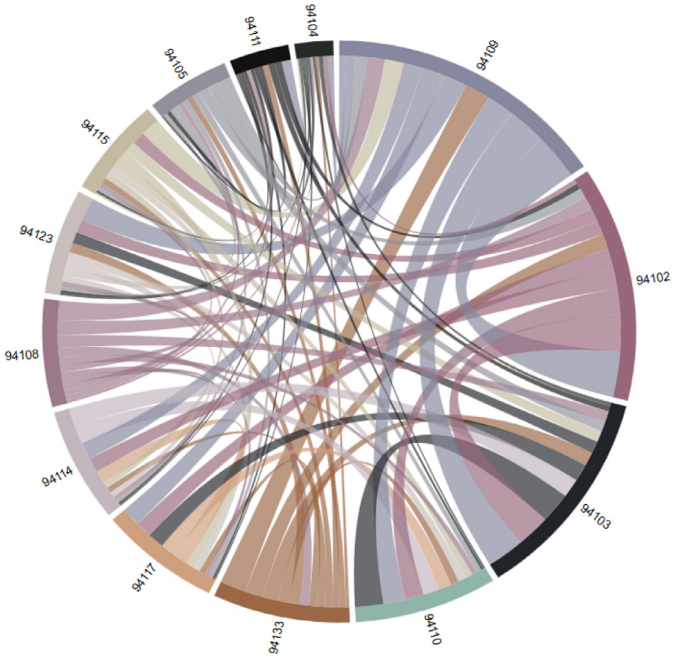

Fig. 2. San Francisco OD matrix as a chord diagram among the 13 central ZIP areas of the city (real cab flows).

Fig. 1, the traces are reported on the city map (for a particular day in the time range 8:00 a.m.–9:00 a.m.).

The total data set consists of 446,079 traces based on about 11.2 million of single GPS points collected by cab movements, not only in the downtown of San Francisco, but spanning the city's neighbourhoods. The areas at higher density are those in the downtown, coherently with what you could have from the movements of pedestrians in the city. This area is covered by about 13 zip code areas.

In order to perform an effective data analysis and visualization, some web tools for viewing and comparing flows in different scenarios were developed. At this regard, OD matrix and thus flows among the zip areas are represented with a chord diagram, to put in evidence single and aggregate contributions to the total flow count among the various city zones (in Fig. 2, the chord diagram is reported for the central part of the city with 13 zip code areas). An interactive version of the produced chord based tool is accessible at http://www.disit.org/6694. The user can select a time interval in the day to visualize the corresponding chord diagram, which is constituted by circular sectors, each of them representing a city area; passing the mouse over a sector provides additional information about the traffic counts originated from it towards other zip areas. In this manner, it is possible to depict in a compact and intuitive way the traffic flows among the various zones. Additionally, it is also possible to remove a circular sector to simplify the diagram and make easier the analysis of the flows of interest.

In the case of San Francisco data, the structure of the city and the position of the APs in the downtown is known (see Fig. 3). The positions of AP in the area have been taken from OpenWiFiSpots (http://www.openwifispots.com). They consist of 494 Wi-Fi APs providing city services, from a total of 983 APs at disposal (also located in coffee shops, hotels, restaurants, libraries, bars, bookstores, grocery stores). Therefore, we may suppose to use the Wi-Fi network to estimate the people flow in the city produced by mobile devices, according to their MAC address or to hash code of the MAC address and other features of the mobile device.

This solution could be implemented by collecting the events of connection and release of mobile devices with respect to the APs (each event reports date, time, device ID to give internet access, and AP identifier). Each AP streams the collected data to a central server which anonymizes the MAC addresses, records the data, and streams the combined multi-streams to the data analytics. In alternative, some of the APs or aggregators of APs may compute the anonymization algorithm, based on a hash code of the identifiers. Once detected the passages of devices on the APs, the OD matrix as well many other information can be can be derived.

This means that, we can exploit these data filtering out the traces matching with hypothetical position of AP and observing if the obtained OD is still valid to represent the whole OD matrix depicting the actual situation calculated on the basis of all traces. This means to produce the AP positioning maximising the correlation of the estimated OD with respect to the actual.

3. Methodology for AP positioning

As a first approximation, we assumed to have the possibility of detecting the flows by using the present APs distribution, by capturing the real traces passing within a distance of 25 m from the AP position. The proposed approach can be viewed as a sort of partial simulation based on real data about traffic flow, that is more realistic than producing fully simulated data. It is obvious that the real data captured by the APs would be probably only a part of the real traffic of people passing close to them. On the other hand, it is reasonable to verify that the simulated measures are strongly correlated to the real effective numbers.

As a general consideration, only 1470,091 trajectories were found to intersect with the real APs positions, which in the downtown are 1418,207 with respect to 494 APs. Therefore, in this manner, we assessed the available distribution of Wi-Fi APs in San Francisco, in order to collect people flows related data through mobile devices. Once obtained the observations by finding the intersections of the traces with the APs, an estimated OD matrix has been produced, as reported by the chord diagram in Fig. 4a. In Fig. 4b, the matrix of difference between the OD matrix of Fig. 2 and that of Fig. 4a is reported; the differences between back and forward flows are not perceivable.

The difference matrix of Fig. 4b give the evidence of the difference from the real traffic flow with respect to the flow that is estimated by using the present APs distribution in the city. The differences are reported with a grayscale (the higher the difference, darker is the matrix element). The two OD distributions are uncorrelated (a correlation of 0.12 has been measured, see Table 1). This result demonstrates the unsuitability of the present distribution of APs in San Francisco for collecting and modeling traffic flows. On the other hand, their placement was not made with the aim of measuring and observing people flows.

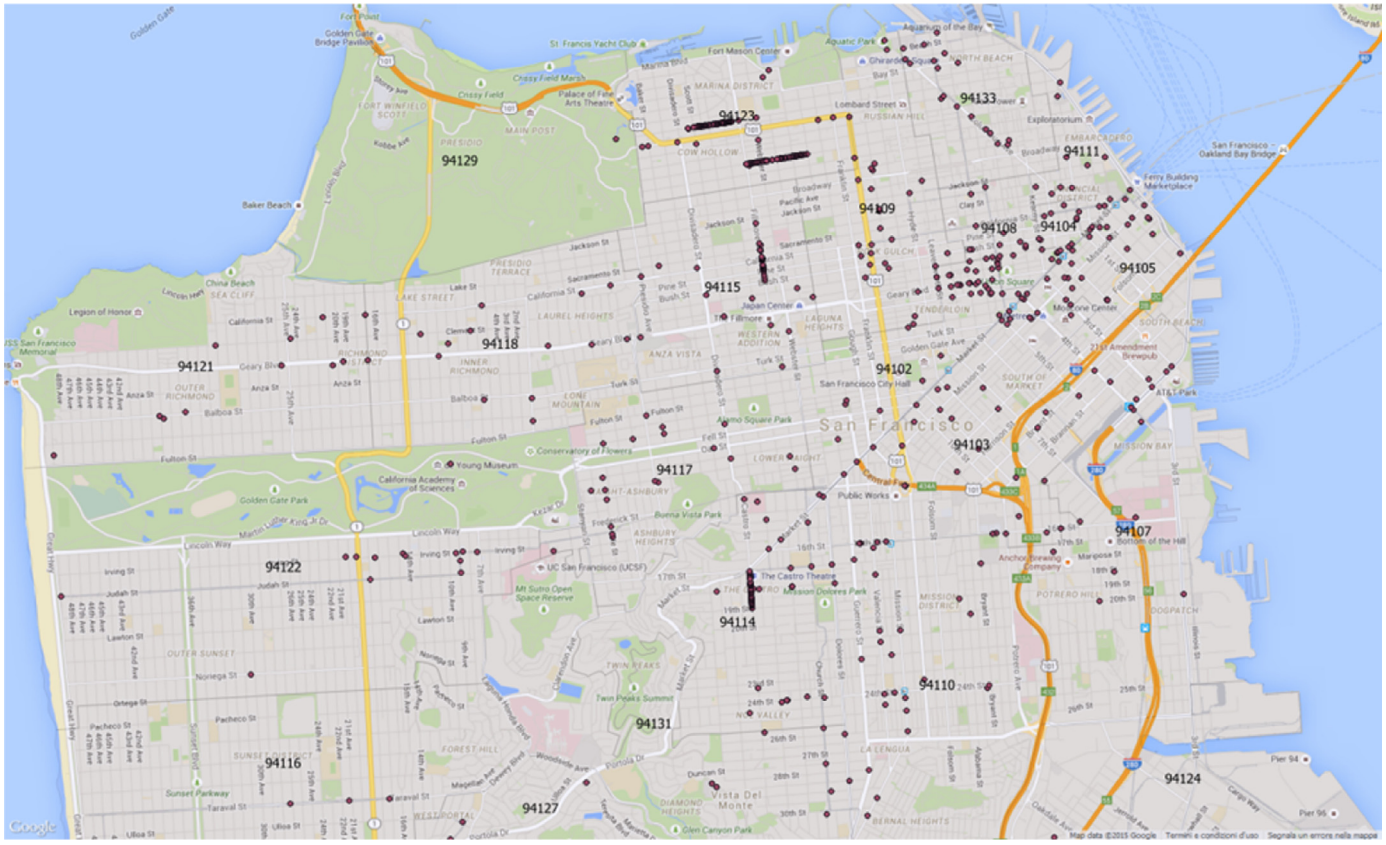

Fig. 3. Distribution of real Wi-Fi APs in San Francisco.

