Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 3136471, 24 pages
https://doi.org/10.1155/2018/3136471

Research Article

Smart Behavioral Analytics over a Low-Cost IoT Wi-Fi Tracking Real Deployment

Javier Andión, José M. Navarro, Gregorio López, Manuel Álvarez-Campana, and Juan C. Dueñas

Departamento de Ingeniería de Sistemas Telemáticos, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040 Madrid, Spain

Correspondence should be addressed to Javier Andión; j.andion@upm.es

Received 3 August 2018; Revised 22 October 2018; Accepted 11 November 2018; Published 2 December 2018

Guest Editor: Jorge Lanza

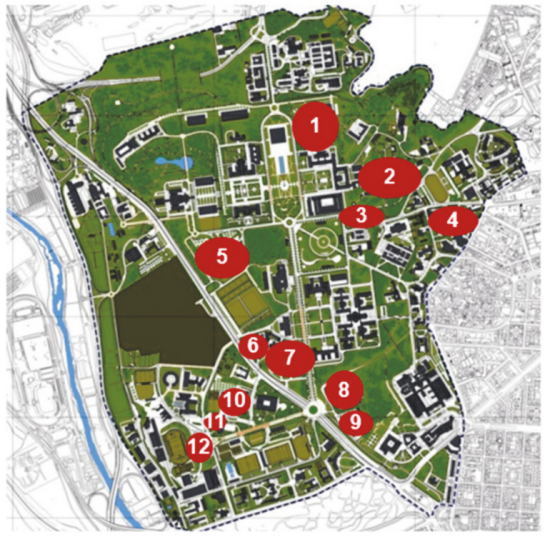
In a more and more urbanized World, the so-called Smart Cities need to be driven by the principles of efficiency and sustainability. Information and Communications Technologies and, in particular, the Internet of Things will play a key role on this, since they will allow monitoring and optimizing all the municipal services that exist and shall exist. People flow monitoring stands out in this context due to its wide range of applications, spanning from monitoring transport infrastructure to physical security applications. There are different techniques to perform people flow monitoring, presenting pros and cons, as in any other engineering problem. Typically, the options that provide the most accurate results are also the most expensive ones, whereas there are cases where presence detection in given areas is enough and cost is a limiting factor. The main goal of this paper is to prove that a minimal deployment of sensors, combined with the adequate analysis and visualization algorithms, can render useful results. In order to achieve this goal, a dataset is used with 1-year data from a real infrastructure composed of 9 Wi-Fi tracking sensors deployed in the Telecommunications Engineering School of Universidad Politécnica de Madrid, which is visited by 4000 people daily and covers 1.8 hectares. The data analysis includes time and occupancy, position of people, and identification of common behaviors, as well as a comparison of the accuracy of the considered solution with actual data and a video monitoring system available at the library of the school. The obtained insights can be used for optimizing the management and operation of the school, as well as for other similar infrastructures and, in general, for other kind of applications which require not very accurate people flow monitoring at low cost.

1. Introduction

The World is going tremendously urbanized. Based on the latest revision on the World urbanization prospects from the United Nations (UN) [1], nowadays 55% of the global population live in urban areas and such a percentage is expected to increase up to 68% by 2050. In addition, the number of so-called megacities (cities with more than 10 million inhabitants) around the World has gone from 10, in 1990, to 28, in 2014, and it is estimated that there will be 43 in 2030.

As a result, efficiency and sustainability become the key principles for the so-called Smart Cities, so that they can accommodate such an amount of inhabitants guaranteeing high levels of comfort. Information and Communication Technologies (ICT) and, in particular, the new paradigm of the Internet of Things (IoT) are key for Smart Cities in that they will allow monitoring city services, ranging from traffic management to waste collection, and running optimizations based on the huge amount of gathered data.

One of the main challenges when considering deploying new Smart City services is that there are many platforms, technologies, and protocols available and that they typically involve a huge number of devices, so the associated investment is remarkable. Therefore, it is especially interesting to have testbeds available for experimentation, but they need to be representative enough so that the conclusions obtained from them are significant. In this context, university campuses appear as ideal places for experimenting and evaluating innovative proposals that can later be scaled to the cities where they are typically integrated, thus arising the concept of living lab [2, 3], which is already a reality in Universities



Engineering schools	Number of sensors	
	Wi-Fi tracking	Environmental
1. Telecommunications	9	3
2. Forestry I	3	2
3. Forestry II	3	2
4. Rectorate	4	2
5. Civil	3	2
6. Agricultural I	3	2
7. Agricultural II	5	2
8. Aeronautical & Aerospace	5	2
9. Naval & Marine	3	2
10. Architecture I	4	2
11. Architecture II	2	-
12. Health Science & Sport	3	2
13. Industrial	5	2
Total	52	25

Figure 1: Summary of the sensors deployed in Smart CEI Moncloa (at July 2018). Figure 1 is reproduced from [8] (2017).

around the World, such as Delft University [4], University of British Columbia [5], or Harvard University [6].

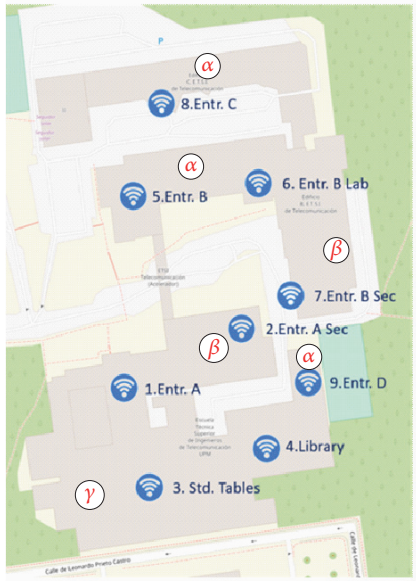
Universidad Politécnica de Madrid and, in particular, the Campus of International Excellence (CEI) of Moncloa presents such a great potential in this sense. This campus is integrated in the metropolitan area of Madrid, spreads across 5.5 Km^2, and counts on a daily flow that goes up to 120,000 people (which is comparable or even greater than many Spanish provincial capitals).

In order to make the most out of this potential, within the UPM City of the Future initiative [7] the IoT platform for Smart City services experimentation Smart CEI Moncloa was deployed [8]. This platform offers currently two pilot services, namely, environmental monitoring and people flow monitoring. The devices used for the environmental monitoring service are based on Arduino and collect measurements of temperature, humidity, luminosity, noise, CO, and NO_2. The devices used for the people flow monitoring service are based on Raspberry Pi and perform Wi-Fi tracking. As Figure 1 shows, for the time being there are 77 devices deployed across the 13 engineering schools of the CEI Moncloa, 52 for the people flow monitoring service, and 25 for the environmental monitoring service. The platform is up and running since 2016, so there is plenty of data available to be analyzed.

This paper focuses indeed on analyzing the people flow monitoring data gathered in the Telecommunications Engineering School (ETSIT) of UPM during 2016. People flow monitoring represents a hot topic nowadays because it presents such a wide range of applications in Smart Cities, spanning from monitoring public transport infrastructure (e.g., metro, airports), private transport infrastructure (e.g., highways), overcrowded scenarios (e.g., demonstrations, concerts), or customer behaviors (e.g., malls), to physical security applications (e.g., presence of unauthorized people in restricted areas). For these purposes, in many cases it is enough with providing presence detection in given areas, instead of more sophisticated and costly solutions to perform very accurate location, which require fingerprinting and very dense sensor deployments gathering data at very high frequencies. This is the case indeed of the people flow monitoring service considered in this paper, which is based on a few low-cost devices that upload data every 15 minutes and that are independent from the institutional network, which allows tracking the users connected to different Wi-Fi networks, if they spatially coexist, or even not connected to any.

Hence, as Figure 1 shows, in the ETSIT, which is one of the biggest schools in the CEI Moncloa, visited by 4000 people daily (3000 students, 500 professors and researchers, and 500 admin and maintenance staff, approximately), there are 9 Wi-Fi sensors covering 1.8 hectares of indoor areas. Figure 2 shows the location of these sensors. Dark areas correspond to the floor of the buildings of the ETSIT (buildings A, B, C, and D). There is a Wi-Fi sensor at the entry of each building. In addition, there are also sensors in the library (4) and student tables (3), as they are large spaces usually crowded by students. As it can be also seen, the area covered by each sensor varies, some of them covering especially large areas, such as the ones in the library (1300m^2) or in the main entrance (1270 m^2).

This paper aims to explore the useful insights that can be obtained from such a cost-effective solution for people flow monitoring. Thus, the paper performs a detailed analysis of the people flow monitoring data, including a temporal analysis, a spatial analysis, and an activity pattern analysis, as well as a comparison of the performance of this solution with a much more expensive one based on video monitoring at the library of the ETSIT. These analyses can be used for optimizing the management and operation of the school, from the work shifts to the proper operation of the lighting to reduce energy consumption and so the carbon footprint. The conclusions can be valid for similar infrastructures, but are also relevant in general for municipalities which will not



Sensor	Description
1. Entr. A	Main entrance of the building A of the school
2. Entr. A Sec	Secondary accesses to building A. It covers most of the classrooms of this building.
3. Std. Tables	Sensor close to "Entr. A", just above an area of tables where students work and study in groups.
4. Library	Inside the library, which is open every day from 9 a.m. to 9 p.m.
5. Entr. B	Main entrance to building B. This building contains mostly laboratories and offices.
6. Entr. B Lab	Secondary entrance of this building, which is the closest entrance to the laboratories.
7. Entr. B Sec	Secondary access to building B covering the classrooms of this building.
8. Entr. C	Main entrance of building C. This building contains only research laboratories and offices, students do not have classes in this building.
9. Entr. D	Entrance of building D, covering backdoor of the library.

Figure 2: Summary of the Wi-Fi sensors deployed in the Telecommunication Engineering School. The map on the left-hand side also shows the location of (α) offices and labs; (β) classrooms; (γ) cantina.

typically be able to afford accurate and costly location systems all over the city.

The rest of the paper is structured as follows. Section 2 presents an exhaustive analysis of techniques currently used for monitoring people flows and identifying behaviors. Section 3 outlines the main characteristics of the sensor infrastructure and the IoT platform that collects the data analyzed in this paper. After a brief sketch of the methodology used in this work, Section 4 presents the analysis performed on data with respect to time and occupancy, position of people, and identification of common behaviors or activity patterns. Then, Section 5 describes the data available as ground truth and the validation of the analyzed system. Finally, Section 6 summarizes the main findings of the analysis and presents some ideas for building similar systems.

2. Related Work

People flow monitoring has always been a useful piece of information. Knowing a person's or a mass of people's position or trajectory allows for the creation of a wide range of different applications, such as crowd monitoring in events or concerts, the discovery of the most common routes in shopping malls, analysis of space usage in public or private infrastructures or security against unusual activities (e.g., presence of unauthorized people in restricted areas). In the last decades, the spread of communication technologies (e.g., the popularization of Wi-Fi networks or the use of smartphones) has become a vast source of data which allows for the improvement or even the automatization of techniques to monitor people.

In most of the cases, positioning in outdoors scenarios can be solved using Global Positioning System (GPS), but it typically presents limitations in terms of accuracy in indoor scenarios [9]. Thus, indoor location or pedestrian location has been a key research topic in the last years. Some works aim to discover, with a high precision, how people move inside buildings by using the mobile network or personal area networks, e.g., [10]. These systems try to help users to discover their path in a building, measure the length of their stay in a mall for commercial purposes, or simply ease people movements by removing obstacles. The most common technologies used in recent years to achieve this kind of pedestrian tracking have been video camera systems, radiofrequency signals, Bluetooth, smartphones sensors, and Wi-Fi networks. These tracking methods can be classified based on two aspects:

(i) Whether they need user intervention, like the usage of a smartphone application or a specific hardware, which would be classified as active, or do not need the cooperation of the users to work, i.e., passive systems.

(ii) Whether or not a sensor network deployment is needed (e.g., by using the existing Wi-Fi access points network).

The usage of video camera systems and computer vision presents some advantages, such as the fact that it is a passive system, which can use existing camera network deployments or provides almost real time feedback. However, it also presents many drawbacks, although there are works that propose ways to mitigate these effects [11, 12], such as its dependency on visual aspects (e.g., poor lighting, obstacles), its low scalability due to deployment complexity and costs, the difficulty of fusing multiple video streams to provide automatic monitoring, or the difficulty of tracking users individually. Reference [13] provides a survey on computer vision techniques for the specific case of crowd scene analysis,

Table 1: Summary of the state of the art of people flow monitoring techniques which are not based on Wi-Fi tracking (P/A stands for Passive/Active).

Technology	Ref	Year	Scenario	Application	P/A	Own deployment
Video camera & Computer Vision	[13]	2015	Indoor Outdoor	Crowd scene analysis	P	No
RF	[14–16]	2013 2014 2015	Indoor	Short range movements (1-10 m) with high accuracy (e.g., elderly care, baby care)	P	Yes
Cellular networks	[18]	2011	Indoor Outdoor	Analyze people flow in a suburban area new NYC with accuracy around hundreds of meters	P	No
GPS	[19]	2015	Outdoor	Prevent critical situations in overcrowded scenarios (e.g., concerts)	A	No
Smartphone sensors (e.g., accelerometer)	[20]	2018	Indoor	PDR. Track individual pedestrian	A	Depends
Bluetooth	[21]	2017	Indoor	Monitor people flow (comparing Wi-Fi with Bluetooth)	A	Yes

covering from existing models and popular algorithms to current research problems and trends.

