

Locations Selection for Periodic Radio Map Update in WiFi Fingerprinting

Germán M. Mendoza-Silva, Joaquín Torres-Sospedra and Joaquín Huerta

Abstract The construction and update of a radio map are usually referred as the main drawbacks of WiFi fingerprinting, a very popular method in indoor localization research. For radio map update, some studies suggest taking new measurements at some random locations, usually from the ones used in the radio map construction. In this paper, we argue that the locations should not be random, and propose how to determine them. Given the set locations where the measurements used for the initial radio map construction were taken, a subset of locations for the update measurements is chosen through optimization so that the remaining locations found in the initial measurements are best approximated through regression. The regression method is Support Vector Regression (SVR) and the optimization is achieved using a genetic algorithm approach. We tested our approach using a database of WiFi measurements collected at a relatively dense set of locations during ten months in a university library setting. The experiments results show that, if no dramatic event occurs (e.g., relevant WiFi networks are changed), our approach outperforms other strategies for determining the collection locations for periodic updates. We also present a clear guide on how to conduct the radio map updates.

Keywords Wifi fingerprinting · Radio map update · Regression · Optimization Genetic algorithm

G. M. Mendoza-Silva (✉) · J. Torres-Sospedra · J. Huerta
Institute of New Imaging Technologies, Universitat Jaume I, Avda. Vicente Sos Baynat S/N,
Castellón, Spain
e-mail: gmendoza@uji.es

J. Torres-Sospedra
e-mail: jtorres@uji.es

J. Huerta
e-mail: huerta@uji.es

1 Introduction

As location-based services have grown in importance during recent years, the indoor positioning has increasingly drawn attention from the research community. The WiFi fingerprinting has been a very popular indoor positioning method for this community. Reasons for its popularity include a large number of WiFi access points (AP) already deployed in many environments, the generalized usage of WiFi-enabled smartphones, and a positioning accuracy that is acceptable for many applications (He and Chan 2016; Yiu et al. 2017). This method, however, has two known drawbacks: the WiFi measurements radio map construction and update.

The radio map construction and update for WiFi fingerprinting usually involve a person, or dedicated receiver, that collects WiFi measurements at some known locations. Thus, the collection process has a cost, either in the time that a paid person employs, or in the cost of deploying and maintaining receivers. The reduction of that cost is referred as mapping, calibration or radio map construction/update effort reduction.

It is acknowledged that, at least to some extent, the larger the number of measurement locations in the target area, the better the accuracy of the WiFi fingerprinting is Kanaris et al. (2016), Wang et al. (2016), Hernández et al. (2017), Yiu et al. (2017), but also the more costly the collection process is. To address this issue, methods that require only a few collection locations have been proposed (Alonazi et al. 2015). Such methods involve regression (interpolation/extrapolation) approaches or turning to collaborative or crowd-sourced approaches. If the collected data reliability is a hard concern, the one option is collecting measurements at all relevant locations. If data reliability is soft concern, another option is to collect measurements only at some location and then estimate measurement values for the remaining locations using a regression approach.

The studies proposing regression approaches generally show that the estimations made by their methods can be used instead of some of the actual measurements without significantly harming the localization accuracy provided by an Indoor Positioning System (IPS). These studies usually specify elimination procedures to drop some of the original locations in order to test their methods. However, those elimination procedures are not to be understood as suggested strategies for determining collection locations. The random locations distribution is a common approach (Ali et al. 2017), despite the locations distribution is very important (Li et al. 2014) for radio map construction. It is also acknowledged that the radio map needs periodic updates so that the positioning method can be robust to changes in the target environment and in the relevant APs (Wang et al. 2016; Hossain and Soh 2015).

The importance of the collection locations is intuitive and has been formally acknowledged in other subjects for other phenomena. Specifically, several papers have addressed the optimal (or quasi-optimal) placement of sensors that best measure a given phenomenon (Rowaihy et al. 2007; Joshi and Boyd 2009). A set of WiFi measurements collected at known locations by a person can be viewed as measurements of a set of sensors. Therefore, choosing the best locations for an individual to

collect the WiFi measurements can be thought as optimizing the placement of a set of sensors.

This paper presents a novel approach for determining the collection locations for periodic WiFi radio map updates. The approach requires initial measurements, taken at a relatively dense set of known locations. The initial measurements are used to determine a set of locations that establishes a compromise between a small set's size and its goodness for estimating the Received Signal Strength (RSS) values at the remaining locations through a regression method. This paper suggests to find such set using a genetic algorithm optimization approach with a specific fitness function. The found set of locations, called the solution set, is proposed to be used as collection locations for the radio map periodic updates.

The proposed approach was tested using SVR as a regression method and a WiFi RSS database collected during ten months at one floor of a university library. The database contains measurements for training and test purposes. The training measurements for the first month were used to determine the solution set. The goodness of the solution set for selecting the measurement collection locations for the periodic radio map updates was explored across the following nine month in terms of: (1) RSS difference between the measurements and the RSS estimations provided by a regression fitted for the solution set, and (2) the effects of using the above RSS estimations for radio map update on the accuracy of a fingerprinting-based IPS, considering the test sets collected at each month. The experiments' results have shown the suitability of using our approach for determining the locations for periodic radio map updates in the tested environment.

In summary, in this paper we propose an alternative to common strategies for locations selection for WiFi radio map update and we experimentally show its benefits. While following those goals, we:

1. Present some drawbacks of the previous common strategies.
2. Describe how to determine a set of locations (solution set) where measurements should be taken in order to obtain fine RSS estimations for the remaining locations through regression.
3. Briefly describe how the proposal can be used to find challenging sets of locations to test regression approaches for WiFi fingerprinting.
4. Experimentally show how to use the estimations obtained from the solution set to update a WiFi radio map.

The remainder of the paper is organized as follows: Sect. 2 provides an overview of fingerprinting calibration efforts reduction, focusing mainly in regression-based approaches. Section 3 presents our proposal for measurement locations determination for WiFi radio map update. Section 4 provides the experimental testing of our proposals. Finally, Sect. 5 summarizes the ideas presented in this paper and proposes its continuation lines.

2 WiFi Radio Map Construction and Update

WiFi fingerprinting is performed in two phases: the offline training phase and the online (operational or query) localization phase (He and Chan 2016; Yiu et al. 2017). In the training phase, WiFi fingerprints are collected in the target area. A WiFi fingerprint is a vector of RSS values of the detected APs measured at a given time. Each training fingerprint is usually labeled with the location at which it was collected. The fingerprints are stored in a training database, which is also called radio map. In the localization phase, an IPS uses the training database to estimate location labels for new, unlabeled fingerprints.

