Applied Geomatics (2019) 11:381–399
https://doi.org/10.1007/s12518-019-00260-z

ORIGINAL PAPER

On cleaning strategies for WiFi positioning to monitor dynamic crowds

Ciro Gioia[1] · Francesco Sermi[1] · Dario Tarchi[1] · Michele Vespe[1]

Received: 12 December 2018 / Accepted: 9 April 2019 / Published online: 7 June 2019

Abstract
Monitoring open crowded areas is fundamental for policy makers to set up the proper measures for people security and safety. Different techniques have been developed to tackle this issue. The most relevant approaches, which are currently available to estimate the number of attenders, are based on the size of the area hosting the event, the count of people passing through specific points, and the employment of satellite images. All these techniques roughly estimate static crowds, but they may be limited by different factors such as the availability of satellite images, the cost of dedicated unmanned vehicles, and the capability to set up multiple counting points. In order to fill this gap, a tool to monitor dynamic crowds, based on WiFi positioning, is presented. The tool allows not only to assess the total number of people attending an event, but also to monitor their spatiotemporal distribution. In particular, the impact of the cleaning strategies on both the estimated number of participants and their spatiotemporal distribution is analyzed. The proposed approach is demonstrated using real data collected during the JRC Open Day 2016. From the results, the need of a clear strategy to identify real users in order to avoid misleading results emerges. Moreover, a proper setting of the thresholds used for the identification criteria is required. Such thresholds need to be set according to the dimension of the site, the geography of the WiFi network, and the duration of the event.

Keywords WiFi · Tracking · Big data · Cleaning

Introduction

Monitoring open crowded areas is fundamental for policy makers to set up the proper measure for people security and safety (Doig 2009; Oberschall 1973; Nardo 1985; Lohmann 1994); of foremost importance is the number of people attending an event (Jacobs 1967; Krewson 2012) and their distribution in time and space. In Cariveau (2006) and Rabaud and Belongie (2006), a review of the common techniques adopted for estimating the number of people attending an event is presented. In McPhail and McCarthy (2004) and Yip et al. (2010), the estimation of the number of people participating to a demonstration is attempted. Many different techniques have been proposed to estimate the dimension of a crowd. The most relevant techniques currently available to assess the number of people attending an event are based on the following:

- the size of the area hosting the event (Jacobs 1967; Krewson 2012): the total number of participant is obtained multiplying the area for a factor depending on the season during which the event takes place.
- the actual count of the people passing through specific points (Watson 2011).
- the employment of satellite images and density-analysis (Sirmacek and Reinartz 2011; Wallace and Parlapiano 2017).

✉ Ciro Gioia
ciro.gioia@ec.europa.eu

Francesco Sermi
francesco.sermi@ec.europa.eu

Dario Tarchi
dario.tarchi@ec.europa.eu

Michele Vespe
michele.vespe@ec.europa.eu

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy

– the employment of images from unmanned aerial vehicles (Choi-Fitzpatrick and Juskauskas 2015; Choi-Fitzpatrick 2014).

All the abovementioned approaches are mainly used for the estimation of static crowds, and they are limited by different factors such as the availability of satellite images, the cost of images from a dedicated unmanned vehicles, and the capability to set specific counting points.

About the dynamic crowd monitoring, several approaches are known to currently be used; each of them presents given benefits and limitations. Two of the most common approaches for dynamic crowd monitoring in outdoor scenarios are mentioned in the following. One of the most adopted is the image-video processing supported by a dense network of cameras (Zhan et al. 2008; Li et al. 2015). Its performance depends on dislocation of the cameras, resolution of the acquired images, characteristics of the adopted image processing algorithm, and applied tracking techniques. Its main limitations are represented by the possible camera occlusion and the weather-environmental conditions, e.g., rain, fog, smoke bombs, and low-light conditions, may limit the applicability of the approach. A comprehensive survey on crowd monitoring based on video and images is available in Lamba and Nain (2017). In Yuan et al. (2013) and Yuan (2014), the authors propose an RF-based crowd density estimation for indoor scenarios, using mobile phones, together with considerations on its extension to outdoor. This approach does not present the limitations of that based on cameras and it provide excellent performance, but it requires data from telecom providers and faces strong privacy constraints, which may limit its application (EC-GDPR 2016).

In this paper, a tool to monitor dynamic crowds is presented; such a tool allows not only the estimation of the total number of people attending an event, but also their spatiotemporal distribution.

The capability to locate and track users exploiting WiFi data has been already proven in several works (Bobescu and Alexandru 2015; Kotaru et al. 2015; Biswas and Veloso 2010). The approach is based on two implicit assumptions (Petre et al. 2017): everyone uses a smartphone whose WiFi is enabled all the time. Starting from these two assumptions, the WiFi positioning and tracking exploit the fact that smartphones repeatedly broadcast probe requests to identify known networks. The probe request contains the device unique identifier: the media access control (MAC) address (Alessandrini et al. 2017). Therefore, in order to be detected, a smartphone does not have to be connected to a WiFi network, but it just needs to be within the range of a WiFi node, when the probe request is sent. Having said that, one can easily imagine that by carefully dislocating WiFi nodes within a given area, it is possible to collect MAC addresses of most of the users passing through that area and, through a proper processing, to track these users.

Nowadays, the applications of WiFi positioning and tracking are numerous and growing, including the analysis of visitor movements and flows, including queue times, dwell times, wait times, and first-time/repeat visitors; the implementation of geofencing systems to define boundaries around areas of interest, triggering alerts when registered mobile devices enter or exit it; and the optimization of security and safety assets dislocation during events that gather critical masses of people.

Obviously, the exploitation of WiFi data is not the only available solution to track users. Among the other techniques, global positioning system (GPS) tracking is the most common and extensively used. Nevertheless, differently from GPS tracking, WiFi tracking is feasible also indoors (Mingkhwan 2006). Moreover, whereas GPS data are owned by telecommunications providers, an ad hoc network of WiFi nodes is both cheap and easily deployable, allowing direct collection of the necessary data. On the other hand, one of the main drawbacks of WiFi tracking is the need of a pre-processing, namely "data-cleaning," which is necessary to identify actual users, but it may limit the real-time application of the method. Another key aspect that favors a tracking method rather than another is how it complies with privacy rules.

Privacy-related aspects are a fundamental topic, when it comes to locate people through WiFi. Some relevant considerations about this subject have been made in Alessandrini et al. (2017), but the discourse on privacy and personal data has definitely grown up within and outside the EU, since the enforcement of the General Data Protection Regulation (GDPR) in April 2018. How does the GDPR affect the use of WiFi data? Primarily, companies will no longer be able to provide free WiFi to consumers in exchange for their browsing data: a very common practice for the profiling and targeting of potential costumers. The GDPR also states that organizations must "implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk" of the provided service. In other words, companies should implement best practices to secure publicly available WiFi networks (Meyer 2018). What is still unclear is how the GDPR will affect WiFi tracking. The whole discourse revolves around the definition of MAC address. In fact, the MAC address, apart from being unique, does not contain any information about the user. Nevertheless, since it identifies a given device, that in the case of smart phone is always kept by the user, the MAC address could be considered as indirect personal information, or "pseudonymous" data (Maldoff 2016). "Pseudonymization" is a new concept introduced by the GDPR, consisting in the separation of data from direct identifiers so that linkage to an identity is not possible

without additional information that is held separately. Much debate surrounds the extent to which pseudonymized data can be re-identified. This issue is of critical importance because it determines whether a processing operation will be subject to the provisions of the Regulation.

The detection of a device using WiFi relies on the uniqueness of its MAC address. Nevertheless, once a device is detected, one needs to identify if that device corresponds to an actual user, i.e., to a physical person, or to a static or fake device. In fact, smartphones are not the only devices using WiFi. There are plenty of new-generation devices that exploit WiFi connection, such as printers, hard drives, music players, surveillance cameras, and smart thermostats; even some digital photo frames use WiFi. Moreover, some devices use a fake MAC address when they broadcast probe requests to identify available WiFi networks. The identification of actual users is the main goal of the abovementioned "data cleaning" procedure. Immediately after the data cleaning, the "positioning" of the detected users comes. The estimation of the user position within a given area can be done according to different methods, which can be based on the strength of the received WiFi signal, the number of times that a WiFi receiver registers a given users, etc.

This paper focuses on both cleaning and user positioning procedures, starting from the data collection of the 2016 Open Day of the Joint Research Centre (JRC) in Ispra (Italy) (Alessandrini et al. 2017) and comparing different approaches on the basis of the results described in Gioia et al. (2017). Data cleaning is conducted following different approaches: a first screening of the device can be made using the number of times a device is registered; the second criterion is based on the number of base stations registering the presence of the user; and a different approach to identify real user is developed considering the dispersion of the estimate user position. The explored positioning strategies are the following: proximity principle based on received signal strength (RSS), proximity principle using the number of records, weighted centroid approach.

The paper is organized in the following way: the next section describes the cleaning (the "Cleaning approaches" section) and the positioning (the "Positioning approaches" section) approaches; the "Experiment" section briefly illustrates the data collection experiment; the results are summarized in the "Results" section , which is followed by relevant concluding remarks in "Conclusions" section.

Methods

In the following sections, the cleaning strategies and the localization algorithms are described.

Cleaning approaches

WiFi data includes static devices (printers, personal computer, etc.), fake devices, and devices outside the test sites. In order to remove the data relative to the not real-users, three cleaning strategies are implemented in the measurement and position domains. The first one is based on the minimum number of times a device is recorded, the second one considers the number of stations by which a device is registered, whereas in the position domain, the distribution of the estimated user positions is analyzed.

A first screening of the devices can be made using the number of times a device is registered:

$$\text{real user} = \text{if num record}_\text{i} \geq \text{th}_\text{rec} \tag{1}$$

where num record_i is the number of records of the ith device and th_rec is the threshold set for the detection of the real users.

The criterion based only on the number of records for a given device allows to screen out devices which are registered for a very limited period of time, for example, devices which were in the proximity of the test field but not entering it. However, this criterion does not allow the exclusions of static devices (printers, PC, etc.) which are registered for all the duration of the experiment. Hence, another selection criterion is adopted, which is based on the number of base stations registering the presence of the given user. Using this approach, a device is classified as "actual user" if its identifier is recorded by a number of access points (APs) higher than a fixed threshold:

$$\text{real user} = \text{if num base}_i \geq \text{th}_\text{base} \tag{2}$$

where num base_i is the number of APs registering the presence of the ith device and th_base is the threshold defining the minimum number of APs required to classify a device as a real user.

A different approach for identifying real user is developed considering the dispersion of the estimate users position. The standard deviation of the user coordinates is computed and compared with a given threshold in order to evaluate the dispersion of the user position:

$$\text{real user} = \text{if } \sigma\ (\text{pos}_\text{user}) \geq \text{th}_\text{pos} \tag{3}$$

where $\sigma\ (\text{pos}_\text{user})$ is the standard deviation of the estimate user positions and th_pos is the threshold used to identify a real user.

