Improving BLE deterministic fingerprinting by using a weighted *k*-NN algorithm over filtered RSSI data *

Fernando J. Aranda, Jose A. Parades , Teodoro Aguilera and Fernando J. Álvarez

Department of Electrical Engineering, Electronics and Automation
Sensory Systems Research Group (https://giss.unex.es/)
University of Extremadura, Badajoz, Spain (06006)
fer@unex.es

Abstract. Fingerprinting systems are a simple and efficient way to build positioning systems for Location Based Services (LBS) applications in smartphones. The technologies that can be used for radio frequency fingerprinting with off-the-shelf smartphones are Wi-Fi, which has been deeply studied, and Bluetooth Low Energy (BLE), which is receiving more attention in the last years. This work presents an improvement in deterministic fingerprinting using a filtering process and a new function to compute the Weighted k Nearest Neighbours (Wk-NN) location algorithm. Results with a BLE data set show that these improvements reduce the localization error and give a simple way to build a fingerprinting positioning system for smartphones with an average precision of less than a meter.

Keywords: Location-Based Services (LBS), Fingerprinting, Bluetooth Low Energy (BLE), k-Nearest Neighbours (k-NN).

1 Introduction

Fingerprinting positioning technique has become a promising approach for indoor localization with smartphones [2]. It obtains the signal strength at certain positions and creates a radio map for further comparison with the real time user measurements. Fingerprinting can use deterministic or probabilistic approaches to process the signal, being the first more easy to implement but less precise than second one [3]. Deterministic approaches, which are the topic of this work, are useful to implement in real LBS applications for smartphones.

This works presents an improvement over the deterministic fingerprinting method with the k-Nearest Neighbours (k-NN) algorithm presented in works such as [6, 9, 11] This improvement is twofold. First, a filtering of the raw RSS data acquired at each position is proposed. Second, a new weighting function is employed to perform a weighted average in the k-NN method. Results show a substantial improvement when compared with those obtained by previous works in the same area. The remainder of this work is organized as follows. In Section 2 the theoretical aspects of fingerprinting and k-NN are presented, pointing the issues that take part in this work. In the first part of Section 3 the experimental data set is introduced and in the second part different simulations are performed over the previous dataset. Finally, results are discussed in Section 4 and the main conclusions are drawn in Section 5.

* This work was supported in part by the Spanish Government and the European Regional Development Fund (ERDF) through the Project MICROCEBUS under Grant RTI2018-095168-B-C54, and in part by the Regional Government of Extremadura and ERDF-ESF under Project GR18038.

CEUR-WS.org/Vol-2498/short56.pdf

2 Fundamentals of Fingerprinting and *k*-NN algorithm

WLAN and WPAN fingerprinting are based on the Receive Signal Strength Indicator (RSSI) which measures, in a logarithmic scale, the signal power attenuation [8]. RSSI measures are unstable between consecutive measurements due to the presence of walls, water and other objects like furniture or human presence [4]. Thus, the RSSI distribution among a continuous set of measurements does not follow a Gaussian distribution or any other known distribution [9]. If we are using BLE, the comunication protocol and antenna gains in both receiver and emitter also affects the signal, since the antenna response is not uniform through protocol frequency channels. A description of these effects can be found in [5]. These problems have fostered different approaches to the creation of fingerprinting systems. There are two main conventional approaches for fingerprinting localization: deterministic and probabilistic [9].

Probabilistic approaches are based on a Maximun A Posteriori (MAP) position estimation. For the fingerprint distribution two main approaches can be used: parametric and non-parametric estimations. Parametric estimation tries to map the data to know analytical distributions, such as Gaussian or log-normal [7]. Meanwhile, non-parametric estimations do not assume any known distributions for the RSSI fingerprints and use a histogram matching to estimate the distribution [13]. These probabilistic approaches are far more complex to implement and need a higher computational time. In return, they typically provide more accurate results. [1].