Radio maps with measurements collected at relatively dense sets of locations provide higher positioning accuracies than those with measurement collected at sparse locations (Kanaris et al. 2016; Wang et al. 2016; Hernández et al. 2017; Yiu et al. 2017). Additionally, periodic radio map updates are needed because WiFi signals are prone to changes, due to either changes in the environment or in relevant APs (including reallocation, replacement and transmission power reconfiguration) (Hossain and Soh 2015; Wang et al. 2016). The effort reduction on radio map construction and the methods robustness to environment's changes has been targeted by WiFi fingerprinting researchers for over 10 years, with many of the attempts included in reviews like Hossain and Soh (2015), Pei et al. (2016), Wang et al. (2016). Some examples of the attempts are found in Yang et al. (2013), Alonazi et al. (2015), Majeed et al. (2016), Gu et al. (2016a). The study of Yang et al. (2013), instead of directly using the RSS values, used order relations between AP's RSS values. The authors in Alonazi et al. (2015) collected WiFi measurements at a few reference points (RPs) located at the ends of corridors and later enriched the radio maps with user-supplied new RSS values. In Majeed et al. (2016), the authors combined a small calibration set, the coordinates of all target locations and several simultaneous operational RSS measurements using semi-supervised alignment of manifolds to estimate the operational measurements' locations. Gu et al. (2016a) used the AP intensity order as similarity score to deal with the changes in relevant APs and mobile device diversity, and tested its approach with a database collected during 6 months.

The above solutions for effort reduction differentiates on whether the measurements are collected by (1) collaborative/crowd-sourced means or by (2) a dedicated collector. Each approach have its own benefits and drawbacks (Pei et al. 2016). In the first approach, the cost is almost negligible, but quality and completeness are concerns. In the second approach, the cost is reduced by making collection at only a few locations and then estimating (mainly performing a regression) the RSS values at the remaining locations.

The collaborative/crowd-sourced approaches include explicit or implicit user collaboration (He and Chan 2016; Wang et al. 2016; Hossain and Soh 2015). In the explicit case, the user is required to label all fingerprints, or at least a subset of them, with the location where they are taken. When there are unlabeled fingerprints, their labels are estimated using techniques that consider additional information, such as readings from other sensors (e.g., using pedestrian dead reckoning (PDR) Xiao et al.

2015) or environment knowledge. The environment knowledge may, for example, indicate the likely corresponding path segments or the intrinsic relations between neighboring fingerprints using models like Markov-chain (Lin et al. 2016). Also, floor plans and APs locations knowledge can be used to generate each AP radio map using propagation models (Ali et al. 2017). In the implicit case, location hints are opportunistically used to label WiFi measurements with the location without the user interaction. The location hints may come from other sensors, like a GPS sensor, or through estimations such as those used for unlabeled fingerprints in the case of the explicit user collaboration. The collaborative/crowd-sourced approaches are also used for radio map update. These approaches have a well known challenge: the labels quality (Wang et al. 2016).

The approaches that do not rely on collaborative/crowd-sourced contributions try to reduce the amount of locations required for constructing the initial radio map. Fingerprints are collected at a small amount of locations and the RSS values at the remaining target locations are estimated using regression (interpolation and extrapolation). The following subsection deepens on this subject.

2.1 Collection Effort Reduction for Fingerprinting Using Regression Approaches

Regression for RSS radio map enrichment is applied as follows. An initial, small set of locations with known coordinates $L_{n \times 2}$ is chosen for the target area. Then, if s measurements are made for each location and m wireless networks are detected in the whole campaign, the initial database is the set $D_{n \times m \times s} = \{r_{ijk}\}$, where r_{ijk} is the RSS value measured at the i th location, for the j th AP, and in the k th location sample. For each wireless network a , the regression method fit a function $f_a(L) = R_a$, with $R_a = \{r_{iak}\}$. Each function f_a is then used to predict RSS values for locations \hat{L} . If the points in \hat{L} lie inside the convex hull of L , the estimation is usually called interpolation, and if they lie outside, it is called extrapolation. Extrapolation methods (extrapolation functions) are known to be less accurate, and thus more challenging and less used than the interpolation ones (Talvitie et al. 2015).

The regression methods has been used for reducing the calibration effort in fingerprinting for more than 10 years (Krumm and Platt 2003; Li et al. 2005). Among the methods found in literature are: linear interpolators (Talvitie et al. 2015), radial basis interpolators (Krumm and Platt 2003; Ezpeleta et al. 2015), Gaussian Process regression (Yiu et al. 2017), and Support Vector Regression (SVR) (Hernández et al. 2017). Some studies particularly focused on the spatial relations of measurements and the spatial characteristics of the environment for regression. They included methods widely used in spatial analysis like Inverse Distance Weighting (IDW) and Kriging (Li et al. 2005; Liu et al. 2015; Jan et al. 2015), Voronoi Tessellation (Lee and Han 2012), Sparsity Rank Singular Value Decomposition (SRSVD) (Gu et al. 2016b)

and other particular heuristics (Bong and Kim 2012). Studies like Zhu et al. (2014) have also taken into account the time dimension for regression.

In the cited studies, the authors first collect a relatively dense dataset of RSS measurements, and, through elimination strategies, produce new datasets. Their regression methods are then applied to the new datasets in order to obtain estimated RSS measurements for the removed collection locations. The regression goodness is usually evaluated as (1) the difference in RSS values between measurements and estimations and (2) the difference in localization error of some IPS, between using dataset with a high percentage of removed points and the original dataset for training. The elimination strategy is an important factor in the results obtained in such evaluations (Talvitie et al. 2015). The regression performance found in literature varies significantly, from discrete but reasonable results of 50% location reduction (Ezpeleta et al. 2015) to astonishing results of 5% locations reduction (Gu et al. 2016b) with very little RSS or localization error difference.

Most of studies found in literature indicate the percentage of collection locations (with respect to all target locations) required for their regression methods to provide proper localization accuracies. However, they do not mention a methodology for determining the number of collection locations for a given environment (though it has been shown to be very important (Li et al. 2014)), or how to determine where those locations should be. An intuitive approach is to choose the amount of collection points as a function of the target area size and randomly determine their positions in that area. Some studies have used similar approaches.

In Kanaris et al. (2016), the authors proposed an algorithm that suggested a collection's sample size given a small preliminary set of measurements. They suggested the definition of a grid of locations in a target area and randomly choosing locations in the amount determined by the sample size calculation. Specifically, for the case of database update, collecting measurements at random locations in a target area is a common approach (Ali et al. 2017). Indeed, depending on the update frequency, the collaborative, crowd-sourced or opportunistic approaches can be also considered strategies of collecting update measurements at random locations.

The elimination strategies used for evaluating the goodness of regression methods found in the research literature have hinted on possible strategies for determining the locations for training set collection. The work of Krumm and Platt (2003) proposed an elimination strategy consisting in running a k-means clustering algorithm, and selecting only the k locations nearest the k cluster centroids. Other studies have resembled in their proposed elimination strategies the types of collection absences that may happen in regular collection processes, like random isolated absent points, zones with higher or lower percentage of elimination (Ezpeleta et al. 2015) or random blocks of absent points (Talvitie et al. 2015).

The following section presents the approach we propose to determine the set of locations where fingerprints for WiFi radio map update are to be taken.

3 Locations Set Determination for Radio Map Update

As seen in Sect. 2, studies found in literature have hinted possible approaches for choosing the collection locations. These approaches, however, have some drawbacks that are experimentally shown in Sect. 4. It is almost intuitive that neither the number of locations, nor their actual distribution, should be chosen randomly without any restriction. In addition, a uniformly spaced locations distribution may not take into account obstacles influencing the WiFi signals propagation. Therefore, a person with experience in WiFi-based indoor localization generally chooses the amount and distribution of the collection locations. Regardless of this person expertise, the previous task is not a trivial one.

