Analysis of crowdsensed WiFi fingerprints for indoor localization

Zhe Peng, Philipp Richter, Helena Leppäkoski, and Elena Simona Lohan
Tampere University of Technology
Tampere, Finland
zhe.peng@student.tut.fi, {philipp.richter,helena.leppakoski,elena-simona.lohan}@tut.fi

***Abstract*—Crowdsensing is more and more used nowadays for indoor localization based on Received Signal Strength (RSS) fingerprinting. It is a fast and efficient solution to maintain fingerprinting databases and to keep them up-to-date. There are however several challenges involved in crowdsensing RSS fingerprinting data, and these have been little investigated so far in the current literature. Our goal is to analyse the impact of various error sources in the crowdsensing process for the purpose of indoor localization. We rely our findings on a heavy measurement campaign involving 21 measurement devices and more than 6800 fingerprints. We show that crowdsensed databases are more robust to erroneous RSS reports than to malicious fingerprint position reports. We also evaluate the positioning accuracy achievable with crowdsensed databases in the absence of any available calibration.**

I. State-of-the-art and motivation

Mobile crowdsensing is a widespread mechanism nowadays to collect various data from users' mobile devices. Typically, this is done on a volunteer basis and via informed consent, when a user gives his/her consent of certain data to be collected from his/her mobile device when installing a certain application on that particular device. Sometimes, such information is collected automatically, some other times it requires the explicit user's feedback or his/her manual inputs.

The data collected via crowdsensing, i.e., the 'crowdsensed data', can encompass a wide variety of formats and types, ranging from data pertaining to the usage of cellular network by a certain user (e.g., time and duration of calls or sms-s, off-line or roaming durations, caller and receiver identities and/or geographical areas, etc.) to data pertaining to the user's location, such as GPS data or WiFi data.

Crowdsensing in the field of mobile positioning is already used by most location service providers and most location-based applications on mobile devices. Our paper focus is on crowdsensed data that is related to the indoor location of the users, and more specifically to WiFi data. WiFi or Wireless Local Area Networks data relevant for positioning includes some spatio-temporal stamps where data is taken, the Media Access Control (MAC) addressed of all the Access Points (AP) or WiFi transmitters in range of the mobile, and the Received Signal Strength (RSS) from each of the APs in range. The network operators and the location service providers rely on such WiFi data to build the so-called 'fingerprinting databases', which can be used later on for locating the users and offering them various location-based services. For such a localization to be possible, the data in the fingerprinting databases should contain some location stamps (e.g., latitude-longitude-altitude coordinates) or geographical labels (e.g. Room 212 in Building A, or address B), that would allow a clear identification of the geographical spot where the user is located at a particular time. The fingerprinting principle in positioning relies on a comparison between the current user data and the data already stored in the fingerprinting database until the best match (or fingerprint) is found. The comparison can be done based on various metrics, such as Euclidean distances, Mahalanobis distances, rank-based metrics, among others [1]. Good overviews of fingerprinting can be found for example in [2].

However, research studies on the quality and robustness of such data collected for positioning purposes are still rare in the present literature, especially for indoor scenarios, to the best of the authors' knowledge.

The authors in [3] described an algorithm to to build accurate WiFi radio signal map with heterogeneous devices in outdoor environments. Their measurements relied on three different Android smartphones in a $5000\,\mathrm{m}^2$ outdoor urban area. Our approach is different in the sense that we concentrate on larger indoor spaces and more heterogeneity in devices (we did measurements in an area of about $22{,}500\,\mathrm{m}^2$ and in a 5-floor university building with 21 different Android devices) and we look not only at the RSS statistics and at the relationship between the measurements reported by different devices, but also at the cumulative distribution functions of the positioning errors.

The studies in [4] also deal with crowdsensing WiFi data for positioning but they focus only on the privacy aspects and propose compressed sensing methods of data collection which preserve the privacy of the users' traces. Thus, our work is complementary to the studies in [4], as the privacy aspects are not considered here, but we focus on the statistical characteristics of the crowdsensed data and on the impact on the positioning accuracy of the heterogeneity of crowdsensing devices and software and of the presence of interference in the collected data.

Another study related to crowdsensed fingerprints was presented in [5] and it focused on how to detect rogue or malicious AP attacks in a crowdsensed database. We also look in here at the impact of malicious APs on the crowdsensed data, but our approach is very different from the one in [5].

Because we look into both, the changes in the reported RSS and in the reported 3D locations and, compared to [5], we use different distributions of the malicious access nodes.

The authors of [6] described the development and use of a crowdsourced WLAN fingerprinting system over the course of a year. They focus on the creation and temporal evolution of their fingerprinting database, which was created by over 200 users and which contains more than 8700 measurements on over 300,000 square feet. They also evaluated briefly the positioning performance, but positioning performance issues due to the quality of the radio map or the use of different devices were not further explored.

Park et al. [7] developed a large, crowdsourced WiFi fingerprinting positioning systems for a multi-floor area with 1373 map spaces (rooms or similar spatial partitions) with more than 100,000 fingerprints. They studied the positioning accuracy during the creation of the database for a nine day period and used RSS clustering to detect and mitigate erroneous fingerprint location inputs. The issue of erroneously reported RSS values was not considered. In [8] they investigated the influence of device heterogeneity on the localization accuracy. They found a linear correlation of RSS between devices and showed that a linear transformation is not enough to compensate for it, but that a wide smoothing of the RSS improves localization across devices. Our methodology to investigate device heterogeneity relies on empirical cumulative distribution functions (CDF) and power maps.

Also Laoudias et al. [9] use CDFs of RSS, with the objective to compensate for the RSS differences reported by different devices. This study is complementary to their work, as our analysis of RSS of different devices focuses on the quality of the fingerprinting database. Additionally, we compare a larger variety of devices and base our findings on more measurement.

Crowdsourcing for indoor localization has also been studied in [10]. The measurement space for the studies in [10] was limited to a single-floor building of about $1600\,\mathrm{m}^2$ area, where only 26 AP were present. In our studies, we rely on a much larger space (as mentioned above) and we detected a total of 992 APs. One AP means here one MAC address, with the observation that several MAC addresses can come from the same physical location of a WiFi transmitter, due for example to multiple BSSID addresses or to multiple-antennas of the WiFi transmitters.