Table 1
AP models, cc = city center, bn = zip boundaries (within 300 m).

Model		Coefficient	Std. Error	t-statistic	p-value	Correlation	# APs
Real APs	β	280393.858	19874.972	14.108	0.000	0.446	983
	α	9.448	0.543	17.400	0.000		
Real APs (cc)	β	1598664.580	116546.825	13.717	0.000	0.12	494
	α	1.714	1.141	1.502	0.135		
(a) Random APs (cc)	β	690144.338	75267.849	9.169	0.000	0.835	400
	α	52.921	2.813	18.816	0.000		
(b) High Traffic APs (cc)	β	684144.945	52950.289	12.921	0.000	0.915	804
	α	10.942	0.389	28.114	0.000		
(c) High Traffic APs (bn, cc)	β	1101641.803	86354.599	12.757	0.000	0.687	448
	α	13.586	1.159	11.727	0.000		
(d) High Traffic APs 400 (cc)	β	810743.094	70801.471	11.451	0.000	0.835	400
	α	24.429	1.297	18.829	0.000		
(e) Real augmented APs withHigh Traffic APs (bn, cc)	β	748987.390	58260.615	12.856	0.000	0.892	400
	α	39.960	1.634	24.453	0.000		

3.1. Adopting AP positioning models

On the other hand, a more efficient AP positioning scheme should achieve better correlations and smaller standard error, and thus better precision for the estimation of OD matrix (and indirectly people flows in the city). To this purpose, similarly to Fei et al. [21] a set of heuristics have been identified to find a compromise from precision and OD estimation. Thus, a number of different methods for AP positioning and thus for flow observations have been adopted and tested, taking them from the literature of the classical traffic flow observations strategies by humans. We then started by creating a uniform distribution grid of APs, ideally placed at the middle of each street. In all cases, each AP was considered as a circular area with 50 m of diameter.

The resulting APs set, consisting of 14,959 APs (a number of devices that is surely too high to be affordable), was further reduced using different strategies as reported in the following. Moreover, the reduction is also reasonable since a uniform distribution in all the zones of the city is not feasible. There are many zones in which the flows are very low, at least in the simulated data taken into account. On the other hand, the positioning of the APs in low flow areas is not efficient.

Also, a flow prediction strategy should be able to tell where to place traffic sensors, and how many sensors to use, providing a tuning strategy for selecting the required set of sensors, with the aim of minimizing the number of traffic sensors and the costs of periodic maintenance of the monitoring infrastructure. In this section, we provide some alternative strategies of AP placement, in order to minimize the number of APs, and to obtain a satisfactory match (i.e., statistically significant) between the real cab data and the data registered by the APs. The possible scenarios for AP distribution are the following.

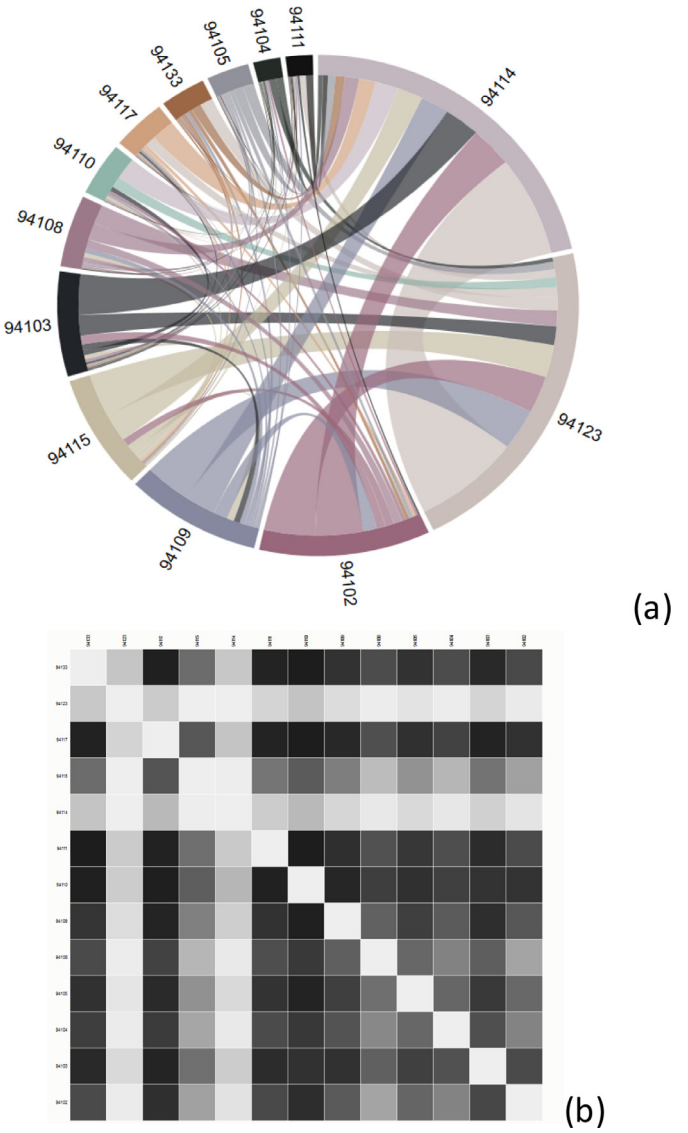

Fig. 4. (a) Chord diagram of flow counts with real Wi-Fi APs in the city center; (b) Difference matrix among OD matrices of real flows and estimated with real Wi-Fi APs in the city center.

a) *Random APs*: identification of the streets with the highest trace flow rate (those that have at least 3000 traces) and then random selection of 400 APs from the AP grid described above (see Fig. 5a for the OD matrix). This set of APs is a subset of the set described in case (b).
b) *High Traffic APs*: identification of the streets with the highest trace flow rate (those that have at least 3000 traces) and then selection of all the APs intersecting those traces, thus resulting in 804 APs (see Fig. 5b for the OD matrix).
c) *High Traffic APs (zip boundaries)*: identification of the streets with the highest trace flow rate (those that have at least 3000 traces) and then, starting from the 804 APs of case (b), selection of those within 300 m from the zip boundaries, thus resulting in 448 APs (see Fig. 6c for the OD matrix). This set of APs is a subset of the set selected in case (b).
d) *High Traffic APs (top 400)*: identification of the streets with the highest trace flow rate (those that have at least 3000 traces) and then, starting from the 804 APs of case (b), selection of the top 400 APs (see Fig. 5d for the OD matrix). This set of APs is a subset of the set selected in case b.
e) *Real augmented APs with selected high traffic APs* (Fig. 5e): the real distribution of the AP in San Francisco's downtown was integrated with the top 300 AP from case (d) with the highest traffic rate. This set was then cleaned up by removing those APs that were found to be at a distance less or equal than 50 m from the real APs, and removing also intersecting APs, thus resulting in 400 APs (221 real APs, 179 high traffic APs).

The resulting OD matrix for these distributions of APs has been estimated by computing the intersections between the real cab measures with the placed APs, according to a capturing range of 25 m radius. The OD matrix for this configuration was generated by evaluating the traffic counts among the various APs, grouped by the zip code they belong to. The chord diagrams of these scenarios are reported in Fig. 5.

3.2. Assessing AP positioning models

A comparative analysis of traffic flows was conducted, using the above cited set of cab traces, consisting of 11,219,955 unique detections from 536 cabs, with respect to the above described scenarios. With the above assumptions, the real set of APs placed in the city centre was used to sample the original data set, by calculating the APs intersections with the cab traces. The OD matrix was calculated from the sampled data set (considering each city zip code as a separate area), reporting the traffic counts among every city's area. This procedure was repeated by choosing the APs with a pseudo random technique, and by placing the APs only in the roads with the biggest amount of traffic. After that, a comparative statistical analysis was conducted for each configuration (see Table 1). The traffic flow outcome is predicted with a linear regression, finding the parameters that best fit the data in the linear model

$$y = \alpha x + \beta \tag{3}$$

where x is the dependent variable or predictor (i.e., traffic counts as registered by the sensors), and y is the outcome (i.e., predicted traffic counts). Building the model in (3) using the set of real APs gives a correlation of 0.446 (0.120 using the real APs in the city's downtown) with respect to the real traces. A number of cases have been assessed following the placement strategies described in Section III. In case (b), the APs have been placed on the roads with the highest traffic rate, producing a model with a correlation of 0.915, and of 0.835 using only the top 400 APs, as described in case (d); using random APs of case (a) gives a correlation of 0.835; using the APs only within 300 m from the areas' boundaries, described in case (c), gives a correlation of 0.687. It is clear from this data that using the real APs set produces noise and doesn't produce a reliable model for flows prediction. Randomly distributing the APs gives a better correlation with the cab traces, while reducing the number of APs and considering only those in the proximity of each area, gives a good correlation while maintaining a limited number of APs. The set of real APs of case (e), integrated with some other APs and cleaned up from some not useful or redundant elements (i.e., mutually intersecting APs), gives a correlation of 0.892. To visualize the results of the various OD models a web interface was developed, with the possibility to view the chord diagram for each computed configuration. The interactive versions of the chords diagrams in which it is possible, for each couple of locations, to see the effective flows (in a way and in the other, for a given time slot of the day) are accessible at http://www.disit.org/6694. This approach allowed to identify which are (i) the positions of the new APs to be added (i.e., 179) and (ii) the minimum set of APs already in place that must be used for data acquisition (i.e., 229). The second point allows keeping limited both the network bandwidth and the workload for the estimation of the OD matrix.

