2020 International Conference on Localization and GNSS (ICL-GNSS), 2–4 June 2020, Tampere, Finland

New Cluster Selection and Fine-grained Search for k-Means Clustering and Wi-Fi Fingerprinting

Joaquín Torres-Sospedra*,†, Darwin Quezada-Gaibor†, Germán M. Mendoza-Silva†, Jari Nurmi‡, Yevgeni Koucheryavy‡ and Joaquín Huerta†
*UBIK Geospatial Solutions S.L., Castellón, Spain
†Institute of New Imaging Technologies, Universitat Jaume I, Castellón, Spain
‡Electrical Engineering Unit, Tampere University, Tampere, Finland

***Abstract*—Wi-Fi fingerprinting is a popular technique for Indoor Positioning Systems (IPSs) thanks to its low complexity and the ubiquity of WLAN infrastructures. However, this technique may present scalability issues when the reference dataset (radio map) is very large. To reduce the computational costs, k-Means Clustering has been successfully applied in the past. However, it is a general-purpose algorithm for unsupervised classification. This paper introduces three variants that apply heuristics based on radio propagation knowledge in the coarse and fine-grained searches. Due to the heterogeneity either in the IPS side (including radio map generation) and in the network infrastructure, we used an evaluation framework composed of 16 datasets. In terms of general positioning accuracy and computational costs, the best proposed k-means variant provided better general positioning accuracy and a significantly better computational cost –around 40% lower– than the original k-means.**

***Index Terms*—Wi-Fi Fingerprinting; Clustering; RSS.**

I. Introduction

The user's position is key for many current applications and services [1]. While GNSS receivers embedded in modern smartphones enable positioning outdoors, GNSS-denied scenarios such as indoors –where humans spend more than 80% of their time [2, 3] – require other technological solutions.

Wi-Fi fingerprinting is a popular technique for position estimation due to its low deployment costs and the simplicity of the positioning algorithm [4]. The notion behind this technique is that a fingerprint – the Received Signal Strength (RSS) from the nearby Access Points (APs) – is representative of the position where it was taken. For a fingerprint taken at an unknown position (operational fingerprint), its position can be computed using the kNearest Neighbour (NN) algorithm and a dataset with reference fingerprints taken at known positions.

Although this solution is widely used, the distance to all the reference fingerprints must be calculated to get the k nearest fingerprints and estimate the final position. Thus, it might suffer from scalability problems if the positioning algorithm is run in a low-profile device (e.g., a smart watch) or provided by a server accessed by multiple concurrent users. Some authors have applied clustering models to group similar fingerprints of the radio map [5, 6, 7, 8]. Later, the computation of the nearest neighbors is split into two searches: the coarse search and the fine-grained search. The coarse search is devoted to calculate the similarity of the operational fingerprint to all the clusters representatives, whereas the fine-grained search is devoted to calculate the similarity of the operational fingerprint to respect all the reference fingerprints belonging to the selected cluster.

Some alternative approaches to clustering use knowledge on the radio signal propagation to filter the radio map on the fly and reduce the computational costs. Some approaches identify the strongest AP in the operational fingerprint and then restrict the comparison to either the reference fingerprints where that AP was detected [9, 10, 11] or the reference fingerprints where that AP was also the strongest one [12]. In general, those filters present a trade-off between the accuracy and cost dimensions. i.e., the smaller the reduced/filtered radio map is, the worse the positioning error is. Current IPSs require solutions that provide better compromises between the two dimensions.

Although k-Means provides a good trade-off between the two dimensions, we identified two main problems. First, computing the similarity to all the clusters – coarse search – for every positioning request is inefficient if the number of clusters and the environment area are both too large [11]. Second, the fingerprints might not be equally distributed among the clusters. The fine-grained search in clusters much larger than the rest may degrade the benefits obtained from clustering.

We introduce three new more computationally efficient variants of k-means clustering based on knowledge about signal propagation. The main contributions of this paper are:

- A new computationally-efficient way to reduce the clusters in the coarse search.
- Two new computationally-efficient ways to further reduce the reduced radio maps in the fine-grained search.
- A reproducible evaluation that comprises an extensive comparison on different scenarios.

The remaining of this paper is organized as follows. Section II briefly reviews related works on clustering and Wi-Fi fingerprinting. Section III describes the integration of the k-means clustering algorithm in Wi-Fi fingerprinting and our proposed variants. Section IV introduces the experimental setup and shows the empirical results. Section V draws the main conclusions of this work.

Corresponding Author: J. Torres-Sospedra (torres@ubikgs.com) The authors gratefully acknowledge funding from Ministerio de Ciencia, Innovación y Universidades (INSIGNIA, PTQ2018-009981); European Union's H2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement No.813278 (A-WEAR, http://www.a-wear.eu/); and Universitat Jaume I (PREDOC/2016/55).