However, in deterministic approaches position estimation is achieved through the selection of a set of poins, known as Reference Points (RP), with fingerprints similar to the RSSI measure in a localization phase and a positioning algorithm. In this section the basic formulation of a deterministic fingerprinting with the k-NN algorithm presented in works such as [9, 11, 12] is studied, pointing at the same time the proposed improvements of these works.

In the area of interest a set of points, the RPs, whose positions have to be perfectly known, must be chosen. These points will be denoted as r_i with $i = 1, 2, .., M$ and determine the system ground truth. The RPs, as well as the beacons location, must be scattered enough to characterize the system properly. The radio map creation begins once the beacons have been placed and the RPs have been chosen. This stage of the process is called offline phase. Using a smartphone as a receiver, the RSSI of all available signals in each RP are measured. There could be points where a beacon signal is missing or not strong enough to read the RSSI [10]. This issue must be taken into account when operating with the RSSI measures. Let $\psi_{r_i,j} = (RSS_1, RSS_2, ...RSS_N)$ be the j-s measurement in position r_i. This vector belongs to the RSSI vector space whose dimension is the number of beacons used, N and should not be confused with the real Euclidean space of the positions. At the end of this stage there must be a matrix DB in which each column correspond to a beacon and each row to a different measurement.

As mentioned above, the RSSI measurements are highly unstable, so for the same position and signal, values can be very different. Thus, it is required to take more than one measurement in each position to compensate for this effect. A filtering process is needed to take a representative measure for each AP in each RP. The most common filtering process are the mean, the maximum and minimum RSSI value.

The proposed filtering process performs the average of all the RSSI measurements over the same RP, despite they come from measurements taken at different times but with the same beacon's setup and positions. The mean value for all the signals in a reference point is denoted as $\overline{RSSI}_{r_i} = \sum_j \psi_{r_i,j}$, the maximum value is M_{r_i}, the minimum is m_{r_i} and let $\delta_{r_i} = \frac{M_{r_i}+m_{r_i}}{2}$ be the average of the latest two. The final value taken into the data base is:

$$\Psi_{r_i} = \frac{\delta + \overline{RSSI}_{r_i}}{2} \tag{1}$$

In a second stage, called online phase, an user in an unknown position, r_q, asks for a localization by taking a RSSI vector using. This consult vector is compared with all the information stored in the data base using a metric distance function. Most of the related works use the Euclidean distance but it has been proved that other metrics, such as the Sorensen distance, perform better [11]. From now on, the distance between two RSSI vectors Ψ_i and Ψ_j will be denoted as $\mathrm{d}(\Psi_i, \Psi_j)$ or just $\mathrm{d}_{i,j}$, where $d()$ is the distance function. Using this metric, all the vectors in the data base are sorted from the closest to the farthest. Finally, a positioning algorithm is used to calculate a final location using the k first vectors in the sorted database. This algorithm is the k-NN, which returns the average value of all the positions associated with these k first RSSI vectors, as shown in the following equation:.

$$r_{out} = \frac{1}{k} \sum_{i=1}^{k} r_i \tag{2}$$

An improvement of this algorithm is done using a weighted average, which leads to the Weighted k-Nearest Neighbours (Wk-NN) method. In Wk-NN not all elements contribute the same in the final position computation, then a function is needed to calculated a weight for each RP. This function will be called the weight function equation. Eq. (3) shows the Wk-NN output where w_{r_i} is the weight associated with position r_i.

$$r_{out} = \frac{\sum_{i=1}^{k} w_{r_i}\ r_i}{\sum_{i=1}^{k} w_{r_i}} \tag{3}$$

The weight is a function of the metric distance and must fulfill two conditions: elements with a closer distance to the minimum must have a weight close to one and those furthest away must tend to zero. Thus, the output is more similar to those with a smaller distance, penalizing those further away. In this work two different weight functions, Eq. (4) and Eq. (5) introduced from [24] and [25] respectively, are proposed. In both equations $d_{q,k+1}$ is the distance between the query vector and the $k+1$ element in the database once it has been ordered, and $d_{q,1}$ is the closest one. A new combined weight function, is proposed by the authors in Eq. (6). Both in k-NN and Wk-NN, the best number of neighbours to be considered is undetermined beforehand.