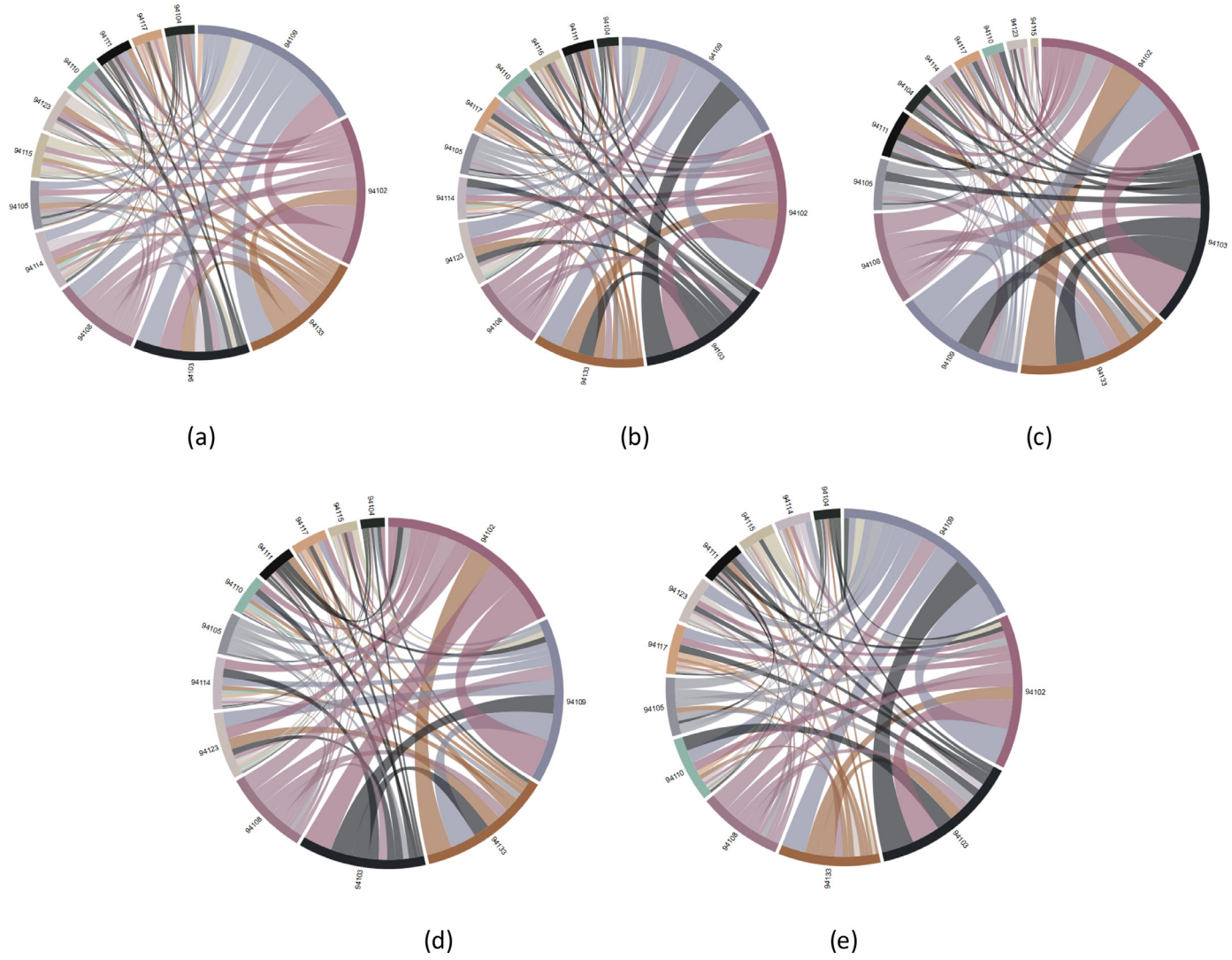

Fig. 5. Chord diagram of flow counts. Cases as described in Table 1: (a) Random APs; (b) High traffic APs; (c) High traffic APs (zip boundaries); (d) High traffic APs (top 400); (e) Real augmented APs.

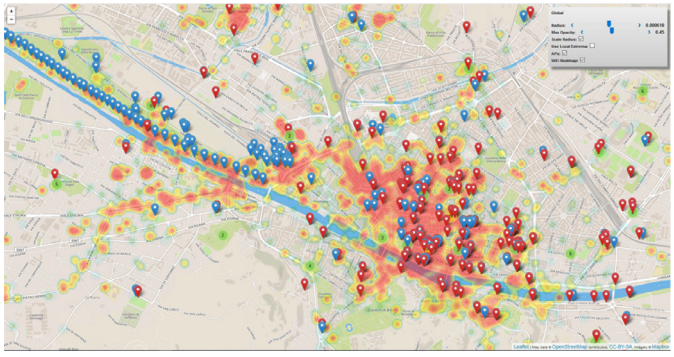

Fig. 6. Heat-map comparing city users' most frequented places vs the position of the 1500 Wi-Fi APs of the whole network (using a colour gradient scale to discriminate between different densities of measures). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A fully mathematical approach could be applied for the identification of the best AP in San Francisco having dense traces, but it would not be suitable for the re-computing it in a new fresh area (without data). In substance if a position of APs is identified in San Francisco just minimising the error, the position of the AP would not follow any rule that could be re applied in a different city to position the APs or select the APs to be reconfigured. Thus we decided to test a set of heuristics and select the best, and thus to use the identified approach to position/select the AP in Florence.

4. City user's behaviour analysis

The above described AP placing methodology has been exploited in the city of Florence (Italy), for selecting the AP needed for the estimation of city users' behaviour.

Typically, it can be supposed to derive users' behaviour from data collected from the telecom operators. On the other hand, the mobile operators are not authorized in reselling data reporting the fine tracking of their users, even if the mobile/user ID is anonymized. In most cases, mobile operators provide data collected every 15 min, reporting the number of users for each cluster of their cells and without tracking the movements from one cell/cluster to another. Some of them provide OD matrixes statistically estimated starting from the described data and thus providing a limited precision in space and time, and not in real time. These facts limit the possibility to use those data to perform a city users'

behaviour analysis, area clustering and the usage of data for early warning.

On the contrary, the usage of Wi-Fi network can be used for tracking city users' behaviour with the needed resolution (in space and time), by accessing to data anonymously and exploiting them according to an informed consent with the users when they connect to the Wi-Fi. The above presented methodology for AP placement has been used on the Firenze Wi-Fi infrastructure to identify the suitable APs to be considered for the analysis, with the aim of reconstructing city users' behaviour in space and time. At this regard, Florence offered a free Wi-Fi network (Firenze Wi-Fi) consisting of about 1500 APs. One relevant issue is that Firenze Wi-Fi APs were installed with the aim of providing a good Wi-Fi coverage in the city's centre and in relevant city services as hospital and university.

As a first step, we identified the most active places and areas to the monitored, on which the above presented methodology would be applied. This action has been performed by interviewing the municipality and by using data collected from mobile App (Florence, Where, What?), available for Android, iOS and Windows Phone stores [11]. That App work with smart city API based on Km4City [4] and provides general information to the city users almost uniformly in the city and on multi-domain since it provides information and suggestions on: public and private mobility, culture, energy, accommodation, restaurant, tourism, free Wi-Fi, bus lines, car parking, pharmacies, ATMs, events, etc. These services are accessible with geo information.

Fig. 6 reports the heat-map derived from the city users' movements in the city by using the App with overlapped the position of the 1500 AP of the Wi-Fi network. Considering the architectural and environmental constraints of the historical centre of Florence (that is part of the UNESCO World Heritage list), you cannot place APs wherever you want: in most cases, we have to switch on the nearest AP to the predicted one, rather than effectively place the desired AP. The resulted analysis allowed us to select the best points and from these about 345 candidates APs to be configured and used as probes, selected from more than 1500 AP located in the city. The data related to the user behaviour tracking via Wi-Fi has been collected in the period from May 2016 to December 2016. They consist of about 56 Million of events of connection and disconnection. Typically, the 60% of connected users are excursionists that stay in the network only for less than 24 h. In the last 6 months, about 1.15 Million distinct users have been detected, which means about 2.3 million of distinct user per year in a city with about 14 million of new arrivals per year and 350.000 inhabitants. So that we tracked about the 16% of people flow. We compared the predictions from the positioning methodology with the existing APs data, finding the APs to be added and those that were useless for the study. According to the selected AP, the resulting heat-map describing the distribution of measures performed by the AP is reported in Fig. 7. The developed tool allows customizing the provided map, for example varying the radius and the opacity of the heat spots.