To summarize the discussion so far, there are several open questions regarding the quality and robustness of the crowdsensed WiFi fingerprints for indoor positioning, such as:

- What is the distribution of the RSSs collected via crowdsensing (e.g., many users and many devices reporting measurements) and how this distribution compares with the case when RSS is collected via a single device (e.g., dedicated data collection from people specifically hired by the location service provider)
- What is the accuracy of a power map constructed with multiple devices, compared with a power map built with a single device?
- How much the positioning error is influenced by the presence of malicious data in the database, such as coming from devices which report wrong locations or labels?
- Can we rely on heterogeneous fingerprinting (i.e., fingerprinting based on multiple devices with different hardware and software versions) to achieve accurate and robust positioning estimates?
- How to deal with intentional or unintentional faulty RSS measurements?

This paper aims to address most of the above-mentioned questions. The novelty of our papers is two folds: first, we present our methodological approach to investigate such research questions, and secondly, we present novel statistics regarding the crowdsensed fingerprinting WiFi data and the accuracy of the indoor location estimation based on such crowdsensed data. In addition, our results are based on a large database of fingerprints collected with heterogeneous devices and softwares, in a multi-floor multi-room building where close to one thousand APs were detected.

The rest of the paper is organized as follows. Section II describes the main procedures and challenges in collecting WiFi fingerprints in a crowdsensed mode. Section III presents our methodology, from data collection to data analysis. Section Section V summarizes the ideas and presents the conclusions.

II. Crowdsensing WiFi fingerprints in indoor spaces

People spend most of their times indoors. Indoor positioning applications, such as fast navigating inside a commuting hall or a hospital or finding promptly a free parking place in a shopping mall, are therefore of high interest for business and service providers as well as for city planners and urban councils. While satellite-based navigation offers nowadays good availability and reliability of positioning and tracking solutions outdoors, navigating in indoor environments is still a challenging issue.

Most indoor navigation solutions nowadays rely on some form of RSS-based fingerprinting, typically based on WiFi signals or other signals available indoors, such as BLE, RFID or cellular. The two biggest challenges, in authors' opinion, of RSS-based fingerprinting indoors are to get access to indoor maps (which are often proprietary, inaccurate or in a format difficult or tedious to convert in a digital application) and to be able to build a dynamic and up-to-date fingerprint database. The scarcity of indoor maps and solutions to overcome this challenge are outside the scope of this paper. Interested readers are referred to [19] and [20] to read more on this issue. The well-known solution to the second challenge, the one of building a dynamic and continuously available fingerprint database is to rely on crowdsensed information from volunteer users. A large scale collection of such a data (e.g., at country or continent level) poses the additional challenge of data transfer and storage. Solutions to these problems have been addressed in [16], [17] for example.

The other challenges related to crowdsourced indoor data for positioning are related to the fact that users employ different devices and software to report such measurements and the impact of this inherent heterogeneity of devices on the database accuracy and positioning achievable accuracy are still poorly understood.

The concept of crowdsensing WiFi fingerprints is illustrated in Fig. 1. Volunteer users located at various floors inside a building have a fingerprint application on their mobile devices including the building map. They report their location inside the building and the information about the AP in range, namely their MAC addresses and their RSS values, to a server. The location information can be a label, like a room number, or coordinates, for example $x_f^{(u)}, y_f^{(u)}, z_f^{(u)}$, with $u = 1, \ldots, U$, U being the total number of volunteer users, and $f = 1, \ldots, F^{(u)}$, $F^{(u)}$ being the total number of fingerprints collected by u-th user. In Fig. 1, $RSS_{f,a}^{(u)}$ is the RSS value measured by u-th user, in the f-th fingerprint, from the a-th AP, $a = 1, \ldots, N$, where N denotes the total number of AP in that particular building. The information collected by the server is then stored in a database of fingerprints, called crowdsensed- or training database. The training database will thus be able to create a power map for each AP in the building, as shown in Fig. 1. Some of those power maps will have many points (e.g., the top power map in the figure, which shows that AP was heard by many users) and some other will have fewer points (e.g, the middle power map in the figure, which probably corresponds to the case when the AP was only heard in a certain part of the building, where not many users have access)

III. Our methodology

Two proprietary Android software applications developed within various projects of the authors were used to collect the data used in our analysis. Both applications are sensing the environment and look for all the APs in range. The MAC and RSS values of each AP in range is stored with a time-stamp value. One of the Android applications (Android app 1), used for the crowdsensed data, relies on a cloud service to report manually the location on a map and it is less accurate than the second one, as the indoor maps included in it are of lower resolution (e.g., errors about 1-3 m are expected in the users' feedback). In the first Android application, only point values are allowed, i.e., the position needs to be input manually at every point. A total of 4648 fingerprints were collected with the first application and a total of 2220 fingerprints were collected with the second application.

The second Android application (Android app 2), which is more accurate, relies only on the storage space of the device where it is installed. The users can select very accurately (down to 0.5 m accuracy) their position on the map, and they can measure a full track (i.e., several points in a row) at a time, just by moving from one location to another, in a straight line. The positions in-between the start and end point are interpolated linearly. The main reason to use two different applications was to attain a heterogeneity of software, which is one of the main assumptions here. The heterogeneity of devices has been achieved by installing these applications on several Android devices. To better verify the heterogeneity assumption, we also took the following approach: the data collected with the first Android app and 21 heterogeneous devices was used as training data, and the data collected with the second application and 3 heterogeneous devices was used as test data. The number of measurements per device is illustrated in Fig. 2.

The measurements were taken in the 5 upper floors of a 6-floor university building in Tampere, Finland (the basement floor was not accessible). Compared to existing studies in the current literature, our measurement space is a huge space, with a large surface and multi-floor multi-room environment. The measurement environment is summarized as follows:

Nr. of AP:	992	Nr. of rooms:	882
Nr. of fingerprints:	6868	Nr. of inner walls:	435
Building area:	22,500 m^2	Nr. of outer walls:	86
Number of floors:	6	Number of floors: with measurements	5

The position estimation was based on a log-Gaussian likelihood method [18]. Let's denote by O_a the observed RSS value from the a-th AP in the estimation track. This value is compared to the training dataset fingerprint by fingerprint, according to the following log-Gaussian metric:

$$G_{a,f}^{(u)} = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \left(-\frac{O_a - RSS_{f,a}^{(u)}}{2\sigma^2}\right) \qquad (1)$$

where f is the fingerprint index in the training dataset and σ is a constant value standing for the shadowing standard deviation, which can either be fixed in the algorithm, or pre-computed based on the training data (in our studies we used a fixed value $\sigma = 7$ dB). The above metric is computed for all those access points commonly heard in the observation sample and the training sample. An overall log-Gaussian metric is then formed by combining the metrics from all commonly heard access points and all the crowdsensed training data from various users:

$$G_f = \sum_a \sum_u G_{a,f}^{(u)}, \qquad f = 1, \ldots, F \qquad (2)$$

with $F = \sum_u F^{(u)}$. These G_f metrics are then sorted from highest to smallest, and the k highest metrics are then selected together with their corresponding locations. The final position estimate is obtained as the average over those k locations (in our paper $k = 3$).