$$w_{r_i} = \frac{\mathrm{d}_{q,k+1} - \mathrm{d}_{q,i}}{\mathrm{d}_{q,k+1} - \mathrm{d}_{q,1}} \tag{4}$$

$$w_{r_i} = \frac{\mathrm{d}_{q,k+1} - \mathrm{d}_{q,i}}{\mathrm{d}_{\mathrm{d},k+1} - \mathrm{d}_{q,1}} \cdot \frac{\mathrm{d}_{q,k+1} + \mathrm{d}_{q,1}}{\mathrm{d}_{q,k+1} + \mathrm{d}_{q,i}} \tag{5}$$

$$w_{r_i} = \sqrt{\frac{\mathrm{d}_{q,k+1} - \mathrm{d}_{q,i}}{\mathrm{d}_{\mathrm{d}_{k+1} - \mathrm{d}_{q,1}} \frac{\mathrm{d}_{q,k+1} + \mathrm{d}_{q,1}}{\mathrm{d}_{q,k+1} + \mathrm{d}_{q,i}}}} \cdot \left(\frac{\mathrm{d}_{q,k+1} - \mathrm{d}q, i}{\mathrm{d}_{q,k+1} - \mathrm{d}_{q,1}} \right)^3 \tag{6}$$

3 Results and algorithm comparison

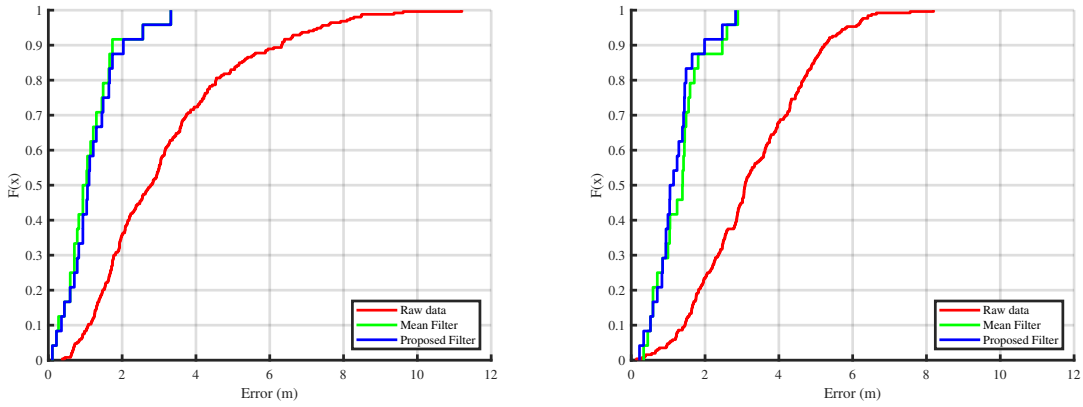
3.1 The BLE UJIIndoorLoc database

Since data collection for fingerprinting environment creation is a very resource and time consuming task [5], this work make use of the database developed by the Geospatial Technologies Research Group (GEOTEC) and available on the Zenodo repository [22]. This Group chose the IBKS 105 beacons from Accent Systems which are able to broadcast with different emission

powers and different BLE advertising protocols, such as Eddystone and iBeacon. The receivers were three Android smartphones: a BQ Aquaris X5 plus, a Samsung Galaxy S6 (SM-G920F) and a Samsung Galaxy A5 2017 (SM-A52OF). These smartphones were denoted as BQ, S6, A5 respectively. The experiments were carried out in a library (151 m^2) and a research laboratory (176 m^2) of the UJI.

3.2 Simulation results

For the fingerprinting simulation, the RPs are randomly divided into training and test subsets, using 60% and 40% for each group respectively. This data split process is done once for the library and for the laboratory, and kept during all the experiments. The results of these simulations are shown using a Cumulative Distribution Function (CDF) which shows the distribution of the error obtained in each localization. When no filtering process has been applied to measurements will be indicated as raw data. The effects of using raw data against filtered over the same RPs are now analysed. A comparative result using $k = 72$ for raw data and $k = 6$ for filtered ones is depicted in Fig. 1 for the BQ (a) and A6 (b) phones. Similar results are obtained using other transmission powers in the laboratory data set. Note that the same value of k cannot be used in both cases since the number of elements in the database is reduced when data is filtered, so after data filtering the maximum value for k is the number RPs. Fig. 1 shows that using average values over RP gives an overall better result. The error distribution median has been reduced by 1.5 m compared with $k-$NN over unfiltered data and the maximum error in more than 9 m.