The data analysis allows identifying the hottest places (in terms of events on the APs) as reported in Fig. 8, where the names of the locations and the precise latitudes and longitudes have been truncated for safety reasons. On the other hand, they are also well known location to everybody in the world.

Similarly, a number of visual analytics graphs are produced, such as: the numbers of distinct users during the day, the average connection time per AP, the number of working APs in the last minutes, the regency (percentage of new users with respect to the already seen users) and frequency of users. This last view is of particular importance since it allows estimating the number of new users coming into the city. Indeed, it is worth noting that for cultural cities like Florence, newcomers are typically tourists (excursionists) or business people that stay in the city only for a few hours and days.

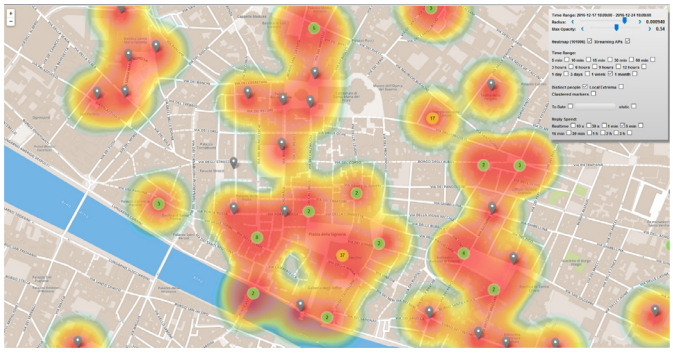

Fig. 7. Segment of the heat-map reporting the hottest places detected by using selected Firenze Wi-Fi APs, in Florence downtown.

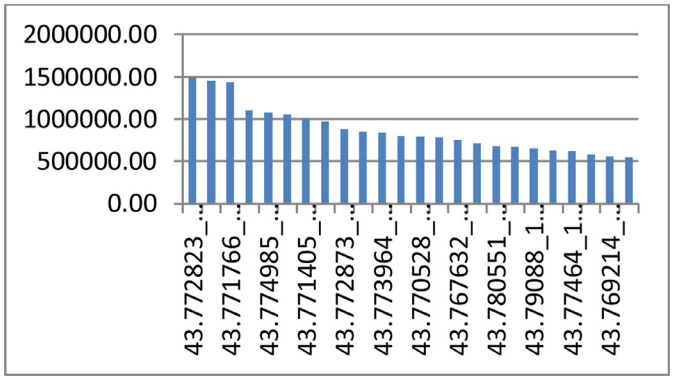

Fig. 8. Distribution of hottest places in the city (truncated series), number of Wi-Fi accesses in last 180 days.

Every working day the network identifies about 34.000 distinct users and among them, about the 10% are new users for the network in the period. For the present analysis, we assumed that new users exploit the city up to 10 days before leaving, while old users continue to exploit the city beyond that limit.

Fig. 9 reports the users regency found in the range 1–28 days. Every column in the histogram shows the number of distinct users (y-axis) that at most returned in the city within a defined number of days (x-axis). It is evident from this pattern, that most of the users using the Wi-Fi network are exploiting the city for a few days before leaving. This kind of analysis can be performed at large scale (i.e., considering the whole city) or simply by observing the user behaviour in some zones of interest. For example, the analysis of regency in the historical city centre (which is normally the most exploited part of the city) can provide valuable insights, since it allows understanding which cultural attractions people prefer to visit, or where and how often they return to them.

4.1. Origin destination analysis for people flow

To better understand the movements in the city, it is mandatory to perform flow analysis to effectively evaluate user's behaviour. Since in the downtown the APs are also overlapped this issue has to be taken into account. The measures performed by the mobile APP (as described in first part of Section 4) have been also used to define a compromising size for each area collecting accesses to the Wi-Fi. On the basis of the tracked city users among the APs of the Wi-Fi network it is possible to computer the OD matrix according to the origin and destination area defined by the distribution of the APs in the city. On the other hand, the OD matrixes are typically

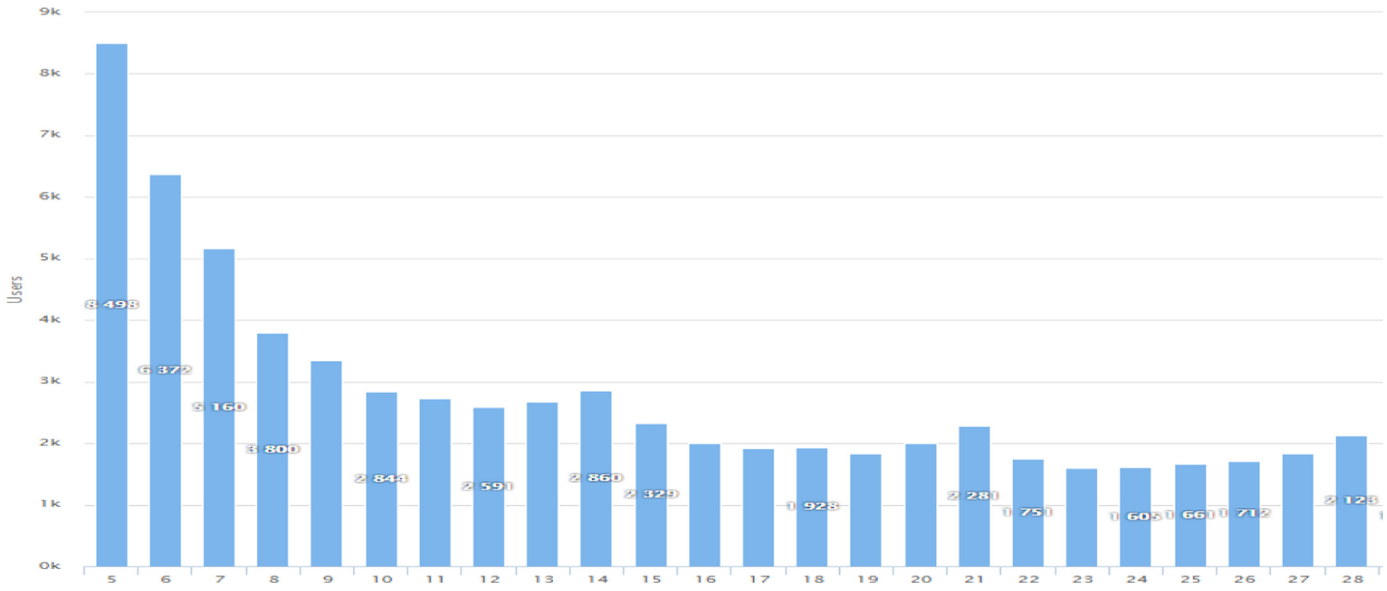

Fig. 9. Number of distinct users accessing to the Wi-Fi network, regency from 1 to 28 days.

quite sparse as one can see in Fig. 10a, where the OD matrix for Florence is reported.

Fig. 10b reports a new approach for depicting and analysing the OD matrixes. It is a visual analytic approach for depicting an OD matrix as what we call *OD Spider Flow* in which the analyst may identify the hottest areas of the city as those with larger and darker points/dots. When a dot is selected the graph reports the major (in/out) flows from that origin to the most probable destinations, also providing the percentage of probability on the destination dots. Every flow is depicted with an arrow and a coloured circle reporting the total number of occurrences and their percentage with respect to the total flows. The analysis can be performed for the whole city users or only for the new arriving users (with respect to the last 10 days), for each time slot of the day or for the whole day, for incoming and outgoing flows, and at different level of resolution (zoom). Zooming in/out the map redraws the flows with a different cluster zone, making possible to depict more detailed or aggregated flows between the various zones. The classical OD matrix can be shown as well from the same tool, also calculated with a customizable range within the city's centre, for the chosen flow configuration (i.e., cluster area's size, hour of the day, user profile). This kind of derived information can be used for running the services in the city, to plan the cleaning, to distribute the security people, etc.

4.2. Understanding city usage from AP data

From the analysis of the OD matrices and/or OD Spider Flows it is evident that different parts of the city are differently used by different city users. AP presents different kind of trends in the usage of the Wi-Fi along the 24 h and in the different days of the week [28]. For example, we may have some areas by which the people typically arrive (station) in the morning and leave in the afternoon while they are less accessed at lunch time. For example, some APs could have a huge workload only during mornings or evenings (when people go/back to/from work), others only on late evenings (when people go out for entertainment), others only of festive days etc.

In Fig. 11, an example of trend for a certain AP along the 24 h of the day. The trend of Fig. 11 has been estimated by computing the averaged value per time slot of a certain AP every working day, extracting data from the 56 million of data described above.

In Florence, as in many other touristic cities, the issue is much more complex, since a lot of different city users' kinds (with different aims) use the city at the same time during the working days, and as well as on Saturday and Sunday.