This study harnesses the similitudes between (1) choosing a subset of measurement locations for estimating the values at remaining locations through regression and (2) choosing the placement of sensors for field estimation. The problem of sensor placement, related to sensor selection (activation), is a well-known problem that has long been addressed for wireless sensor networks. The sensor selection problem can be stated as choosing a set of k measurements from a set of m possible sensor measurements, which minimizes the error in some parameters estimation (Joshi and Boyd 2009). We suggest that the approaches for solving the previous problem can also be applied to finding the set of k locations from m possible ones, where the WiFi measurements will be collected so that the WiFi signal intensities for the remaining locations can be obtained through regression with a small error. What is more, we do not consider a fixed number of locations, but instead, obtain a compromise between the location set's size and the goodness of the regression.

The approaches to deal with the sensor placement/selection problem vary depending on the usage of the sensor measurements (Rowaihy et al. 2007). Specifically, some studies have proposed approaches for the case of using the sensor measurements for estimating a field of values (Joshi and Boyd 2009; Ranieri et al. 2014; Roy et al. 2016). The combinatorial nature of the problem (Joshi and Boyd 2009) makes it unfeasible to explore the whole solution space. If the total number of locations is very small, e.g., six, it is feasible to manually determine fine sets of locations where the measurements are to be taken. However, if a target environment has a (still small) set of 24 locations, and measurements are to be taken at 12 of those locations, the number of different possible sets of locations is $\binom{24}{12} = 2,704,156$. If the number of measurement locations is not already decided, the number of possible combinations rises to $2^{24} = 16,777,216$.

This paper determines the set of locations in a way simpler than those presented in Joshi and Boyd (2009), Ranieri et al. (2014), Roy et al. (2016) for sensor placement. Those studies have harnessed some property of the target problem or forced some form for the solution. We have used an optimization strategy based on genetic algorithms. Sensor placement optimization has already been addressed using genetic algorithms (Yao et al. 1993; Macho-Pedroso et al. 2016), even for indoor acoustic localization (Macho-Pedroso et al. 2016).

3.1 Genetic Programming for Locations Set Determination

The approach proposed in this study uses a genetic optimization algorithm to find a set of locations that includes only a small number of locations and the goodness of the regression obtained using these locations should be similar to the one obtained using the whole set of possible locations. The explanation presented here for genetic algorithms, as well as the library used in the experiments, are based on Mitchell (1998).

The genetic algorithms try to efficiently find solutions to problems that have huge spaces of candidate solutions. Each candidate solution for a problem is called an individual. Commonly, an individual is encoded as a bit string, where each bit represents the presence ('1') or absence ('0') of a trait. These algorithms start by considering a population of random individuals, and iteratively evolves it. The population of each iteration is called a generation. The following generation is the result of applying genetic operators on the current generation. The selection operator selects pairs of individuals whose traits are combined using a crossover operator to produce offspring. A fitness value is computed for every individual in a generation and those with higher fitness values are more likely to be chosen by the selection operator. A mutation operator is applied to the offspring to produce subtle changes in the resulting traits. Some of the new individuals can be randomly discarded, but the population size is maintained.

In this paper's proposal, the set of all locations $L = \{l_1, \dots, l_n\}$ from the initial, dense collection represents the possible traits that each individual may have. The location set L have associated WiFi RSS measurements $D = \{r_{ijk}\}$. Assume a function $fmap(A, B) \rightarrow C$ so that A is a set of RSS values, B is a set of locations and C is the set of RSS values in A associated to locations in B. Then, $D_{l_p} = fmap(D, \{l_p\}) = \{r_{pjk}\}$ are the RSS measurements associated to location l_p . An individual represents a subset L_I of L . The size of the population, as well as the number of generations considered for population evolution are parameters of the algorithm that are presented in Sect. 4. We have designed the fitness value calculation of an individual so that larger subsets and differences between measured and estimated RSS values are penalized. Specifically, the fitness computation steps are:

1. Fit regressions f_a , for every detected access point a , using L_I and their associated measurements $fmap(D, L_I)$.
2. Use regressions f_a to estimate RSS values $E = \{\hat{r}_{ia}\}$ for locations of $\hat{L}_I = L - L_I$.
3. Compute the AP-wise and location-wise RSS absolute differences between E and $fmap(D, \hat{L}_I)$. Let MRD be the maximum value of those differences.
4. The individual's fitness is $(MRD + 2MRD \frac{ab}{tb})^{-1}$, where ab and tb are the number of '1' bits and the total number of bits, respectively. If for some reason the target number of locations is already predefined, say k , the individual's fitness can become $(MRD + 2MRD|ab - k|)^{-1}$.

After a given number of generations, the individual with higher fitness value, called the elite individual, could be chosen as the set of locations where WiFi

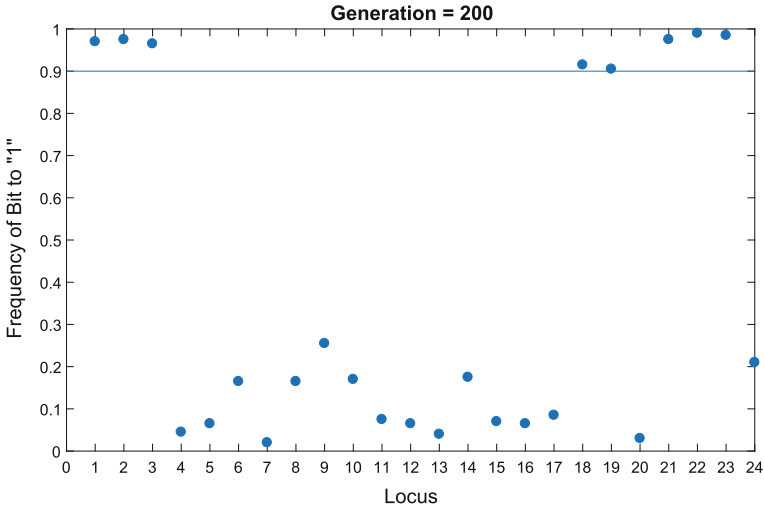


Fig. 1 Locations (bits) frequency. Blue dots represent how often a location has been included in individuals of generation 200. The blue line represents the frequency threshold

measurements are to be collected. The elite individual represents the set that has so far achieved the best compromise between a small number of locations and little degradation of the regression goodness. The genetic algorithm does not guarantee that the elite individual would be the optimal solution for a given problem, but is a fair alternative to an exhaustive search given the combinatorial nature of the problem.

This paper’s main goal is not selecting the best locations for making a one-time regression. The main goal is determining the suitable locations for conducting periodic WiFi radio map updates so that the new RSS measurements help in estimating RSS values for remaining, target locations. The elite individual may represent a solution that is over-fitted for the initial measurements. Therefore, we propose to look at the occurrence frequency of each location in the final population and select only those with high frequency. We call this set of highest frequency locations the *solution set*. Figure 1 shows an example of the location’s frequency for a population of (200) evaluated sets of locations after 200 iterations. The number of traits, i.e., the number of locations in the initial, dense collection is 24. The bit frequency represents how often a location is found in sets of locations. If we chose a high frequency threshold of 0.9, the solution set would be {1, 2, 3, 18, 19, 21, 22, 23}.