IV. Fingerprints analysis

A. RSS distributions

To assess the different modes of creation of a fingerprinting database, we compare the fingerprints collected by users in crowdsensed mode with the fingerprints recorded systematically by a trained person. We follow the approach in [9] and combine the RSS of all fingerprints that were collected by the

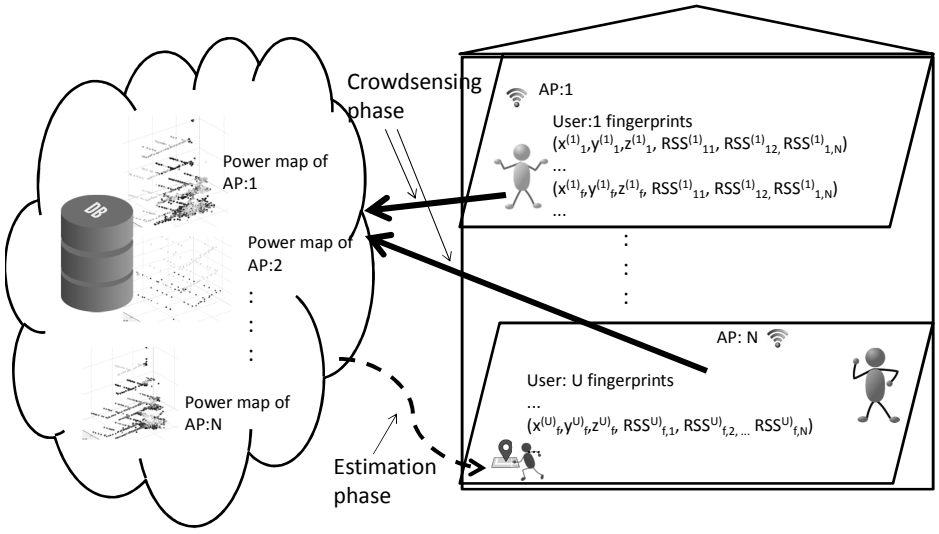

Fig. 1. Crowdsensing principle.

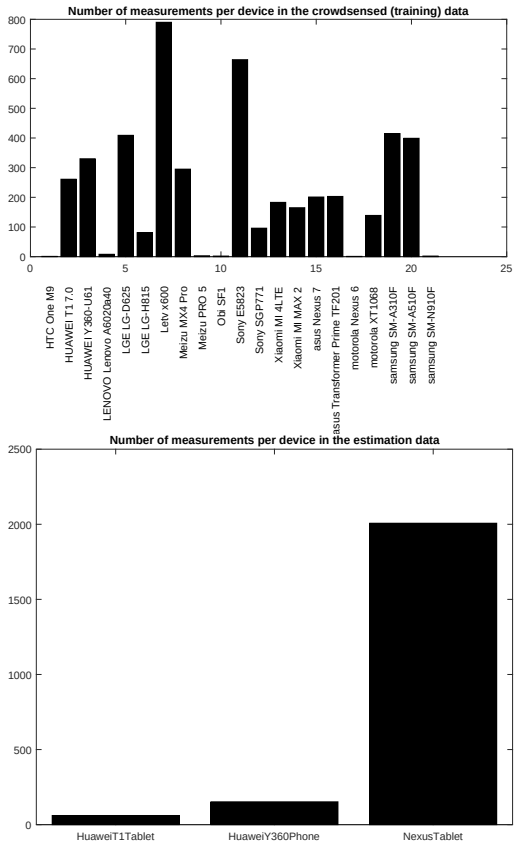

Fig. 2. Number of measurements per device in the training (upper) and estimation (lower) datasets.

corresponding user(s)/device(s), irrespective of its later use as training or test data and irrespective of the AP from which the signals were emitted. Fig. 3 shows the CDF of RSS, (i) collected systematically by a trained person, with Android app 2, covering the whole building (Huawei Y360, #FP 2508), (ii) collected uncoordinated via crowdsourcing by two different users, with Android app 1, (LGE LG-H815, #FP 81; motorola XT1068, #FP 139), (iii) collected via crowdsourcing, with Android app 1, but combining the measurements from all users (all devices, #FP 4648). The CDF of RSS collected by a

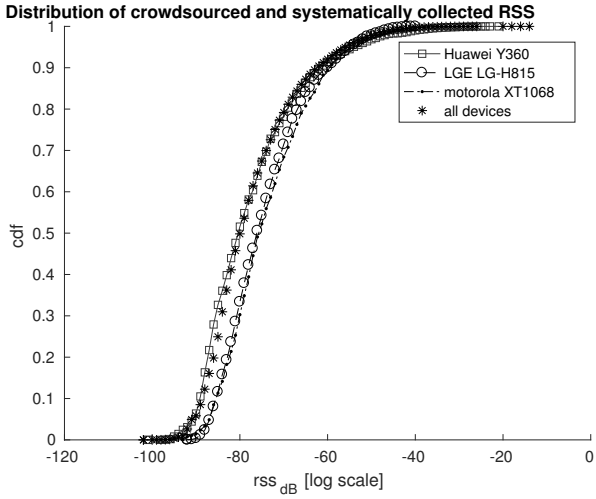

Fig. 3. CDF of RSS collected in a systematic manner or by crowdsourcing.

trained person with the Huawei Y360, which covers the whole building, resembles the CDF of RSS collected by volunteer users (LG-H815 and motorola XT1068), even though their fingerprint coverage is rather small, c.f. [9]. The range of RSS from the Huawei Y360 is larger than the range from the LG-H815 and motorola XT1068. However, this range depends on many factors, especially the dynamic range of the wireless network interface card (NIC) and also the spatial distribution of the fingerprints and their distances to the APs. We found a median RSS difference comparing RSS of all devices, either systematically collected or crowdsourced RSS, of 3.5 dB in average and of 7 dB maximum. This is the same order of magnitude as the standard deviation of RSS, thus, linear compensation to improve positioning, as proposed by many studies (see [11], [9] and references therein), is likely to be insufficient [8] unless the RSS are smoothed. Considering the CDF of the crowdsourced, multi-device training database, that is the CDF of RSS from all 21 devices, one observes that it is almost identical to the CDF of RSS from the single devices that covered the whole building systematically. That suggests that creating fingerprinting training databases by crowdsourcing is an appropriate replacement for dedicated data collection by professionals, as long sufficient data is collected.