Therefore, in order to tune the services in the city (security, cleaning, transport, etc.), it is very important to infer patterns and analyse city user's behaviour. In the present scenario, the major interest is related to understand how the city is used by city users which in turn can be re-conduced to the problem of understanding how APs work and are used. The idea is to exploit some data mining techniques clustering AP on the basis of their normalized temporal pattern. This will allow grouping them in areas and put in evidence the flows and the service exploitation in the different city's zones. Clustering the APs' behaviours can help to understand if there are zones having a similar usage and exploitation and hence similar flow patterns, and needs in terms of services.

According to the data collected from the Wi-Fi network described at the beginning of Section 4, the averaged trend along the 24 h of the day, for each AP, for each day of the week has been computed. Since the main interest is to find the similar patterns for each AP a Scale Factor and the normalized averaged pattern (from 0 to 1) has been computed. This resulted in 345 APs, on 7 days, on 48 time slot for the day (one every 30 min) (from 00:00 to 00:30, from 00:30 to 01:00 and so on until 23:30). A preliminary analysis of AP patterns showed a marked difference between festive and ferial days. For this reason, we chose to cluster the time series by keeping track of their respective day of week, thus considering working days, Saturdays and Sundays as three distinct groups. From the statistical point of view, the temporal pattern for each AP presents an average and an interval confidence for each time slot as depicted in the examples reported in Fig. 13.

Since we are interested in finding similar patterns for the APs, a clustering approach has been adopted to find similarities in time series as in the Dynamic Time Warping [42], and by using different clustering algorithms and metrics to evaluate both the better ranked clustering algorithm and the proper number of clusters.

Among the clustering algorithms we compared the results obtained by using: k-means clustering algorithm minimizes the within-class sum of squares for a given number of clusters [26,33] hierarchical clustering [39], density-based clustering or subspace clustering. Unlike k-means clustering, hierarchical cluster-

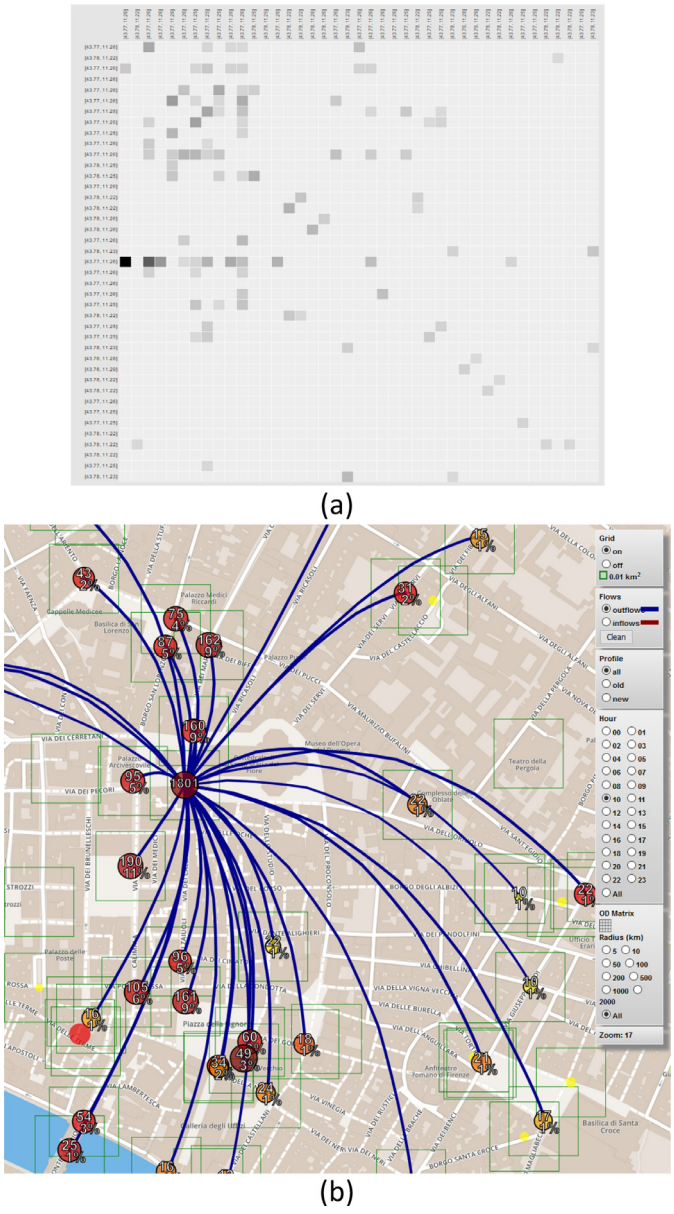

Fig. 10. OD Matrix for Florence downtown: (a) classical view; (b) advanced proposed view. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

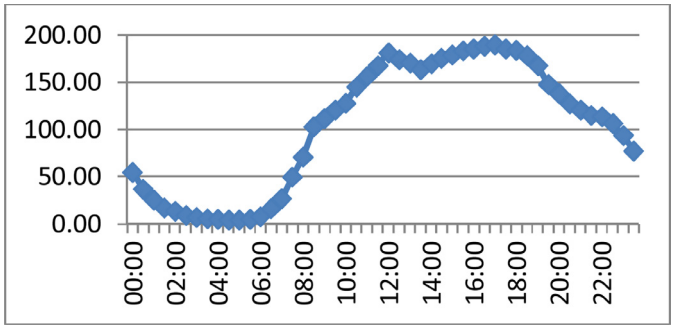

Fig. 11. Typical AP trend in terms of number of connections along the 24 day, a working day.

Table 2
Geometric characteristics of mixture models.

Model	Distribution	Volume	Shape	Orientation
EII	Spherical	Equal	Equal	–
VII	Spherical	Variable	Equal	–
EEI	Diagonal	Equal	Equal	Coordinate axes
VEI	Diagonal	Variable	Equal	Coordinate axes
VEE	Ellipsoidal	Variable	Equal	Equal
VVE	Ellipsoidal	Variable	Variable	Equal

ing builds a bottom-up hierarchy, and does not need to specify the number of clusters. For the clustering, the closeness of cluster elements can be determined by using (a) complete linkage clustering (i.e., finds the maximum distance between points of two clusters), (b) single linkage clustering (i.e., finds the minimum distance between points of two clusters), (c) mean linkage clustering (finds all pairwise distances for points of two clusters, calculating the average), (d) centroid linkage clustering (i.e., finds the centroid of each cluster and then calculate the distance between the centroids of two clusters).

4.3. AP clustering experimental results

In this section, the comparative analysis among some of the above mentioned different clustering methods is reported. It should be noted that, different clustering techniques and, even for the same algorithm the selection of different parameters or the presentation order of data objects may greatly affect the final clustering partitions. Thus, the adoption of rigorous evaluation criteria is mandatory to trust the cluster results: selection of model and clusters number.

As first step, we have tested cluster tendency i.e., the hypothesis of the existence of patterns in the data using the Hopkins statistics [8]. Hopkins statistic has been used to assess the clustering tendency of the dataset by measuring the probability that a given dataset is generated by a uniform data distribution (i.e., no meaningful clusters). Hopkins statistic is equal to 0.2186, thus the data is clusterable.

As a second step, two clustering techniques have been adopted and compared using the above described observations and data sets of AP patterns in the 24 h (Monday-Friday, Saturday and Sunday). The first technique was a sort of K-means clustering algorithm, partitioning around medoids (PAM) which are the most representative elements in the cluster instead of the centroid as in the k-means. PAM approach is also called K-medoids [29]. The second approach is the Model-based Expectation- Maximization algorithm or EM algorithm (EM method) [18],[34]. It is a generalization of the k-means approach that uses an iterative process to find the maximum likelihood (or the maximum a posteriori estimates of parameters, MAP). The algorithm's iteration consists of two steps: the expectation step (E) which, using the parameters' current estimation, calculates a function for the expectation of the respective log-likelihood; and a maximization step (M) which calculates the parameters maximizing the expected log-likelihood from step (E). The estimated parameters are used to calculate the distribution of latent variables in the next iterative step E.

A model-based method was used to evaluate the number of clusters/groups and the BIC criteria to determine the best model [22]. Under this approach, each mixture component represents a cluster, and group memberships are estimated using maximum likelihood [18]. The maximum likelihood estimator (MLE) of a finite mixture model is usually obtained via the EM algorithm [18],[34]. In the multivariate setting, the volume, shape, and orientation of the covariances can be constrained to be equal or variable across groups. Table 2 reports six possible models with the

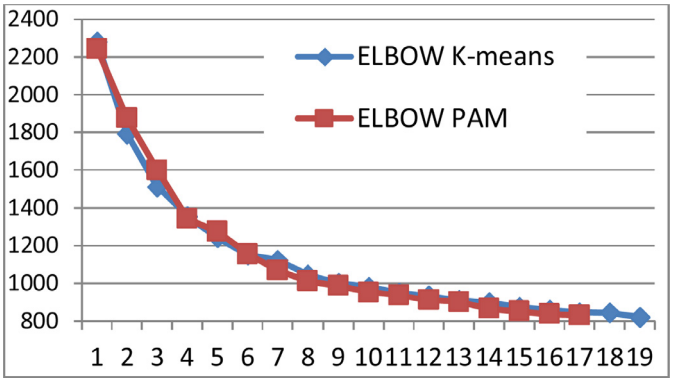

Fig. 12. Optimal number of AP clusters via Elbow criteria (comparing K-means and PAM): within sum of square function.