In summary, the steps needed for selecting the locations where the periodic update measurements are to be collected are:

1. Collect a relatively dense WiFi RSS training database.
2. Use a genetic algorithm, such as the one described in pages eight and nine of Mitchell (1998), using the fitness function described above in this section, to determine the locations’ frequency in the population of sets of locations.

3. Choose as the solution set the locations with frequencies above a certain threshold.

Section 4 also provide a guide on how to use new measurements collected at the solution set for updating the radio map. We advise applying our approach independently for clearly unrelated zones, i.e., zones that belong to different buildings or different floors.

Besides suggesting a very good placement for the measurement locations, the proposed approach can be also used for testing the performance of regression methods. By computing an individual's fitness as $MRD + 2MRD_{ib}^{ab}$, the genetic algorithm would determine a compromise between a large number of locations and a high RSS absolute difference. The set resulting from choosing the n highest fitness sets of locations can be used as a challenging test for evaluating the performance of regression methods.

4 Experiments

The approach proposed in Sect. 3 was tested using a WiFi RSS database collected in a university library during ten months (30 days of separation time, approximately). The database contains training and test sets for each month. Figures 2 and 3 show the collection locations for the training and test sets, respectively, using colored circles. The location label for each fingerprint is expressed in local coordinates in a 2D

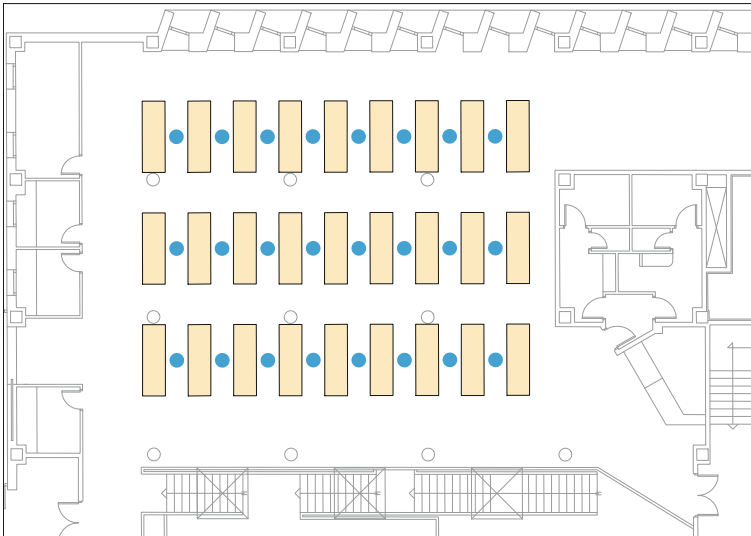


Fig. 2 Collection locations for the training sets. The colored rectangles represent the bookshelves

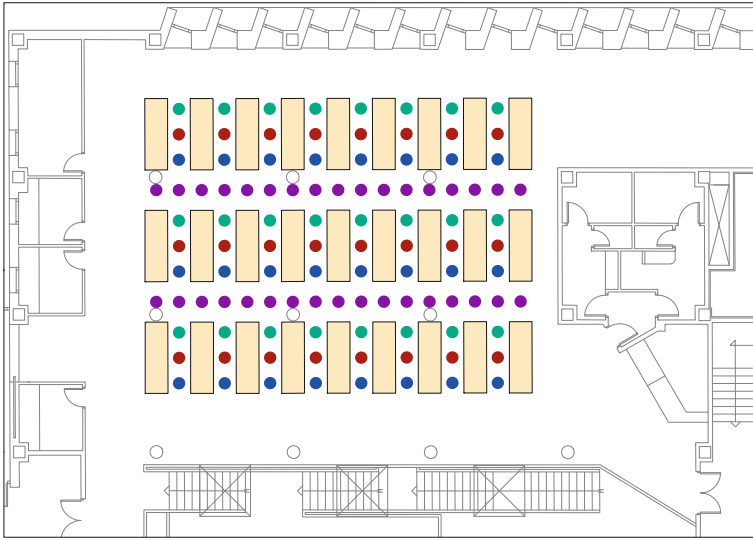


Fig. 3 Collection locations for the test sets. A circle's color identifies to the group to which it belongs: red are groups 1 and 5; blue is group 2; green is group 3; and violet is group 4

Euclidean space. The collection locations are among bookshelves in the third floor of the library building. The database is part of a larger effort to gather data for studying short and long term RSS variations and for developing positioning method robust to those changes. Twelve fingerprints were collected at each location. The tests sets were divided into five groups for their collection. Each group had a particular location distribution and collection directions. Most of the experiments presented in this section used only the training sets. The test sets were only used for the evaluation using the kNN fingerprinting presented at the end of this section.

For the training sets and the groups 1, 2 and 3 of the test sets, the collector (a trained person) faced the “up” direction when collecting the first six fingerprints of each point, and the “down” direction when collecting the other six fingerprints. For groups 4 and 5 of the test sets, the faced directions were “right” and “left” instead of “up” and “down”, respectively. For data dimensionality reduction, the APs detected in less than 5% of the fingerprints were removed. The device used for collection was a Samsung Galaxy S3 smartphone.

Section 3 defines the set determination without establishing any explicit restriction for the regression to use. However, an implicit restriction exists: The regression method should enable both interpolation and extrapolation, because there may be target measurement locations lying outside the convex hull of the locations in the solution set. This implicit restriction is also important because the extrapolation usage is mandatory for environments that, at collection time, contain areas where measurements cannot be taken (e.g., because of a meeting in an office).

As providing recommendations on regression methods for WiFi fingerprinting was not among the goals of our study, we tested only a few regression methods: IDW (interpolation and extrapolation), radial basis function interpolators like those in Ezpeleta et al. (2015), combinations of interpolation (linear, nearest, and natural) and extrapolation (linear, nearest) as provided by MathWorks® (2017a), and SVR as provided by MathWorks® (2017b). We chose SVR, using a Radial Basis Function (RBF) kernel and performing predictor data standardization, as regression method as it provided the best results regarding RSS absolute differences between RSS measurements and estimations and because it has been successfully used in previous studies (Hernández et al. 2017). We suggest to perform the regression method evaluation for a given environment before making a choice. Guides regarding interpolation and extrapolation can be found in Talvitie et al. (2015).

For evaluating the goodness of each set of locations, we have used a metric defined as the maximum value of the AP-wise RSS absolute differences between the original RSS measurements and the estimated ones. We have preferred the maximum difference over other measures (e.g., the mean) that may mask high RSS differences that are significant for distance-based techniques like kNN-based fingerprinting.

4.1 Evaluation for the Initial Month

Section 2 hinted on approaches for determining where to collect the RSS measurements to be used for fitting a regression. This subsection shows the evaluation results of three strategies for determining the collection locations. The strategies, which were applied to the training set corresponding to first month of our WiFi RSS database, are:

1. Random Sets of Points,
2. Manually-defined Sets of Points,
3. Optimized Set of Points.

The first approach considers differently sized sets of random locations. The second approach uses sets manually defined by an expert. The third approach finds a set of locations that establishes a compromise between the set's size and the regression goodness. The following subsections provide more details about each approach and its evaluation.