The criterion represented in Eq. 3 is very general: it can be applied to the single coordinates separately

$$\begin{aligned} &\text{real user} = \text{if } \sigma\left(\text{pos}_{\text{user},x}\right) \geq \text{th}_{\text{pos},x} \\ &\text{real user} = \text{if } \sigma\left(\text{pos}_{\text{user},y}\right) \geq \text{th}_{\text{pos},y} \end{aligned} \tag{4}$$

or it can be applied using both conditions in Eq. 4 (see first equation in Eq. 5), or on the horizontal component (second equation in Eq. 5)

$$\begin{aligned} &\text{real user} = \text{if } \left(\sigma\left(\text{pos}_{\text{user},x}\right) \geq \text{th}_{\text{pos},x} \;\&\; \sigma\left(\text{pos}_{\text{user},y}\right) \geq \text{th}_{\text{pos},y}\right) \\ &\text{real user} = \text{if } \sigma\left(\text{pos}_{\text{user,hor}}\right) \geq \text{th}_{\text{pos,hor}}. \end{aligned} \tag{5}$$

Positioning approaches

In this section, the methodologies developed to compute the users position are described. Specifically, three different strategies have been adopted to track users:

- proximity received signal strength indicator (RSSI)-based (Manandhar et al. 2008; Dempster 2009);
- proximity occurrence-based;
- Weighted centroid (WeC) (Wang et al. 2013; Borio et al. 2016; Gioia and Borio 2014a, b).

These approaches are commonly used with simultaneous measurements, i.e., the object to be localized is simultaneously connected to two or more nodes (Manandhar et al. 2008; Dempster 2009; Borio et al. 2016; Gioia and Borio 2014a). During the performed test, the objects to be localized are usually seen only by one node at a time; this is due to the size of the site where the experiment has been carried out and to the typical area coverage of the APs adopted for the experiment, together with their geographical displacement within the site. Hence, the traditional algorithms have been modified to compute the position of the tracked object after accumulating measurements during a given time interval. The time interval used to estimate the user position is 3 min; this value has been selected considering the following factors:

- The size of the site, the geometry of the network, and the typical dynamic of the users: during the experiment, the mean distance between the APs along the East direction is almost 250 m and some 160 m along North, assuming that a pedestrian moves at approximately 4 km/h (1.1 m/s), this covers about 200 m in 3 min.
- Heterogeneity of the device, in particular considering the different data rates. A fundamental element to set the value is the update rate of the measurements; the cumulative distribution function of the number users as a function of the update rate is shown in Fig. 1. From

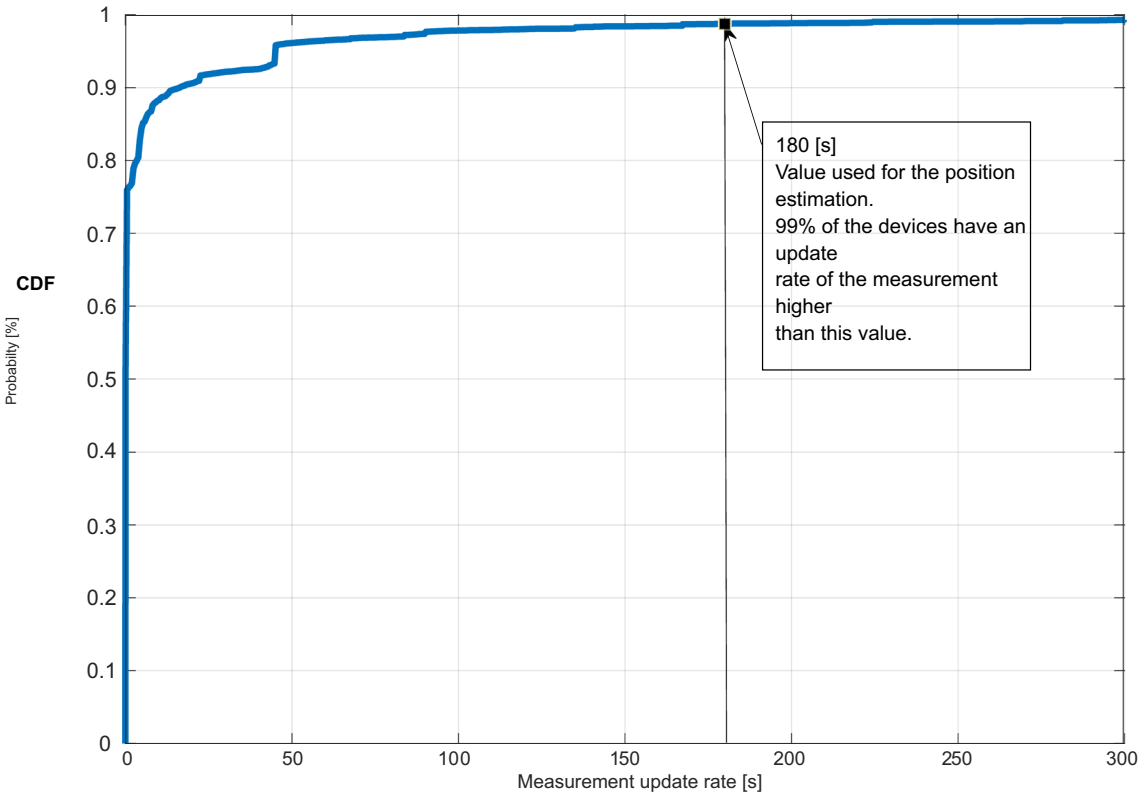

Fig. 1 CDF of the number of users as a function of the update rate

the figure, it can be noted that 99% of the devices have an update rate higher than 3 min.

In Fig. 2, the three positioning approaches are shown together with the relative inputs. The input of the occurrence-based positioning method is only the list of the stations which recorded the user presence in the time interval, whereas the other two approaches exploit also the power of the received signal.

As mentioned above, three approaches have been implemented: two derived the from proximity concept and the third exploits the centroid concept.

Proximity RSSI-based

The first algorithm is based on the proximity principle and exploits the RSS measurements. The position of the tracked object is associated with the position of the station recording the signal of the user with the highest RSSI in the specific time period ΔT:

$$\mathbf{pos_u}\,(\Delta T) = \mathbf{pos_{s_{MaxRSSI}}}\,(\Delta T) \tag{6}$$

where $\mathbf{pos_u}$ is the vector containing the user coordinates and $\mathbf{pos_{s_{MaxRSSI}}}$ is the vector containing the coordinates of the station that registered the signal of the user with the maximum RSSI.

Proximity occurrence-based

The RSS is strongly affected by multipath and fading phenomena; these effects are intrinsic characteristics of the propagation environment and they can amplify or reduce the received signal power. The multipath-induced variation of the RSS could lead to erroneous object localization; in order to fill this gap, an algorithm based on the proximity principle, but exploiting the number of times a users is recorded by a station during a specific time period ΔT, is proposed.

In this case, the position of the tracked object is associated with the position of the node that registers more times the presence of the user.

$$\mathbf{pos_u}\,(\Delta T) = \mathbf{pos_{s_{MaxNumPres}}}\,(\Delta T) \tag{7}$$

WeC

The third approach is the WeC, which is an extension of the proximity principle; in this case, the user position is a linear combination of the node coordinates. The mathematical expression of the WeC is that of Borio et al. (2016) and Gioia and Borio (2014a):

$$\mathbf{pos_u}\,(\Delta T) = \frac{\sum_{i=0}^{N-1} w_i \mathbf{P_{s,i}}}{\sum_{i=0}^{N-1} w_i}, \tag{8}$$

where $\mathbf{P_{s,i}} = (x_i, y_i)$ is the vector containing the coordinates of the ith station and w_i is the weight associated to the ith node. In this work, the weights are related to the RSSI of the received signal, in particular the following weighting function is adopted:

$$w_i = 1/\left(2 \cdot 10^{(-\mathrm{RSSI})_i/10}\right), \tag{9}$$

where $(\mathrm{RSSI})_i$ is the RSS of the ith received signal expressed in dB.

The user position is obtained as the WeC of the nodes coordinates. If the time interval is reduced and only one measurement is obtained within the considered interval, the WeC solution converges to the RSSI-based proximity solution.

In Table 1, the three implemented positioning algorithms are summarized, together with their main pros and cons.

Experiment

In this section, the experimental setup is briefly described; a more comprehensive description of the experiment is available in Alessandrini et al. (2017).

The experiment was carried out on 28 May 2016, during the JRC open day event (JRC Open Day 2016). Thanks to the large attendance (more than 7500 participants) and to its duration (some 11 h), the event was a unique opportunity to collect a large amount of data. For the experiments, 20 APs were placed within the JRC Ispra site, as shown in Fig. 3, where the AP locations and the relative identification number are reported. The figure shows also the theoretical

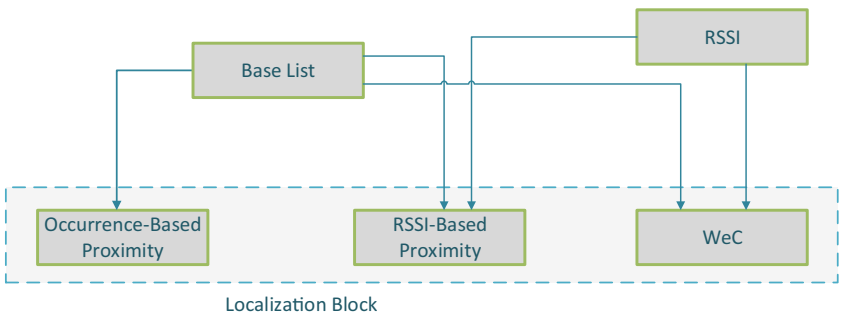

Fig. 2 Block diagram of the positioning procedures

Table 1 Positioning algorithms implemented

Method	Expression	Pros	Cons
Proximity occurrence-based	$\mathbf{pos_u}\,(\Delta T) = \mathbf{pos_{s_{MaxNumPres}}}\,(\Delta T)$	Not affected by multipath-induced RSSI anomalies	Limited dynamics
Proximity RSSI-based	$\mathbf{pos_u}\,(\Delta T) = \mathbf{pos_{s_{MaxRSSI}}}\,(\Delta T)$	More resilient to anomalous RSSI observations than WeC	Limited dynamics
WeC	$\mathbf{pos_u}\,(\Delta T) = \frac{\sum_{i=0}^{N-1} w_i \mathbf{P_{s,i}}}{\sum_{i=0}^{N-1} w_i}$	More dynamics	Potentially affected by anomalous RSSI measurements; a proper weighting function needs to be adopted.

coverage (the effect of obstacles limiting the range of the APs is not considered) of the APs. The blue markers identify APs located in proximity to the main entrance: two devices (AP numbers 1 and 2) were placed close to the gate reserved to the general public, which remained closed until the official opening of the event (9:00 AM), whereas the AP 3 was placed close to the gate reserved for the access of volunteers, which was open from 7:30 to 9:00. The yellow markers show the positions of the APs in the central area of the site; among the yellow markers, the AP 7 was located in the Brebbia gate, which was reserved to volunteer entrance and was open only 1 h and half before the official opening. Finally, the red markers are used to identify APs close to the exit gate, located in the west part of the site.

The event was a unique opportunity, not only to collect a huge amount of data, but because of the nature of the site: only one access and one exit for the general public, with the security staff counting the accesses, the security report allows to partially verify the results obtained (Sousa 2016). This gave us the possibility to perform a "qualitative" check of the results obtained through the proposed methods. The results check could be only "qualitative" for two main reasons:

– Despite the security office reported 7623 accesses during the entire event (Sousa 2016), there is no way to know how many of the participants held a mobile device with WiFi enabled. In fact, the two implicit assumptions "everyone use a smartphone whose WiFi is enable all the time" are not entirely true: among the participants, there were children and elderly people who might not had a mobile; and it is reasonable to assume that there were people who's mobile had the WiFi switched off. Therefore, the estimation of the number of participants carried out with the proposed method can be compared with the actual count of accesses reported by the security service only through an assumption on the share of participants without a mobile or with WiFi disabled.