People localization and tracking based on radiofrequency (RF) measurements has been also widely addressed in literature. References [14–16] represent some recent remarkable research work on this topic. These solutions are based on antennas that transmit RF signals and are able to locate and track people based on body radio reflections. In consequence, they can be classified as passive systems. They provide very accurate results, allowing tracking forward and backward movements or body part movements (including breathing), and are able to even track several people under certain configurations. However, it is also difficult to track a fairly high number of users individually based on this kind of techniques (even if we assume that a person may have a certain type of body radio reflection signature, it would not be different enough between similar people and might change depending on the distance between the person and the vantage points). They are mainly applied in short range movements (1-10 meters), so they could only be applied in small rooms, at the cost of a large and specific deployment.

People flow monitoring can be also performed exploiting the ubiquity, communications capabilities, and integrated sensors of smartphones [17]. In [18], for instance, Call Data Records (CDR) are used to locate the base stations the smartphones are attached to and thus analyze people flow in and out of a suburban city near New York City. The main drawbacks of this way of locating and tracking people are its low accuracy (around hundreds of meters) and that the terminal has to be active (e.g., make or receive a call or send or receive an SMS) in order to be monitored, although this technique would be classified as passive given that the needed user activity is not aimed to contribute to the positioning.

Reference [19] proposes a solution to prevent critical situations in overcrowded scenarios based on a smartphone application that transmits its GPS location data. Although it yields good results in terms of accuracy, the main drawbacks of this approach are that it is an active technique, since it requires the installation of the app, the impact of the consumption of the app on the autonomy of the terminal, and it may also present problems to work properly indoors.

Smartphone sensors can be also used to perform Pedestrian Dead Reckoning (PDR), which is a method that tries to estimate a pedestrian's position based on their past position and the output of smartphone's sensors, e.g., accelerometer, gyroscope, etc. This technique is usually supported by other positioning methods [20], but in most cases it does not need a specific sensor network other than the smartphone itself. It is an active system, and it is typically used to track individual pedestrians instead of flocks, but the main drawback of this technique is that it suffers tracking error accumulation and needs an extra location measure frequently.

Bluetooth has also been used to monitor people flows in indoor environments. Some works such as [21] perform a comparison between the usage of Bluetooth and other techniques, such as Wi-Fi, and its conclusion is that its capabilities are below other options, although it can be used in combination with other positioning systems to improve their accuracy. The main disadvantage of this tracking systems is that most Bluetooth devices only react to Bluetooth signals when the users make them visible to the network. Also, these implementations require a specific deployment of Bluetooth sensors which usually cannot be used for other tasks. Table 1 summarizes the previous research work reviewed so far, without considering Wi-Fi tracking based works.

Lastly, people tracking systems based on Wi-Fi have been a hot topic from more than fifteen years and it is still so. This is indeed the technique used in this paper. Thus, Table 2 is exclusively devoted to summarizing and comparing previous research work based on this technique.

As it is shown in Table 2, works related to Wi-Fi tracking techniques can be focused on different objectives: some try to obtain users' positions as accurately as possible [22–32], others analyze the trajectories followed by pedestrians [33–35], or flocks [36–38], and, finally, others study the

Table 2: Summary of the state of the art of people flow monitoring techniques based on Wi-Fi tracking (P/A stands for Passive/Active).

Ref	Year	Scenario	Application	P/A	Own deployment
[22]	2003	Indoor	One of the earliest approaches on precise positioning using Wi-Fi (precision 2.6m)	A	Yes
[23]	2006	Indoor	Wi-Fi fingerprint to identify the general location and applying logistic regression to distinguish between finer-grained locations.	A	Yes
[24]	2006	Indoor, office building	Precise positioning. PDR combined with Wi-Fi to reduce the accumulated error	A	Yes
[25]	2007	Outdoor and indoor	Creation of Wi-Fi map. Positioning comparing with the created map	A	Yes. Own hardware. Offline analysis
[26]	2007	Indoor, campus	Comparison of positioning and tracking methods using Wi-Fi	P	No. Offline
[27]	2008	Indoor, campus	Estimate the position using Wi-Fi and tracking with PDR	A	Yes
[28]	2009	Indoor, campus	Real time Wi-Fi positioning, web portal to check user's positions	A	Yes
[39]	2009	Indoor, campus	Count of users in different buildings. Analysis of users' mobility between buildings	P	No. Institutional APs
[57]	2010	Indoor, campus and office building	Coarse position with Wi-Fi and Bluetooth. Graph of user co-occurrence.	A	No. Offline
[43]	2010	Indoor, campus and office building	Calculate of stay length based on Wi-Fi positioning. Analysis of favorite locations	A	No. Offline
[44]	2011	Indoor, campus	Extension to [39]. User characterization based on their mobility patterns	P	No. Institutional APs
[29]	2011	Indoor, tunnel in construction	Precise positioning in real time of workers inside a constructing tunnel using Wi-Fi (precision 5m)	P	Yes. Own AP deployment
[30]	2012	Indoor, campus	Creation of Wi-Fi fingerprint map. Map usage to positioning with smartphone application	A	No
[36]	2012	Indoor, campus	Study of crowd movement Wi-Fi based. Analysis of mobility patterns, users' arrivals and departures from campus	P	No. Institutional APs
[58]	2013	Indoor	Wi-Fi path analysis in real time.	A	No. Institutional APs
[33]	2014	Indoor and outdoor, campus	Analyze pedestrian destination frequencies in an area of 55 hectares of a university campus during 5 weekdays.	P	No. Institutional APs and Radius server
[59]	2014	Indoor, campus	Localization and tracking system exploiting particle filters to combine dead reckoning, Wi-Fi RSS-based analyzing and knowledge of floor plan together. (precision 0.7m)	A	
[60]	2015	Indoor, shopping mall	Wi-Fi Channel State Information analysis to detect shopper activities	P	Yes. Own AP deployment
[31]	2015	Indoor	Precise positioning based on sensor fusion combining Wi-Fi, PDR and landmarks. Smartphone application. (Positioning 1m)	A	No. Smartphones
[32]	2015	Indoor, parking	Precise positioning combining Wi-Fi RSS and electromagnetic field map		
[37]	2015	Outdoor, concert. Indoor, campus	Portable Wi-Fi based user count. Analysis of crowds in concert and in campus	P	Yes. Raspberry Pi based
[61]	2015	Outdoor	Creation of Wi-Fi map using GPS	A	
[62]	2016	Indoor	Precise positioning combining Wi-Fi and PDR	A	No. Smartphones
[34]	2016	Indoor, airport	User path detection. Combining Wi-Fi, GPS, PDR and Bluetooth to create a multilevel map and study of user's trajectory prediction	A	No. Smartphones
[45]	2016	Indoor, campus	Analysis of users' activities. User tagging based on activities registered	P	No. Institutional APs
[46]	2016	Indoor and outdoor, campus	Analysis of user movements to different food points to predict the operation of new stores based on price and location	P	No. Institutional APs

Table 2: Continued.

Ref	Year	Scenario	Application	P/A	Own deployment
[63]	2017	Indoor	Crowdsourcing positioning based on Wi-Fi fingerprint	A	No. Institutional APs
[41]	2018	Indoor	Coarse positioning, room level precision, based on probabilistic Wi-Fi fingerprint. Usage of Hidden Markov chain models to analyze user movement.	P	No. Institutional APs
[35]	2016	Indoor	Trajectory analysis based on Hidden Markov chain models	P	No. Institutional APs
[42]	2017	Indoor	Estimate the number of participants and their space and time evolution in an area of about 167 hectares during 2016 Open Day of the European JRC	P	No. Institutional APs
[47]	2016	Indoor Outdoor	Study mobility-related activities in a campus of 440 hectares based on the 2700 APs of the institutional network and additional opt-in smartphone application	A	No. Institutional APs
[53]	2014	Indoor	Classify users in a hospital (e.g., patient, doctor, administrative) by checking the number of hours and the positions of a user over time based on the institutional Wi-Fi network	P	No. Institutional APs
[38]	2012	Indoor	Identify flocks walking in a building and their behavior based on signal strength from the institutional Wi-Fi and using clustering techniques	P	No. Institutional APs
[45]	2016	Indoor	Analyze users' occupation (based on Markov models) as well as regular and irregular hours in a university campus	P	No. Institutional APs
[40]	2016	Indoor	Analyze room utilization and people tracking providing heat maps. Analyze device statistics	A	No. Institutional APs
[48]	2017	Indoor Outdoor	Analyze people mobility monitoring and tracking in Smart Cities and traffic in a highway (e.g., driving behavior, traffic forecasting)	P	Yes. Raspberry Pi based
[49]	2018	Indoor Outdoor	Provide user localization, user profiling, and device classification	A	Yes. Raspberry Pi based
[54]	2005	Indoor Outdoor	Analyze Wi-Fi tracking records gathered during more than one year in Madeira to classify users as tourists or locals and identify touristic spots	P	Yes. Based on TP-Link MR3240v2 home router
[56]	2017	Indoor	Obtain semantic trajectories. Classify users based on their locations. Analyze the probability of a user going to a specific shop based on their history and propose the creation of a recommender based on the whole dataset	-	-

occupation of different zones [39–42] and obtain behavior patterns [36, 43–49].

Wi-Fi tracking of a specific user is usually done by analyzing the collected records related to a specific MAC address, i.e., tracking users is equivalent to tracking their MAC address. This technique, in general, presents issues related to tracking people individually and privacy. Although it is true that a user carrying several devices (e.g., smartphone, tablet, laptop) with the Wi-Fi interfaces on would be at first identified as several users, after a reasonable period of time this information can be correlated to fix the problem [50]. In the case of the work presented in this paper, the files associated with the measurements of different sensors of the same building are compared in order to avoid counting the same mobile/person several times (e.g., due to overlapping Wi-Fi cells). The resulting file allows for the analysis of people flow at a building level, identifying the total stay time or the frequency of the visited places. However, as it is explained below on the position analysis subsection, this problem appears in the collected data and it is necessary to perform a postprocessing of the data to deal with these collisions. Users may also use so-called MAC spoofing (i.e., replacing their actual MAC address by the MAC address of another device), what could be seen as a kind of attack. However, this may yield connectivity problems and it is a negligible behavior in the scenario considered in this paper.

Regarding privacy, several proposals to protect Wi-Fi communications by means of MAC address anonymization have arisen in recent years. First, these proposals appeared as apps for smartphones (allowing performing kind of MAC spoofing), but, recently, smartphone manufacturers have started including these techniques in the latest versions of their Operating Systems (OS) (e.g., iOS, Android, and Windows). Such MAC address anonymization techniques aim to avoid using the actual MAC address until the device gets connected to the Wi-Fi network (i.e., they use a fake MAC address in their probe frames). The specific solution for this problem depends on the manufacturer and OS. In the case of iOS, the solution involves sending locally administered MAC addresses in the probe frames, randomly selecting the

three less significant bytes of the MAC address. This can be easily detected just by inspecting the first byte of the MAC address. In the case of Android, some manufacturers have decided to use random MAC addresses in the probe frames from the MAC address ranges assigned by the IEEE to them. Nevertheless, even with these techniques in place it is possible to end up obtaining the actual device information [51, 52]. In addition, MAC randomization is not actually a relevant problem in the scenario considered in this paper since most of the devices are connected to the available Wi-Fi networks (e.g., Eduroam) and, to connect to a Wi-Fi network, devices must use their actual MAC addresses.

Wi-Fi tracking systems can be classified into two main groups: those that use the enterprise Wireless Local Area Network (WLAN) itself and those that use a dedicated low-cost passive Wi-Fi infrastructure, which is indeed the case of the actual deployment considered in this paper. One of the main drawbacks of the systems that use the enterprise WLAN is that they can only track the users of such networks; whereas independent dedicated low-cost passive Wi-Fi infrastructures allow tracking the users connected to different Wi-Fi networks, if they spatially coexist, or even not connected to any, if MAC randomization is not used.

As examples of works that use the enterprise WLAN or the existing infrastructure of access points (APs), [42] presents the 2016 Open Day of the European Joint Research Center (JRC), where 8000 people participated within an area of about 167 hectares, as a case study where the Wi-Fi infrastructure of the event was used to estimate the number of participants and their space and time evolution based on properly processed MAC addresses. Reference [47] presents MobiCamp, a large-scale testbed, composed of around 2700 APs, to study mobility-related activities, which combines user mobility traces based on Simple Network Management Protocol (SNMP) data with enriched data (e.g., gender, age) provided by an opt-in smartphone application.

Reference [53] represents yet another example of the analysis that can be made with this kind of information. Its scenario is a hospital, and by checking the number of hours and the positions of a user over time they can classify that user according to a role, e.g., patient, doctor, administrative, etc. Reference [38] identifies flocks walking in a building and their behaviors applying clustering techniques to the signal strength measurements provided by the institutional WLAN.

Reference [33] presents a campus scenario where, by using the university network infrastructure, a detailed profile of the user's activity can be obtained. Users tracked are those logged into the university network, which provides extra information about the user, such as their role, gender, etc. Combining that information with a detailed map which contains thousands of Point of Interest (POIs), the authors can extract an activity log that shows the different user's activities with a minute precision. The main drawback of this work is that it totally depends on the users' profiles database and the POI's map and both are resources complicated to gain access to or create. In reference [45], employing the university network infrastructure and the location of each AP in the university campus, each sensor record only stores the closest AP. With this simple information the authors can create an activity profile similar to the one showed in [33]. By analyzing the basic results obtained, they are able to extract new information (e.g., a count of irregular hours) or detect patterns of anomalous events (e.g., periods of exams or holidays).