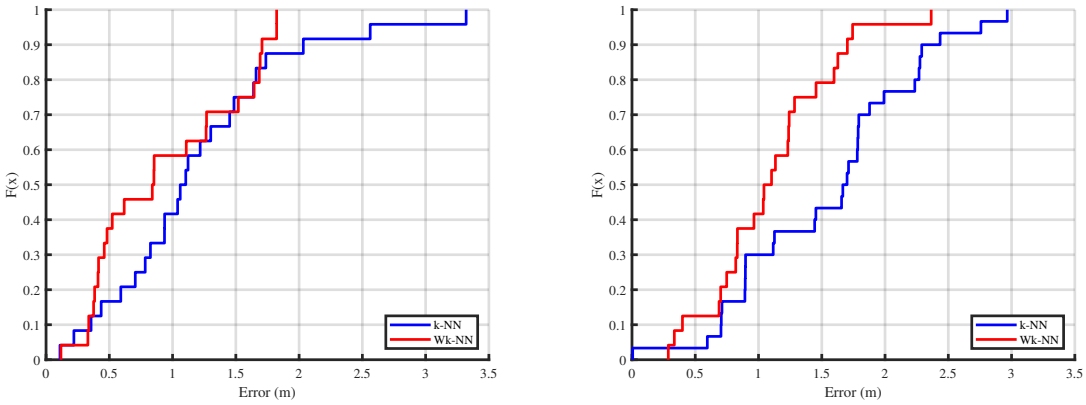


(a) Bq (b) A5

Fig. 1: Cumulative error distribution between filtered (blue) with and raw data (red) with k-NN. Simulation using the library data set and the BQ (a) and A6 (b) smartphone with $k = 10$.

Each one of these k closer vectors contributes the same to the final output position. The higher this value, the closer the output position will be to the center of mass of RPs distribution when k-NN is used over filtered data. Above simulations are compared now with a weighted average, Wk-NN as described in Section II.

Fig. 2 illustrates the performance of the two algorithms. As can be seen, Wk-NN performs better than the k-NN over filtered data. The output result using k-NN gets closer to the center



(a) Bq (b) A5

Fig. 2: Cumulative error distribution between k-NN (red) and Wk-NN (blue) with filtered data. Simulation over the library data set with the BQ (a) and A6 (b) smartphones with $k = 10$. The blue lines are the same as Fig. 1. Weights for Wk-NN were built using Eq. 4.

of mass of the RPs distribution when the k parameter increases, meanwhile, this effect is reduced with the weighted average because the most distant elements (in the RSSI vector space) have a smaller weight. A new comparative between the proposed filter and the mean filter are shown in Fig. 3 where it can be seen that the proposed filter performs better.

This simulations are repeated with all phones available in the library data set and all transmission powers in the laboratory obtaining the similar results. The effect of the k parameter in the Wk-NN is presented in Fig. 5. The above simulation has been repeated for the A5 and BQ phones for different k values, from 1 (cell ID) to 9. Results for both phones are shown in Fig. 5.