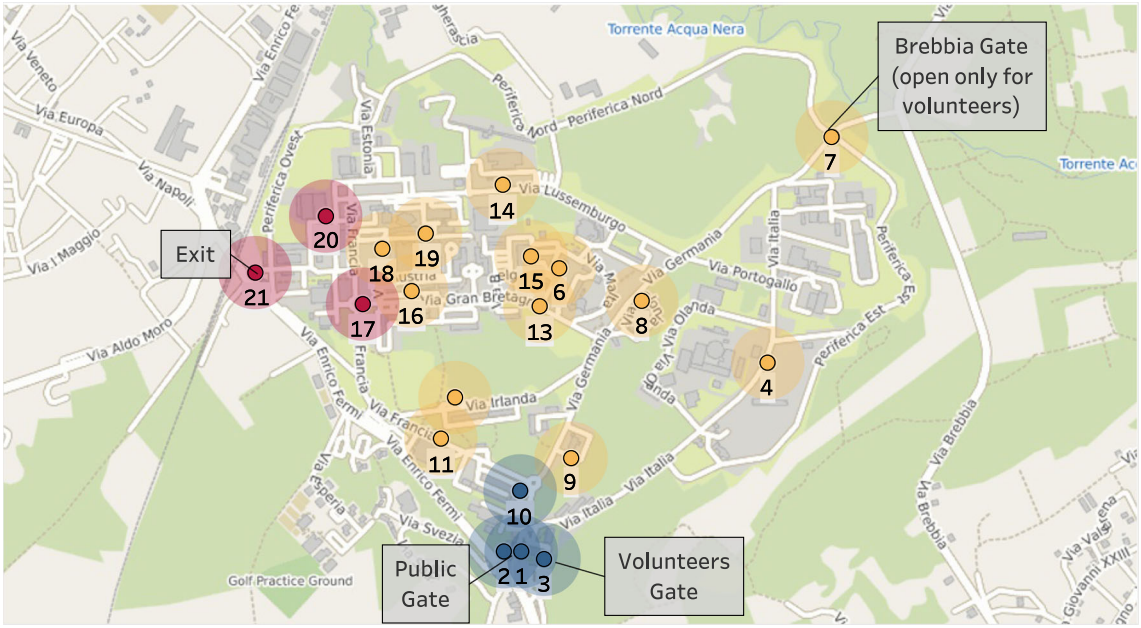

Fig. 3 Location of the APs used for the experiment during the JRC open day 2016 and their theoretical coverage

– Analogously, the participants were not actually tracked with a different system during the event; therefore, the results relative to their flow within the site can be only qualitatively interpreted according to the schedule of the event, the geography of the site, and the location of the different exhibitions.

However, having said that, the impossibility to perform a more accurate ("quantitative") validation of the results with the ground truth does not represent a limitation to the potential of the proposed method as a crowd-monitoring tool, especially when it is compared with the currently exploited techniques and with their accuracy.

Results

In this section, the results obtained combining the diverse cleaning strategies and positioning approaches are presented. The results are at first analyzed in terms of estimated number of real users; then, the concentration of the identified users is shown; finally, the movements of the users among the nodes of the network are considered.

Cleaning based on the number of base stations and records

The estimated number of real users using the criteria described by Eqs. 1 and 2 is discussed in this section.

In Fig. 4, the estimated number of real users as a function of the threshold value is shown. In the upper box, the criterion based on the number of stations which have recorded a device is considered. From the analysis, it can be noted that the number of real users decreases exponentially: the grey line identifies the exponential trend. The relation between the estimated number of users and the minimum number of stations which recorded the device is:

$$\ln(\text{number of real users}) = a * \text{th}_{\text{base}} + b \tag{10}$$

where $a = -0.245$ and $b = 11.26$.

The R-squared value (Everit and Skrondal 2010; Rawlings et al. 1998) of the model is 0.88. The total number of

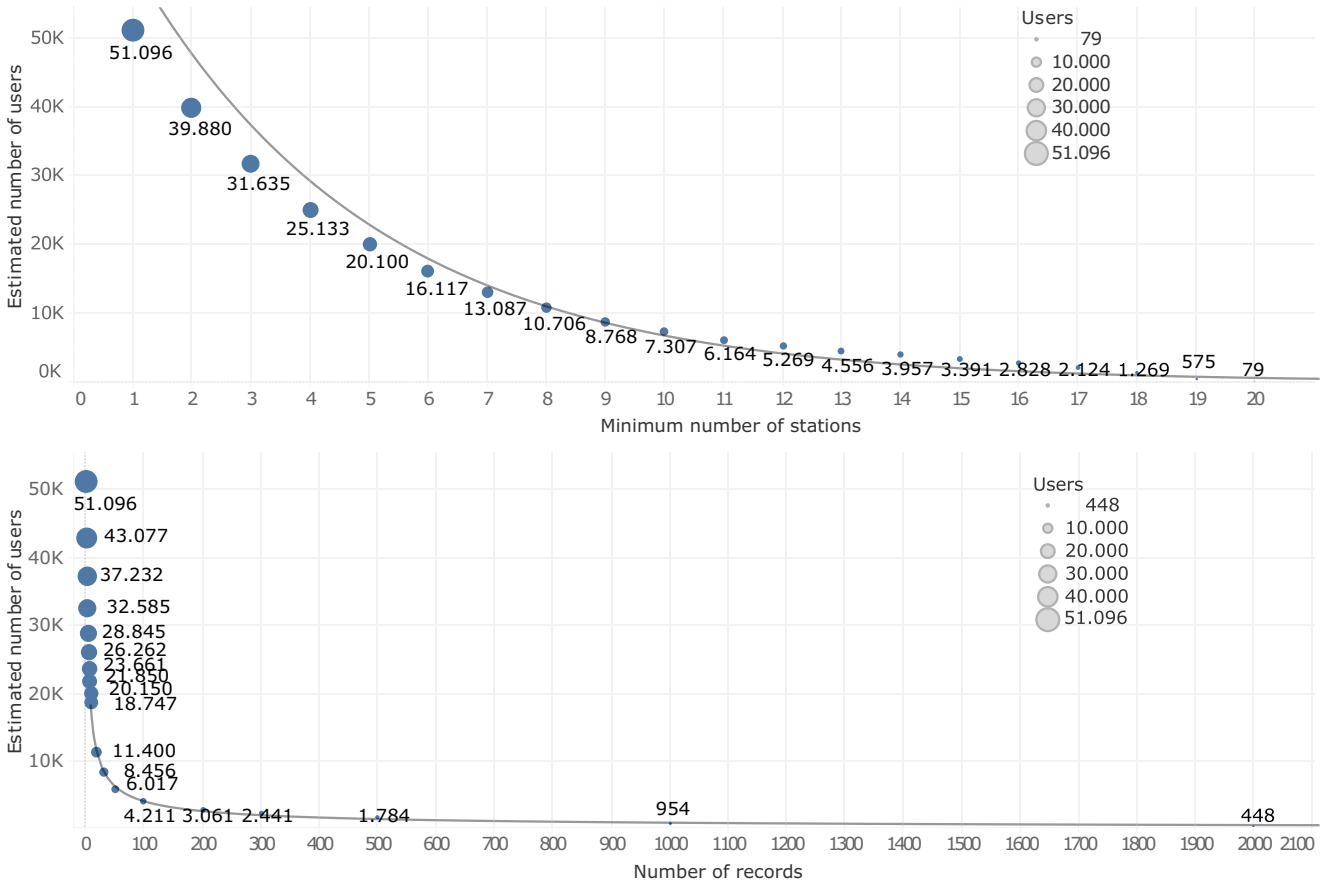

Fig. 4 Estimated number of real users as a function of minimum number of stations recording the presence of the device (upper box) and minimum number of records of a single device (lower box)

unique devices is 51,096, which is reduced by some 12,000 if a user is defined as a device recorded at least by two stations (number of devices recorded by at least two stations 39,880). Increasing the threshold, the number of users is reduced; only 79 users were recorded by all the APs.

The estimated number of users, as a function of th_{rec}, is shown in the bottom graph of Fig. 4, where the number of users for 21 different values of th_{rec} is explicitly reported. The number of users decreases when increasing th_{rec}, according to the function:

$$\ln(\text{number of real users}) = c * \ln(th_{rec}) + d \tag{11}$$

with $c = -0.622$ and $d = 11.2$; the R-squared is 0.989.

From the the same graph, it can be noted that there were some 8000 devices $(51,096 - 43,077)$ which were recorded only once. These records are probably due to devices generating fake identifiers before connecting to a node (Zebra Technologies 2015). It can also be appreciated that only 448 devices out of 51,096 where recorded more than 2000 times.

As mentioned in the "Cleaning approaches" section, the two methods, individually considered, are not able to exclude all the devices which are not associated to real users. In order to enhance the exclusion capability of the algorithm, the two criteria need to be used together; the results obtained combining the two criteria are shown in Fig. 5. In the figure, the size of the square indicates the number of real users identified by the combination of the two criteria. From the graph, it can be appreciated that the squares with the largest size are in the upper left, indicating a low exclusion capability. The number of exclusion increases passing from up to down and from left to right; in the two corners, upper left and lower right, the extreme values recorded are 51,096 and 307 respectively. Using $th_{base} = 3$ and $th_{rec} = 50$, a total number of

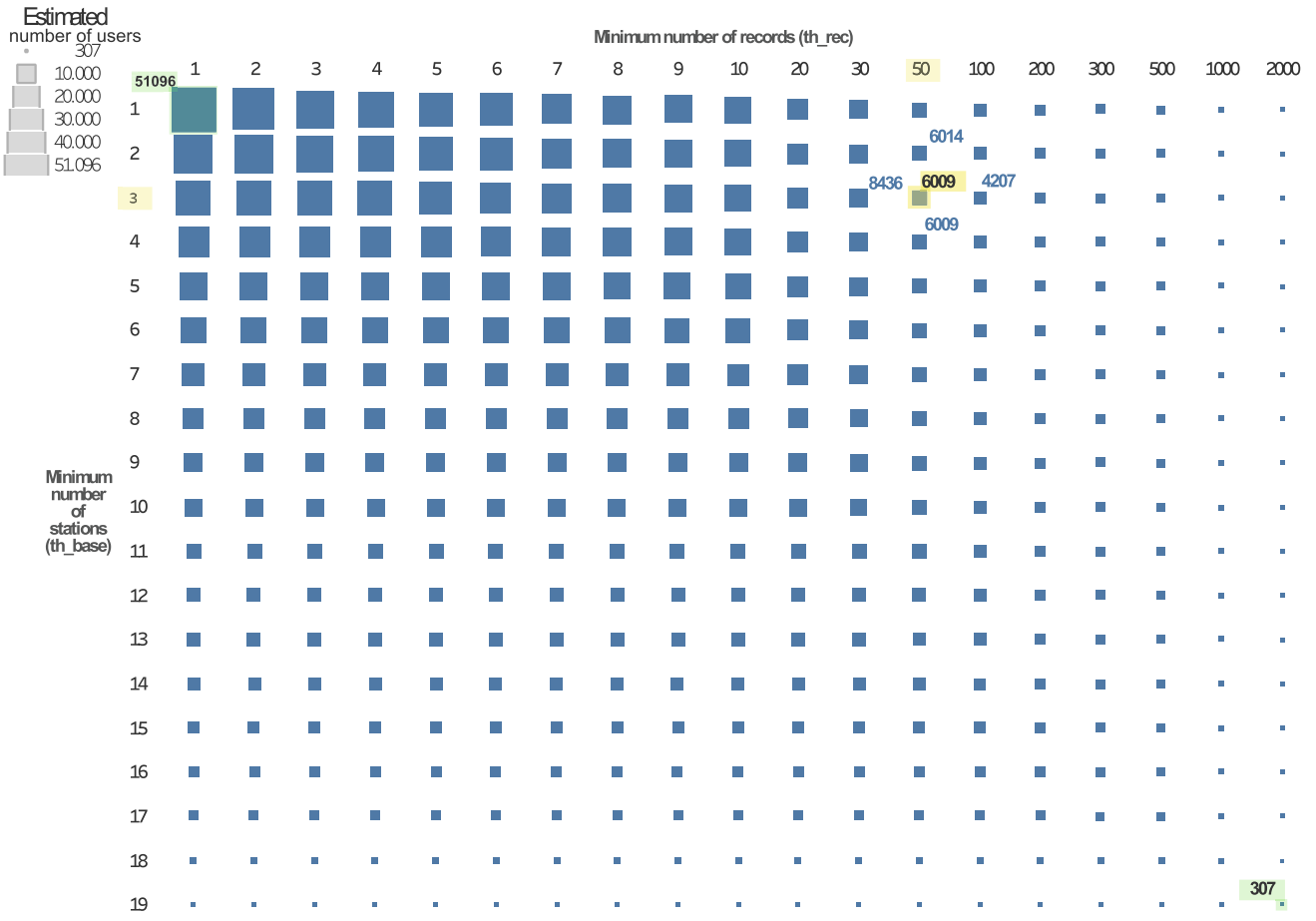

Fig. 5 Estimated number of real users considering two criteria together: a device is classified as real user if it is recorded by a minimum number of stations and for a minimum number of times