The RSS values shown in Fig. 3 originate from different APs and they were sampled at different time instances at different locations. In addition, RSS are non-stationary in general, both temporally [12] and also spatially, as suggested by [13]. Temporal stationarity can be assumed for shorter time lags (shorter than changes in the environment due to e.g. people or actual changes of the environment). Studies about spatial stationarity of RSS are lacking, nevertheless, we hypothesize that RSS are stationary over spaces with a homogeneous transmission medium. For modelling purposes, the intrinsic hypothesis (a weaker form of second-order stationarity) is commonly assumed [13], [14], [15]. Thus, the RSSs used in Fig. 3 exhibit different statistics, or even stem from different random processes. Computing the CDF of RSSs measured at different time instances, at different fingerprint locations is therefore inconsistent. Nevertheless, it provides a simple method to visualize certain features of a training database in a relative manner, such as device heterogeneity, the amount of fingerprints or the coverage. The RSS CDF does not allow conclusions about the spatial extend of the fingerprints, i.e. fingerprints could be dense in a local region or rather distributed sparsely in the whole area or about the influence of the wireless NIC.

B. Power maps differences

The differences between the power maps reported by different datasets (training versus estimation) and different devices have been analysed in two different manners. The first approach is illustrated in Figures 4 and 5 and consisted of a floor-by-floor visual investigations of the power maps under various circumstances. For example, if we look at the AP heard in most of the crowdsensed measurements (here AP $\#492$) and we compare the power maps obtained in the training and in the estimation data, we see the results from Fig. 4 (only second floor is shown here for clarity, but similar results have been obtained across the floors). We first notice that the coverage areas of this AP is slightly different in the two datasets, no doubt due to different devices and different users reporting the measurements. A second observation is that the strongest power level (shown in white in the plot) happens in similar regions in both plots of Fig. 4, i.e., around $x = 90\,\text{m}$ and $y = 20\,\text{m}$. A third observation, by looking at the colour bar from Fig. 4 is that the range of reported RSS values is also slightly different in the training and estimation datasets, as well as the RSS fluctuations with the location of the fingerprint. This can be explained by the effect of shadowing, which obeys spatio-temporal variations, as described for example in [16].

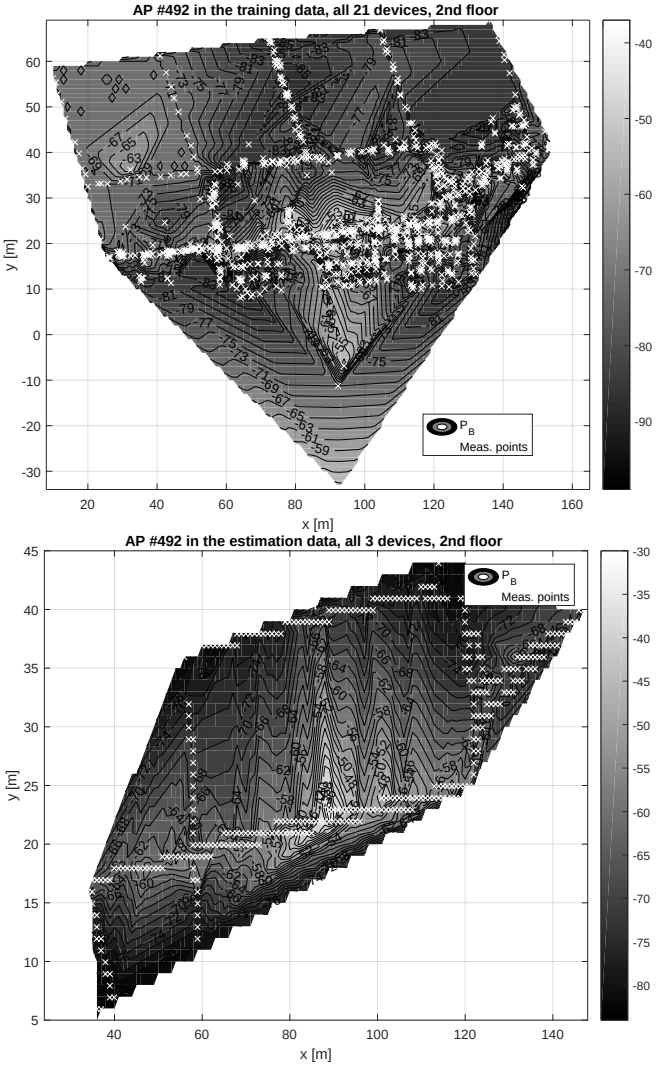

Fig. 4. Example of the differences of a power map between the crowdsensed training data and the estimation data. Top: training data; bottom: estimation data.

Fig. 5 does a similar comparison for the same AP (#492) between different devices used to collect the crowdsensed data. Two devices with high number of measurements per device

were selected for the comparison: Sony E5823 device, which reported 664 measurements and Letv-x600 device, which reported 790 measurements. Similar observations as above also hold here: different devices have different coverage areas, different RSS ranges and different shadowing profiles, but the strongest reported RSS values happen in similar regions for both devices (also around $x = 90\,\text{m}$ and $y = 20\,\text{m}$ point).