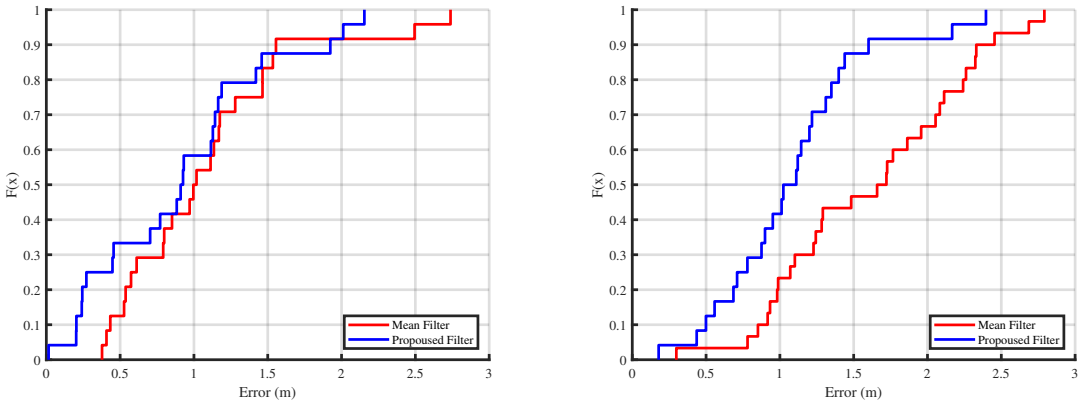
This figure shows that the error for $k = 1$ is bounded, thus the output is one of the RPs and the error can never be greater than the distance among two consecutives RP. When k is increased, the function moves to the right and the average quality of the location decreases. For both phones it is found that the best result is obtained when k is around 5, but it is not exactly the same for both. For instance, the best result is given by $k = 5$ for the A5 and by $k = 6$ for the BQ.

Finally the effects of the weighted function are studied. The weighted functions shown in Eq. 4, 5 and 6 are now compared. Simulations are repeated for the library data set and the BQ phone. A comparative study is shown in Fig. 4 for $k = 6$ and 7, which shows that the new proposed function (W-3) works better than the other two 100% of the time with $k = 7$ and around 60% with $k = 6$.

4 Discussion

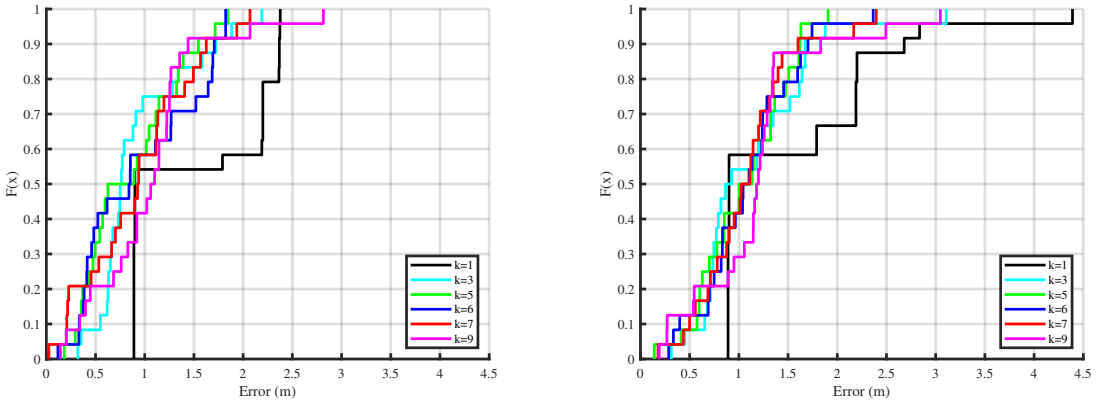
The results obtained in the previous simulations provide useful information about how to build a deterministic fingerprinting system for WLAN and WPAN networks.

The key in the radio map construction is the characterization between the real positions and the vectors in the RSSI space. Due to smartphone RSSI measurements are highly unstable just one is not enough to characterized the system. Thus, measurements must be repeated in each



(a) Bq (b) A5

Fig. 3: Cumulative error distribution between mean filter (red) and propoused filter (blue). Simulations over the library data set with the BQ (a) and A6 (b) smartphones with $k = 6$.