6009 potential real users is estimated. The estimate strongly depends on the setting of the thresholds, which has to be set taking into account both the geometry of the site and the length of the phenomena to monitor. In our case, considering the geometry of the site, three can be chosen as a reasonable threshold for the minimum number of the stations, because in the proximity of the main gate, three APs were installed.

In order to further investigate the impact of the minimum number of stations, the estimated number of users as a function of the minimum number of stations considering a minimum number of times is shown in Fig. 6. The lines represent the estimated number of users as a function of the minimum number of stations. From the graph, it is evident that in the considered case (more than 3 base and at least 50 times), there is only limited impact passing from 1 to 4 stations (i.e., the number of users is reduced only by 8). If the minimum number of records is reduced, a larger impact of the minimum number of stations can be appreciated. In all the curves, a plateau is present in the first part of the line, the plateau becomes longer as the minimum number of records is increased; this is reasonable because a user registered for longer time is more likely registered by a larger number of stations. Hence, the dominant criterion in the considered case is the minimum number of records. If a different set of thresholds is used, the effect of the two criteria varies.

Cleaning based on the analysis of users movements

The potential number of users can be estimated with a criteria based on the user movements, as described in the "Cleaning approaches" section. In this section, the results relative to the aforementioned criteria are discussed; the localization of the users has been performed using the three positioning strategies described in the "Positioning approaches" section.

In Figs. 7, 8 and 9, the estimated number of potential real users is shown; the positions of the users are computed using WeC, proximity occurrence-based, and proximity RSSI-based respectively. In the upper boxes, the estimated number of users is obtained applying the exclusion criterion at the East coordinates, while in the central boxes, the criterion is applied on the North coordinates, finally in the lower boxes, the number of users is obtained from the intersection of the users satisfying both criteria (East and North). In the three figures, the estimated number of users is plotted as a function of $\mathrm{th}_{\mathrm{pos}}$.

For all the cases, a linear relation (gray line) can be seen between the estimated number of users and the threshold values:

$$\text{number of real users} = m * \mathrm{th}_{\mathrm{pos}} + q. \tag{12}$$

The values of m and q for the different cases are reported in Table 2.

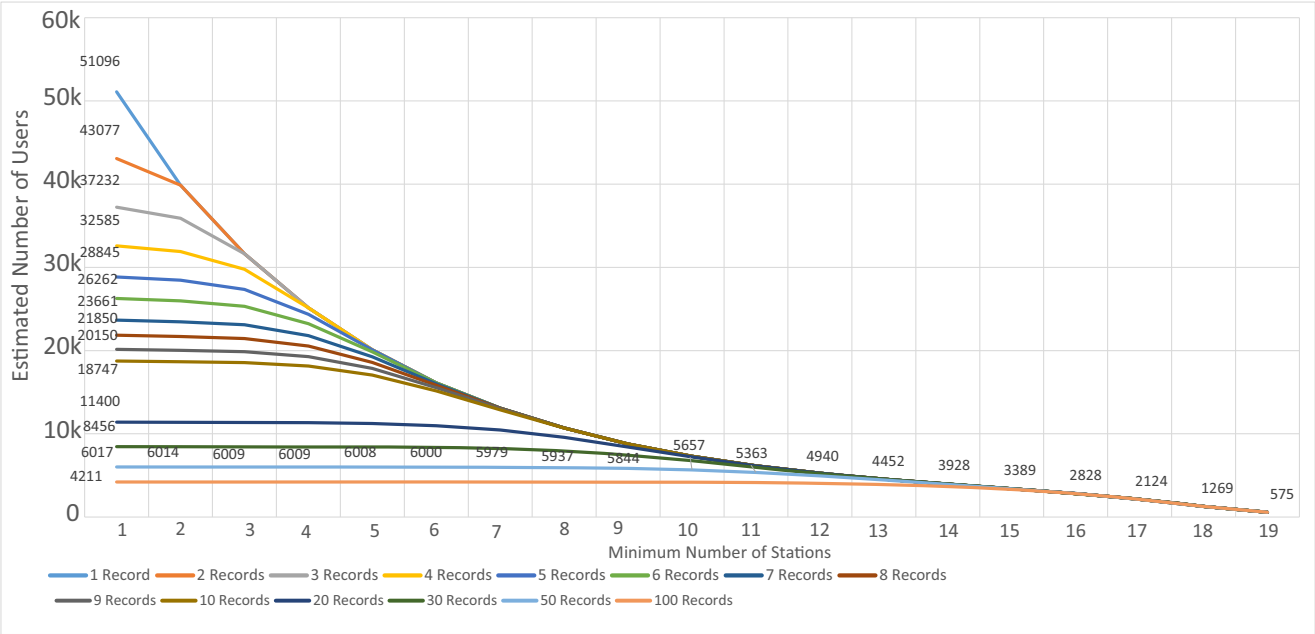

Fig. 6 Estimated number of real users as a function of the minimum number of stations considering a minimum number of records. In the considered case (more than 3 base stations and at least 50 records) there is only a limited impact passing from 1 to 4 stations (i.e., the number of users is reduced only by 8)

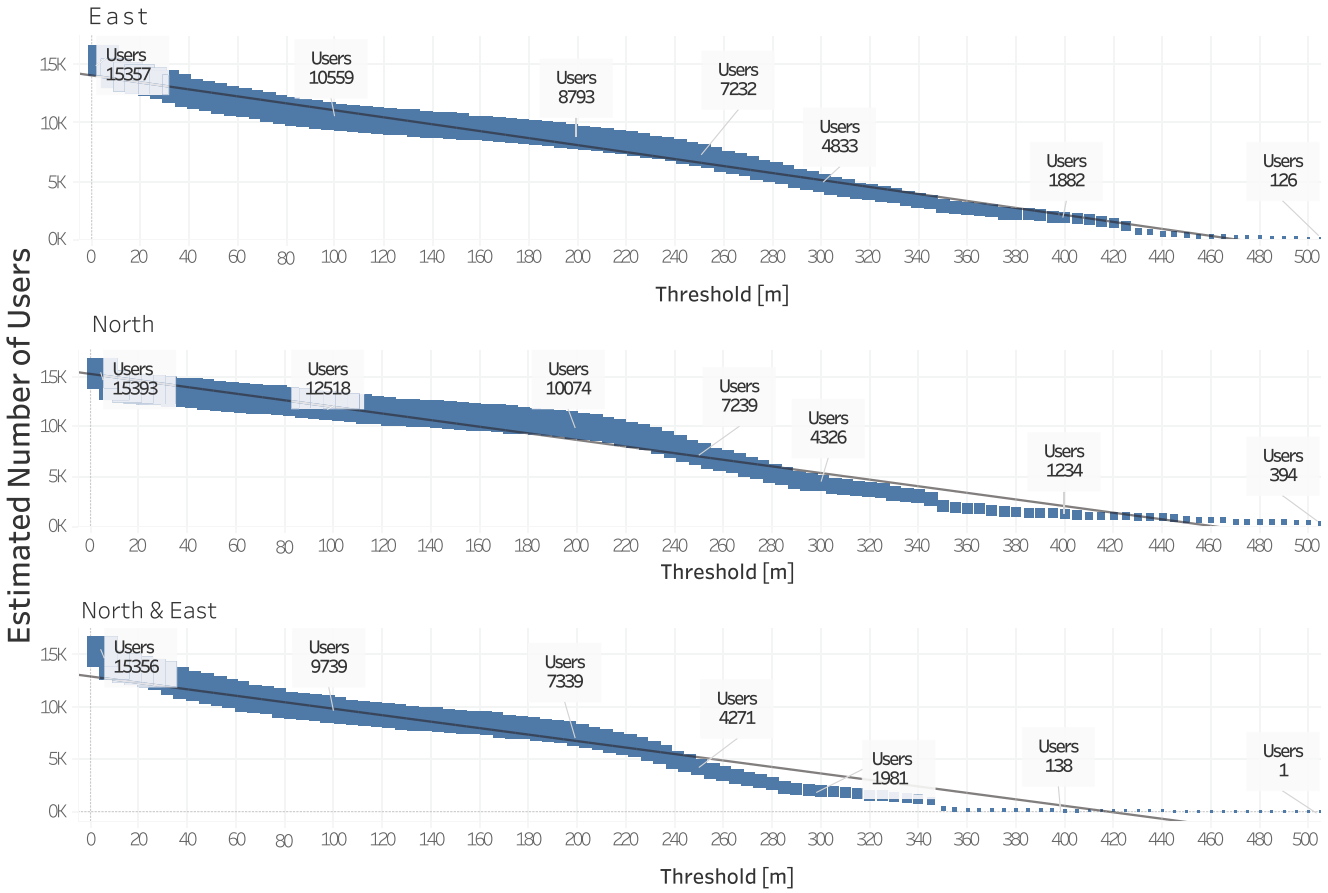

Fig. 7 Estimated number of real users as a function of position dispersion along East (upper box), North (central box), and East and North (lower box). The position of the user is computed using the WeC approach

The criteria allow the exclusion of the static devices and of the devices registered only one time. In fact, when the value of th_{pos} passes from zero to 5 m, the total number of devices reduces from 51,096 to some 13,000, for the RSSI-based and occurrence-based, and to some 15,000 for the WeC positioning.

The threshold for the screening criteria has to be set in accordance with the size of the site to monitor and to the geometry of the monitoring network. In the specific case, a threshold of 300 m (corresponding at some half of the mean distance among the nodes of the network) was adopted. From the results, it can be noted that the criteria based on the East and North components were too much stringent and the results obtained are not consistent. For example, using the proximity RSSI-based algorithm (Fig. 9), a total of 6596 real users were estimated using the criterion based on the East coordinate, whereas only 4595 were the users satisfying the condition on the North component. Finally, only 1673 devices satisfied both criteria. This relatively high inconsistency is due to the fact that the criteria do not properly represent the users motion: usually, a user does not move only along a single axes or within a square.