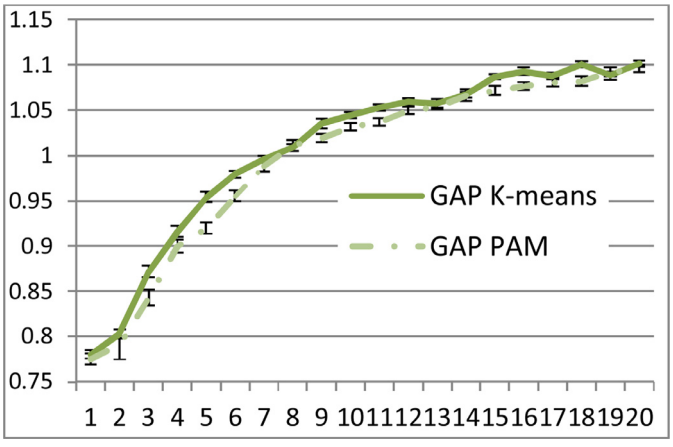

Fig. 13. Optimal K number of clusters via Gap curve (comparing K-means and PAM).

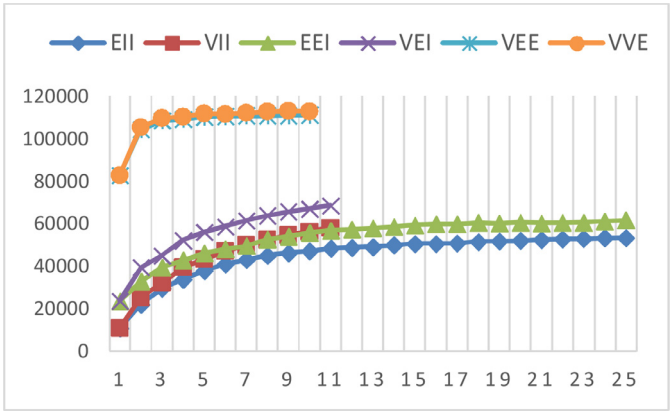

Fig. 14. Average BIC for mixture models vs K number of cluster, higher values are better, the curves are truncated at the best value for K they found.

Table 3
Standard deviation and population for AP clusters. W: Working days, Sa: Saturday, Su: Sunday.

Cluster Id	Avg. std. dev.	Population
1	0.2379	W: 172, Sa: 23, Su: 24
2	0.0849	W: 23, Sa: 43, Su: 43
3	0.0882	W: 8, Sa: 42, Su: 34
4	0.1820	W: 3, Sa: 30, Su: 26
5	0.1059	W: 20, Sa: 15, Su: 14
6	0.0822	W: 38, Sa: 15, Su: 8
7	0.1311	W: 9, Sa: 57, Su: 34
8	0.1374	W: 2, Sa: 23, Su: 55
9	0.1226	W: 4, Sa: 32, Su: 38
10	0.1460	W: 52, Sa: 12, Su: 3
11	0.2487	W: 11, Sa: 13, Su: 21
12	0.1617	W: 1, Sa: 28, Su: 31

corresponding distribution structure type, volume, shape, orientation, and associated model names.

With the Elbow method (as reported in Fig. 12), the solution criterion value (within groups sum of squares) will tend to decrease substantially with each successive increase in the number of clusters: after 8 clusters the observed difference in the within-cluster dissimilarity is not substantial. Consequently, we can say with some reasonable confidence that the optimal number of clusters to be used seems to be 7. Note that identifying the point in which a "kink" exists is not a very objective approach and is very prone to heuristic processes. For these reasons, we computed the Gap statistics [40] to assess the optimal number of clusters in the data. From this analysis reported in Fig. 13, the estimated number of clusters $K = 12$.

Finally, Fig. 14 shows the average BIC (Bayesian Information Criteria) values for six different mixture models using the model-based approach over a range of different numbers of clusters [34]. With the VEE mixture model, the maximum average BIC score is reached at 10 clusters. In addition, the VVE mixture model also achieves higher BIC values than the VEE model up to 10 clusters. Therefore, the model-based approach favors the diagonal model which produces higher quality clusters. The BIC analysis selects the VVE model at 10 clusters. Note that although the BIC analysis does not select the best model, it allowed selecting the better number of clusters in this data set.

We used the Dunn index [20] as a measure to assess the validity of cluster techniques. Dunn index is based on inter-cluster distance and the diameter of cluster hypersphere. It can be seen that PAM clustering performs the best with 12 clusters (Dunn index for PAM is equal to 0.0798, for K-means is equal to 0.0730 and for Model-based is equal to 0.0478).

As a final result, the EM algorithm with 12 clusters has been adopted for massive and continuous computing. On this regard, Table 3 reports the average standard deviation and the related population of each AP cluster.

In Fig. 15, the distribution of clustered AP in the Florence map for day kind: Monday-Friday, Saturday and Sunday in which AP of the identical color belong to the same cluster disregarding the day kind. From Fig. 15, it can be noticed that group of APs located in the Cascina park (black) is enlarging passing from working days to Sunday, while the cluster of downtown (dark red) is losing some of its APs passing from working days to Sunday. While some of them remain stable: mainly those located in the major attractions for tourists. The maps reported in Fig. 15 can be easily accessed by a real time tool accessible for the municipality of Florence. Each cluster has a different color, and clicking on an AP opens a popup with detailed data about the specific AP and the cluster at which it belongs to (i.e., cluster id, maximum, minimum, average flow and standard deviation). In this way, we are able to see in an intuitive manner if there are adjacent zones that show similar AP daily patterns.

Fig. 16 reports the normalized shapes of the 12 clusters identified which resulted from the best clustering algorithm, the EM. It can be noticed that the second cluster presents APs with relevant activity during the morning and afternoon respecting a break for lunch. Moreover, some clusters provide an evident activity in the afternoon with respect to the morning or vice versa, but with different proportions. A few of them present significant activity also after dinner and in the first hours of the night, as clusters number

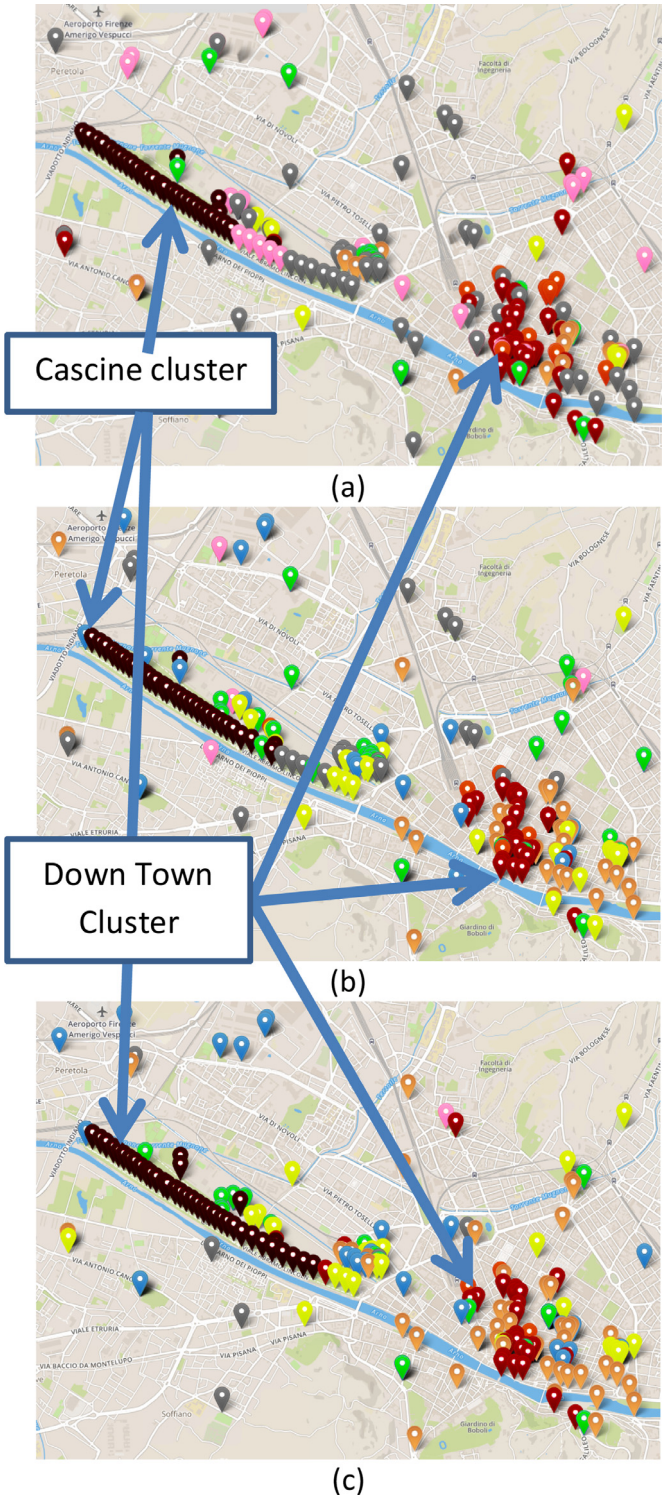

Fig. 15. Map of AP clusters: (a) Monday-Friday, (b) Saturday, (c) Sunday. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

1 and 9. So that, it is evident where the city is active during the night.