II. Related work

Given that Wi-Fi fingerprint matching and large radio map sizes account for important computational loads [13, 14], several authors applied approaches that solve the load issue while also maintaining or improving the positioning accuracy. Some authors tackled the issue using general-purpose unsupervised learning models. They applied the divide and conquer approach and, somehow, broke down the whole radio map into smaller pieces. This is the case of clustering approaches like k-means [5, 6] and Affinity Propagation [7, 8]. In contrast to clustering, other authors proposed optimization heuristics based on their knowledge on signal propagation and Wi-Fi fingerprinting [11, 12, 15, 16]. Most of the heuristics are based on the fact that the RSS value somehow indicate the distance of the measurement device (e.g., smartphone or smartwatch) to the AP.

Shin *et al.* [5] proposed a tracking system that automatically builds a labeled topological map and estimates the users' location. In their place learning stage, they applied k-means to automatically organize the spaces in an unknown environment. According to the authors, the clustered topological radio map could determine the division of the operational area.

Abdullah *et al.* [17] slightly modified the k-means model by applying the Bregman divergence as distance for clustering formation, but still used the Euclidean distance for cluster determination in the online phase. The authors tested their proposal in terms of positioning accuracy against the original k-means and Affinity Propagation in a medium sized area.

Cramariuc *et al.* [18] tested k-means using Euclidean distance in the coordinate space and Affinity Propagation using Log-Gaussian distance in the feature (RSS) space for clustering formation in large multi-floor environments. They stated that the Affinity Propagation based on Log-Gaussian RSS distance obtained the largest time reductions while the k-means based on Euclidean coordinates distance obtained the best error, when compared among them and to non-clustering weighted k-NN approach.

Park *et al.* [19] tested k-means using on Euclidean distance in the feature space for clustering formation in a small environment. The cluster determination in the online phase used a probability distance.

Anuwatkun *et al.* [20] tried k-means using on Euclidean distance in the feature space for clustering formation in a small environment. Instead of using the RSS values directly, the authors used the strength difference among the APs.

In contrast to the previous works, k-means has also been used for coordinate-based clustering [18, 21], floor-wise fingerprint clustering [6] and, even, to cluster the positions of the list of nearest neighbors provided by the k-NN algorithm [22].

All the previous papers have something in common, the knowledge on the radio signal propagation seems not to be fully exploited to, for instance, reduce the computational load on the selection of the best cluster.

III. k-means and proposed variants to reduce the computational load in the on-line stage

The k-means method [23] automatically divides the feature space into k non-overlapping regions (clusters) represented by their centroids (the mean of the cluster's fingerprint vectors). The clusters generation starts with random centroids, which are iteratively adapted by minimizing the intra-cluster distances. The algorithm minimizes the variances of the samples that fall within the cluster.

In this work, we used the enhanced cluster initialization procedure proposed in Arthur *et al.* [24] rather than the completely random one. Note that the improved initialization is also stochastic and the resulting clusters depend on the initial cluster representatives.

The information from the clusters is integrated in Wi-Fi fingerprinting using two phases:

- The off-line phase, which executes k-means over the reference fingerprints, obtaining k clusters. We could say that k-means provides a local version of the radio map for every cluster.
- The on-line phase, which finds the reference fingerprints most similar to the operational fingerprint in two steps. The first step selects the cluster whose centroid is the most similar to the operational fingerprint. The second step performs a fine-grained search on the selected cluster's fingerprints.

Under ideal conditions (uniform distribution of samples among the clusters) and choosing $k = \sqrt{n}$, the best asymptotic computation time of cluster-based fingerprinting method is $\mathcal{O}(\sqrt{n})$, where n is the number of samples in the radio map as shown in Figure 1.

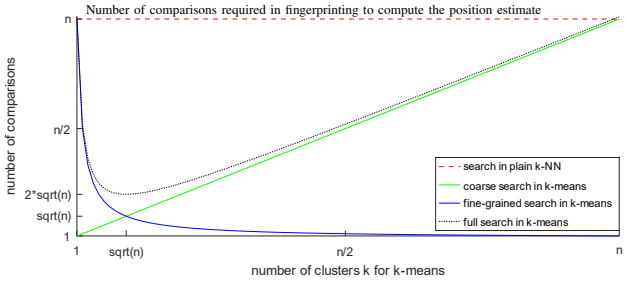


Fig. 1. Computational load as number of vector comparisons of Wi-Fi fingerprinting with and without k-means clustering

Although k-means and k-NN are commonly used together, the meaning of the variable k in both models is quite different. It stands for the number of nearest neighbors to perform a supervised classification/regression in k-NN, whereas it stands for the number of clusters generated by the unsupervised algorithm in k-means.

In the offline stage, the three variants we propose determine the clusters (and their centroids) using k-means. In addition, they analyse the clusters to find information that is relevant for improving the search times in the on-line stage.