In order to remove the dependence from the direction, a criterion based on the horizontal component (second equation in Eq. 5) is applied; the estimated number of real users is shown in Fig. 10. The mean horizontal distance among the stations is some 600 m, hence 300 m has been identified as a suitable value for th_{pos} for the horizontal case. In correspondence of such a value, the estimated number of users varies between 5534 (WeC yellow markers) and 8233 (proximity RSSI–based red markers).

In order to complement and compare the results shown in Figs. 8, 9, and 10, a sample of the plotted data is provided in Table 3.

In order to further investigate the impact of the value of the threshold on the number of estimated real users, the cumulative distribution function (CDF) of the excluded device as a function of the horizontal position dispersion

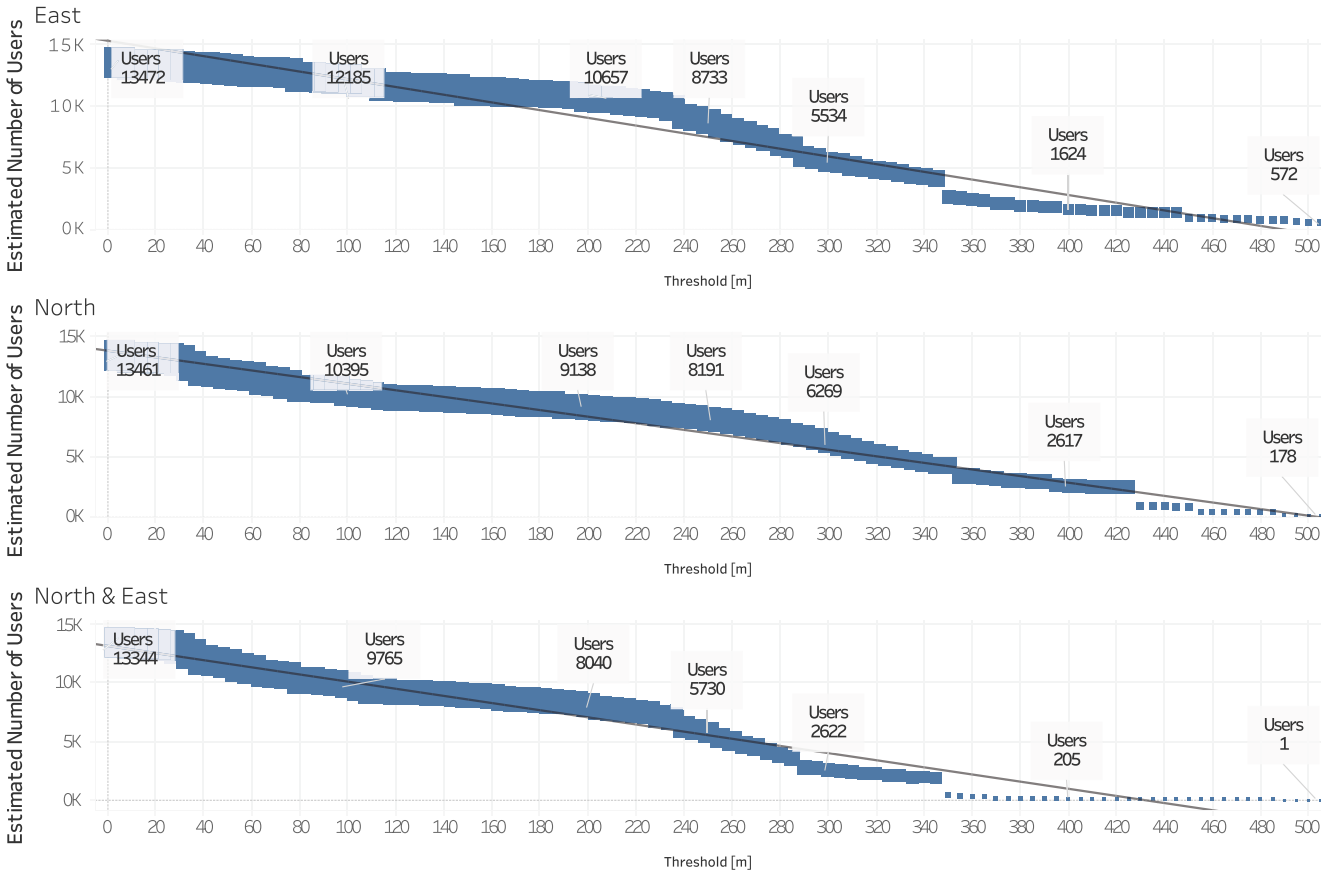

Fig. 8 Estimated number of real users as a function of position dispersion along East (upper box), North (central box), and East and North (lower box). The position of the user is computed using the proximity occurrence-based approach

threshold is shown in Fig. 11. From the figure, the impact of the threshold values clearly emerges, and considering the current value, the percentage of excluded devices is 86, 84, and 83 for the three positioning approaches. Considering the approaches based on RSSI and on the number of occurrence, some 75% of the total number of device where completely static; this percentage is a bit reduced (some 70%) for the case of the WeC. This discrepancy is due to the nature of the positioning approach used. For example, if a small oscillation in the RSSI of a device is present, the methods based on the RSSI and number of occurrence are robust with respect to this phenomenon, while the same anomaly produces small variations in the positions estimated using the WeC.

The estimated number of real users is a fundamental figure; however, it is important to verify if the different cleaning procedures identify the same devices as real users; hence, the intersection among the possible user lists, using the three different positioning methods, is computed and plotted as a function of th_{pos} in Fig. 12.

The number of users estimated using the proximity RSSI-based includes the users identified using the WeC approach; this evidence clearly emerges from the fact that the blue continuous line is almost flat and close to one (users identified using the proximity RSSI-based are also identified using the WeC) and the blue dashed line is lower than the blue one (not all the users identified using the proximity RSSI-based are included in the user list obtained using the WeC approach). However, a significant overlapping among the set of real user can be appreciated, since all the curves are higher than 50%.

User concentration and node connections

In order to evaluate the users distribution after the exclusions of the devices not classified as real users, heat maps of the user concentration are shown in Fig. 13. The heat maps are computed using the user position estimates obtained exploiting the three localization algorithms and two cleaning approaches, one based on the horizontal

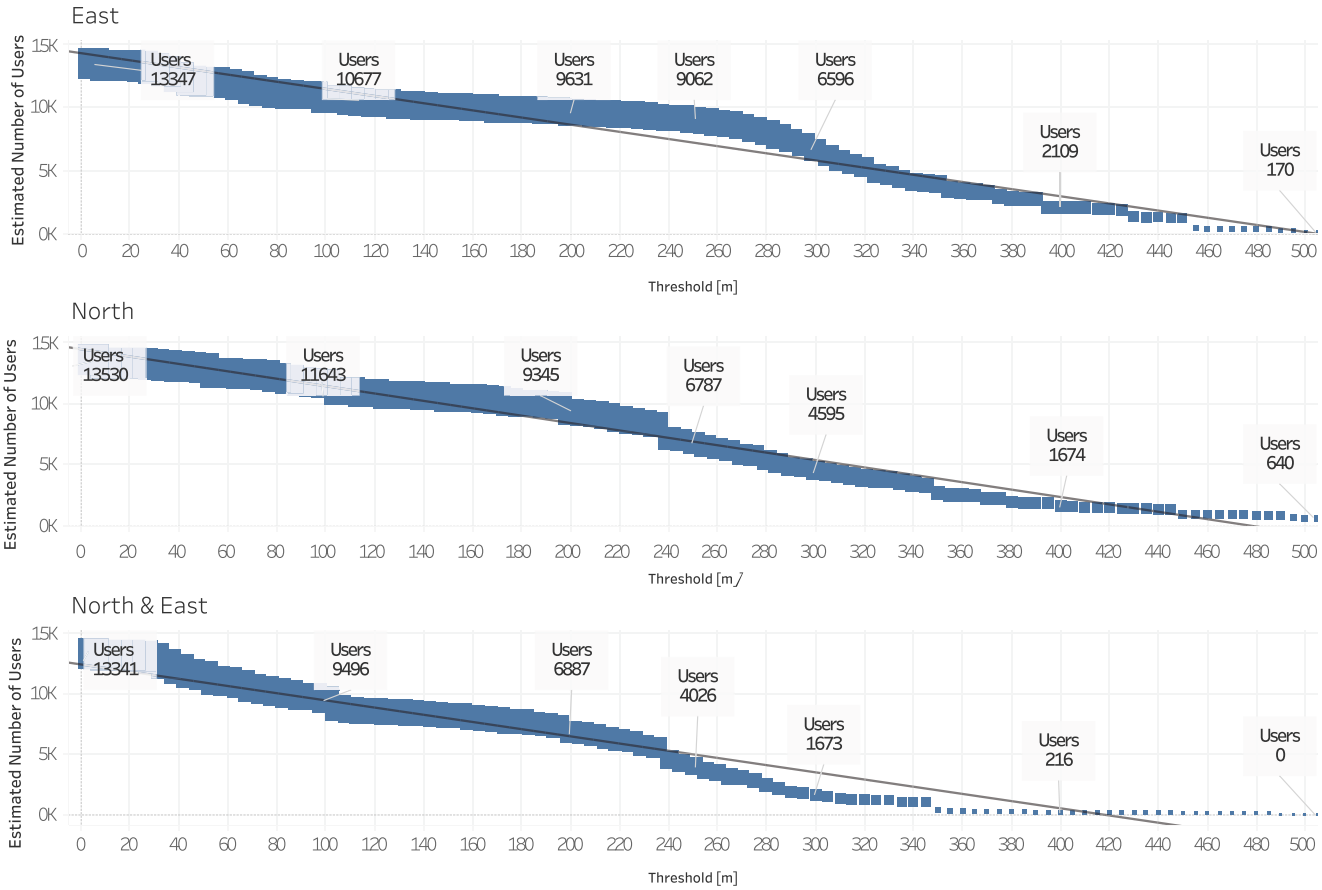

Fig. 9 Estimated number of real users as a function of position dispersion along East (upper box), North (central box), and East and North (lower box). The position of the user is computed using the proximity RSSI–based approach

position dispersion and the one obtained by combining the minimum number of base stations and the minimum number of records. The thresholds used for the heat map generation are $th_{pos,hor} = 300$, $th_{base} = 3$, and $th_{rec} = 50$. The heat map has been generated considering a time span of 1 h between 11:00 and 12:00.

Table 2 Parameter values of the linear model connecting estimated number of real users and th_{pos}

Case	m	q	R^2
East WeC	–29.81	14,060	0.99
North WeC	–33.12	15,309	0.97
N&E WeC	–30.84	12,929	0.94
Hor WeC	–31.23	15,287	0.95
East Occ	–31.23	15,286	0.95
North Occ	–27.40	13,842	0.98
N&E Occ	–30.20	13,094	0.95
Hor Occ	–27.40	13,842	0.98
East RSSI	–28.15	14,238	0.96
North RSSI	–30.06	14,408	0.97
N&E RSSI	–29.63	12,435	0.94
Hor RSSI	–22.89	14,130	0.97

The combinations of positioning methods and cleaning strategies show that the users are mainly concentrated in the central area of the site; only small differences can be appreciated considering the same localization algorithm and a different cleaning strategy (i.e., comparing the heat maps in the same columns). On the other hand, a sensible difference can be appreciated when comparing the heat map obtained using different localization algorithms (i.e., comparing the heat maps in the same rows): the user concentration is very similar using the WeC amd the proximity RSSI–based approaches, whereas a different user concentration is obtained when using the proximity occurrence-based. This is probably due to the effect of the environment on RSSI measurements.