4.1.1 Random Sets of Points

This is an intuitive approach for the selection of the collection locations. The algorithm proposed in Kanaris et al. (2016) may allow determining the number of measurement locations. We instead decided to explore several numbers of locations, ranging from 6 to 18 points, which accounts for 25–75% of the 24 total locations in the target area, and they represent reasonable effort reductions. Table 1 presents

Table 1 Minimum and maximum values of RSS error metric for sets of randomly chosen locations

Set Size	Metric Min (dBm)	Metric Max (dBm)
6	19.13	44.74
8	15.71	43.01
10	12.12	41.35
12	11.18	42.10
14	11.99	35.58
16	10.52	35.22
18	9.56	28.59

the maximum and minimum of the RSS error metric previously defined. The experiment for each amount of points was repeated 200 times.

Table 1 shows two main facts. First, the more points are used for fitting the regression, the better the estimations are. Second, and more important, the RSS estimation quality heavily depends on the distribution of the randomly chosen locations, as absolute differences between the maximum and minimum metric values are up to 30.92 dBm.

4.1.2 Manually-Defined Sets of Points

As previously seen, selecting random points creates much uncertainty in the quality of the RSS estimations. A logical alternative is to manually define the set of locations. Better choices are done when the extent of the collection locations and the influence of the building layout and the furniture are taken into account. This subsection presents six alternative sets we considered that are likely choices and could provide fine RSS estimations through regression. The process of determining the tentative sets of locations is time-consuming, and it is especially cumbersome due to the large number of alternatives for each set’s size. Table 2 presents the value of the RSS error metric for each alternative set. The ID of each set indicates its amount of locations. Figure 4 shows the location distribution of each set.

Table 2 Values for RSS error metric for manually defined sets of locations

Set ID	Metric (dBm)
6A	21.39
8A	23.45
8B	19.63
12A	23.13
12B	11.90
14A	12.20

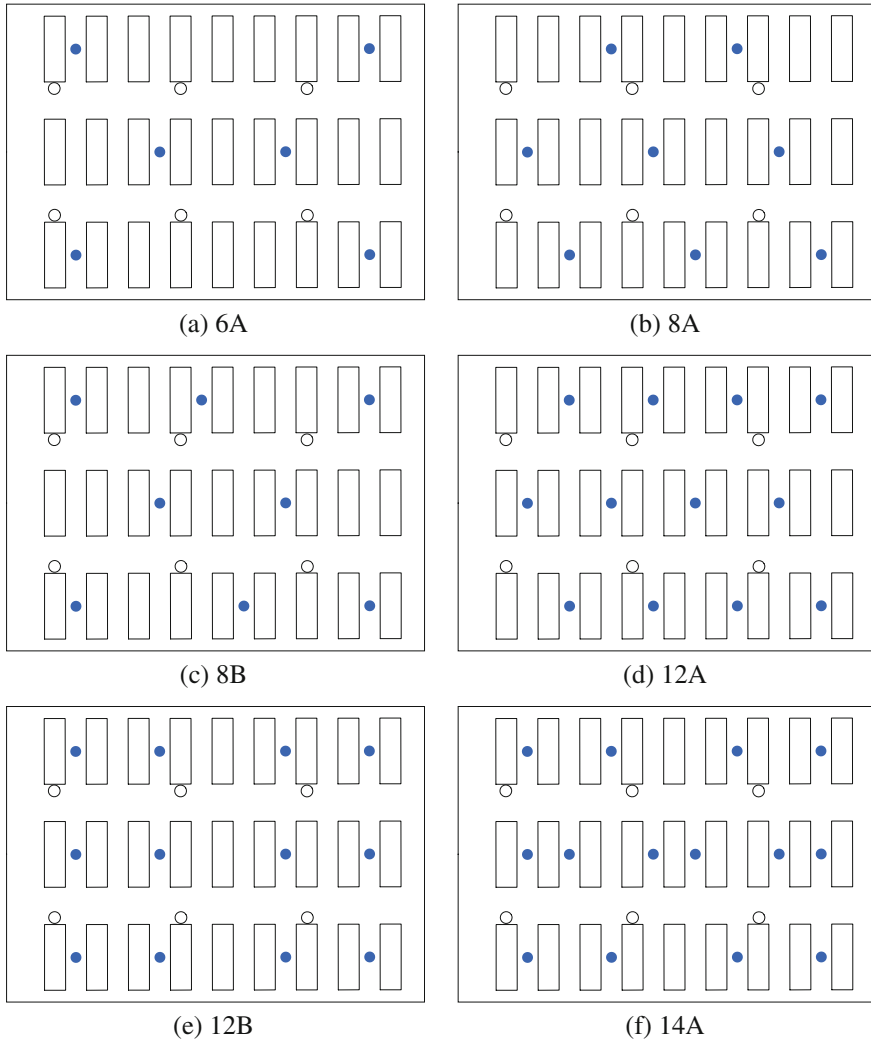


Fig. 4 Manually chosen sets of locations

The results presented in Table 2 reinforce the importance of the distribution of the collection locations. The estimation quality does not strictly decrease with the increase of the number of locations used for regression fit, as seen when comparing the set 6A with set 8A, set 8B with set 12A, and set 12B with set 14A. The locations distribution of each set, as shown in Fig. 4, sheds some light on the previous fact. The convex hulls of sets 6A and 8B include more of the target area than those of sets 8A and 12A, respectively. Nevertheless, the convex hulls of sets 12B and 14A are the same, and the set 12B provide better estimations than set 14A despite having

a smaller number of locations. The above facts lead to conclude that even a well-designed set of locations may not be the best choice. Additionally, a set of locations that is optimal for a given environment, may not be optimal for another environment, a fact that we leave unproven because is beyond the focus of this paper.

4.1.3 Optimized Set of Points

As described in Sect. 3, with an optimization strategy based on a genetic algorithm it is possible to search for fine locations for fitting the regression. Specifically, the genetic algorithm implementation provided in Burjorjee (2009), which is in turn based on Mitchell (1998) was used for the experiments. We defined a population size of 200 individuals, used the fitness function proposed in Sect. 3, and run the algorithm for 100 generations. After testing several values, the numbers of 200 and 100 for population size and algorithm generations were the ones that provided higher stability (reproducibility) in the outputted solution. The obtained elite individual (best set of locations found for fitting a regression) and the solution set (described in Sect. 3) using a higher frequency threshold of 0.9, are depicted in Fig. 5. The value of the RSS error metric for the elite set (11 locations) was 8.8453 dBm, which is lower than any of the values obtained using the previous two strategies. The metric value for the solution set (eight locations) is 11.18 dBm, which is still lower than most of the values obtained using the previous two strategies.

Figure 5 shows two important facts. First, all target locations are contained in the convex hull of both optimized sets, which avoids the usage of extrapolation. Second, and more important, the location distributions of the optimized sets do not resemble those of the strategies explored in the Sect. 4.1.2, nor they are intuitive. Therefore, the locations chosen for fitting a regression should not be random, and determining a small set of locations that provides good estimations when used for fitting a regression, is not a trivial problem.