5. Predicting access point connections

The data collected from the Firenze-Wi-Fi network have been analysed to derive a number of information and knowledge: heat map as most frequent places and hottest areas in the city, daily user behaviour patterns in the city area to understand how the city is used, OD matrix to extract people movements. In this section, the development of a model for predicting the number of connections of each specific AP in the city is presented. The number of connections of an AP is directly related to people presences. And thus, it can be used for planning in advance, as well as, it poses the basis to be used as an instrument for early warning: that is for detecting dysfunctions as un-expected patterns in the city users' behaviour. To this end, the autoregressive integrate moving average approach (ARIMA), have been adopted as solutions to set up accurate predictive models in order to detect dysfunctions. The autoregressive part (AR) of model creates the basis of the prediction, which can be improved by a moving average modelling for errors made in previous time instants of prediction (MA). The order of ARIMA modelling is defined by the parameters (p,d,q): p is the order of autoregressive model; d is the degree of differencing, and q is the order of the moving average part, respectively. The predictive model has been developed by using Box-Jenkings methodology as ARIMA modelling [12], and the solution has been compared in terms of performances with a set of other models.

We chose to consider the time series by dividing the week in three distinct groups, thus considering working days, Saturdays and Sundays, in order to maintain consistency with the cluster analysis.

For the analysis, we have applied different predictive ARIMA models for each 30 min interval, for each groups of days and for each AP. Note that, for each time interval we estimates the best ARIMA model according to the AIC, Akaike Information Criterion [1]. In most cases, the best predicting model has been an ARIMA [5,1,0] (it is an ARMA model), meaning that the model takes into account of 5 observations from the past, and by the difference of the last two observations. The best AICs have been obtained in the range of 1000–1300 in different time slots. Better predictive results have been obtained for the AP in which a significant number of accesses are typically present.

The ARMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. For ARMA [5,1] we have:

$$\begin{aligned} y_t - y_{t-1} = {} & \varepsilon + ar_1(y_{t-1} - y_{t-2}) + ar_2(y_{t-2} - y_{t-3}) \\ & + ar_3(y_{t-3} - y_{t-4}) + ar_4(y_{t-4} - y_{t-5}) + ar_5(y_{t-5} - y_{t-6}) \end{aligned}$$

Where: ar_1, ar_2, ar_3, ar_4, ar_5 are determined during the identification of the model minimising the root square error during the learning period, ε is an independent variable with normal distribution and zero mean.

In Fig. 17, two examples of AP time series with prediction are reported. Each of them reports: in blue line the average value of the cluster at which the AP belong; the light blue bound describes the interval confidence of the reference cluster of the AP; the red line the actual value of the day; the orange bound describe the interval confidence obtained by the distribution of the value of the AP in the past; finally, the RED segment (second part segment) is the effective prediction by using the ARIMA model. Please note that, the adopted ARIMA model does not take into account the value collected by the same AP in the day, since we would like to use the predictive model for detecting dysfunctions and not to follow the most probable next values. The detection of critical situation can be obtained making the difference from those two approaches/ estimations.

6. Conclusions

Understanding and predicting city user behaviour is one the major topics in the context of Smart City to optimize and tun-

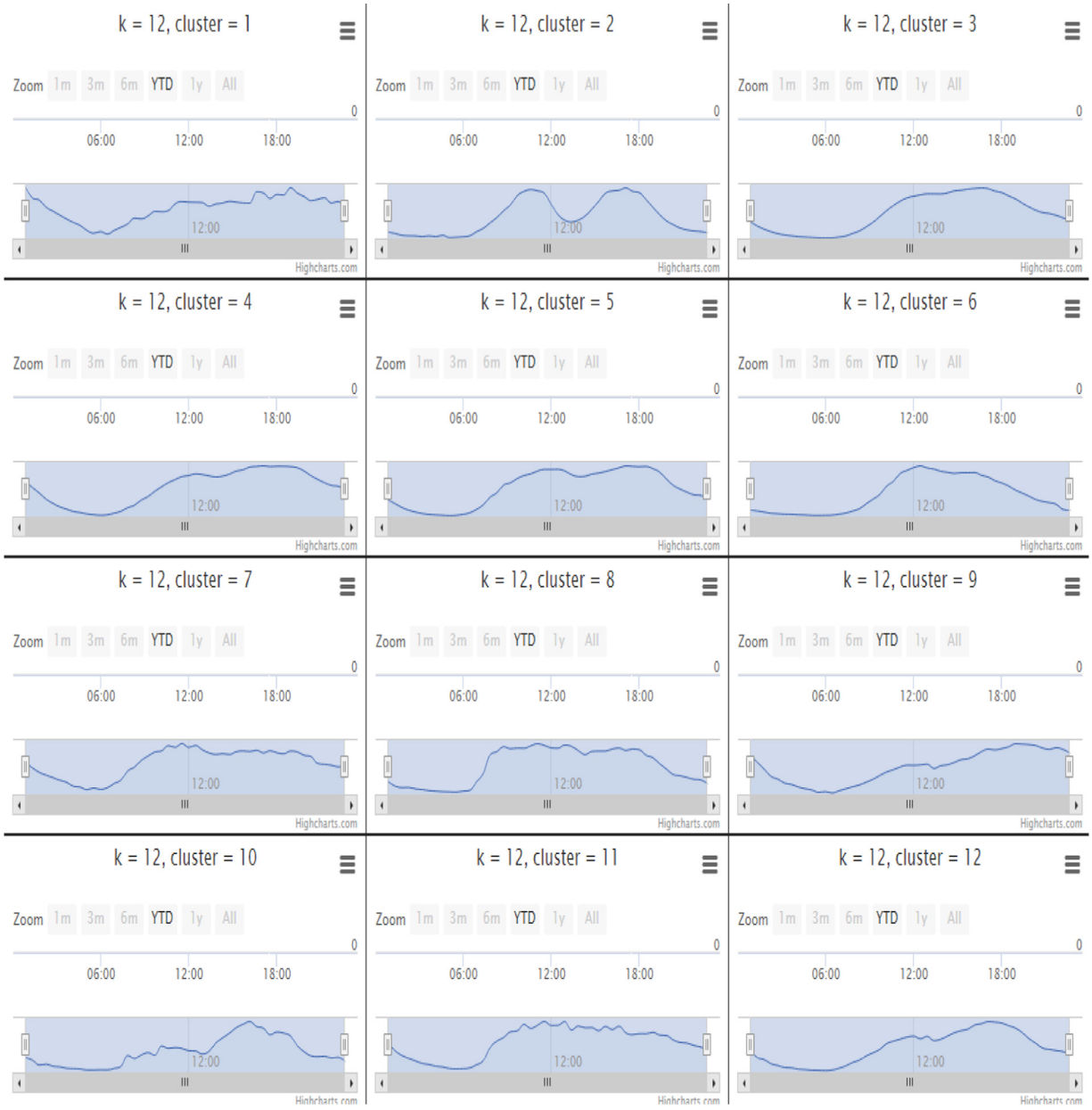

Fig. 16. The shapes of the AP clusters with k = 12 and EM clustering algorithm.

ing city services (security, clean, transport,..) and to be ready in reacting via anomaly detection. In this paper, we presented a method for AP placement, and a number of algorithms, techniques and solution for estimating city user behaviour: heat-map, OD Spider Flow, clustering of AP usage in the city, predictive model.

The proposed methodology is general and can be applied to different urban scenarios, in the context of Smart City people flow assessment and management. It makes use of Wi-Fi AP distributed in the city. Comparative analysis has shown that is possible to have a reasonable precision in assessing city behaviour by AP positioning and collecting data from Wi-Fi, as demonstrated by a validation based on real data. The proposed approach allows identifying which are the needed APs to be added, with respect to the APs that are already in place in the city, to exploit the whole infrastructure of Wi-Fi, also for people flow monitoring and assessment.

The proposed methodology has being applied to identify significant APs in the city of Florence (Italy). Wi-Fi. Thus collected data were analysed to produce usage metrics and studying AP usage. To this end, several clustering techniques have been adopted to identify the better clustering approach for grouping city users' usage trends in the day for each city area. The results have shown that about 12 different major clusters/patterns have been identified. Each AP can be classified with respect to a cluster trend and provides its specific own scale. The corresponding AP data, trend, and cluster can be used for predicting number of accesses and thus city usage, as a well as for detecting unexpected trends incepting in the different places of the city (they may be due to programmed events as well as to detect anomalies as early warning tool). We performed our analysis by using various clustering algorithms and calculated different informative criterions to select the best and assess their objective quality. The resulting model proves to be effective for connection to AP forecast in the entire Wi-Fi network and potentially for early warning.