In Fig. 14, the heat maps of the device exclusions are shown; the concentration has been calculated considering the whole duration of the event. The heat maps are built on the exclusion performed using the configurations described above. An high consistency among the heat maps can be noted: almost all the heat maps show a high number of excluded devices in the correspondence of the AP numbers

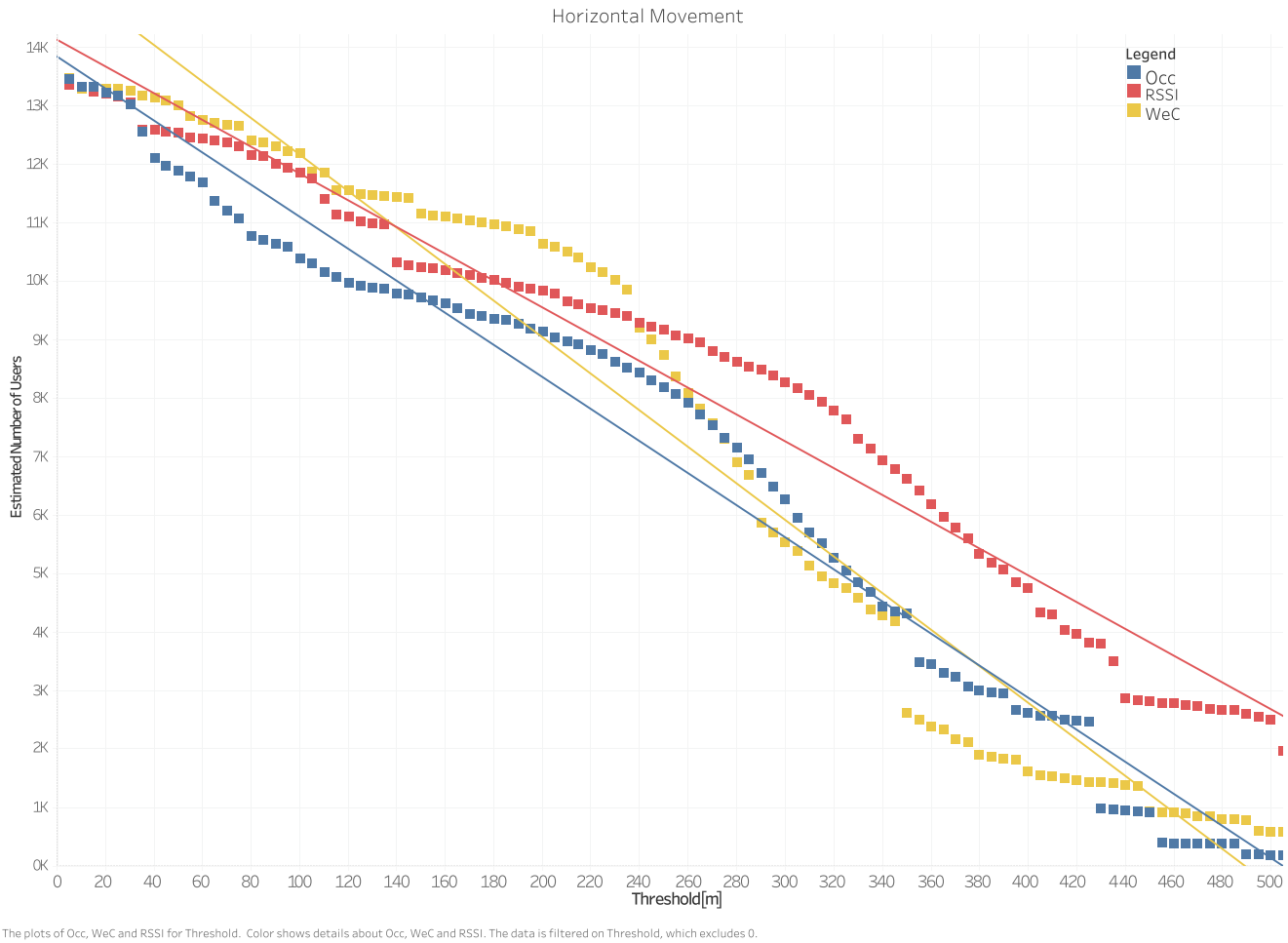

Fig. 10 Estimated number of real user as a function of the user horizontal position dispersion

Table 3 Comparison of the results in terms of estimated number of users obtained with the different methods as proposed in Figs. 8, 9, and 10

Thresh (m)	East			North			North and East		
	WeC	Occ	RSSI	WeC	Occ	RSSI	WeC	Occ	RSSI
5	15,357	13,347	13,347	15,393	13,461	13,530	15,356	13,344	13,341
10	14,217	13,324	13,324	14,190	13,333	13,256	14,036	13,281	13,241
20	13,732	13,257	13,257	13,821	13,228	13,254	13,477	13,175	13,172
50	12,198	12,016	12,016	13,166	11,898	12,821	11,640	11,564	11,504
100	10,559	10,677	10,677	12,158	10,395	11,643	9739	9765	9496
150	9724	10,032	10,032	11,097	9719	10,471	8673	8921	8231
200	8793	9631	9631	10,074	9138	9345	7339	8040	6887
250	7232	8998	8998	7239	8191	6787	4271	5730	4026
300	4833	6596	6596	4326	6269	4595	1981	2622	1673
350	2933	3928	3928	2051	4307	2656	403	473	442
400	1882	2109	2109	1234	2617	1674	138	205	216
450	455	1292	1292	656	918	1022	90	184	191
500	126	170	170	403	178	647	1	1	0
505	119	170	170	394	178	640	1	1	0

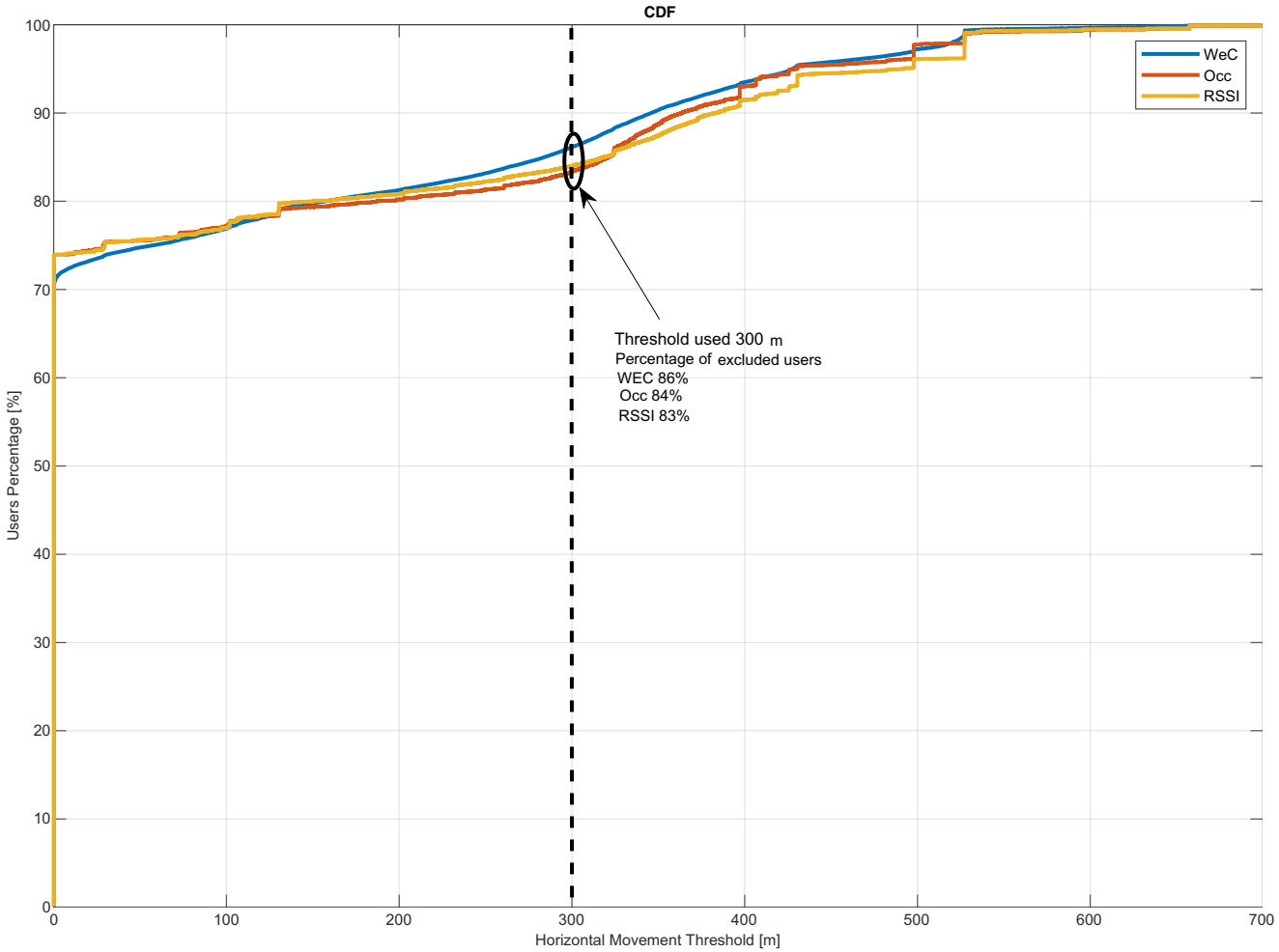

Fig. 11 CDF of the excluded device as a function of the horizontal position dispersion threshold

10 and 21, whereas only few devices are excluded in the proximity of the other APs. The most probable explanation for these results is given by the fact that, whereas AP 10 is located in the core of the administrative zone of the site (where most of the remote devices as printers etc. is present), AP 21 is closed to the west fence, which separates the site from the most busy road in the area (therefore, it is possible that part of the users registered by this AP never enters the site).

The user flow between the nodes of the network is plotted in Figs. 15 and 16. Each node is colored differently from blue to yellow; if a line connects two nodes, it means that at least one user moves from one node to the other. The color of the line is the same color of the node from which the user moved, while the width of the line represents the total number of users moving between the nodes.

Both figures are built considering 1 h of data; specifically, Fig. 15 refers to the time frame 8:00–9:00 (before the official opening) and the second figure refers to the time frame 12:00–13:00. The configurations used to build the graphs are obtained combining positioning proximity approaches with exclusion criteria based on the horizontal position dispersion and the one considering the number of stations and records. The color of the line identifies the starting node.

From Fig. 15, it emerges that real users are moving from only three nodes, specifically numbers 3, 7, and 10: presumably, the identified users were volunteers accessing the site before the official opening of the event. The first two APs (3 and 7) were located in the proximity of the gates reserved for volunteers, while AP 10 was at the main gate. The connections of the APs 3 and 10 are almost identical, because almost all the devices passing from AP 3 had to pass also in front of the AP 10. The nodes 3 and 10 are connected with almost all the nodes of the network and the curves are wider than those departing from node 7. This is due to the fact that the AP of the Brebbia gate registered only 47 users, as described in Gioia et al. (2017).