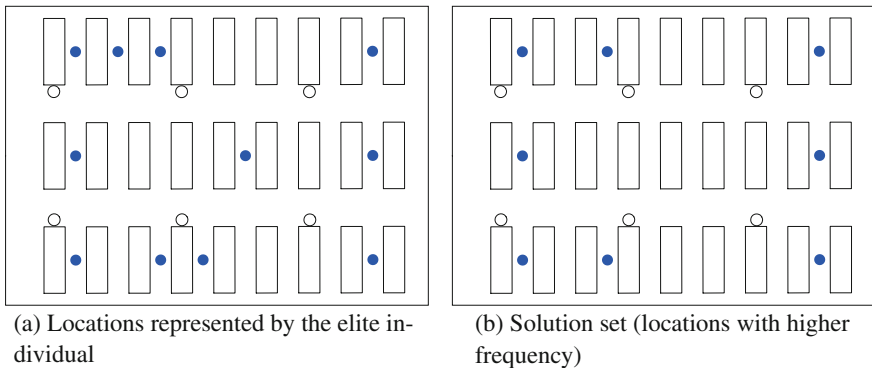


Fig. 5 Set of locations obtained through optimization using a genetic algorithm

The following subsection explores the usage of the solution set for WiFi radio map update, following the procedure presented in Sect. 3.

4.2 Usage of the Solution Set for RSS Radio Map Update

The Sect. 4.1.3 presented the locations that our approach suggested for conducting the periodic updates to the WiFi radio map. This section presents the results of experiments that explored the goodness of those updates along 9 months (month 2–10) in terms of RSS difference between estimations and real measurements, and in terms of the accuracy of a fingerprinting-based IPS.

Regarding RSS differences between estimations and real measurements, the experiments tried three sets of locations and three RSS difference metrics. Table 3 shows the results of these experiments. Each table header indicates the usage of a particular set of locations and a specific RSS difference metric.

To explore the suitability of a set of locations for each month, a regression was fit using their associated measurements of the month training set. Besides the solution set (GA), sets 8A and 8B were also used for regression fitting. The sets 8A and 8B, previously introduced in Sect. 4.1.2, are now used for baseline comparisons. As RSS difference metrics, the experiments used:

1. MRD: The MRD value introduced in Sect. 3.1 for the fitness function definition.
2. Mean: Its value is computed in a way similar to MRD, but the mean value is used instead of the maximum. This metric is included because it is frequently used in the literature for evaluation of WiFi RSS regression methods.
3. MeanP: It is calculated as: Compute the AP-wise RSS absolute differences between RSS measurements and estimations. Compute per each location the mean of those differences. Take the maximum of those mean values. This metric is included because indicates how much the RSS difference may affect a RSS distance-based method like kNN.

Table 3 Values for RSS differences (dBm) according to metrics MRD, Mean and MeanP for sets 8A, 8B and GA

Month	MRD8A	MRD8B	MRDGA	Mean8A	Mean8B	MeanGA	MeanP8A	MeanP8B	MeanPGA
02	21.1	21.2	19.4	1.8	1.5	1.7	5.2	3.7	3.8
03	19.9	20.8	16.9	1.6	1.4	1.5	5.3	3.7	3.0
04	20.5	21.0	18.7	1.7	1.5	1.5	4.6	3.8	3.0
05	19.7	23.8	17.2	1.5	1.3	1.3	4.9	4.6	2.9
06	23.9	21.8	21.9	1.6	1.4	1.5	4.7	3.6	3.7
07	21.5	24.9	26.6	1.5	1.4	1.5	4.8	4.2	3.3
08	26.5	30.0	18.9	1.6	1.4	1.4	5.0	4.3	3.5
09	22.9	25.4	20.2	1.5	1.3	1.2	4.5	3.4	3.3
10	23.7	22.0	18.0	1.8	1.6	1.6	5.3	4.3	4.1

The results presented in Table 3 indicate that the solution set is a better choice than the other two sets as a set of collection locations for periodical updates. Regarding the MRD metric, the solution set provides the best result for most months. It is noticeable that for month number seven, the value for the solution set is 5.1 dBm worse than the one for the set 8A. Some insights on that behavior will be later provided when analyzing the set effects on fingerprinting-based IPS accuracy. Regarding the Mean metric, the solution set is consistently better than the set 8A, and slightly worse than the set 8B for some months. As for the MeanP metric, the solution set is much better than the 8A set. In comparison with the set 8B, the solution set is notably better for five of the months, and only slightly worse for two of them.

The experiments also explored how the localization accuracy of an IPS is affected by the usage of the proposed update approach, i.e., by taking the training RSS measurements of each month only at the solution set and using regression to estimate the RSS values for the remaining locations. As IPS, we tested a kNN fingerprinting approach. Given a training set of fingerprints with known location labels, a query fingerprint, and two parameters specified by the value of k and a distance metric on the fingerprint space, the kNN method finds the k fingerprints in the training set that are closest to the query fingerprint. The location label is estimated as the centroid of the location labels of the selected k closest fingerprints.

To measure the accuracy of an IPS, a test set of query fingerprints is usually used. The location labels are also known for the test set fingerprints, so that, for each fingerprint, the location estimation provided by the IPS and its original location label are used to compute a positioning error distance. In this paper, the positioning distance has been calculated using the Euclidean distance and the localization accuracy has been explored using the 75 percentile of the computed distances for test set.

The tested kNN used the RSS Euclidean distance as fingerprint distance metric. The k parameter value was experimentally determined using the training and test sets of the first month of the WiFi RSS database. Figure 6 shows the resulting localization accuracies. The value of k that provides the best metric value is nine, and it is the one used for kNN in the remaining experiments. This value may appear large, but it is a reasonable value given that 12 fingerprints were taken at each location and no aggregation operation was performed for fingerprints with the same location label.

For comparisons, the experiments included an evaluation of the radio map update at each month using all the training measurements collected at that month. Two updates strategies were tested: Replacement and addition. With the replacement strategy, all training fingerprints collected at one month replaced all fingerprints from the previous month in the WiFi radio map. With the addition strategy, the fingerprints of each month were added without any replacement or deletion from the previous months' fingerprints. The kNN method was used to estimate, for each month, the locations associated to the fingerprints of the test set of that month. Figure 7 shows the behavior of each update strategy along the time.