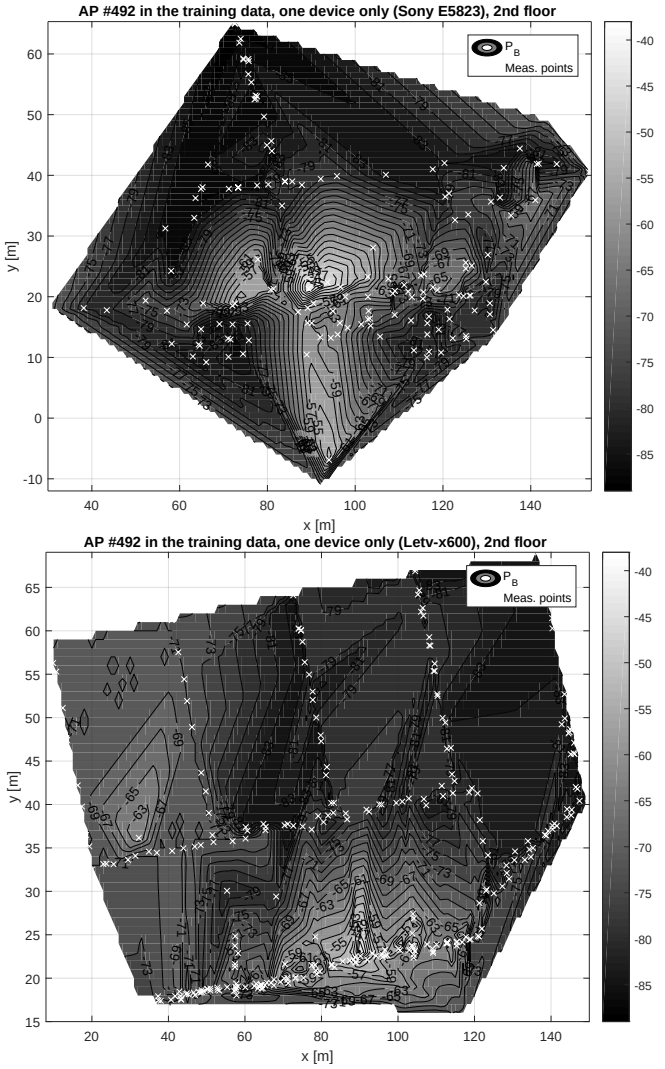

Fig. 5. Example of power maps provided by two different devices in the crowdsensed training data. Top: Sony E5823 device; bottom: Letv-x600 device.

The second approach to compare the power maps was to look at the distribution of the power map differences. This was done by selecting the same floor and the same AP (one by one) from different datasets, building an interpolated and extrapolated power map for it to cover exactly the same spatial area and then looking at the histogram of the power map differences (in dB scale) and comparing it with 11 theoretical distributions, namely: Gaussian, Exponential, Log-normal, Extreme value, Rayleigh, Gamma, Weibull, Logistic, Burr type XII, and Generalized extreme value distributions. An example of the power map differences between Sony E5823 and Letv-x600 devices is illustrated in Fig. 6.

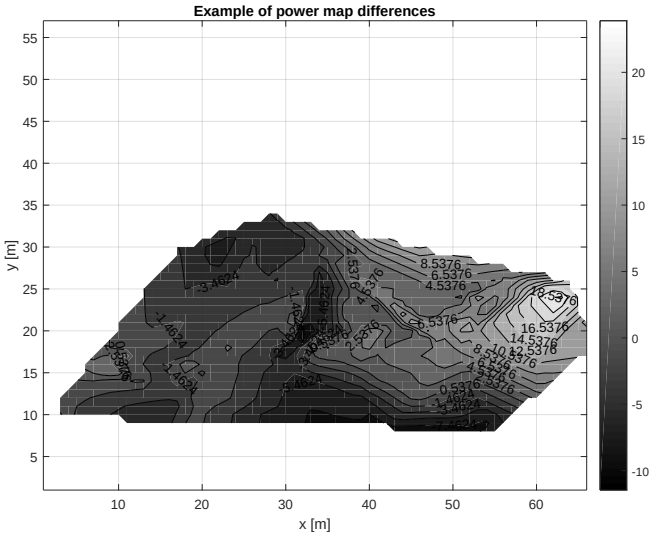

Fig. 6. Example of the differences of the power maps between two different devices in the crowdsensed training data (Sony E5823 power map minus Letv-x600 power map for one AP).

The best-fit distribution among the tested ones proved to be Burr distribution, as shown in Table I and example of the measurement histogram and best-fit distribution is given in Fig. 7. The power differences would ideally be constant

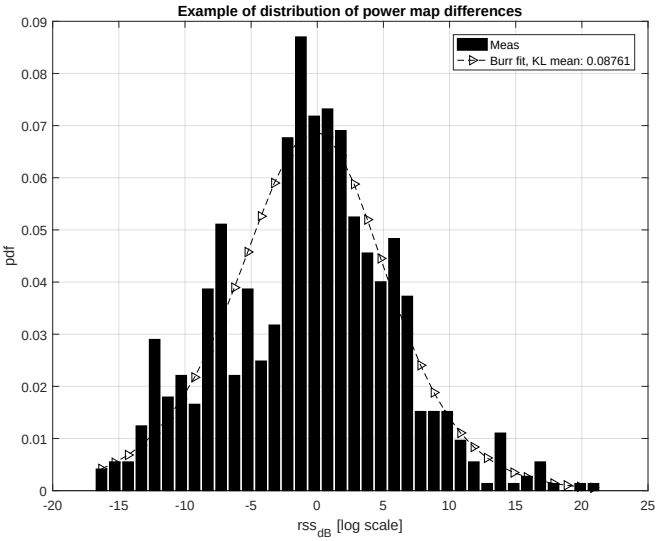

Fig. 7. Example of the distribution of power map differences between two datasets (Sony E5823 versus Letv-x600 datasets).

according the approximately linear relation of RSS of different devices, but as RSS are subject to noise, which is reflected in this distribution. The distribution is symmetric, around zero and heavy tailed. RSS differences do not obeys a Gaussian distribution, which means that calibration between different devices could be difficult. That is why, our next analysis about the achievable positioning accuracy with heterogeneous

TABLE I
BEST-FIT DISTRIBUTION (ON AVERAGE) FOR POWER MAP DIFFERENCES

Compared datasets	**Best-fit distribution**	**Distribution main parameters (average values)**
Crowdsensed (Android app 1) and Huawei T1 (Android app 2)	Burr	$a = 4.65;\ \ c = 1.37;\ \ k = 1.50$
Crowdsensed (Android app 1) and Nexus (Android app 2)	Burr	$a = 13.41;\ \ c = 1.43;\ \ k = 1.88$
Sony E523 (Android app 1) and Letv-x600 (Android app 1)	Burr	$a = 2.00;\ \ c = 1.91;\ \ k = 2.35$
Nexus (Android app 2) and Huawei T1 (Android app 2)	Burr	$a = 9.51;\ \ c = 1.81;\ \ k = 1.44$

devices, in the absence of calibration, is an important steps towards a better understanding of crowdsensing in positioning.