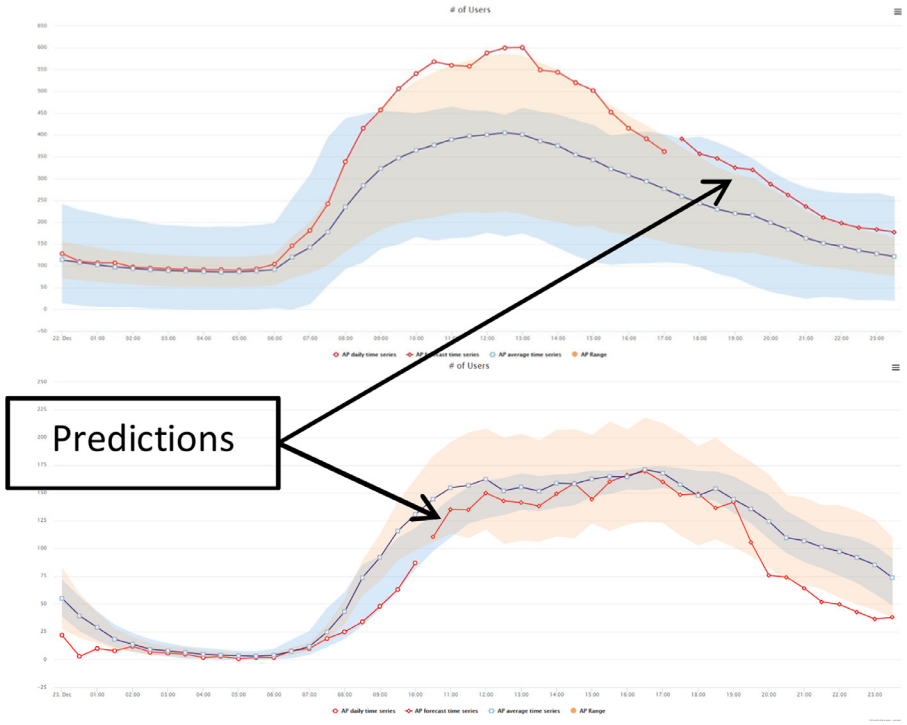

Fig. 17. APs time series with their respective cluster ranges (see details in the text).

Acknowledgements

The research work described in this article has been conducted for RESOLUTE project. The authors would like to thank the Municipality of Florence for the collaboration in collecting data, and to the European Commission for funding the RESOLUTE H2020 project Grant Agreement n. 653460 (http://www.resolute-eu.org).

References

[1] H. Akaike, Factor analysis and AIC, Psychometrika 52 (3) (1987) 317–332.
[2] A. Alessandrini, et al., WiFi positioning and Big Data to monitor flows of people on a wide scale, Navigation Conference (ENC), 2017 European, IEEE, 2017.
[3] K. Ashok, M.E. Ben-Akiva, Alternative approaches for real-time estimation and prediction of time dependent origin-destination flows, Transp. Sci. 34 (2000) 21–36.
[4] C. Badii, P. Bellini, D. Cenni, G. Martelli, P. Nesi, M. Paolucci, Km4City Smart City API: an integrated support for mobility services, in: 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016), St. Louis, Missouri, USA, May 2016, pp. 18–20.
[5] C. Badii, P. Bellini, D. Cenni, A. Difino, P. Nesi, M. Paolucci, Analysis and assessment of a knowledge based smart city architecture providing service APIs, Future Gener. Comput. Syst. (2017). Elsevier http://dx.doi.org/10.1016/j.future.2017.05.001 .
[6] X. Ban, R. Herring, J.D. Margulici, A.M. Bayen, Optimal sensor placement for freeway travel time estimation, 18th International Symposium on Transportation and Traffic Theory (ISTTT), July 2009.
[7] X. Ban, L. Chu, R. Herring, J.D. Margulici, Sequential modeling framework for optimal sensor placement for multiple intelligent transportation system applications, J. Transp. Eng. 137 (2) (2011).
[8] A. Banerjee, R.N. Dave, Validating clusters using the Hopkins statistic„ in: 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542), vol. 1, 2004, pp. 149–153.
[9] X. Bao, H. Li, L. Qin, D. Xu, B. Ran, J. Rong, Sensor location problem optimization for traffic network with different spatial distributions of traffic information, Sensors (Basel). 2016 Nov; Vol.16, no. (11): 1790.
[10] M. Bartolozzi, P. Bellini, P. Nesi, G. Pantaleo, L. Santi, A smart decision support system for smart city, in: IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, IEEE press, Cina, December 2015, pp. 117–122.
[11] P. Bellini, M. Benigni, R. Billero, P. Nesi, N. Rauch, Km4City ontology building vs data harvesting and cleaning for smart-city services, J. Vis. Lang. Comput. 25 (6) (2014) 827–839.
[12] G.E. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2011 Sep 20.
[13] A. Caragliu, C. Del Bo, P. Nijkamp, Smart cities in Europe, 3rd Central European Conference in Regional Science CERS, 2009.
[14] Report on the State of Play of the Outcome of the Work of the High Level Group, October 2014 European Commission.
[15] E. Cascetta, M.N. Postorino, Fixed point approaches to the estimation of O/D matrices using traffic counts on congested networks, Transp. Sci. 35 (2001) 134–147.
[16] S. Contreras, P. Kachroo, S. Agarwal, Observability and sensor placement problem on highway segments: a traffic dynamics-based approach, IEEE Trans. Intell. Transp. Syst. 17 (3) (2016).
[17] A. Danalet, M. Bierlaire, B. Farooq, Estimating pedestrian destinations using traces from WiFi infrastructures, Pedestr. Evacuation Dyn. 2014 (2012) 1341–1352 Springer International Publishing.
[18] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Society Ser. B. 39 (1977) 1–38.
[19] J. Doblas, F.G. Benitez, An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix, Transp. Res. Part B 39 (2005) 565–591.
[20] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (3) (1973) 32–57.
[21] X. Fei, H.S. Mahmassani, P. Murray-Tuite, Vehicular network sensor placement optimization under uncertainty, Transp. Res. Part C 29 (2013) 14–31.
[22] Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. (17) (2001) 126–136.
[23] Y. Fukuzaki, et al., A pedestrian flow analysis system using Wi-Fi packet sensors to a real environment, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM, 2014.
[24] R. Giffinger, C. Fertner, H. Kramar, R. Kalasek, N. Pichler-Milanovic, E. Meijers, Smart Cities Ranking of European Medium-Sized Cities, 2007 Available at http://www.smartcities.eu/.
[25] Z. Gong, Estimating the urban o-d matrix: a neural network approach, Eur. J. Oper. Res. (106) (1998) 108–115.
[26] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a k-means clustering algorithm, J. R. Stat. Society Ser. C 28 (1) (1979) 100–108.
[27] J. Ivanchev, H. Aydt, A. Knoll, Information maximising optimal sensor placement robust against variations of traffic demand based on importance of nodes, IEEE Trans. Intell. Transp. Syst. 17 (3) (2016).
[28] S. Jiang, J. Ferreira, M.C. González, Clustering daily patterns of human activities in the city, Data Min. Knowl. Discov. 25 (3) (2012) 478–510.
[29] L. Kaufman, P.J. Rousseeuw, Partitioning around medoids (program pam), in: Finding Groups in Data: An Introduction to Cluster Analysis, 1990, pp. 68–125.
[30] H. Kim, S. Beak, Y. Lim, Origin-destination matrices estimated with a genetic algorithm from link traffic counts, Transp. Res. Record 1771 (2001) 156–163 Transportation Research Board, Washington, D.C.
[31] E. Lovisari, W.C. Canudas de, A.Y. Kibangou, Optimal sensor placement in road transportation networks using virtual variances, IEEE 54th Annual Conference on Decision and Control (CDC), 2015.

[32] J.T. Lundgren, A. Peterson, A heuristic for the bilevel origin-destination matrix estimation problem, Transp. Res. Part B 42 (2008) 339–354.
[33] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: L.M.L. Cam, J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[34] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[35] Y. Nie, H.M. Zhang, W.W. Recker, Inferring origin-destination trip matrices with a decoupled GLS path flow estimator, Transp. Res. Part B 39 (2005) 497–518.
[36] P.H. Patil, A.A. Kokil, WiFiPi-tracking at mass events, Pervasive Computing (ICPC), 2015 International Conference on, IEEE, 2015.
[37] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser, A parsimonious model of mobile partitioned networks with clustering, in: Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International, Jan. 2009, pp. 1–10.
[38] L. Schauer, M. Werner, P. Marcus, Estimating crowd densities and pedestrian flows using wi-fi and bluetooth, in: Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.
[39] R. Suzuki, H. Shimodaira, pvclust: Hierarchical Clustering with P-Values via Multi-scale Bootstrap Resampling, 2014 R package version 1.2-2.
[40] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Society (63) (2001) 411–423.
[41] F. Ting, X. Hong, Discovering meaningful mobility behaviors of campus life from user-centric WiFi traces, in: Proceedings of the SouthEast Conference, ACM, 2017.
[42] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, K. Keogh, Experimental comparison of representation methods and distance measures for time series data, Data Min. Knowl. Discov. (2010) 1–35.
[43] M. Wenyuan, et al., Detecting pedestrians behavior in building based on Wi-Fi signals, 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), IEEE, 2015.