The strategy of addition provides values for the localization error metric that are smaller and smoother than those provided by the replacement strategy. The metric values for the strategy of addition ranges from 3.25 to 2.84 m. For the replacement strategy, however, the localization error metric ranges from 4.10 to 3.14 m. The

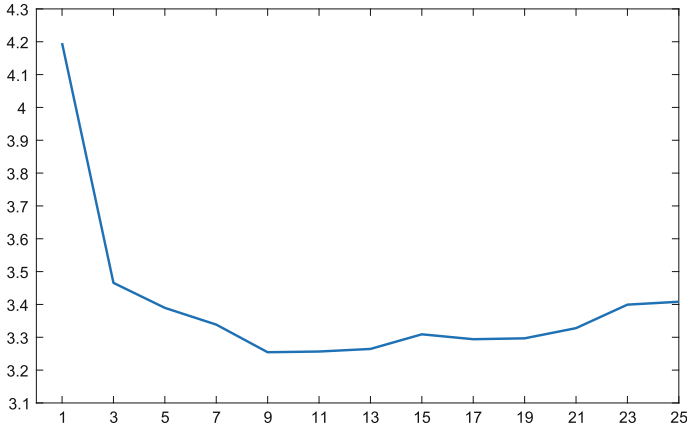


Fig. 6 75 percentile of positioning error using kNN for the first month

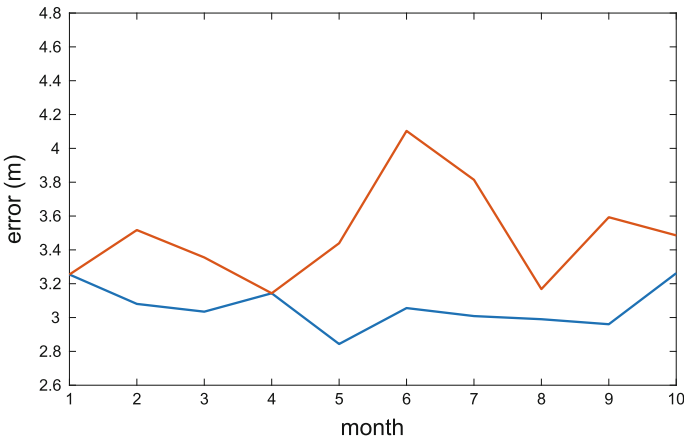


Fig. 7 Comparison of the strategies of replacement (red) and addition (blue). Measurements for all locations are available

months 6 and 7 have the highest metric values, which may indicate that the training values for those months were not as good (representative) as they were for other months.

The evaluation of using the solution set for radio map update was conducted as follows. For each month, the solution set was used to fit a regression, and the RSS values were estimated for the rest of locations. However, the estimation provides one fingerprint per location, while the training and test sets in the database contain 12 fingerprints per location. Additionally, the k value determined above for the kNN fingerprinting is the best under the assumption that there are 12 fingerprints per point. Therefore, we decided to create 12 fingerprints per location using the one fingerprint per location obtained through regression estimation and adding a random value.

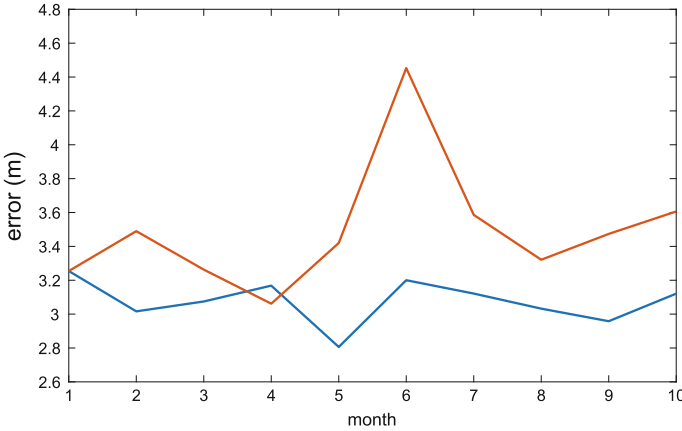


Fig. 8 Comparison of the strategies of replacement (red) and addition (blue). Measurements are available only for the solution set

In the training set from first month, the AP-wise standard deviations values were less than 6 dBm in 80% of cases. The added random value is then uniformly chosen in the interval $[-6;6]$. The random value addition is specific to the evaluation presented in this study and will not be needed for an IPS radio map update, for which it may be desirable to collect only one fingerprint per point. The fingerprints newly estimated for each month were considered for radio map update following the strategies of replacement and addition described above. Figure 8 presents the localization accuracy metric values for both strategies.

The results obtained using the strategy of addition and the RSS estimations from the solution set are very similar to those using that strategy and the measurements available for all locations. The localization error metric for the strategy of addition ranges from 3.25 to 2.81 m, which is the same interval obtained when using the RSS measurements for all locations. The strategy of replacement showed larger metric values, with higher variations, than the addition strategy. When compared to using the same strategy and the measurements from all locations, the usage of the estimations from the solution set caused larger variability, with the metric values ranging from 4.45 to 3.06 m, having a steeper variation for month 6.

The above results suggest that the usage of the solution set as collection locations for WiFi radio map update is a reasonable choice for the tested environment. The approach of determining the solution set is automatic, so that the specialized and cumbersome task of manually determining a proper set of collections locations is avoided. The MeanP value, i.e., the maximum of the location-wise mean RSS differences between estimations and measurements, was lower than the detected AP-wise standard deviation for all months. Additionally, the accuracy of the tested kNN fingerprinting had a similar behavior when using measurements from all location and estimations to when using the estimations obtained from the measurements of locations in the solution set.

The previous results have been obtained for an initial radio map collection month and across nine months of radio map updates. During the ten months period, no drastic changes in the presence or power configuration of the APs were observed, apart from wireless networks with very low presence in the data. Therefore, we advise the usage of our proposal for environments that allow an initial relatively dense collection and as long as no drastic change happens to the detected APs. Such changes could be detected, for example, by reviewing the appearance or disappearance of AP with strong RSS values in a significant number of fingerprints. We acknowledge that the required initial collection is costly, but if the cost is assumed, the knowledge of the locations for performing periodic updates could translate into a better IPS performance.

5 Conclusions

This paper has presented an approach for determining a subset of the target locations in a goal area where RSS measurements are suggested to be collected for periodic radio map updates. The measurements at the remaining locations are proposed to be obtained through regression. The subset, called the solution set, is determined from initial measurements taken at all target locations through an optimization approach based on a genetic algorithm. The proposed approach was tested using a database collected over ten months in a university library. The regression method tested in the experiments was SVR regression. The experiments' results support the suitability of using the estimation determined using the measurements of the solution set for periodic WiFi radio map updates. The suitability has been shown in terms of the RSS estimation accuracy and in terms of its effects on the localization accuracy of a fingerprint-based IPS.

We consider that the proposed approach may be of particular interest for future efforts devoted to automate the fingerprint collection process using dedicated devices or robotic agents. Future continuation lines of this study include (1) testing our approach in a larger and less densely collected environment, (2) explore the effects of drastic changes in the environment APs, and (3) explore the variant to our approach proposed in Sect. 2 for finding challenging sets for testing regression approaches used in WiFi fingerprinting.

Acknowledgements Germán M. Mendoza-Silva gratefully acknowledges funding from grant PRE-DOC/2016/55 by Universitat Jaume I.