As last example of systems that use the institutional WLAN, reference [40] presents a web application in which the occupation of different rooms on a campus is shown in real time. The number of people in the room is calculated using the number of Wi-Fi devices detected by the APs. Using the signal strength measurement of each of them, a heatmap is drawn that shows the user distribution in the room. The collected data are analyzed offline to make reports of utilization of the different rooms and to obtain conclusions from the detected patterns.

On the other hand, [37, 48, 49, 54] represent some examples of works which use independent dedicated low-cost passive Wi-Fi infrastructures, as it is the case of the deployment considered in this paper. In the case of [48], a network of devices called MOBYWIT, based on a Raspberry Pi and two wireless USB dongles, are used to track people and vehicle's movement, sniffing not only Wi-Fi but also Bluetooth signals emitted by smartphones and vehicle hand-free calling systems. In the case of [49], a passive Wi-Fi infrastructure, based on low-cost devices that combine a Raspberry Pi and a TP-LINK Wi-Fi dongle, is used to provide user localization, user profiling and device classification based on the properly processed MAC addresses captured from the IEEE 802.11 probe request frames. Reference [37] also uses this approach to count people in a concert and, in reference [54], the considered scenario is a whole island (Madeira, Portugal), where the records gathered all over there are analyzed to classify users as tourists or locals, as well as to identify touristic spots.

One of the main features that make the work presented in this paper to stand out compared to previous work is that one-year data from an actual Wi-Fi tracking system deployed in a real-life environment is analyzed. Most of previous works consider hours or a few days (e.g., weekdays) or weeks. Only the work presented in [54] covers a similar period of time (being even larger), but the analysis is much broader, being far away from the level of detail provided in this paper. The considered period of time allows analyzing seasonality effects and other patterns that, although may be seen as common knowledge, do bring value since they represent numerical evidences that support decision making (e.g., someone can think that the Wi-Fi access in a given area does not work properly because it is always overcrowded, but numerical evidences are needed to appropriately justify the investment of increasing the number of AP of the corporate WLAN in that given area to improve the service). In addition, such well-known patterns, when obtained automatically by processing the available data, become models which can be used to detect anomalies or atypical situations, as it is common practice in nonsupervised machine learning. It is also worth to mention the use of clustering to improve the data analysis and interpretation (as in previous works, e.g., [38]), as well as the application of the semantic trajectory concept [55], which combines positioning data with an external source of

(a) (b) (c)

Figure 3: (a) Raspberry Pi; (b) TP-LINK USB Wi-Fi dongle; (c) developed Wi-Fi listening device.

information to classify the different positions according to the activity carried out in the area (e.g., users positioned in the cantina will be eating), and which has not been extensively explored in previous works (e.g., [56]).

Another strong point of the work presented in this paper compared to previous work is the validation of the Wi-Fi tracking technique to estimate the occupation of the library of the Telecommunications School and the comparison with a video camera system. Only a few previous works, such as [40] or [49] perform a similar validation (in [40] the number of people detected in the room is compared with the attendance list of the seminar taking place there and in [49] students are requested to turn on the Wi-Fi interfaces of their devices and provide the MAC addresses of their smartphones and laptops through an anonymous web form to serve as ground truth for device classification). However, again the period considered in this paper is much larger and it is proved that the Wi-Fi tracking system outperforms the more expensive video camera system. As a matter of fact, the Wi-Fi tracking system is actually used in a day-to-day basis by the library staff, which illustrates the value that this kind of IoT infrastructure can bring to real-life problems and services.

3. Data Acquisition Infrastructure

Figure 3 shows the Wi-Fi listening device/sensor developed for the people flow monitoring sensor network of the Smart CEI Moncloa. It is based on very common and cheap hardware, namely, a Raspberry Pi board [64], one of the most widely used hardware for IoT, and a TP-LINK USB Wi-Fi dongle [65] configured in monitor mode. As a result, the cost of this solution is in the order of tens of euros (around 80€ when manufactured on 2016), which represents a remarkable cost reduction compared to other solutions available in the market.

As Figure 4 illustrates, broadly speaking, these Wi-Fi sensors scan each of the Wi-Fi channels from both the 2,4 GHz and the 5 GHz bands during a configurable amount of time (currently, 250 ms), read the header of the radio IEEE 802.11 packets (e.g., data packets or probe requests) in its region of coverage, and record the sender MAC addresses. As these MAC addresses are unique per device, counting them is a good indicator of the number of devices available in the surroundings of the Wi-Fi sensors (although there may be issues related to the fact that a single user can bring several devices, as already pointed out in Section 2), and they allow for temporal correlation analysis, thus obtaining useful information such as stay time, availability patterns, etc.

Regarding how this information is carried all the way up to the backend server and processed there, Figure 5 shows an overview of the communications architecture and protocol stack of the people flow monitoring service of the Smart CEI Moncloa.

As it can be seen, the Wi-Fi sensors are directly connected to the backend via the Ethernet network of the UPM. Communications are protected end-to-end by the use of Transport Layer Security (TLS) on top of Transport Control Protocol (TCP)/Internet Protocol (IP). Measurements are periodically sent using Message Queue Telemetry Transport (MQTT) [66]. The publish/subscribe mechanism provided by MQTT allows the Wi-Fi sensors not only to send measurements (i.e., events) periodically, but also to receive commands (e.g., to reboot them or to perform a remote firmware update).

Taking advantage of the hierarchical structure of the MQTT topics, all the publish events follow the structure SERVICE/ID/EVENT(/TIMESTAMP). Thus, the publish events from the Wi-Fi sensors start by Wi-Fi, followed by the MAC address of its Ethernet interface, which is used as unique ID. The format of the content published under the different topics is Comma Separated Value (CSV), which is a lightweight solution especially appropriate when the data structure is fixed, since the meaning of each field of the subsequent lines is explained only in the header at the beginning of the file.

Privacy issues have been also considered carefully: the developed Wi-Fi sensors apply an irreversible hash MD-5 function with salt to the MAC address, which avoids brute-force attacks with precomputed tables. In addition, as Figure 5 shows, once anonymized, the data are carried securely up to the platform servers where they are handled in an aggregate manner, instead of individually.

Furthermore, the software of the developed sensors has been modified in order to avoid that the MAC anonymization mechanisms presented in Section 2 affect the obtained measurements. Thus, the Wi-Fi frames with locally administered

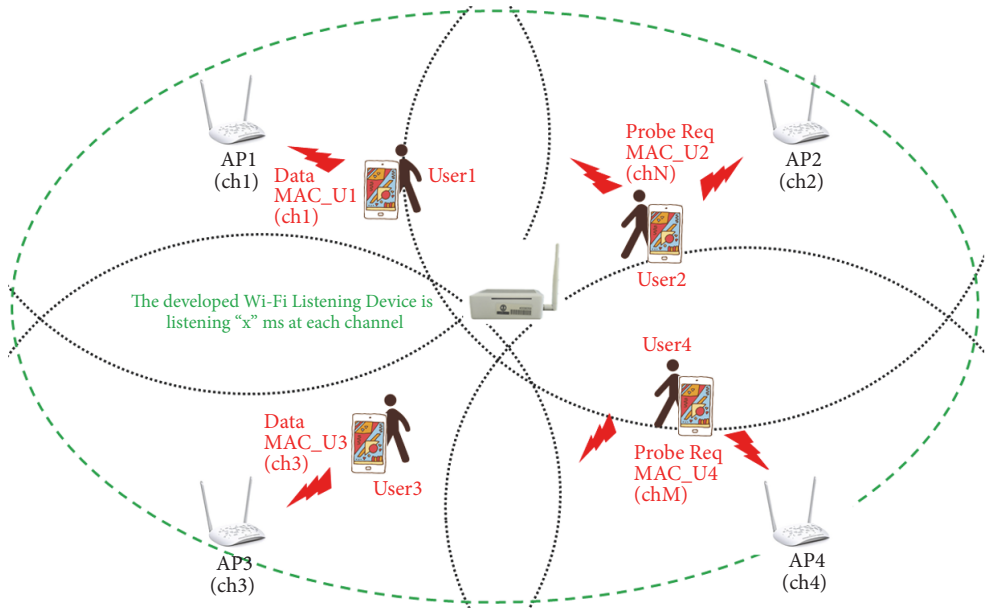


Figure 4: Sketch of how the developed Wi-Fi listening devices obtain the MAC addresses of the users surrounding them.

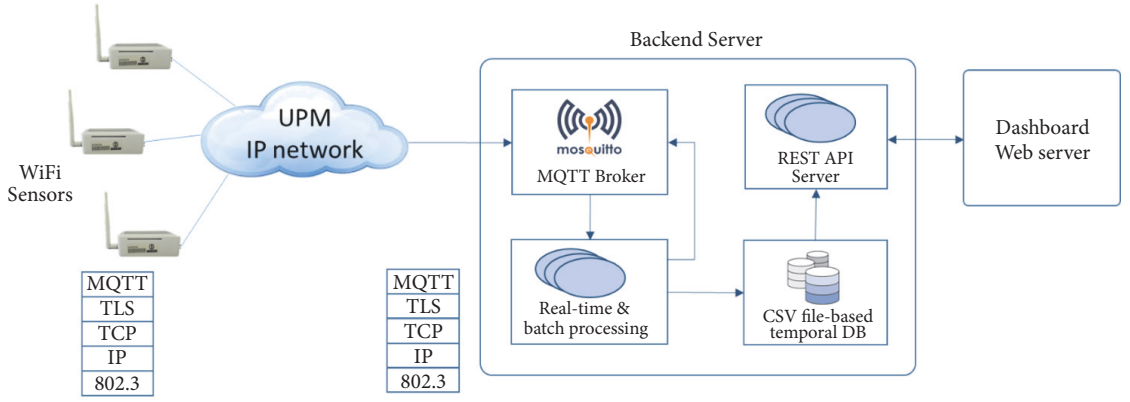


Figure 5: Communications architecture and protocol stack for the people flow monitoring service of the Smart CEI Moncloa.

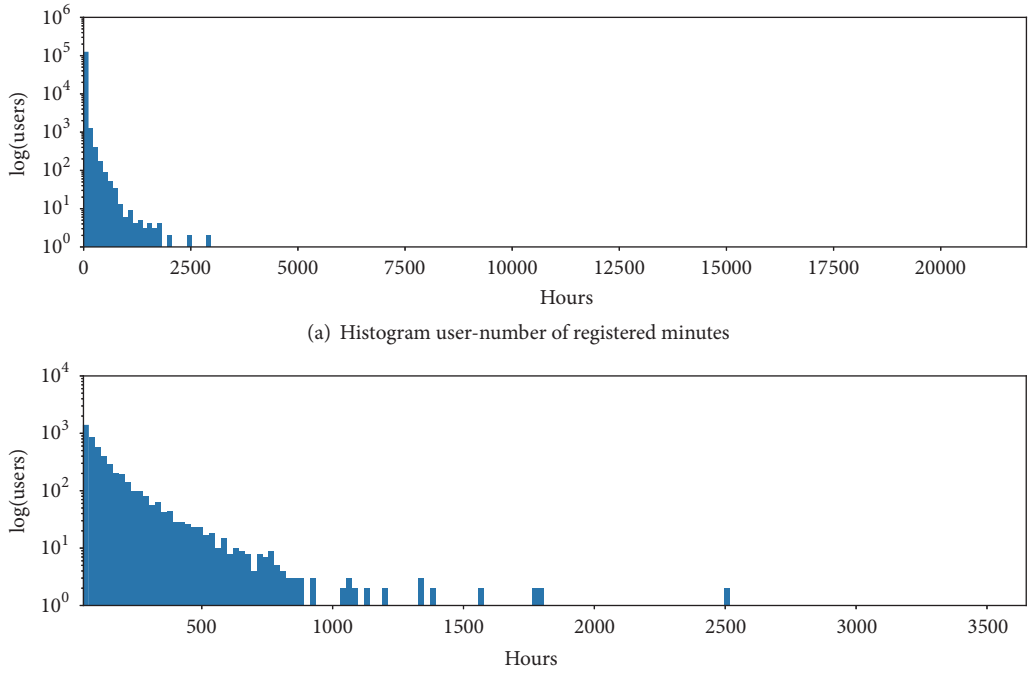
MAC addresses or including special MAC address ranges are discarded, so these devices are not considered. Anyway, as it has been already mentioned in Section 2, MAC randomization is not actually such a big deal in our case, since most of the users are connected to the Eduroam free Wi-Fi access, so their smartphones end up using their actual MAC address.

After gathering the CSV files for a large time span, we moved to an offline analysis platform on a private cluster. The Apache Spark 2.2 software platform was selected as the data processing tool due to its optimized capabilities to work with large amounts of structured data. This drives into large datasets which common centralized system will struggle to process, but distributed systems, such as Spark, can handle easily.

The private cluster mentioned is composed by eight HP ProLiant SL250s Gen 8 machines with two Intel Xeon e52630v2 2.6GHz (6 cores each) and RAM 32 GB. In addition to Spark, this platform also runs an Apache Hadoop Filesystem, where the dataset and the results are stored. This is a distributed filesystem which allows that all the machines access to the stored data in parallel. The usage of such platform is advisable in order to speed up analysis, but not mandatory, since the algorithms we are going to describe are available in many other software platforms (such as those provided with R or Python).