In Fig. 16, the time interval 12:00–13:00 is considered; here, all the nodes are connected, because of the presence

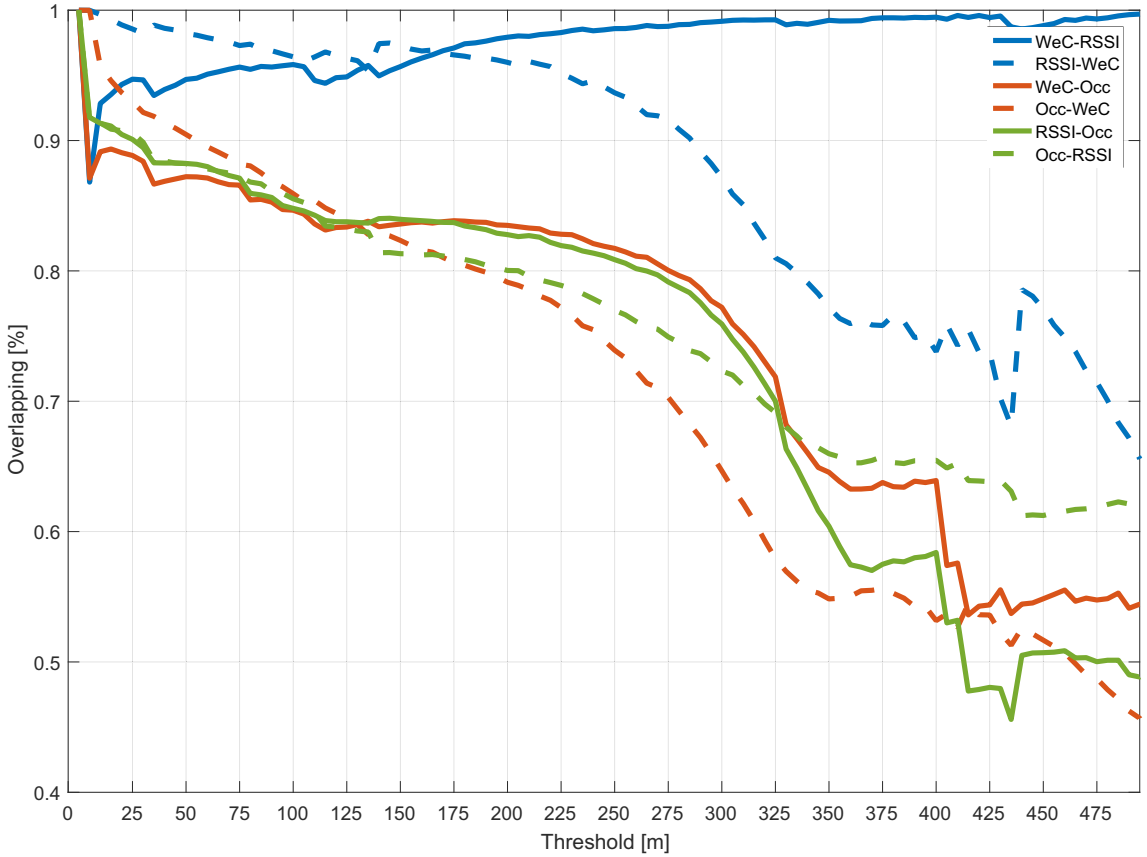

Fig. 12 Intersection among the lists of the real users obtained using the different localization algorithm and the cleaning procedure based on the horizontal position dispersion

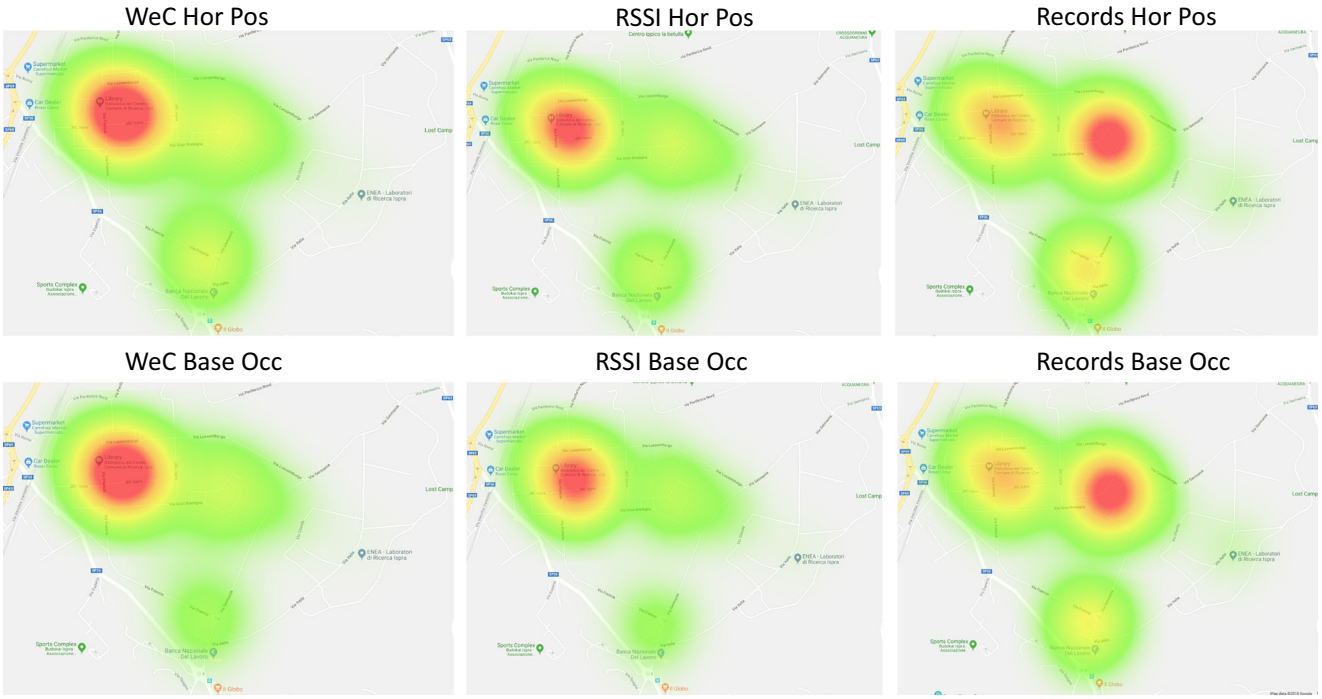

Fig. 13 Heat maps relative to the estimated users distribution between 11:00 and 12:00. The six heat maps are obtained after the exclusions of the devices not classified as real users obtained the following thresholds $th_{pos,hor} = 300$, $th_{base} = 3$, and $th_{rec} = 50$. Different positioning approaches are compared horizontally, i.e., between heat maps on the same line, whereas different cleaning strategies are compared vertically

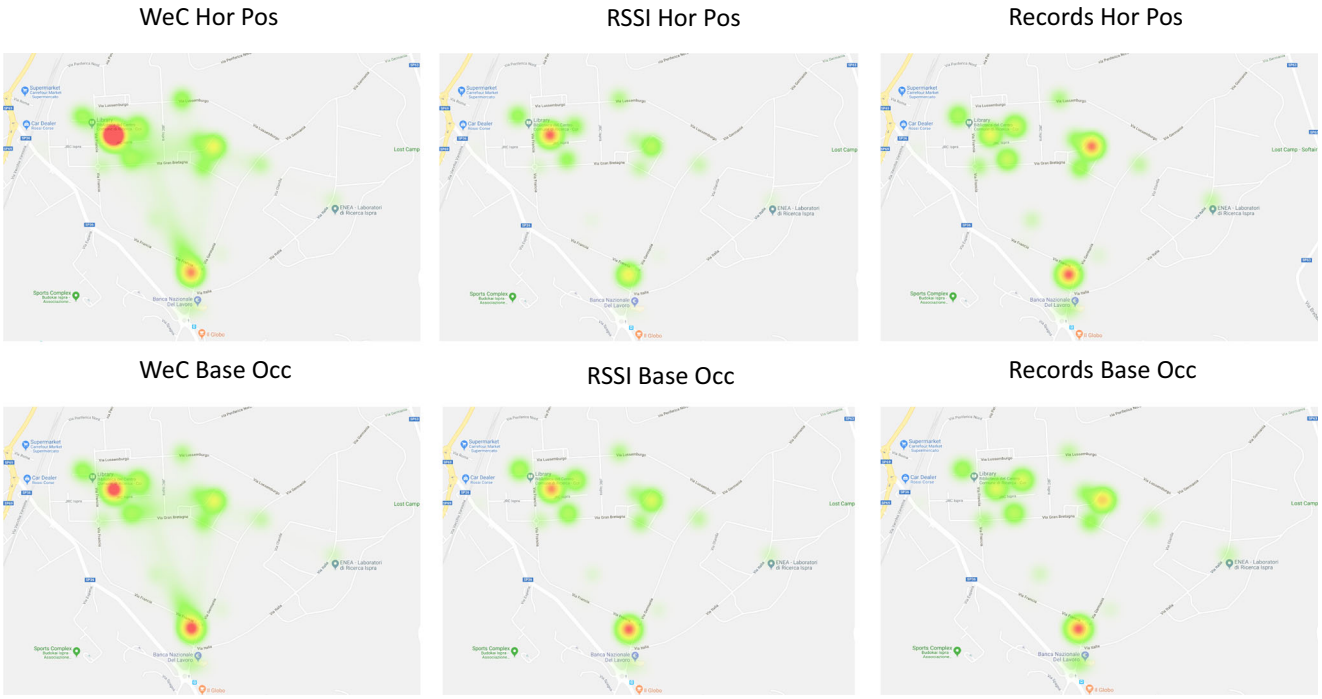

Fig. 14 Heat maps of the excluded devices. The entire duration of the event is considered here. As for the previous figure, the thresholds used are $th_{pos,hor} = 300$, $th_{base} = 3$, and $th_{rec} = 50$. Different positioning approaches are compared horizontally, i.e., between heat maps on the same line, whereas different cleaning strategies are compared vertically

of visitors moving within the site, from one exhibition to the others. For clarity of representation, only the connection starting from two nodes (the nodes with the highest number of connections) is shown. It is worth noting that the thickest lines connect nearby nodes. For example, a thick connection is visible between nodes 6 and 15, which were placed on opposite sides of the same plaza.