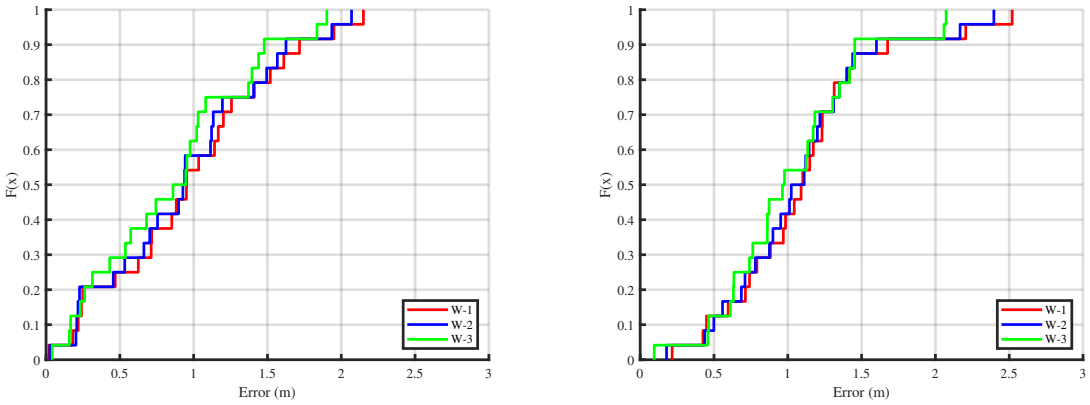


(a) Bq (b) A5

Fig. 4: CDF for different k with Wk-NN. Simulation from library's data with Bq (a) and A5 (b) smartphone. Weights built with Eq. 4.

position. However, making more measurements in all the RP drastically increase the time spent in radio map construction and consequently the battery drain through this process. As mentioned in Section 2, the RSSI distribution does not follow a Gaussian bell, with a well-defined average and median values. Since the characterization of the RSSI distribution remains unsolved, the correct filtering process used cannot be optimized. The proposed filtering process takes into account the dispersion of the measurements when the distribution is not symmetrical.

The second major issue considered in this work is the difference between k-NN and Wk-NN for final location computation. k-NN algorithm performs poorly compared with the weighted version once the data have been filtered however, results are no very different when both are



(a) Bq (b) A5

Fig. 5: Comparative results between weights of Eq. 4 (red), Eq. 5 (blue) and Eq. 6 (green) for $k = 6$ for the Bq (a) and A5 (b) smartphones.

used with raw measurements. This is caused by the stochastic nature of the RSSI measurements, which makes individual measurements not to have enough information. Furthermore, since the final output of the proposed system is a linear combination of the RPs, how these points are choose highly affect the final output, because different selections can change the set of accessible points and possible correlations between coordinates.

The number of neighbours (k) for the algorithm is different in each situation (k-NN and Wk-NN) but also in simulations between different locations and phones. It is not possible to determine beforehand what is the value providing the best result in each location and with each experimental setup. The weighted functions proposed in this work provide a better result in all the experiments although some issues must be pointed. First, the result is no exactly the same for all the k values. There is a relation between the metric distance and weighted distance that has not been explored completely yet. The effect of the weighted function is less important than the filtering process effect and the Wk-NN use.

It must be said that in most fingerprinting experiments, the used technology -Wi-Fi or BLE- is not designed for that purpose. Nor BLE neither the smartphones (hardware and software) are designed to make fast RSSI reading for localization.

5 Conclusion and future works

In this work new considerations about how to improve an indoor localization system with Bluetooth Low Energy (BLE) and deterministic fingerprinting have been presented. A data filtering process and the Weighted k-Nearest Neighbours (Wk-NN) instead of the k Nearest Neighbours (k-NN) have been proposed, with specific proposals for the weight functions.

Simulations have shown that using a filtering process over measurements taken in the same RP, even at different times, can significantly reduce the error in fingerprinting positioning. Maximum errors have been reduced in 65% and the median error in 40%. Results have shown that a weighted average performs better when a k-NN algorithm is used, but only when data have been filtered. This improvement reduces the error by 60% compared with the k-NN. The error in the

simulations with Wk-NN and filtered data has been reduced by more than 80% for both median and maximum error when compared with the k-NN and raw data fingerprinting simulations.

As part of the currently ongoing work, the authors are working in an improvement of the algorithm based on the physical characteristics of the RSSI to correctly design the mathematical steps involved. Mainly, the RSSI distribution over measurements to correctly define a filtering process and the combination between the metric and weighted function used in the k-NN algorithm.