4. Data Analysis and Results

4.1. Data Processing. Although this work is not a proper data mining process, given that we are not using those kinds of algorithms and analyses, the necessary steps previous to the actual analysis are the same that in a KDD—Knowledge Discovery in Databases—process [67]. For our analysis we took the aforementioned CSV format files, each one containing the data collected by a single sensor during a period of 15 minutes. The observation period used in our analysis is a full



(a) Histogram user-number of registered minutes

(b) Histogram of hours per user during whole year-filtered

Figure 6: Histogram user-number of registered hours: (a) raw analysis; (b) first threshold applied.

year, from January 2016 to December 2016. In a one-year time lapse, 9 sensors, each generating a file every fifteen minutes, would create around 315K files, a theoretically maximum because a file is only created if the sensor is activated and detects at least one device during the period. The sensors were deployed at the beginning of 2016, but until March the deployment was not totally finished. There were also some holidays during 2016 when ETSIT was fully closed. During those days some of the sensors (although there are always security staff somewhere in the covered area) detected no devices, so they did not generate a file. As a result, the number of files, which we actually have for the analysis, is 246K.

The first step in preprocessing these data was to merge all the files into a single one, resulting in a 2 GB CSV plain text file, where each row represents the observation of a device during the associated time period including, among others, the anonymized MAC address, the sensor name, and the detection of the device for a given minute in the observation timespan. With this transformation the resulting dataset contains 63 million rows, each one representing the detection of a single device in a single minute by a single sensor.

A total of 128K unique devices were detected during the whole year. It is known that the number of people that regularly go to ETSIT is around 4K, so not all of the devices detected can be considered actual users of interest, consequently the data needed to be filtered. Only those of people that perform activities related to the place, such as students, professors, etc. should be taken into account. Thus, it is necessary to filter out devices keeping only the ones that can be labelled as users. We apply filters based on the observations of each device.

So, for each device we count the number of minutes it was detected in the entire dataset. A device can be seen by more than one sensor during the same minute, so, to generate this measure, we considered that the repeated minutes are counted only once per device. Grouping the resulting count, a histogram (Figure 6) is obtained, on which it is possible to make a classification of the devices based on the total time recorded during the whole year.

To facilitate the analysis of the chart, the horizontal axis has been expressed in hours and the count, in the vertical axis, is shown in logarithmic scale. Around 95% of the devices were seen for less than 48 hours during the whole year, in average less than four hours per month. This group is mainly composed by people passing near the school buildings, momentarily entering the coverage area of the sensors, without accessing ETSIT. Figure 6(b) shows the histogram applying a lower threshold of 48 hours and an upper threshold of 3650, an average of 10 hours per day. In this chart it can be observed some isolated peaks in the tail of the graphic, starting around 1000 hours in the horizontal axis. A detailed analysis of these peaks revealed that they were devices that remained connected continuously for several days, like servers. So, we applied the label "user" only to those devices that registered a number of hours during the year between these two thresholds. As shown in Table 3, from the 128K detected devices only 4653 were classified as users, over which we will perform the rest of the analysis.

The last step before the proper analyses is to merge this dataset with the information about the position and name of

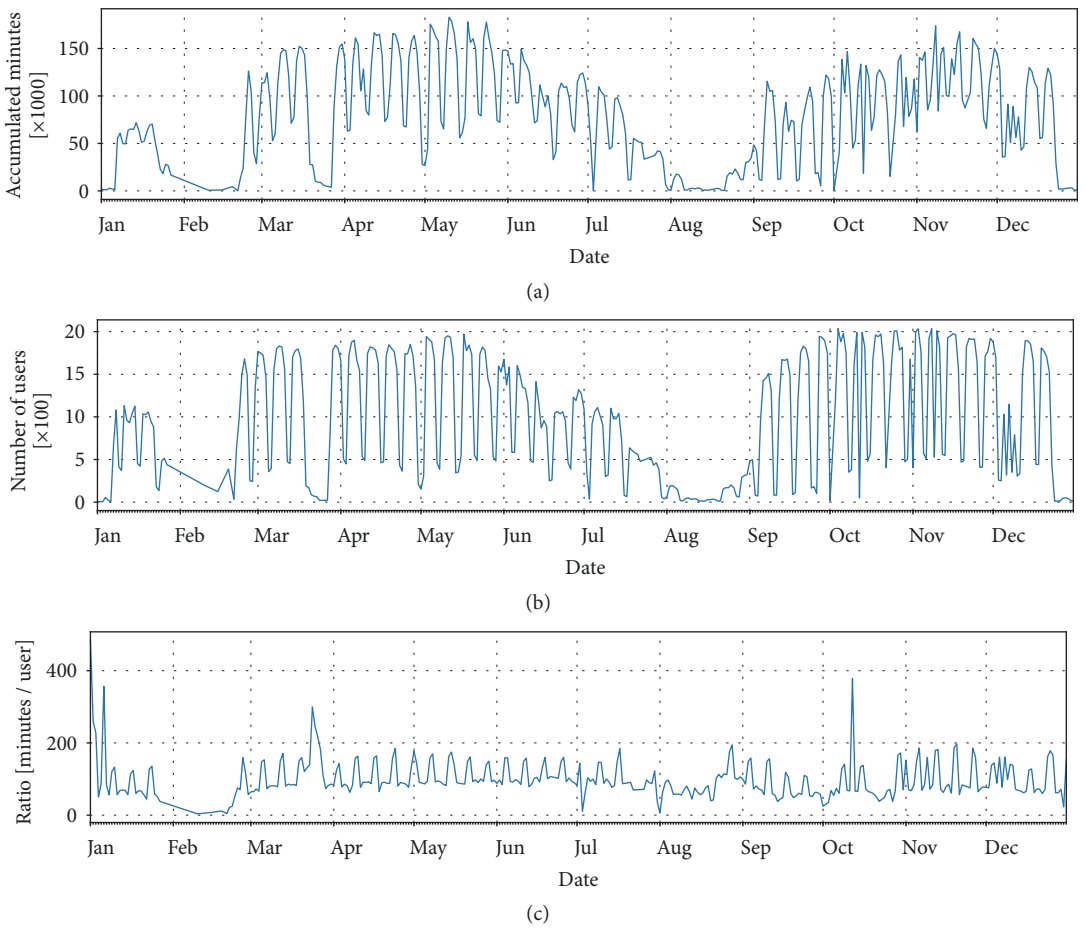


Figure 7: Daily analysis aggregated during 2016, (a) minutes accumulated; (b) unique users registered; (c) ratio between minutes and users.

Table 3: Dataset statistics after and before preprocessing.

	Rows	Devices
Before preprocessing	63427489	128188
After preprocessing	41294344 (65%)	4653 (3.6%)

the sensors. Additionally, only the data columns with relevant information for the analyses are kept in order to reduce the in-memory cost:

(1) timestamp: the detection time, measured in minutes,

(2) user: the detected MAC address device,

(3) sensor: the MAC address of the sensor which made the detection,

(4) sensorName: the name of the sensor which made the detection,

(5) accessPoint: the MAC address of the SSID to which the device is connected, if any,

(6) sensorLat: latitude of the sensor's position,

(7) sensorLng: length of the sensor's position.

4.2. Temporal Analysis. Once the data was ready, we performed a temporal study, aiming to know whether the captured data allowed for the identification of significant periods of activity or trends in user behavior in ETSIT. As one year is too long for a minute-level analysis, we aggregated our data by days and by hours.

Figures 7–10 are classified into groups of two charts. The first one, accumulated time chart, will show the number of activity minutes registered by each sensor during a time slice. The second chart presents the number of unique users seen by each sensor during the time slice. In addition, Figures 7-8 include a third chart with the ratio of accumulated time over unique users; this provides a hint on how stationary users are. This idea can be observed more precisely in the ratio chart: peaks represent moments when users are still (e.g., students in class) and dips are associated with transition times (e.g., students arriving at the school).

First, Figure 7 shows an overview of these metrics throughout the year. In this figure the measurements of the 9 sensors are aggregated into a single line. In the case of the Figure 7(a), the result is not exactly equal to the sum of activity minutes each sensor accumulates, since a user can be detected in the same minute by different sensors, and these occasions are represented as single instant in this line.

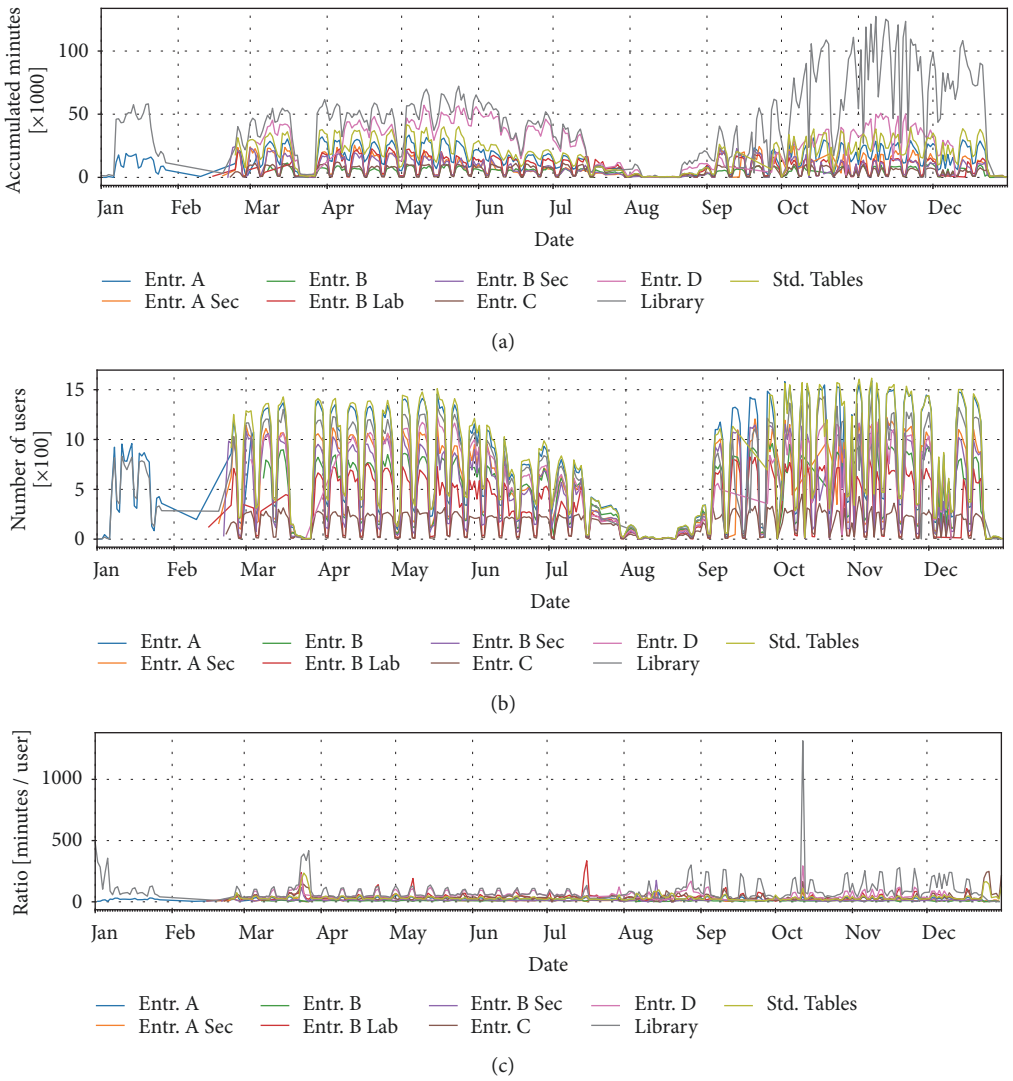


Figure 8: Daily analysis during 2016, (a) minutes accumulated by each sensor; (b) unique users registered by each sensor; (c) ratio between minutes and users by each sensor.

The most obvious observation that can be drawn from Figure 7 is the presence of a strange behavior during the months of January and February. The reason is that, as we previously discussed, the sensor network began to be deployed in January 2016 and was not completely operative until March. We keep this data in the analysis because it can be used to compare an anomalous situation with a regular one, also, it is much more intuitive to present a whole year range instead of nine months. Nevertheless, data collected in those months are not used to raise conclusions.

Holiday periods are clearly shown in the unique user's chart when the line falls, for example in March (Easter), summer holidays or some isolated holidays. These periods can also be seen in the ratio chart, since it increases because in those dates many fewer users attend to the school but usually spend long periods of time in the library. A remarkable point is October 12nd, in which the ratio chart reaches the highest peak and the number of users is almost zero. That day even the library was closed, so only the security staff was in the school. We can also observe the effect of weekends on each chart, both the number of users and the accumulated minutes decrease, but the ratio increases for the same reason exposed for the holidays: users will study at the library during the weekends. Finally, between the months of May to June, a decreasing trend can be observed both in the unique users and in accumulated minutes, going up in mid-June for the examination period.