References

- Ali MU, Hur S, Park Y (2017) Locali: calibration-free systematic localization approach for indoor positioning. *Sensors* 17(6). <https://doi.org/10.3390/s17061213>
- Alonazi A, Ma Y, Tafazolli R (2015) Less-calibration wi-fi-based indoor positioning. In: 2015 IEEE international conference on communications (ICC), pp 2733–2738. <https://doi.org/10.1109/ICC.2015.7248739>
- Bong W, Kim YC (2012) Fingerprint wi-fi radio map interpolated by discontinuity preserving smoothing. In: International conference on hybrid information technology. Springer, pp 138–145
- Burjorjee KM (2009) SpeedyGA: a fast simple genetic algorithm. <https://es.mathworks.com/matlabcentral/fileexchange/15164-speedyga--a-fast-simple-genetic-algorithm>
- Ezpeleta S, Claver JM, Pérez-Solano JJ, Martí JV (2015) Rf-based location using interpolation functions to reduce fingerprint mapping. *Sensors* 15(10):27, 322–27, 340
- Gu Y, Chen M, Ren F, Li J (2016a) HED: handling environmental dynamics in indoor WiFi fingerprint localization. In: 2016 IEEE wireless communications and networking conference, pp 1–6. <https://doi.org/10.1109/WCNC.2016.7565019>
- Gu Z, Chen Z, Zhang Y, Zhu Y, Lu M, Chen A (2016b) Reducing fingerprint collection for indoor localization. *Comput Commun* 83:56–63. <https://doi.org/10.1016/j.comcom.2015.09.022>
- He S, Chan SHG (2016) Wi-Fi fingerprint-based indoor positioning: recent advances and comparisons. *IEEE Commun Surv Tutor* 18(1):466–490. <https://doi.org/10.1109/COMST.2015.2464084>
- Hernández N, Ocaña M, Alonso JM, Kim E (2017) Continuous space estimation: increasing wif-based indoor localization resolution without increasing the site-survey effort. *Sensors* 17(1)
- Hossain AKMM, Soh WS (2015) A survey of calibration-free indoor positioning systems. *Comput Commun* 66:1–13. <https://doi.org/10.1016/j.comcom.2015.03.001>
- Jan SS, Yeh SJ, Liu YW (2015) Received signal strength database interpolation by kriging for a wi-fi indoor positioning system. *Sensors* 15(9):21, 377–21, 393. <https://doi.org/10.3390/s150921377>
- Joshi S, Boyd S (2009) Sensor selection via convex optimization. *IEEE Trans Signal Process* 57(2):451–462
- Kanaris L, Kokkinis A, Fortino G, Liotta A, Stavrou S (2016) Sample size determination algorithm for fingerprint-based indoor localization systems. *Comput Netw* 101:169–177. <https://doi.org/10.1016/j.comnet.2015.12.015>
- Krumm J, Platt J (2003) Minimizing calibration effort for an indoor 802.11 device location measurement system. Microsoft Research, November
- Lee M, Han D (2012) Voronoi tessellation based interpolation method for wi-fi radio map construction. *IEEE Commun Lett* 16(3):404–407. <https://doi.org/10.1109/LCOMM.2012.020212.111992>
- Li B, Wang Y, Lee HK, Dempster A, Rizos C (2005) Method for yielding a database of location fingerprints in wlan. *IEE Proc—Commun* 152(5):580–586. <https://doi.org/10.1049/ip-com:20050078>
- Li L, Shen J, Zhao C, Moscibroda T, Lin JH, Zhao F (2014) Experiencing and handling the diversity in data density and environmental locality in an indoor positioning service. *ACM—Association for Computing Machinery*
- Lin K, Chen M, Deng J, Hassan MM, Fortino G (2016) Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings. *IEEE Trans Autom Sci Eng* 13(3):1294–1307. <https://doi.org/10.1109/TASE.2016.2543242>
- Liu C, Kiring A, Salman N, Mihaylova L, Esnaola I (2015) A kriging algorithm for location fingerprinting based on received signal strength. In: 2015 sensor data fusion: trends, solutions, applications (SDF), pp 1–6. <https://doi.org/10.1109/SDF.2015.7347695>
- Macho-Pedroso R, Domingo-Perez F, Velasco J, Losada-Gutierrez C, Macias-Guarasa J (2016) Optimal microphone placement for indoor acoustic localization using evolutionary optimization.

- In: 2016 international conference on indoor positioning and indoor navigation (IPIN), pp 1–8. <https://doi.org/10.1109/IPIN.2016.7743609>
- Majeed K, Sorour S, Al-Naffouri TY, Valaee S (2016) Indoor localization and radio map estimation using unsupervised manifold alignment with geometry perturbation. *IEEE Trans Mob Comput* 15(11):2794–2808. <https://doi.org/10.1109/TMC.2015.2510631>
- MathWorks® (2017a) Extrapolating scattered data, in MATLAB® R2017b. <https://es.mathworks.com/help/matlab/math/scattered-data-extrapolation.html>
- MathWorks® (2017b) Support vector machine regression, in MATLAB® R2017b and statistics and machine learning toolbox™. <https://es.mathworks.com/help/stats/support-vector-machine-regression.html>
- Mitchell M (1998) An introduction to genetic algorithms. MIT press
- Pei L, Zhang M, Zou D, Chen R, Chen Y (2016) A survey of crowd sensing opportunistic signals for indoor localization. *Mob Inf Syst* 2016
- Ranieri J, Chebira A, Vetterli M (2014) Near-optimal sensor placement for linear inverse problems. *IEEE Trans Signal Process* 62(5):1135–1146
- Rowaihy H, Eswaran S, Johnson M, Verma D, Bar-Noy A, Brown T, La Porta T (2007) A survey of sensor selection schemes in wireless sensor networks. *Proc SPIE* 6562:A1–A13
- Roy V, Simonetto A, Leus G (2016) Spatio-temporal sensor management for environmental field estimation. *Signal Process* 128:369–381
- Talvitie J, Renfors M, Lohan ES (2015) Distance-based interpolation and extrapolation methods for rss-based localization with indoor wireless signals. *IEEE Trans Veh Technol* 64(4):1340–1353. <https://doi.org/10.1109/TVT.2015.2397598>
- Wang B, Chen Q, Yang LT, Chao HC (2016) Indoor smartphone localization via fingerprint crowdsourcing: challenges and approaches. *IEEE Wirel Commun* 23(3):82–89. <https://doi.org/10.1109/MWC.2016.7498078>
- Xiao Z, Wen H, Markham A, Trigoni N (2015) Robust indoor positioning with lifelong learning. *IEEE J Select Areas Commun* 33(11):2287–2301. <https://doi.org/10.1109/JSAC.2015.2430514>
- Yang S, Dessai P, Verma M, Gerla M (2013) Freeloc: calibration-free crowdsourced indoor localization. In: 2013 proceedings IEEE INFOCOM, pp 2481–2489. <https://doi.org/10.1109/INFOCOM.2013.6567054>
- Yao L, Sethares WA, Kammer DC (1993) Sensor placement for on-orbit modal identification via a genetic algorithm. *AIAA J* 31(10):1922–1928
- Yiu S, Dashti M, Claussen H, Perez-Cruz F (2017) Wireless rssi fingerprinting localization. *Signal Process* 131:235–244
- Zhu JY, Zheng AX, Xu J, Li VOK (2014) Spatio-temporal (s-t) similarity model for constructing wifi-based rssi fingerprinting map for indoor localization. In: 2014 international conference on Indoor positioning and indoor navigation (IPIN), pp 678–684. <https://doi.org/10.1109/IPIN.2014.7275543>