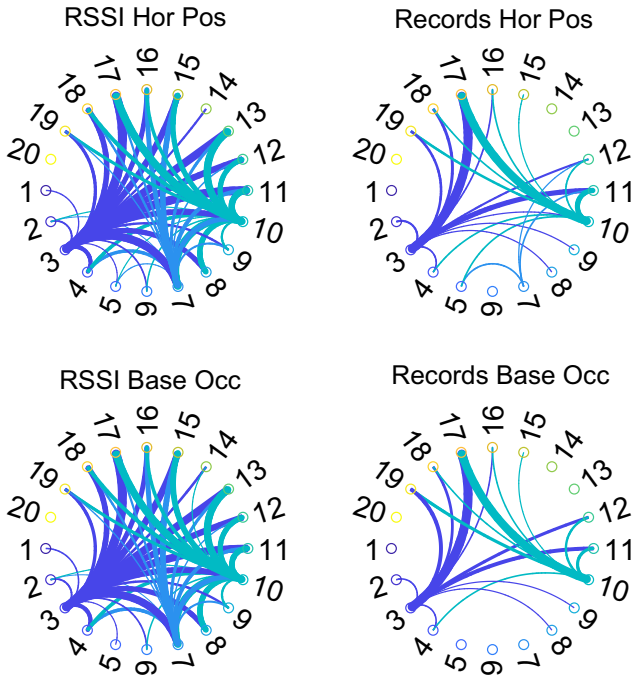

Fig. 15 User flow between the nodes of the network, before the official opening. Only APs 3, 7, and 10 registered real users. The connections of the APs 3 and 10 are almost identical because almost all the devices passing from AP 3 had to pass in front of the AP 10

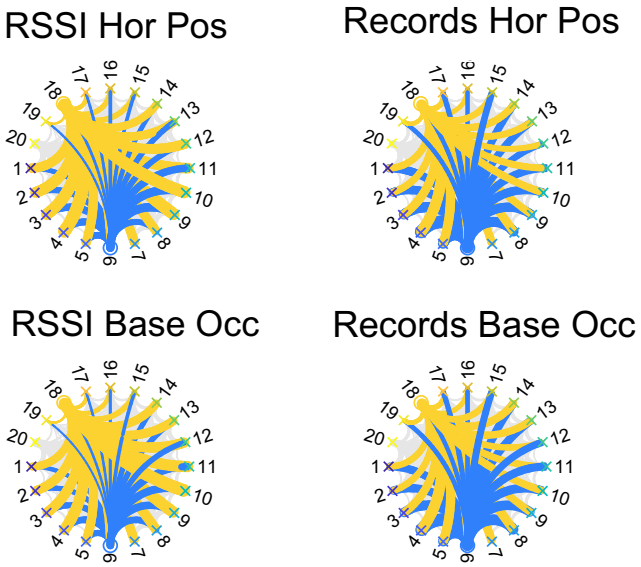

Fig. 16 User flow between the nodes of the network, after the official opening (between 12:00 and 13:00). For clarity of representation, only the connections departing from the busiest nodes (6 and 18) are represented. Thickest lines connect nearby nodes: large connection is visible between nodes 6 and 15 which were placed on opposite sides of the same plaza

Conclusions

Monitoring open crowded areas is fundamental for people security and safety, and of foremost importance is the number of people attending an event and their distribution. In this paper, a monitoring tool exploiting big data and WiFi positioning is presented. One of the issues related to this type of activity is the identification of the devices belonging to real users. This aspect is fundamental to avoid misinterpretation of results; for example, devices broadcasting fake identifiers before connecting to the network can generate false users; analogously, static devices, such as printers and PC, which continuously connect to the monitoring network, can generate a false overcrowded area, masking the real users concentration.

To fill this gap, an approach to identify real users among a set of devices is proposed. The developed approach combines seven exclusion criteria with three WiFi localization algorithms. The exclusion criteria are based on the following:

- the minimum number of APs recording a device;
- the minimum number of times a device is registered by the network of APs;
- combination of minimum number of APs and minimum number of records;
- user East-West position dispersion;
- user North-South position dispersion;
- user East-West and North-South position dispersion;
- horizontal user position dispersion.

Results discussion

The proposed localization algorithms developed are based on the proximity and WeC principles. The traditional algorithms exploit measurements collected simultaneously, but this condition is seldom verified when the tracking of an object is performed within a wide area. To fill this gap, the traditional proximity and WeC algorithms have been modified. In the proposed versions, the algorithms estimate the users positions using measurements collected during a given time interval; the time interval adopted is 3 min. The three developed localization approaches are as follows:

- proximity occurrence-based: the user is located in correspondence of the AP which recorded the user more times during the time interval;
- proximity RSSI-based: the user position is associated with that of the AP which recorded the signal of the users with the highest power, during the reference time interval;
- WeC: the user position is estimated as a linear combination of the positions of the APs recording the presence of the user in the time interval.

The screening algorithm has been tested using a unique data set, which was collected during the JRC open day 2016. More than 7500 people attended the event and almost 11 h of data was collected using 20 WiFi APs; the data set is unique because of the extension and restricted access nature of the site and of the availability of a program of the event, which allows to verify the results, at least in a qualitative way. This is usually one of the weak points of big data analysis.

Main considerations

From the results, it can be concluded that:

- In the measurement domain, the two criteria taken individually are not able to properly identify the real users; hence, their combination should be adopted. In addition, the threshold of the criteria has to be set according to the distribution of the nodes of the monitoring network and to the duration of the event.
  - For the specific case hereby analyzed, the threshold for the minimum number of station was identified as 3 (corresponding to the number of station in the proximity of the main entrance gate), while the minimum number of records for a given device was set to 50.
  - Using these thresholds, a total number of 6009 users were identified, some 75% of the actual number of people attending the event (7623 according to the report of the security office).
- In the position domain, the threshold for the screening criteria has to be set in accordance with the size of the site and with the geometry of the monitoring network.
  - In the specific case, a threshold of 300 m (corresponding at some half of the mean distance between the nodes) was adopted.
  - The criteria based on the East-West and North-South components individually were much too stringent and the results obtained are not consistent. For example, using the proximity RSSI–based algorithm, a total of 6596 real users were identified using the criterion based on the East coordinate, whereas only 4595 were the users satisfying the condition on the North component. Finally, only 1763 devices satisfied both criteria. This is due to the fact that the criteria do not properly represent the user motion; usually, a user does not move only along a single axes or within a square.
  - A possible solution is to consider the horizontal user position distribution criterion, using such screening method and the three

positioning approaches, the estimated number of real users was 5534, 6269, and 8283 for the WeC, proximity occurrence-based, and proximity RSSI-based respectively. The overestimation of the real number of users is probably due to the interference effect on the RSSI measurements.

After the screening procedure, the distribution of the users was computed. From the distribution of the excluded users, it emerges that the main part of the excluded devices was concentrated in the proximity of two APs; this result is consistent among all the approaches used. Finally, the connections among the nodes of the network were analyzed: a high consistency can be noted among the diverse methods adopted.

In conclusion, the feasibility of crowd monitoring through WiFi positioning has been demonstrated. Different cleaning strategies have been adopted and the relative results compared. A general rule, applicable to all possible WiFi positioning scenarios, cannot be stated. In fact, the particular geography of the AP network within the area to be monitored and the coverage range of the available APs are fundamental to the setting of thresholds necessary for a proper data cleaning.

Another point to be noted is that the results obtained are derived by a posteriori processing, which has been carried out on the whole data collection. Therefore, the real-time implementation of the methods still remains to be investigated. This will most probably be the direction of our future research.

References

Alessandrini A, Gioia C, Sermi F, Sofos I, Tarchi D, Vespe M (2017) WiFi positioning and Big Data to monitor flows of people on a wide scale. In: Proceedings of the European navigation conference (ENC)

Biswas J, Veloso M (2010) Wifi localization and navigation for autonomous indoor mobile robots. In: IEEE International conference on robotics and automation

Bobescu B, Alexandru M (2015) Mobile indoor positioning using WI-FI localization. Review of the Air Force Academy

Borio D, Gioia C, Baldini G (2016) Asynchronous pseudolite navigation using C/N0 measurements. J Navig 69:639–658

Cariveau D (2006) Crowd size estimation. http://course1.winona.edu/cmalone/promotion/UndergraduateResearch/Cariveau%20Report.pdf

Choi-Fitzpatrick A (2014) Drones for good: technological innovations, social movement and the state. International Affairs

Choi-Fitzpatrick A, Juskauskas T (2015) Up in the air: applying the Jacobs crowd formula to drone imagery humanitarian technology, science, systems and global impact

Dempster A (2009) QZSS's indoor messaging system GNSS friend or foe? Inside GNSS 4(1):37–40

Doig S (2009) How big will the inaugural crowd be? Do the math. Nbc news

EC-GDPR (2016) European commission EU directive 95/46/ec - the data protection directive; tech. rep., European Commission

Everit B, Skrondal A (2010) The Cambridge dictionary of statistics 4th edn. Cambridge University Press

Gioia C, Borio D (2014a) Asynchronous pseudolites and GNSS hybrid positioning. In: Proceeding of international conference on localization and GNSS (ICL-GNSS). https://doi.org/10.1109/ICL-GNSS.2014.6934165

Gioia C, Borio D (2014b) Stand-alone and hybrid positioning using asynchronous pseudolites sensors

Gioia C, Sermi F, Tarchi D, Vespe M (2017) Crowd monitoring through WiFi Data: the JRC Open Day Experiment. Coordinates

Jacobs H (1967) To count a crowd. Columbia Journalism Review

JRC Open Day (2016) JRC Open day programm

Kotaru M, Joshi K, Bharadia D, Katti S (2015) SpotFi: decimeter level localization using WiFi. ACM Special Interest Group on Data Communication (SIGCOMM)

Krewson A (2012) Estimating crowds: size matters. Columbia Journalism Review

Lamba S, Nain N (2017) Crowd monitoring and classification: a survey. Advances in Intelligent Systems and Computing

Li Teng et al (2015) Crowded scene analysis: a survey. IEEE Trans Circ Syst Vid Technol 25.3:367–386

Lohmann S (1994) The dynamics of informational cascades: the monday demonstrations in Leipzig, East Germany, 1989-91. World Politics

Maldoff G (2016) Top 10 operational impacts of the GDPR: Part 8 - Pseudonymization. International Association of Privacy Professionals (iapp)

Manandhar D, Okano K, Ishii M, Torimoto H, Kogure S, Maeda H (2008) Development of ultimate seamless positioning system based on QZSS IMES. In: Proc of the 21st international technical meeting of the satellite division of the institute of navigation (ION GNSS 2008)

McPhail C, McCarthy J (2004) Who counts and how: estimating the size of protests. Contexts

Meyer D (2018) What the GDPR will mean for companies tracking location. International Association of Privacy Professionals (iapp)

Mingkhwan A (2006) WI-FI tracker: an organization WI-FI tracking system. In: IEEE - Canadian conference on electrical and computer engineering

Nardo JD (1985) Power in numbers: the political strategy of protest and rebellion. Princeton University Press, Princeton

Oberschall A (1973) Social conflict and social movements. Prentice-Hall, Englewood Cliffs

Petre AC, Chilipirea C, Baratchi M, Dobre C, van Steen M (2017) Chapter 14 - WiFi tracking of pedestrian behavior. In: Xhafa F, Leu FY, Hung LL (eds) Smart sensors networks. Intelligent data-centric systems. Academic Press, pp 309–337, https://doi.org/10.1016/B978-0-12-809859-2.00018-8

Rabaud V, Belongie S (2006) Counting crowded moving objects. IEEE Computer Society Conference Proceedings

Rawlings J, Pantula S, Dickey D (1998) Applied regression analysis a research tool. Springer

Sirmacek B, Reinartz P (2011) Automatic crowd analysis from very high resolution satellite images. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences

Sousa R (2016) Event final numbers JRC open day

Wallace T, Parlapiano A (2017) Crowd scientists say women's march in Washington had 3 times as many people as Trump's inauguration. New York Times

Wang Y, Yang X, Zhao Y, Liu Y, Cuthbert L (2013) Bluetooth positioning using RSSI and triangulation methods. In: Proceeding of 2013 IEEE consumer communications and networking conference (CCNC)

Watson R (2011) How many were there when it mattered? Estimating the sizes of crowds. Significance

Yip P, Watson R, Han K, Lau E, Chen F, Xu Y, Xi L, Cheung D, Ip B, Liu D (2010) Estimation of the number of people in a demostration. Australian & New Zealand Journal of Statistics

Yuan Y (2014) Crowd monitoring using mobile phones. In: Sixth international conference on intelligent human-machine systems and cybernetics. Hangzhou, pp 261–264

Yuan Y, Zhao J, Qiu C, Xi W (2013) Estimating crowd density in an RF-based dynamic environment. IEEE Sensors J 13(10):3837–3845

Zebra Technologies (2015) White paper from Zebra Technologies. Analysis of iOS 8 MAC randomization on locationing

Zhan Beibei et al (2008) Crowd analysis: a survey. Mach Vis Appl 19.5–6:345–357