Figure 8 shows a second set of graphs that correspond to a daily analysis of the whole year representing each sensor. One of the first conclusions that can be drawn is the difference between the proportions in unique user's chart and accumulated minutes chart. The difference between the number of registered users per sensor is not as remarkable as the difference between the number of accumulated minutes. Again, this is due to the fact that users spend much longer periods of time in the library than in other areas. Observing

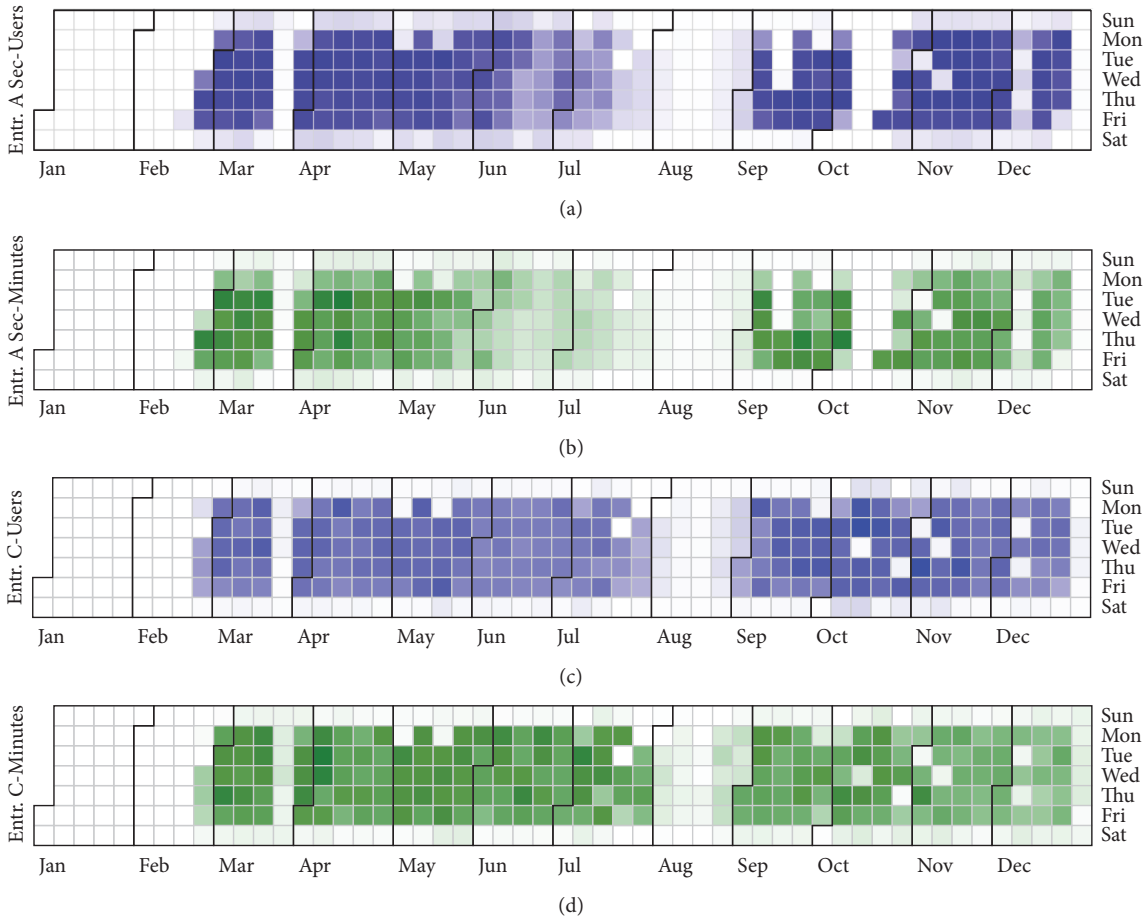


Figure 9: Calendar view, (a) unique users registered by sensor Entr. A Sec; (b) minutes accumulated by sensor Entr. A Sec; (c) unique users registered by sensor Entr. C; (d) minutes accumulated by sensor Entr. C.

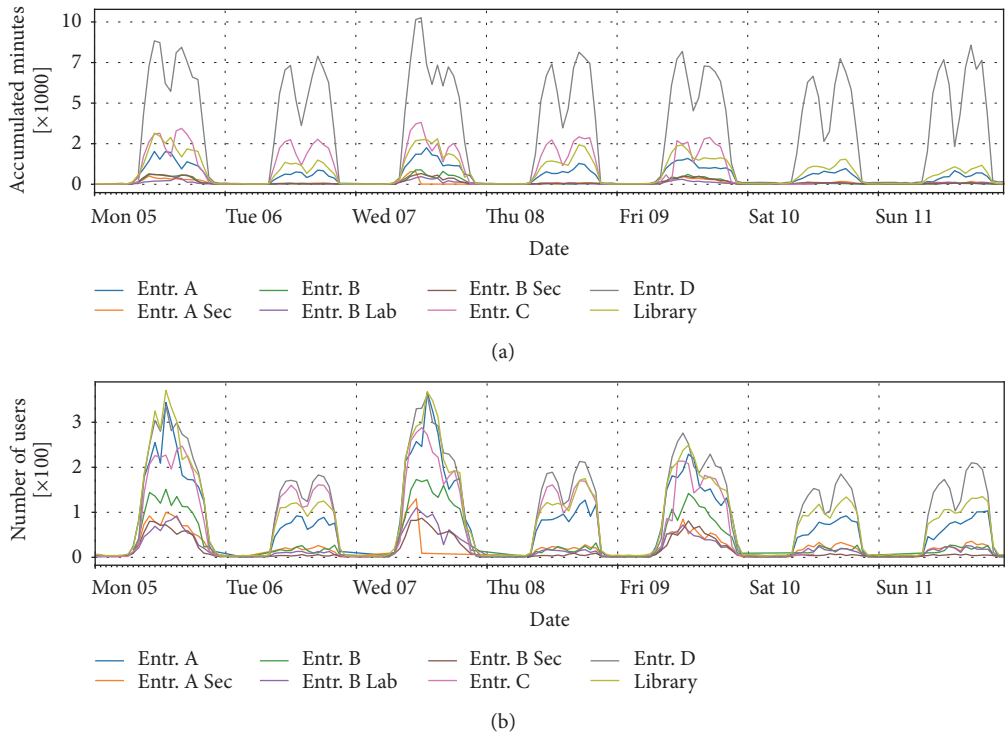


Figure 10: Hourly analysis during first week of November, (a) minutes accumulated by each sensor; (b) unique users registered by each sensor.

Table 4: Sensor records collisions.

Collisions	Count	Total	Percent
9	3	27	0.00%
8	45	360	0.00%
7	421	2947	0.01%
6	5371	32226	0.08%
5	59222	296110	0.72%
4	560994	2243976	5.46%
3	2525718	7577154	18.42%
2	7268825	14537650	35.35%
1	16433862	16433862	39.96%

the line of the sensor of building C in the accumulated minutes chart, it can be seen that the magnitude is maintained throughout the year. This sensor covers a building where there are professors' offices and research laboratories, i.e., this sensor registers mostly users who work at school, and they maintain a more regular schedule throughout the year than students who go to classes or to the library.

The line charts can be difficult to interpret for periods of time as long as a whole year. To ease the visual analysis, a new calendar visualization is offered in which the same data can be studied in a more intuitive way. Figure 9 presents the full year view in calendar format, each row representing one day of the week and the months appear delimited in black. The color intensity of the cell is proportional to the data it represents. Figure 9 is formed by 4 subfigures, which corresponds to the count of users and number of minutes by the sensors "Entr. A Sec." and "Entr. C". These visualizations are relative and can be used to obtain general conclusions. For a detailed study, both visualizations with absolute measurements and a data examination are still necessary.

Figure 9(a) represents the number of users detected by the sensor "Entr. A Sec.", which covers the classrooms of this building, and Figure 9(b) the number of minutes accumulated by the same sensor. It is clear that the first two months of the year this sensor was not operational, on Saturdays and Sundays (last and first row) this sensor does not register almost any activity, and holiday periods are clearly identified, such as Easter in March or summer holidays in July and August. But this visualization reveals other information that is more difficult to observe in a line graph, such as that Mondays are the days that the users spend less time in this area. They also highlight some blank cells in the last months of the year revealing that the sensor did not work during those days.

Figures 9(c) and 9(d) represent the information of the number of users detected and accumulated minutes by the "Entr. C" sensor. As mentioned in the description of Figure 8, this sensor includes very stable measures, because it covers the research laboratories and the workers' schedules are not affected by the school calendar, as it is the case of the activity seen by other sensors like "Entr. A Sec".

Figure 10 presents a different view, plotted at the hour level during a week in November. This eases the identification of activity hours, which span from 8 in the morning to 22 in the evening. Another notorious effect is the valley in the middle of the day, corresponding to the lunch break, when users move to the cantina (see Figure 2) or go out from the buildings. Finally, it is observed that the users leave the school gradually during the afternoon.

4.3. Position Analysis. Before getting into the details and insights obtained from the analysis of the one-year gathered data from the spatial perspective, it is worth to mention that a set of tests was carried out in a controlled environment during the first stage of the deployment in order to check that the Wi-Fi tracking system worked properly. These tests included tracking a well-known group of MAC addresses throughout the Wi-Fi tracking sensors checking that they appeared in the appropriate ones. It was also checked that the system correctly located and tracked the security staff throughout their nigh security tours.

A spatial analysis provides insight on how the users are distributed throughout the buildings during different times of the year. To reach these conclusions, we have improved the method presented in [68]. The first improvement is related to the event when a user is detected by two or more sensors during the same minute. From now on, this event will be identified as a collision. The number of collisions is a significant one in the case of some sensors that are close to each other (e.g., library and building d) or sensors that cover transition areas (e.g., Entr. A or Std. Tables).

Table 4 shows a study of the number of collisions. The order of collision is the number of sensors that collide for the same minute, and the count, the number of rows in which a collision of that order occurs. Collisions of order 2 and 3 group more than 50% of the data. The solution to this situation was to eliminate these collisions by replacing, for this experiment, all the rows corresponding to a collision by a single row whose latitude and longitude data are the centroid of the positions of the sensors participating in the collision.

The second improvement consists in the incorporation of external information with the approximate position of the access points to which the users are connected. This information has been obtained from the API Mylnikov Geo [69], getting the position of all ESSID registered throughout the year. In the cases the user is connected and the approximate position data of the access point are available, this new position is used instead of the position of the sensor

(a) (b) (c)

Figure 11: Monthly centroids heatmaps. (a) June; (b) August; (c) September.

that generates the row. This allows measuring the user's position in more accuracy. In addition, it allows smoothing the effect of using only the positions of the sensors, which causes that the resulting possible positions are always inside the hull of the polygon formed by the positions of the sensors.

This experiment is based on the user's centroid concept. This centroid is the average position of a certain user along a certain period of time. Representation of these points over the map reveals hints of the users' behaviors over the studied time slice based on the amount of people in each area.

Figure 11 shows heatmaps of three different months. Figure 11(a) is June, a month belonging to the second semester of the scholar course and the month when the final terms take place. Figure 11(b) is August, during summer holidays when there are no classes or exams and most of the professors, researchers, and staff are on vacation. Figure 11(c) represents September, start of semester.

Figure 11(b) confirms one of the facts extracted in the temporal analysis: on holidays the number of users falls and most of the users present in the school keep studying at the library (the warmest point is located over the library and is much smaller than on other months).

Both Figures 11(a) and 11(c) show that building A is the area with the highest concentration of users. Even so there are differences between two figures. In Figure 11(a), the warmest area in the map is over the library. This is explained by the final exam period of June. The same observation can be repeated in other periods of time to discover when the students have exams. In Figure 11(c), the hot spots are on the classrooms. September is the beginning of the school year and there is the greatest attendance to classes. Thus, this observation is an indicator of the level of students' assistance to class.

The same type of analysis can be done with shorter time frames to obtain more detailed behaviors. Figure 12 shows the centroids heatmap grouped by hours on September 2nd.

A detailed analysis by hours like this allows observing the users' movement throughout the day. The first row of maps in Figures 12(a), 12(b), 12(c), and 12(d), shows the evolution of user's centroids at lunch time, between 12:00 and 16:00. In this transition it can be seen that there are hot spots on the classes and library at the beginning. On the next map those centroids move to the cantina (see Figure 2), and in the last map they return to their original positions. The second row of maps in Figure 12 presents the start and the end of the activity time in the school. In Figures 12(e) and 12(f), it can be observed how the first users in the day go directly to classes. By contrast, Figures 12(g) and 12(h), reveal that users tend to be at the library at the end of the day.

Finally, Table 5 presents the count of the different users detected by each sensor throughout the year. Recalling that the total number of users obtained in Section 4.1 is 4653, the data in the table reveals that over the year most users have ever been seen by each sensor at some time. The two exceptions to this fact are the sensor of building C and the sensor of the laboratories of building B. These sensors cover the professor' offices and research laboratories, so they are unusual for students to stay in those areas.

4.4. Behavior Analysis. The third set of experiments we performed deals with the behaviors that each user follows throughout a single day. To obtain them, we grouped the data using a user-day key. For each key a vector of 24 positions - one per hour- is created. In each position of this vector, we determine which one has been the sensor that has detected this user most of the time. This vector represents, therefore, the route that the user followed throughout that day, hour by

Table 5: Sensor annual statistics.