C. Positioning cumulative distribution functions

Fig. 8 shows the CDFs of positioning errors with crowdsensed training data and in the absence of calibration. To better

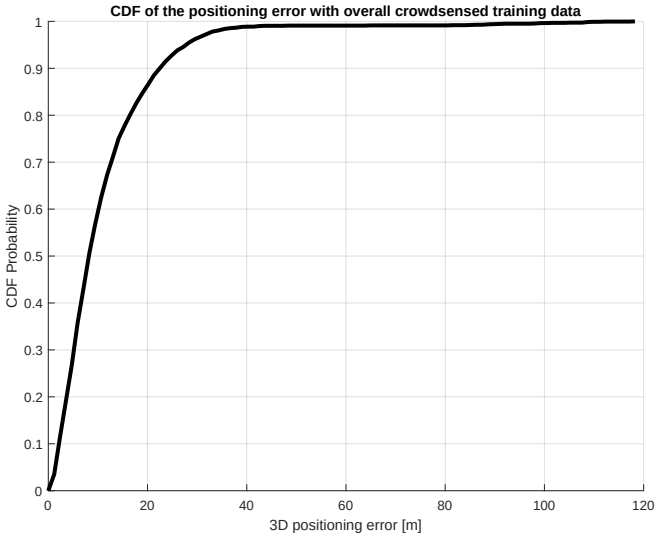

Fig. 8. Cumulative distribution function of the positioning error with crowdsensed training data.

understand the impact of heterogeneity of the devices on the positioning accuracy, Fig. 9 shows the CDF of the positioning error per device, i.e., when a single device is used from the available training database. Of course, the number of training points plays an important role in the results. Devices with low number of training data points give very low accuracy, while devices with a high number of training data points have a higher accuracy. However, a simple increase in the number of training data points does not insure an increased positioning accuracy, as we can see if we compare the best plot in Fig. 9, where an accuracy of less than 10 m error in achieved in almost 90 % of cases, with the plot in Fig. 8, where less than 10 m error accuracy is achieved in less than 70 % of cases. This basically points out the fact that there is an inherent deterioration in the positioning accuracy when several heterogeneous devices and software are used, compared to the case of a single-device single-software approach for data collection. One could further investigate calibration to alleviate parts of this problem, but this is out the scope of our paper.

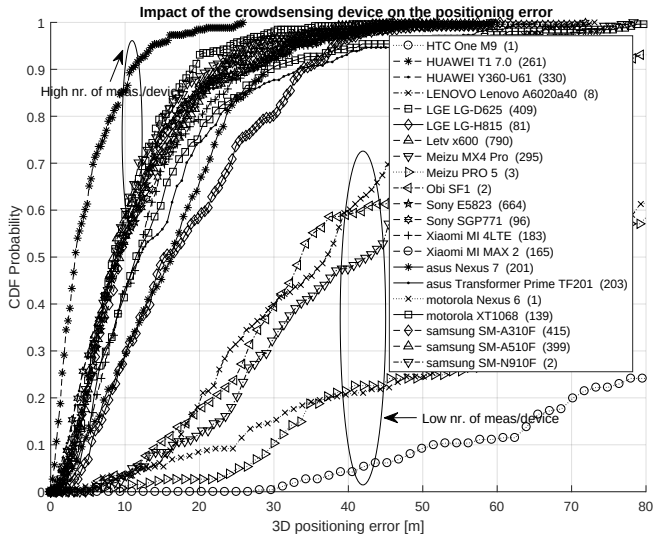

Fig. 9. Impact of the crowdsensing device type and the number of measurements per device on the positioning error.

D. Intentional database deterioration

In this section we analyse the impact of an intentional database deterioration by a certain percentage of the crowdsensing devices. Two case studies were investigated:

- Incorrectly reported positions
- Incorrectly reported RSS values

1) Incorrectly reported positions: To simulate erroneously reported fingerprint locations, we modified the floor number and the coordinate of a fingerprint position: First, to create the floor error, the original data's floor number is changed randomly to another floor number. For example, if one point is measured at the 2nd floor, it will be randomly changed to 1st, 3rd or 4th floor. Secondly, to modify the coordinates (x, y), the maximum and minimum values of all x and y values in the training database are computed. Then the mid coordinates for x and y coordinates are computed and the points assumed to be affected by malicious intent are then modified in symmetry to the mid point.

Fig. 10 shows the CDF of the positioning accuracy when various percentages of the crowdsensed data are affected by a positioning error. These percentages range from 0 % (no error in the database) to 100 % (all database is erroneous).

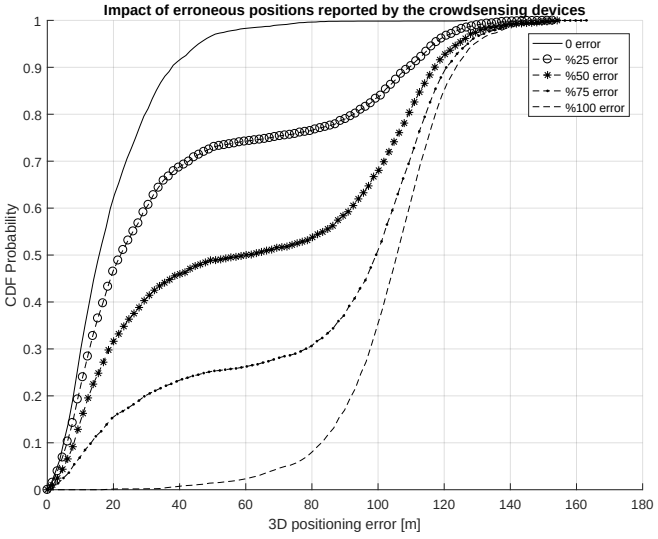

Fig. 10. Impact of the incorrectly reported locations by the crowdsensing devices on the positioning error.