References

1. Brena, R.F., García-Vázquez, J.P., Galván-Tejada, C.E., Muñoz-Rodriguez, D., Vargas-Rosales, C., James Fangmeyer, J.: Evolution of indoor positioning technologies: A survey,. Journal of Sensors **vol 2017**, 1,21 (2017), https://www.hindawi.com/journals/js/2017/2630413/cta/
2. Davidson, P., Piché, R.: A survey of selected indoor positioning methods for smartphones. IEEE Communications Surveys Tutorials **19**(2), 1347–1370 (Secondquarter 2017). https://doi.org/10.1109/COMST.2016.2637663
3. Dawes, B., Chin, K.W.: A comparison of deterministic and probabilistic methods for indoor localization. Journal of Systems and Software **84**(3), 442 – 451 (2011). https://doi.org/https://doi.org/10.1016/j.jss.2010.11.888, http://www.sciencedirect.com/science/article/pii/S0164121210003109
4. Fang, S., Lin, T., Lee, K.: A novel algorithm for multipath fingerprinting in indoor wlan environments. IEEE Transactions on Wireless Communications **7**(9), 3579–3588 (Sep 2008). https://doi.org/10.1109/TWC.2008.070373
5. Faragher, R., Harle, R.: Location fingerprinting with bluetooth low energy beacons. IEEE Journal on Selected Areas in Communications **33**, 1–1 (11 2015). https://doi.org/10.1109/JSAC.2015.2430281
6. He, S., Chan, S..G.: Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons. IEEE Communications Surveys Tutorials **18**(1), 466–490 (Firstquarter 2016). https://doi.org/10.1109/COMST.2015.2464084, https://ieeexplore.ieee.org/document/7174948
7. Honkavirta, V., Perala, T., Ali-Loytty, S., Piche, R.: A comparative survey of wlan location fingerprinting methods. In: 2009 6th Workshop on Positioning, Navigation and Communication. pp. 243–251 (March 2009). https://doi.org/10.1109/WPNC.2009.4907834
8. Jiuqiang, X., Liu, W., Lang, F., Zhang, Y., Wang, C.: Distance measurement model based on rssi in wsn. Wireless Sensor Network **2**, 606–611 (01 2010). https://doi.org/10.4236/wsn.2010.28072
9. Khalajmehrabadi, A., Gatsis, N., Akopian, D.: Modern WLAN fingerprinting indoor positioning methods and deployment challenges. IEEE Communications Surveys Tutorials **19**(3), 1974–2002 (thirdquarter 2017). https://doi.org/10.1109/COMST.2017.2671454, https://ieeexplore.ieee.org/document/7874080
10. Mendoza-Silva, G.M., Matey-Sanz, M., Torres-Sospedra, J., Huerta, J.: Ble rss measurements dataset for research on accurate indoor positioning. Data **4**(1) (2019). https://doi.org/10.3390/data4010012, http://www.mdpi.com/2306-5729/4/1/12
11. Torres-Sospedra, J., Montoliu, R., Trilles, S., Óscar Belmonte, Huerta, J.: Comprehensive analysis of distance and similarity measures for wi-fi fingerprinting indoor positioning systems. Expert Systems with Applications **42**(23), 9263 – 9278 (2015). https://doi.org/https://doi.org/10.1016/j.eswa.2015.08.013, http://www.sciencedirect.com/science/article/pii/S0957417415005527
12. Yiu, S., Dashti, M., Claussen, H., Perez-Cruz, F.: Wireless RSSI fingerprinting localization. Signal Process **131**(C), 235–244 (Feb 2017). https://doi.org/10.1016/j.sigpro.2016.07.005, https://doi.org/10.1016/j.sigpro.2016.07.005
13. Youssef, M.A., Agrawala, A., Udaya Shankar, A.: Wlan location determination via clustering and probability distributions. In: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003. (PerCom 2003). pp. 143–150 (March 2003). https://doi.org/10.1109/PERCOM.2003.1192736