Sensor	Number of rows	Number of users
Library	14166793	4674
Entr. D	6818165	4621
Std. Tables	5625974	4670
Entr. A	4736237	4620
Entr. A Sec	2352019	4579
Entr. B Sec	2258618	4530
Entr. B Lab	2229560	3630
Entr. C	1469026	3093
Entr. B	1248462	4127

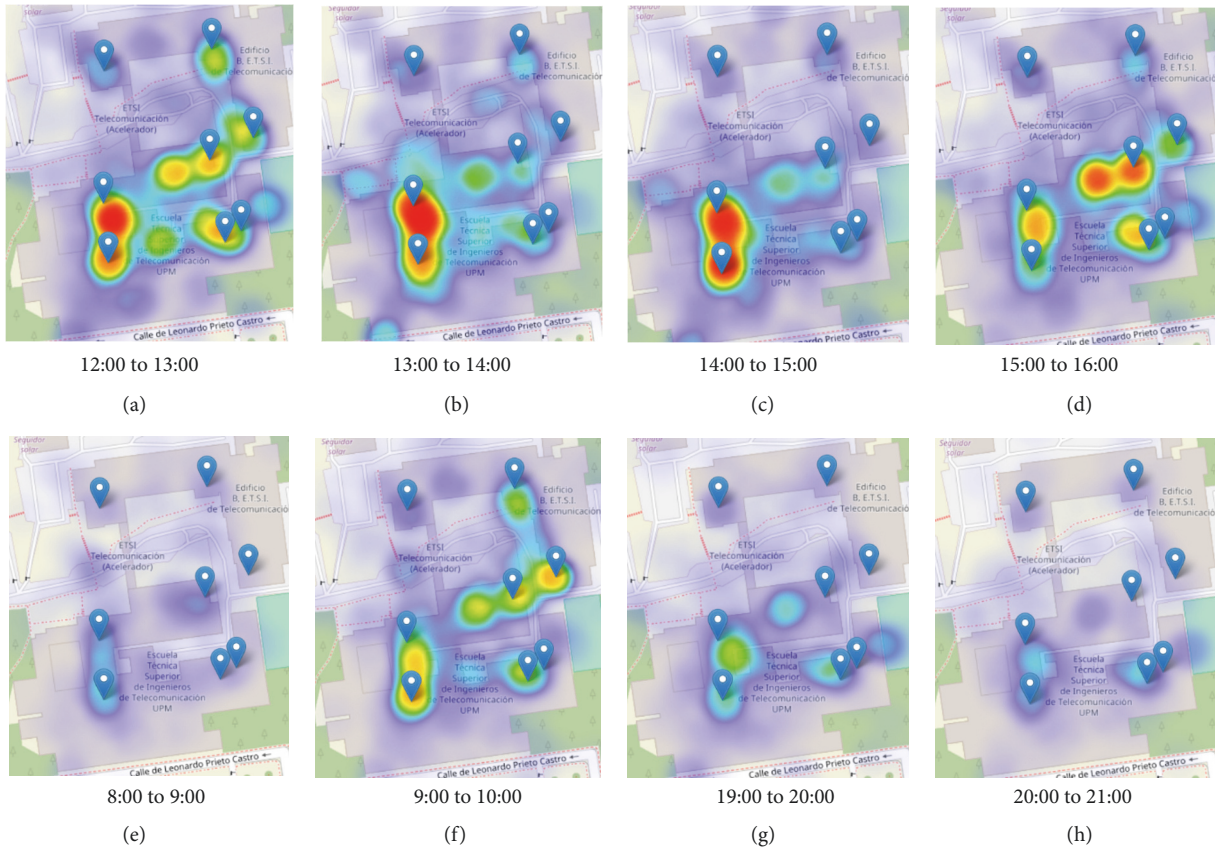


Figure 12: Hourly centroids heatmaps of 2nd September. (a) 12:00 to 13:00; (b) 13:00 to 14:00; (c) 14:00 to 15:00; (d) 15:00 to 16:00; (e) 8:00 to 9:00; (f) 9:00 to 10:00; (g) 19:00 to 20:00; (h) 20:00 to 21:00.

hour. Once the behavior vectors are obtained, the information of the day and the user is discarded to make a count of the most repeated behavior vectors. The dimensionality of theses vectors makes the number of possible behaviors huge, theoretically $2410 \approx 6 * 1013$ or $1410 \approx 289$K millions using only the activity hours, but it is known that users behave similarly, so a much lower number of behaviors can be expected, even though it will still be a high number.

Table 6 presents the 20 most repeated behaviors throughout the year. A total of 285K behaviors are detected, of which 139K are unique. The first 500 most frequent behaviors group 25% of the total behaviors. Table 6 shows only the part corresponding to the activity hours of the school, which, as observed in the temporal analysis, covers from 8:00 am to 10:00 pm. As it can be seen the majority of the most repeated behaviors are periods between two and five hours of stay in the library. It is necessary to expand the analysis to the top 20 to observe the class attendance behaviors. It is easy to appreciate that the different behaviors are usually morning or afternoon, with lunchtime from 1:00 p.m. to 3:00 p.m., which means that the majority of users go either in the morning or in the afternoon, but they do not spend all day at school.

Another quite obvious observation is that many of the behaviors obtained are very similar among them (e.g., going

Table 6: Top 20 most repeated behaviors.

#	8:00 - 9:00	9:00 - 10:00	10:00 - 11:00	11:00 - 12:00	12:00 - 13:00	13:00 - 14:00	14:00 - 15:00	15:00 - 16:00	16:00 - 17:00	17:00 - 18:00	18:00 - 19:00	19:00 - 20:00	20:00 - 21:00	21:00 - 22:00
1		Lib[1]	Lib	Lib	Lib	Lib								
2		Lib	Lib	Lib	Lib									
3				EntA[2]										
4		Lib	Lib	Lib										
5			Lib	Lib	Lib	Lib								
6		Lib												
7						EntA								
8									Lib	Lib	Lib	Lib	Lib	Lib
9	Std[3]	Std	Std	Std	Std	Std	Std	Std	Std	Std	Std	Std	Std	Std
10			EntA											
11					EntA									
12									Lib	Lib	Lib	Lib	Lib	
13				Lib	Lib									
14		Lib	Lib											
15								EntA						
16	EntA													EntA
17		EntA												
18			Lib	Lib										
19		SecA[4]	SecA	SecA	SecA									
20					EntA									

[1] Lib = Library.
[2] EntA = Entr. A.
[3] SecA = Entr. A Sec.
[4] Std = Std Tables.

to the library from 9:00 a.m. to 2:00 p.m. or going to the library from 9:00 a.m. to 1:00 p.m.). In order to lower this redundancy, we performed a clustering procedure on them. Each behavior can be understood as a categorical vector of length 24, in which the categories correspond to the 9 possible sensors of the data set plus the empty category. The chosen clustering algorithm is Proximus [70], due to its simplicity, efficiency, scalability and results' reliability. The algorithm works with binary vectors, and creates clusters based on the Hamming distance (the number of bits that differ between two binary vectors). A vector is chosen as the center of the cluster and other vectors, whose distance to the center is smaller than the maximum cluster radius, are added to that cluster.

The behaviors are expressed in categorical vectors, so it is necessary to transform them into binary vectors to be used in Proximus. The transformation shown in Equation (1) is proposed. This transformation is simple and also reversible, which allows for the recovery of the original behavior drivers after clustering.

$$\begin{aligned} B &= \{h_0, \ldots, h_{23}\} \\ h_i &\in \{\emptyset, a_1, \ldots, a_9\} \end{aligned} \longrightarrow \begin{aligned} \widehat{B} &= \left\{\delta_{0,a_1}, \ldots, \delta_{0,a_9}, \ldots, \delta_{23,a_1}, \ldots, \delta_{23,a_9}\right\} \\ \delta_{i,a_j} &= \begin{cases} 0 & if\ h_i \neq a_j \\ 1 & if\ h_i = a_j \end{cases} \end{aligned} \tag{1}$$

Where $\{a_1, \ldots, a_9\}$ are the nine sensors, B is the behavior vector described above and $\widehat{B}$ is the $24*9$ long transformed vector, made out of 9 samples subgroups, each one associated with each hour. Every sample on each subgroup is 0, except for the index of the active sensor on each hour, which is marked with a 1 on its variable.

Table 7 shows the results of applying the Proximus clustering to the found behaviors. The first column is the center of the cluster, the second one the number of behaviors that are grouped in that cluster, and the third, the number of behaviors that fall within that cluster. The results are presented ordered by the number of behaviors included in the cluster. 45707 clusters are obtained. The first 300 group 50% of the behaviors recorded throughout the year. The most important cluster behaviors in Table 7 can be understood this way:

Table 7: Behavior clusters.

#	Cluster center	Cluster components	Behaviors count
1	Entr. A Sec from 9:00 to 14:00	415	4963
2	Entr. B Sec from 10:00 to 13:00	458	3756
3	Entr. B Sec from 15:00 to 19:00	471	2924
4	Library from 9:00 to 13:00	256	2870
5	Library from 10:00 to 12:00	176	2396
6	Library from 18:00 to 21:00	244	2231
7	Entr. A 13:00	787	1979
8	Library 12:00	563	1894
9	Library from 18:00 to 19:00 and from 20:00 to 21:00	175	1853
10	Entr B. Lab 11:00 to 20:00	568	1831

(i) Clusters 1, 2, and 3: students attending to classes. These sensors cover the main classes in the school, and the intervals matches with the class schedule.

(ii) Clusters 4, 5, 6, and 9: students at the library.

(iii) Cluster 7: users which usually are outside of the sensors coverage areas, arriving or leaving.

(iv) Cluster 10: laboratory equipment and professors. This sensor covers some of the professors' offices and laboratories which usually have some laboratory equipment connected during work hours.

5. Validation of Occupancy Estimation Based on Wi-Fi Tracking

Although, as it has been already pointed out in Section 4.3, at the very first stage of the deployment it was tested that the Wi-Fi sensors properly locate and track well-known MAC addresses, it was still needed to validate the accuracy of the system for estimating occupancy (as it can be distorted by the aforementioned fact that a single user can carry several devices connected to Wi-Fi networks). As no ground-truth data was available for the full set of buildings and only some data was found for the library, we centered our validation efforts in comparing our data with the available ones: if we can trust our results in that area, then we can extend our trust to the rest of the areas for which no well-known data are available.

The library of ETSIT has 408 study sites and it offers a web service to check the number of available seats at a given moment of time [71]. This system is based on two sources of information: a person who counts the empty seats every opening hour from Friday to Sunday, and a video camera located at the main entrance of the library that counts the number of people entering or leaving at 15-minute intervals. The human system provides a ground truth about the number of occupied positions, but this measure is very different from the actual number of people in the library, since, a common situation, especially during examination terms, is that students place their study material at the seat to reserve it while they are not in the library. This situation is a problem for the library staff and therefore they installed the video camera system to count the student's entrances and exits.

This system generates an estimation of the number of people in the library, adding to the previous measure the number of people which are detected entering and subtracting the number of people which are detected leaving. The system is not perfect and, in most cases, it carries an accumulated error that increases in the estimation of the number of people in the library. The total error can be calculated clearly at the end of the day, when the library closes, and the number of people inside is supposed to be zero. In summer there is a situation that aggravates this error and consequently the measure achieved by this system: due to the rise in temperature, the back door of the library is opened to improve ventilation and allows students to exit through it, although they must continue entering through the main door. The camera does not count students leaving through this back entry.

The library staff provided us with the data collected by the two systems (human and camera) between June 5th and 30th. These measurements can be compared by those obtained by the Wi-Fi sensor installed at the library to validate them. For this test, all the data collected by the sensor will be used, without filtering the MACs of the sporadic users, as it has been explained before.

Figure 13 shows the data collected by the three systems on Sunday, June 5th. This is the first day with data from the three sources. Other days in which these three sources are present have the same trends. It is clear that there is a divergence between the human system observation and the rest of the data. The graph of the human system shows that the number of occupied seats increases in the first hours up to the maximum and remains steady until the end of the day, without being affected by the behavior of the users at lunchtime. However, this effect is reflected in the camera system and Wi-Fi tracking measurements. We observed that, in general, the number of people accounted for by the camera system is under the Wi-Fi-tracking system measure, although the proportion is maintained over time. Finally, the figure shows the cumulative error effect of the camera system, which at the end of the day still renders 65 people in the library.

To better study the relation between the camera system and the Wi-Fi-tracking system, we generated a detailed visualization that allows us to observe the data of the whole month in a single figure. Figure 14 is composed of 3 subfigures: each of them is a matrix of colored cells, the lines represent a full

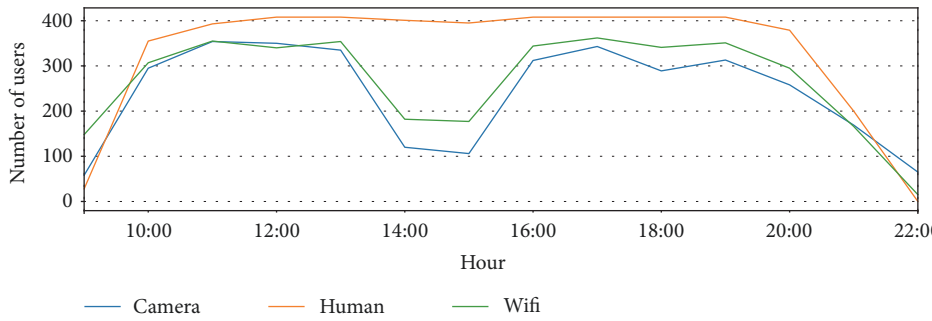


Figure 13: Measure of library occupation using Wi-Fi, camera, and human systems.