2) Incorrectly reported RSS values: The approach to model incorrectly reported RSS values was the following one: in randomly selected measurements points (according to a random percentage varying from 0 % to 100 %) we assume that all the RSS data received from the AP in range has a fake or bogus value a. Two cases were studied: a small a value, when $a = -90\,\text{dB}$, and a high a value, when $a = -40\,\text{dB}$. This modelling corresponds to a situation when there are users reporting fake or garbage data into the database, e.g., by tampering with the application that collects RSS data (malicious intent) or by simply having a faulty device, which is not able to compute correctly the RSS values. The results are shown in Fig. 11. Surprisingly, the impact of the RSS errors is rather small on the positioning accuracy. This apparent contradiction can be explained by the fact that, even if the RSS values are reported incorrectly, the ID of the AP in range is reported correctly. Thus, the log-Gaussian likelihood algorithm of eq. (2) falls back into a rank-based algorithm, which relies on the commonly heard access points in the training and estimation data. It turns out that by simply knowing the APs heard in a certain point (even without their correct RSS values) provides already enough information to be able to locate the mobile user inside a building.

E. Interchanging the training and estimation databases

Fig. 12 shows what happens if we interchange the training and estimation data: one curve shows the results with the crowdsensed training data collected with the first Android application and 2220 estimation points collected with the second Android application; the second curve in this figure shows the results when the data collected with the second

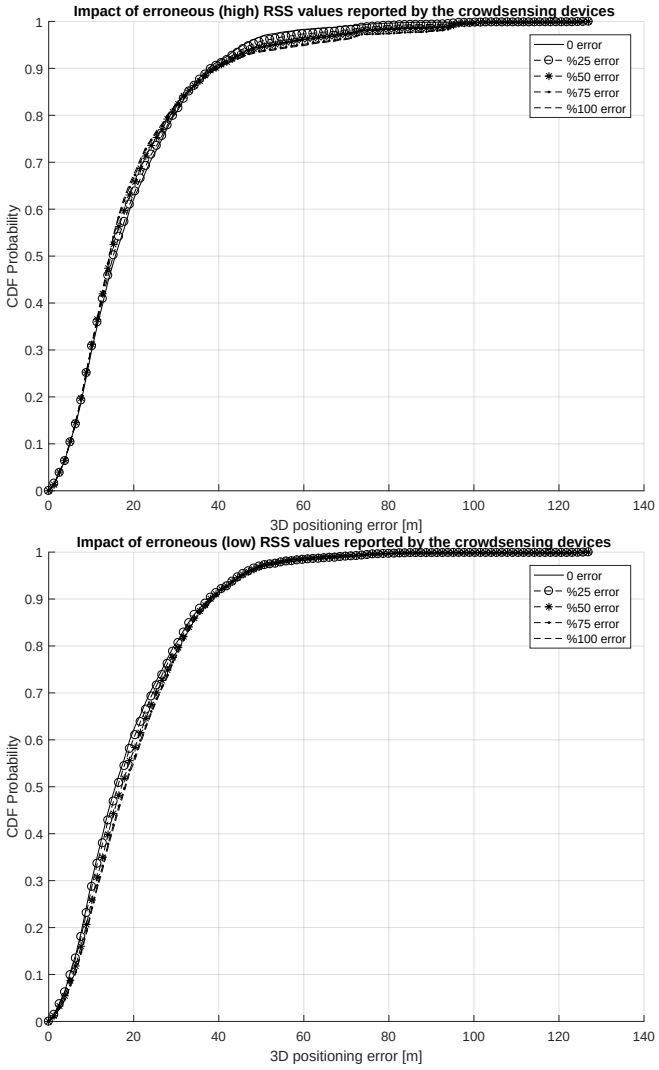

Fig. 11. Impact of the incorrectly reported RSS values by the crowdsensing devices on the positioning error. Top: wrongly reported RSS has a high value, here $-40\,\text{dB}$. Bottom: wrongly reported RSS has a high value, here $-90\,\text{dB}$.

Android application is now used as training data, and the estimation data is based on the crowdsensed 4648 points. The first case gives better results, and this could be explained by two reasons: on one hand, we have more training points in the first case than in the second; and on another hand, the training data collected with a higher number of devices can offer a higher degree of resistance to various calibration issues, and thus, on average, it is expected to work better.

V. Conclusions

This paper analysed several effects related to a crowdsensing approach for indoor positioning based on RSS. An extensive measurement campaign involving several devices, several software applications and several users was conducted in order to collect the data used in our analysis. The analysis looked at the differences between the power maps collected with various devices, at the positioning accuracy in the presence of crowdsensed data and at the impact of various crowdsensed errors

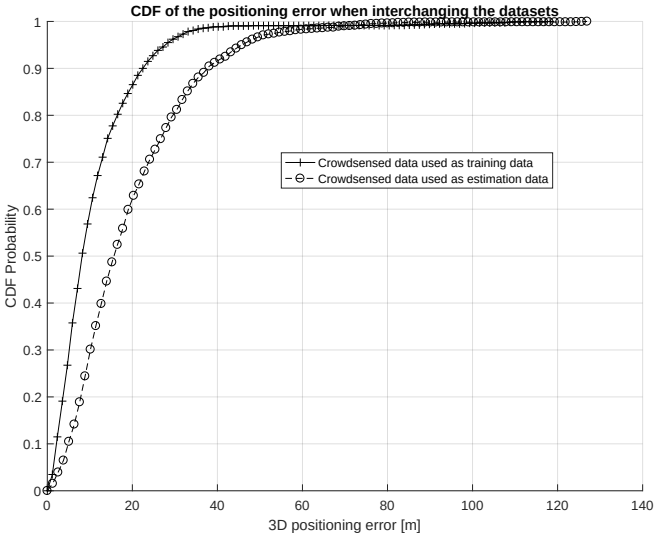

Fig. 12. Cumulative distribution function of the positioning error when we interchange the training and estimation data.

on the location estimate. Our analysis of fingerprints showed that a sufficient high number of crowdsourced fingerprints can yield a fingerprinting database similar to those systematically created by trained personnel. The difference between the RSS of different devices is affected by non-Gaussian noise, which eventually hinders the calibration and makes calibration approaches based on such an assumption inaccurate. It was also shown that crowdsensed databases are more robust to RSS reports than to malicious fingerprint position reports, as long as the malicious reports send correctly the IDs of the AP in range. Our studies relied on un-calibrated data, thus future studies should focus on calibration methods and whether such methods can significantly improve the location accuracy or robustness.