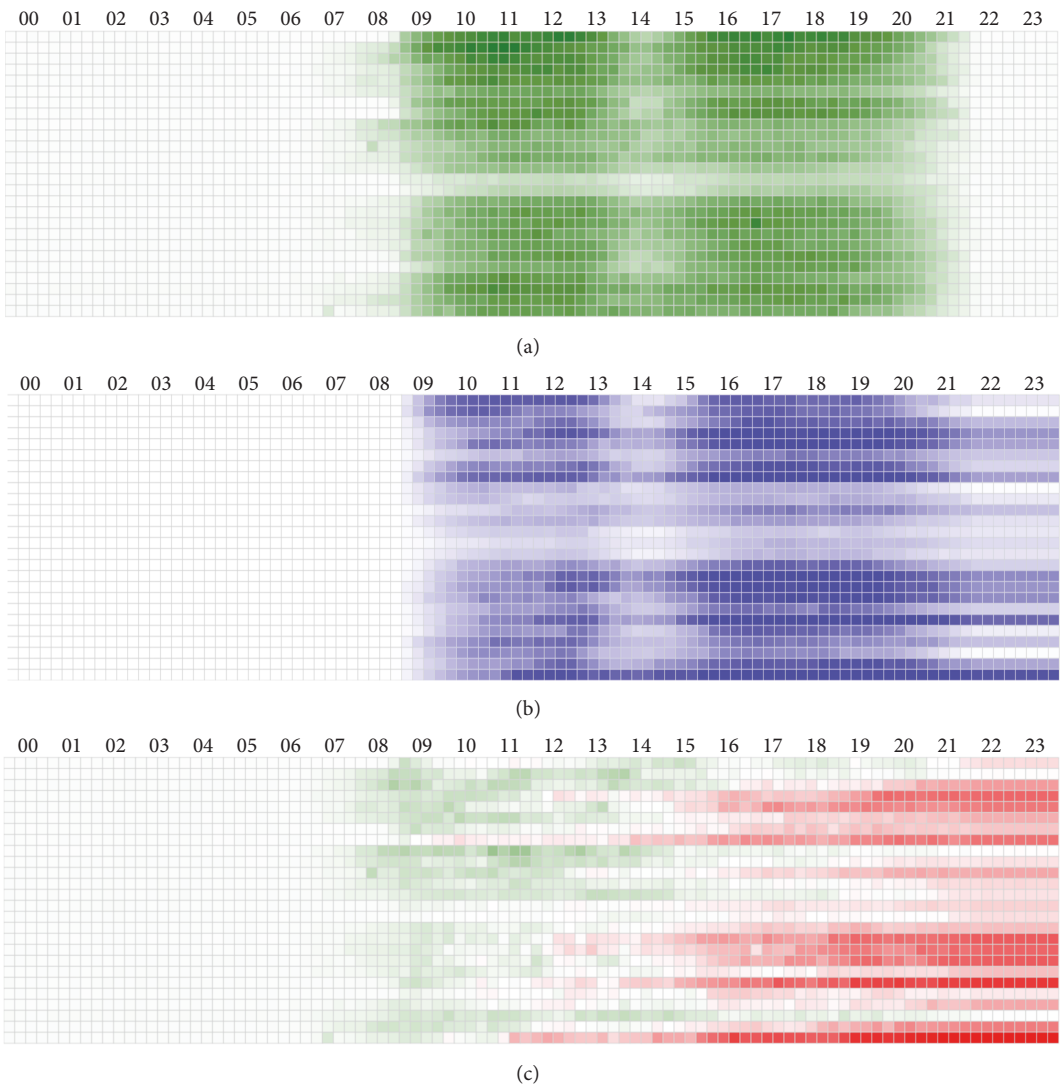


Figure 14: Measures each 15 minutes of library occupation during June: (a) Wi-Fi records; (b) camera records; (c) difference between Wi-Fi and camera.

day, and the columns are divisions of 15 minutes. The color intensity is proportional to the number of users measured in each interval, being more intense when more users are detected. Figure 14(a) shows the measurements of the Wi-Fi sensor. The behavior is the one observed in previous sections of this document. Figure 14(b) shows the measurement of the camera system. During the first half of the day, the same trend as in Figure 14(a) can be observed, but in the afternoon and night the accumulated error begins to be appreciated. This error reaches a maximum of 378 people, with an average of

139 people at the end of the day, while the Wi-Fi-tracking system registers a maximum of 10 and an average of 7 at the end of the day (these are machines that are still turned on when the library is empty of people). Finally, Figure 14(c) shows the difference between the two previous ones, using the green color when the error is positive (the Wi-Fi-tracking system registers a higher value than the camera system), and red color otherwise. This figure validates the measurements obtained by the Wi-Fi-tracking system, since the difference with the measure of the camera system in the first half of the day is small (an average of 27 people, counting the data from 8:00 to 16:00); whereas, in the afternoon the error accumulated by the camera system provokes that the average difference grows to 83 people between 16:00 to 22:00.

The Wi-Fi sensor of the library was not placed for the specific purpose of counting the number of people in the library and, in consequence, its coverage area can detect devices that are outside the library. This explains why in some cases it registers a measurement greater than that of the camera system. In any case, the Wi-Fi sensor, with its limitations, registers a measure very similar to the system of counting people using a camera, even avoiding the cumulative error of this system; being much cheaper (tens of euros against thousands of euros) and less intrusive than a video camera. Currently, since the camera system is already installed, the Wi-Fi sensor measures can be used as a maximum, which would limit the error made by this system. In summary, the usage of Wi-Fi sensor to count people in the library provides accurate results despite the potential errors induced by the sensing period, collisions, ratio between users and MAC seen, and data processing performed. These results invite to trust that the results of our analyses for the rest of the areas in campus are also correct.

6. Discussion and Conclusions

In the execution of this case study we have learned some lessons about the limitations we faced that are worth to be taken into account for similar systems and analyses. First, there is the key issue of the sensors deployment: it is not just a question of density (number or sensors per surface), but of adapting its design to the topology of the place to be controlled. So, for example, it is crucial not only to have sensors in entries and exits of buildings, but also inside them, as in the considered deployment, where there are sensors placed in the entrance and exists of the buildings, but also in other especially relevant places, such as the library or the work-in-group area.

A complete coverage of indoors devices, without collisions and for the complete timespan of the stay indoors, would require a large deployment with many sensors able to cover all corners. This is not suitable in this context; instead we counted with 9 sensors, some in gates, and others in large rooms, which proved to be enough to check if a device is in the reach of any of them. But some assumptions had to be taken: we considered as valid behaviors only those that spent a significant amount of time once they had been seen (to remove transient behaviors and those of people walking out of the buildings); also, we discarded behaviors that did not appear enough times in the yearly timespan. On the other hand, the topology of these buildings did not impose severe restrictions on sensors' coverage. In fact, we found several devices seen by different sensors at once. As a summary, we can conclude that the finer the spatial granularity (more sensors), the larger the set of different behaviors; so getting raw data from sensors would lead to an explosion of states that would render meaningless results.

The gathered data is another key issue: some works in literature got signal strength measurements every few seconds, allowing them to identify indoor trajectories. This was not our case, and in fact this revealed to be very limiting. Thus, the precision of our analysis is limited by the regions covered by each sensor, which hampers us from analyzing any kind of movement within regions. Nevertheless, as it has been seen, relevant results can still be obtained without the increase on energy consumption and the risk of flooding the school communications network that the other approach would entail.

As regards the analysis of data to identify users' behaviors, we have to indicate that a long observation period is a must. Obtaining data for a complete work cycle (in this case, a full school year) helps in discovering common behaviors that happen in a university. Using long observation times can help detecting erroneous or atypical operations on the sensors, as happens in the first months of the year in this case. Getting even larger observation datasets would reinforce the behaviors we have found (e.g., seasonality), but the chance to discover new ones is negligible, due to the expected behaviors in the campus will be periodic with the school year.

The usage of big data platforms for analysis, while not mandatory, eases the management of large datasets and the execution of iterative study on the data. The preprocessing work carried out allowed us to clean and filter our data. For example, some discovered behaviors are too regular and extended over time which may correspond to servers or machines which we could, then, filter out.

The temporal analysis has revealed some behaviors that are maintained throughout the year and others that occur occasionally. The work has focused on the study of behaviors that are repeated and maintained throughout the year. A closer view, such as the analysis per hour, shows the hours of activity, nocturnal patterns, or movements at mealtime. In the daily vision, the difference in activity between work days and weekends or holidays is clearly seen. Visualization has been a key technique in order to detect and understand these temporal patterns.

The spatial analysis revealed that a problem of collision happens in the user detection, but its effect was mitigated by the use of innovative algorithms and techniques, such as the calculation of centroids and the combination of sensors information with external sources of information (i.e., semantic trajectories), so a much more precise positioning of the users is achieved than with the exclusive use of the sensors. The visualization of centroids fostered the visual understanding of complex data such as the position of each user over a month, and the superposition of the centroids in a heatmap allowed knowing the movements of the groups of people and the occupation of the different zones.

Furthermore, we have found dominant users' behaviors as the most repeated behaviors registered by the sensor throughout the year. The number of found behaviors has been huge, but the application of the Proximus clustering algorithm reduced this number to a manageable amount. Then, the study of the obtained clusters has revealed that the most frequent behaviors coincide with what can be expected from a college building: researchers working in laboratories and students attending to classes or studying in the library.

Although some of the discovered behaviors and patterns can be seen as common knowledge, it is worth to stress that they do bring value since they represent numerical evidences that support decision making (e.g., someone can think that the Wi-Fi access in a given area does not work properly because it is always overcrowded, but numerical evidences are needed to appropriately justify the investment of increasing the number of AP of the corporate WLAN in that given area to improve the service). In addition, such well-known patterns, when obtained automatically by processing the available data, become baseline models which can be used to detect anomalies or atypical situations, as it is common practice in nonsupervised machine learning.

Lastly, we have validated the accuracy of using Wi-Fi tracking for occupancy estimation comparing it with the library staff manual counting (considered as ground truth) and with a video camera system installed at the library main entrance. As a main finding, Wi-Fi tracking has proved to be more accurate than the video camera system, in addition to being way cheaper. As a matter of fact, the library staff is currently using, preferably, the occupancy estimation based on Wi-Fi tracking rather than the one based on the video camera system. Nevertheless, the accuracy of the Wi-Fi tracking system can be further improved by correlating MAC addresses detected in same places over fair enough periods of time and considering only one, thus mitigating the issue related to the fact that a single person can bring several devices connected to Wi-Fi networks.

Beside this, the Wi-Fi tracking system is currently used by the library staff to perform more sophisticated studies, such as figuring out the percentage of students from the different schools of the university who come to study to the library of the Telecommunications Engineering School during the weekend. Figure 15 shows the results of such an analysis, which represents a token of how this kind of IoT system can help solving real-life problems and improving the operation of already running services.

To summarize, we have studied a one-year Wi-Fi tracking dataset obtained from a reduced set of low-cost sensors with limited capabilities deployed on an actual university campus that receives around 4000 people every day. We have processed the data in order to identify traces of mobile devices enabled with Wi-Fi, which are identified as people moving in the campus buildings, and then we have extracted people's stays, movements, and common behaviors. The obtained results represent numerical evidences that illustrate how a low-cost Wi-Fi tracking system can be used in real-life conditions to improve or optimize the operation of the monitored premises. These results can allow dimensioning appropriately the WLAN infrastructure or the canteen personnel or detect anomaly situations in real time. Furthermore, the data from the people flow monitoring system is currently being used together with the data from the environmental monitoring system to try to reduce the environmental footprint of the school [72]. In addition, the Wi-Fi tracking system is actually used by the library staff in their day-to-day activity, which illustrates the value that this kind of IoT infrastructure can bring to real-life problems and services.

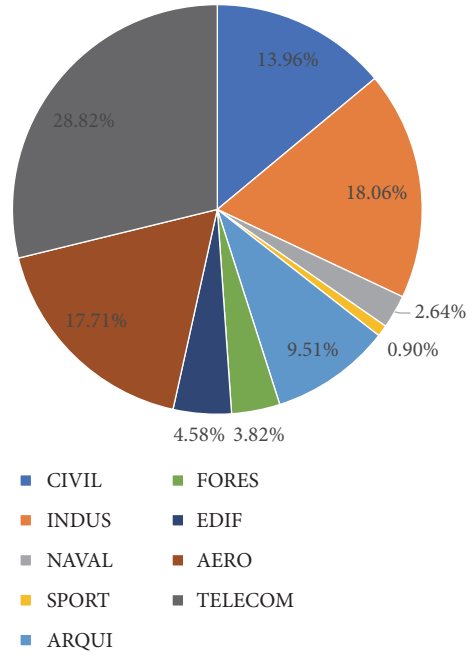


Figure 15: Users who only go to the library of the Telecommunication School on weekends classified by their school.

Data Availability

The dataset with the Wi-Fi sensors records used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been partly funded by Universidad Politécnica de Madrid through the project RES2+U (*Responsables, Sostenibles y Universitarios*) (http://blogs.upm.es/res2masu/). The work performed by José M. Navarro was funded by the Ministerio de Educación, Cultura y Deporte de España under Grant no. FPU 14/03209.

References

[1] "2018 Revision of World Urbanization Prospects — Multimedia Library - United Nations Department of Economic and Social

Affairs," https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html.

[2] M. Pallot, "Engaging Users into Research and Innovation: The Living Lab Approach as a User Centred Open Innovation Ecosystem," https://web.archive.org/web/20120509081658/http://www.cwe-projects.eu/pub/bscw.cgi/1760838?id=715404_1760838.

[3] E. Almirall and J. Wareham, "Living Labs: arbiters of mid- and ground-level innovation," *Technology Analysis & Strategic Management*, vol. 23, no. 1, pp. 87–102, 2011.

[4] "TUDelft Green office," https://www.tudelft.nl/sustainability/.

[5] "UBC Campus as a living laboratory," https://sustain.ubc.ca/our-commitment/campus-living-lab.

[6] "Harvard Sustainability," https://green.harvard.edu.

[7] "UPM City of the Future initiative," http://blogs.upm.es/cityofthefuture-upm/en/initiative/.

[8] M. Alvarez-Campana, G. López, E. Vázquez, V. A. Villagrá, and J. Berrocal, "Smart CEI moncloa: An iot-based platform for people flow and environmental monitoring on a Smart University Campus," *Sensors*, vol. 17, no. 12, 2017.

[9] M. B. Kjærgaard, H. Blunck, T. Godsk, T. Toftkjær, D. L. Christensen, and K. Grønbæk, "Indoor positioning using GPS revisited," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 6030, pp. 38–56, 2010.

[10] A. Belmonte-Hernández, G. Hernández-Peñaloza, F. Álvarez, and G. Conti, "Adaptive Fingerprinting in Multi-Sensor Fusion for Accurate Indoor Tracking," *IEEE Sensors Journal*, vol. 17, no. 15, pp. 4983–4998, 2017.

[11] M. S. Kristoffersen, J. V. Dueholm, R. Gade, and T. B. Moeslund, "Pedestrian counting with occlusion handling using stereo thermal cameras," *Sensors*, vol. 16, no. 1, 2016.

[12] L. Zheng, X. Ruan, Y. Chen, and M. Huang, "Shadow removal for pedestrian detection and tracking in indoor environments," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 18321–18337, 2017.

[13] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[14] F. Adib and D. Katabi, "See through walls with WiFi!," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication, ACM SIGCOMM 2013*, pp. 75–86, China, August 2013.

[15] Z. Kabelac, D. Katabi, and R. C. Miller, "3D Tracking via Body Radio Reflections," in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, pp. 317–329, 2014.

[16] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via RF body reflections," in *Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2015*, pp. 279–292, USA, May 2015.

[17] Z.-A. Deng, G. Wang, D. Qin, Z. Na, Y. Cui, and J. Chen, "Continuous indoor positioning fusing WiFi, smartphone sensors and landmarks," *Sensors*, vol. 16, no. 9, 2016.

[18] R. A. Becker, R. Cáceres, K. Hanson et al., "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 18–26, 2011.

[19] J. E. Mallah, F. Carrino, O. A. Khaled, and E. Mugellini, "Crowd monitoring critical situations prevention using smartphones and group detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9189, pp. 496–505, 2015.

[20] J. Kuang, X. Niu, and X. Chen, "Robust Pedestrian Dead Reckoning Based on MEMS-IMU for Smartphones," *Sensors*, vol. 18, no. 5, p. 1391, 2018.

[21] A. Kurkcu and K. Ozbay, "Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors," *Transportation Research Record*, vol. 2644, no. 1, pp. 72–82, 2017.

[22] A. Kotanen, M. Hannikainen, H. Leppakoski, and T. Hamalainen, "Positioning with IEEE 802.11b wireless LAN," in *Proceedings of the 14th IEEE 2003 International Symposium on Personal, Indoor and Mobile Radio Communications.*, vol. 3, pp. 2218–2222, Beijing, China, 2003.

[23] W. Ho, A. Smailagic, D. P. Siewiorek, and C. Faloutsos, "An adaptive two-phase approach to WiFi location sensing," in *Proceedings of the 4th Annual IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2006*, pp. 452–456, Italy, March 2006.

[24] F. Evennou and F. Marx, "Advanced integration of WiFi and inertial navigation systems for indoor mobile positioning," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 86706, 11 pages, 2006.

[25] J. Rekimoto, T. Miyaki, and T. Ishizawa, "LifeTag: WiFi-based continuous location logging for life pattern analysis," *LNCS*, vol. 4718, pp. 35–49, 2007.

[26] J. A. Besada, A. M. Bernardos, P. Tarrío, and J. R. Casar, "Analysis of tracking methods for wireless indoor localization," in *Proceedings of the 2nd International Symposium on Wireless Pervasive Computing (ISWPC '07)*, pp. 492–497, February 2007.

[27] O. Woodman and R. Harle, "Pedestrian localisation for indoor environments," in *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*, pp. 114–123, Seoul, Republic of Korea, September 2008.

[28] F. Aloul, A. Sagahyroon, A. Al-Shami, I. Al-Midfa, and R. Moutassem, "Using mobiles for on campus location tracking," in *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia, MoMM2009*, pp. 231–235, Malaysia, December 2009.

[29] S. Woo, S. Jeong, E. Mok et al., "Application of WiFi-based indoor positioning system for labor tracking at construction sites: A case study in Guangzhou MTR," *Automation in Construction*, vol. 20, no. 1, pp. 3–13, 2011.

[30] N. Le Dortz, F. Gain, and P. Zetterberg, "WiFi fingerprint indoor positioning system using probability distribution comparison," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 2301–2304, March 2012.

[31] Z. Chen, H. Zou, H. Jiang, Q. Zhu, Y. C. Soh, and L. Xie, "Fusion of WiFi, smartphone sensors and landmarks using the kalman filter for indoor localization," *Sensors*, vol. 15, no. 1, pp. 715–732, 2015.

[32] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao, "Magicol: indoor localization using pervasive magnetic field and opportunistic wifi sensing," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1443–1457, 2015.

[33] A. Danalet, B. Farooq, and M. Bierlaire, "A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 146–170, 2014.

[34] O. Czogalla and S. Naumann, "Pedestrian indoor navigation for complex public facilities," in *Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2016*, pp. 1–8, Spain, October 2016.

[35] L. Schauer, P. Marcus, and C. Linnhoff-Popien, "Towards feasible Wi-Fi based indoor tracking systems using probabilistic methods," in *Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2016*, pp. 1–8, Spain, October 2016.

[36] F. Meneses and A. Moreira, "Large scale movement analysis from WiFi based location data," in *Proceedings of the 2012 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2012*, Australia, November 2012.

[37] B. Bonné, A. Barzan, P. Quax, and W. Lamotte, "WiFiPi: involuntary tracking of visitors at mass events," in *Proceedings of the IEEE 14th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '13)*, pp. 1–6, Madrid, Spain, June 2013.

[38] M. B. Kjaergaard, M. Wirz, D. Roggen, and G. Troster, "Mobile sensing of pedestrian flocks in indoor environments using WiFi signals," in *Proceedings of the 10th IEEE International Conference on Pervasive Computing and Communications (PerCom '12)*, pp. 95–102, Lugano, Switzerland, March 2012.

[39] S. Sendra, M. Garcia, C. Turro, and J. Lloret, "People mobility behaviour study in a university campus using WLANs," in *Proceedings of the 3rd International Conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies, UBICOMM 2009*, pp. 124–129, Malta, October 2009.

[40] J. Scheuner, G. Mazlami, D. Schöni et al., "Probr - A Generic and Passive WiFi Tracking System," in *Proceedings of the 41st IEEE Conference on Local Computer Networks, LCN 2016*, pp. 495–502, UAE, November 2016.

[41] Y. Li, S. Williams, B. Moran, and A. Kealy, "Quantized RSS Based Wi-Fi Indoor Localization with Room Level Accuracy," in *Proceedings of the International Global Navigation Satellite Systems 2018*, 2018.

[42] A. Alessandrini, C. Gioia, F. Sermi, I. Sofos, D. Tarchi, and M. Vespe, "WiFi positioning and Big Data to monitor flows of people on a wide scale," in *Proceedings of the 25th European Navigation Conference, ENC 2017*, pp. 322–328, Switzerland, May 2017.

[43] L. Vu, K. Nahrstedt, S. Retika, and I. Gupta, "Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus," in *Proceedings of the13th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '10)*, pp. 257–265, October 2010.

[44] M. Garcia, S. Sendra, C. Turro, and J. Lloret, "User's Macro and Micro-mobility Study using WLANs in a University Campus," *International Journal On Advances in Internet Technology*, vol. 4, no. 1, pp. 37–46, 2011.

[45] Y. Xu, I. D. G. Groeneveld, R. Sulzer, E. Theocharous, O. T. Willems, and M. S. Tryfona, "Determine activity based on the classified identity of users by using Wi-Fi monitoring," Geomatics Synthesis Group Project Report, 2016.

[46] A. Danalet, L. Tinguely, M. D. Lapparent, and M. Bierlaire, "Location choice with longitudinal WiFi data," *Journal of Choice Modelling*, vol. 18, pp. 1–17, 2016.

[47] M. Zhou, K. Sui, M. Ma, Y. Zhao, D. Pei, and T. Moscibroda, "MobiCamp: A campus-wide testbed for studying mobile physical activities," in *Proceedings of the 3rd International Workshop on Physical Analytics, WPA 2016*, pp. 1–6, Singapore, 2016.

[48] A. Fernández-Ares, A. M. Mora, M. G. Arenas et al., "Studying real traffic and mobility scenarios for a Smart City using a new monitoring and tracking system," *Future Generation Computer Systems*, vol. 76, pp. 163–179, 2017.

[49] A. E. Redondi and M. Cesana, "Building up knowledge through passive WiFi probes," *Computer Communications*, vol. 117, pp. 1–12, 2018.

[50] L. Huang, K. Matsuura, H. Yamanet, and K. Sezaki, "Enhancing wireless location privacy using silent period," in *Proceedings of the 2005 IEEE Wireless Communications and Networking Conference, WCNC 2005: Broadband Wirelss for the Masses - Ready for Take-off*, pp. 1187–1192, USA, March 2005.

[51] J. Martin, T. Mayberry, C. Donahue et al., "A Study of MAC Address Randomization in Mobile Devices and When it Fails," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, pp. 365–383, 2017.

[52] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens, "Why MAC address randomization is not enough: an analysis of Wi-Fi network discovery mechanisms," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*, pp. 413–424, ACM, Xi'an, China, June 2016.

[53] A. J. Ruiz-Ruiz, H. Blunck, T. S. Prentow, A. Stisen, and M. B. Kjaergaard, "Analysis methods for extracting knowledge from large-scale WiFi monitoring to inform building facility planning," in *Proceedings of the 2014 12th IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, pp. 130–138, Hungary, March 2014.

[54] N. Nunes, M. Ribeiro, C. Prandi, and V. Nisi, "Beanstalk - A community based passive Wi-Fi tracking system for analysing tourism dynamics," in *Proceedings of the 9th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS 2017*, pp. 93–98, Portugal, June 2017.

[55] C. Parent, N. Pelekis, Y. Theodoridis et al., "Semantic trajectories modeling and analysis," *ACM Computing Surveys*, vol. 45, no. 4, pp. 1–32, 2013.

[56] C. Wei, "Mining of User Behavioral Features Based on Indoor Semantic Trajectories," Boletín Técnico, ISSN:0376-723X, vol. 55, 2017.

[57] K. V. Long, D. Quang, and N. Klara, *Lessons learned from bluetooth/wifi scanning deployment in university campus*, Urbana, Illinois, USA, 2010.

[58] V. Radu and M. K. Marina, "HiMLoc: indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced WiFi fingerprinting," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '13)*, pp. 1–10, IEEE, Montbeliard-Belfort, France, October 2013.

[59] F. Hong, Y. Zhang, Z. Zhang, M. Wei, Y. Feng, and Z. Guo, "WaP: Indoor localization and tracking using WiFi-Assisted Particle filter," in *Proceedings of the 39th Annual IEEE Conference on Local Computer Networks, LCN 2014*, pp. 210–217, Canada, September 2014.

[60] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Analyzing shopper's behavior through WiFi signals," in *Proceedings of the 2nd Workshop on Physical Analytics, WPA 2015*, pp. 13–18, Italy, 2015.

[61] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using WiFi signals," *PLoS ONE*, vol. 10, no. 7, p. e0130824, 2015.

[62] Z. Tian, Y. Jin, M. Zhou, Z. Wu, and Z. Li, "Wi-Fi/MARG Integration for Indoor Pedestrian Localization," *Sensors*, vol. 16, no. 12, p. 2100, 2016.

[63] B. Zhou, Q. Li, Q. Mao, and W. Tu, "A Robust Crowdsourcing-Based Indoor Localization System," *Sensors*, vol. 17, no. 4, p. 864, 2017.

[64] "Raspberry Pi," https://www.raspberrypi.org/.

[65] "TP-LINK Wi-Fi USB dongle datasheet," http://www.tp-link.com/us/products/details/cat-5520_TL-WN722N.html.

[66] "ISO/IEC 20922:2016 - Information technology – Message Queuing Telemetry Transport (MQTT) v3.1.1," https://www.iso.org/standard/69466.html.

[67] G. Piateski-Shapiro and W. Frawley, *Knowledge Discovery in Databases*, MIT Press, Cambridge, MA, USA, 1991.

[68] J. Andión Jiménez, J. M. Navarro González, M. Álvarez-Campana Fernández-Corredor, and J. C. Dueñas López, "A passive, non-intrusive, cheap method to identify behaviours and habits in the Campus," in *Proceedings of the XIII Jornadas de Ingenieria Telematica - JITEL2017*, vol. 40, no. 47, pp. 10–4995, September 2017.

[69] A. Mylinikov, "Geo project," https://www.mylnikov.org/.

[70] K. Mehmet and G. Ananth, "PROXIMUS: A framework for analyzing very high dimensional discrete-attributed datasets," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, vol. 147, no. 156, pp. 147–156, USA, August 2003.

[71] "UPM Library occupation," http://ceiboard.dit.upm.es/smart-campus/biblio.

[72] C. A. R. Inarejos, A. Rodríguez, G. López, and M. Alvarez-Campana, "Análisis de la huella de carbono de la ETSIT de la UPM y propuesta de mejora basada en datos de la plataforma IoT Smart CEI Moncloa," in *Proceedings of the I Congreso Iberoamericano de Ciudades Inteligentes (ICSC-CITIES 2018)*, 2018.