Open data

Part of our measurement data is also openly available [22].

Acknowledgment

The authors express their warm thanks to the Academy of Finland (project 303576, www.insure-project.org) for its financial support. The authors would also like to thank the anonymous volunteers who helped in the WiFi data collection and to the team of the following students who helped in building the Android software used in the data collection: Jukka-Pekka Venttola, Jeri Haapavuo, Kalle Immonen, Lauri Laaksonen, Matti Ylinevä,and Marko Leppänen.

References

[1] V. Honkavirta, T. Perälä, S. Ali-Löytty and R. Piché, "A comparative survey of WLAN location fingerprinting methods" in *2009 6th Workshop on Positioning, Navigation and Communication*, Hannover, 2009, pp. 243–251. doi: 10.1109/WPNC.2009.4907834

[2] A. Khalajmehrabadi, N. Gatsis and D. Akopian, "Modern WLAN Fingerprinting Indoor Positioning Methods and Deployment Challenges" in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1974–002, thirdquarter 2017. doi: 10.1109/COMST.2017.2671454

[3] X. Fan, P. Yang, C. Xiang and L. Shi, "iMap: A Crowdsensing Based System for Outdoor Radio Signal Strength Map" in *2016 IEEE Trustcom/BigDataSE/ISPA*, Tianjin, 2016, pp. 1442-1447. doi: 10.1109/TrustCom.2016.0226

[4] X. Wu, P. Yang, S. Tang, X. Zheng and Y. Xiong, "Privacy preserving RSS map generation for a crowdsensing network" in *IEEE Wireless Communications*, vol. 22, no. 4, pp. 42–48, August 2015. doi: 10.1109/MWC.2015.7224726

[5] T. Zhou, Z. Cai, B. Xiao, Y. Chen and M. Xu, "Detecting Rogue AP with the Crowd Wisdom" in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, Atlanta, GA, 2017, pp. 2327–2332. doi: 10.1109/ICDCS.2017.31

[6] A. Barry, B. Fisher, M.L. Chang, "A Long-Duration Study of User-Trained 802.11 Localization", in *Mobile Entity Localization and Tracking in GPS-less Environnments: Second International Workshop*, MELT 2009, Orlando 2009, pp. 197–212.

[7] J. Park, B. Charrow, D. Curtis, J. Battat, E. Minkov, J. Hicks, S. Teller, and J. Ledlie, "Growing an Organic Indoor Location System" in *Proc. International Conference on Mobile Systems, Applications, and Services (MobiSys)*, San Francisco, CA, Jun. 2010, pp. 271–284.

[8] Park, J., Curtis, D., Teller, S., and Ledlie, J. "Implications of Device Diversity for Organic Localization", in *IEEE INFOCOM*, Shanghai, China, 2011, 3182–3190

[9] C. Laoudias, R. Piché, C. G. Panayiotou, "Device self-calibration in location systems using signal strength histograms", in *Journal of Location Based Services*, vol. 7, no. 3, pp. 165–181, 2013. doi: 10.1080/17489725.2013.816792

[10] C. Wu, Z. Yang and Y. Liu, "Smartphones Based Crowdsourcing for Indoor Localization" in *IEEE Transactions on Mobile Computing*, vol. 14, no. 2, pp. 444–457, Feb. 1 2015. doi: 10.1109/TMC.2014.2320254

[11] T. Vaupel, J. Seitz, F. Kiefer, S. Haimerl, and J. Thielecke, "Wi-Fi positioning: System considerations and device calibration" in *International Conference on Indoor Positioning and Indoor Navigation* (IPIN), 2010.

[12] K. Kaemarungsi and P. Krishnamurthy, "Analysis of WLAN's received signal strength indication for indoor location fingerprinting" in *Pervasive and Mobile Computing*, vol. 8, no. 2, pp. 292–316, 2012. doi: 10.1016/j.pmcj.2011.09.003

[13] E. M. Delmelle, P. A. Rogerson, M. R. Akella, R. Batta, A. Blatt and G. Wilson, "A spatial model of received signal strength indicator values for automated collision notification technology" in *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5, pp. 432–447, 2005.

[14] B. Li, Y. Wang, H. K. Lee, A. G. Dempster and C. Rizos, "Method for yielding a database of location fingerprints in WLAN" in *IEE Proceedings - Communications*, vol. 205, no. 5, pp. 580–586, 2005. doi: 10.1049/ip-com:20050078

[15] P. Richter and M. Toledano-Ayala, "Revisiting Gaussian Process Regression Modeling for Localization in Wireless Sensor Networks" in *Sensors*, vol. 15, no. 9, pp. 22587–22615, 2015. doi: 10.3390/s150922587

[16] J. Talvitie, M. Renfors, M. Valkama and E.S. Lohan, "Method and Analysis of Spectrally Compressed Radio Images for Mobile-Centric Indoor Localization" in *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1-1. doi: 10.1109/TMC.2017.2741487

[17] E. Laitinen and E.S. Lohan, "Are all the access points necessary in WLAN-based indoor positioning?" in *2015 International Conference on Location and GNSS (ICL-GNSS)*, Gothenburg, 2015, pp. 1–6. doi: 10.1109/ICL-GNSS.2015.7217150

[18] A. Cramariuc, H. Huttunen, E.S. Lohan, "Clustering benefits in mobile-centric WiFi positioning in multi-oor buildings", 2016 International Conference on Localization and GNSS (ICL-GNSS), 2016, pp. 1–6.

[19] D. Gotlib, M. Gnat, "Spatial database modeling for indoor navigation systems", in *Reports on Geodesy and Geoinformatics*, vol. 95, no. 1, pp. 49–63, December 2013.

[20] A.S. Nossum, "Developing a framework for describing and comparing indoor maps", in *The Cartographic Journal*, vol. 50, no. 3, pp. 218–224, 2013.

[21] E.S. Lohan and P. Figueiredo e Silva, "User traces analysis based on crowdsourced data", 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, 2017, pp. 1303–1308. doi: 10.1109/IWCMC.2017.7986473

[22] E. S Lohan, J. Torres-Sospedra, P. Richter, H. Leppkoski, J. Huerta and A. Cramariuc, "Crowdsourced WiFi database and benchmark software for indoor positioning" [Data set], Zenodo. doi: 10.5281/zenodo.889798