A. Proposed Variant 1: Improved coarse search

As in the traditional fingerprint model, a scalability problem may occur if the number of clusters is large. Computing the similarity of the operational fingerprint to all the clusters might be too inefficient. We propose an improved coarse search.

In the off-line stage, this variant finds a function f_1 that maps an AP to the set of clusters that are relevant for it, storing all the mappings. A cluster is said to be relevant for i$^{\text{th}}$ AP if the cluster contains at least one fingerprint $fp = (r_1, \ldots, r_{na})$ for which $|r_{max} - r_i| \leq \rho, 1 \leq i \leq na$, being na the number of detected APs, r_{max} the strongest RSS value of fp and ρ a predefined threshold. The APs that do not map to empty sets are marked as operative.

In the on-line stage, for an operational fingerprint, the operative AP that reports the strongest RSS signal is determined. The function f_1 is then used to get a cluster set for that AP using the pre-calculated mappings. Later, the cluster selection in the coarse search is performed on that cluster set, using the common approach of selecting the cluster whose centroid is the most similar to the operational fingerprint. This variant performs the fine-grained search by applying k-NN directly over the selected cluster's fingerprints.

B. Proposed Variant 2: Soft-filtered fine-grained search

The k-means model does not guarantee that generated clusters are balanced. Therefore, we improved in this variant the fine-grained search for oversized clusters.

The second variant adds to the first variant a filtering step in the fine-grained search. The filtering is applied to oversized clusters whose number of fingerprints exceeds four times $\frac{n}{c}$, where n is the number of reference fingerprints in the entire radio map and c is the number of clusters.

In the off-line stage, this variant determines an additional function f_2 for oversized clusters. This function maps an AP and a cluster to the subset of the fingerprints that are relevant to that AP and belong to that cluster. In this function, a fingerprint is deemed relevant for an AP if it contains a valid RSS value for the AP.

In the online-stage, the AP is determined and a cluster is selected as explained for Variant 1. If the cluster is oversized, f_2 is then used for that cluster and AP to obtain the subset of fingerprints where the fine-grained search is performed, i.e, over which the k-NN is applied. Otherwise, the fine-grained search is applied as explained for Variant 1.

C. Proposed Variant 3: Hard-filtered fine-grained search

The third variant is based on the second one, defining f_2 in a more restrictive way. For this variant, a fingerprint from a cluster is only considered relevant for i$^{\text{th}}$ AP if the fingerprint $fp = (r_1, \ldots, r_{na})$ satisfies that $|r_{max} - r_i| \leq \rho, 1 \leq i \leq na$, being na, r_{max} and ρ as defined for f_1 in Variant 1.

In the online-stage, the coarse and fine-grained searches are applied as explained for Variant 2.

IV. Experiments and Results

A. Experimental Setup

Clustering has been explored many times in the IPS literature. However, the diversity in implementation details, evaluation criteria and evaluation scenarios prevents credible comparisons using the reported results. Thus, we created an experimental setup that includes the k-NN as core IPS, two sets of hyperparameters for k-NN (*Simple Configuration* and *Best Configuration*), 3 variants for k-means, 16 datasets and 10 execution runs. The clusters have been randomly generated ensuring that k-means and the 3 variants share the same initialization for each dataset and execution run.

The hyperparameters for k-NN are the RSS representation, and the k value and the distance function for k-NN [25]. *Simple Conf.* stands for $k = 1$, Manhattan distance and positive data representation. *Best Conf.* stands for the hyperparameter configuration that reported the lowest positioning error for a dataset after evaluating 144 alternatives.

The datasets were collected at the Tampere University [6, 18, 26], University Jaume I [27, 28], University of Mannheim [29], and University of Minho. Supplementary materials, with method implementation and dataset explanation, are available in Zenodo [30] for research reproducibility.

Finally, the results collected for this paper are the mean 3D positioning error (ϵ_{3D}) and the computational time (τ_{DB}) resulting from processing all the operational fingerprints. Due to the heterogeneity of the datasets, we report the normalized values, $\tilde{\epsilon}_{3D}$ and $\tilde{\tau}_{DB}$, against the results from a baseline method –plain k-NN with the *Simple Configuration*. Due to the length limit, we report the average of the normalized values for the 16 datasets. Table I shows in the last row how the average of the two metrics, $\tilde{\epsilon}_{3D}$ and $\tilde{\tau}_{DB}$, is calculated for plain k-NN. The experiments were performed in a computer with Intel Core i7-8700 CPU, 16 GB of RAM and Octave 4.0.3.

TABLE I
Positioning error and computation time for Simple and Best parameter configurations using plain k-NN for each dataset.

	Simple Conf.				Best Conf.			
Database	ϵ_{3D}	τ_{DB}	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	ϵ_{3D}	τ_{DB}	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$
DSI 1	4.95	12.23	1	1	3.79	14.11	0.77	1.15
DSI 2	4.95	5.18	1	1	3.8	15.35	0.77	2.97
LIB 1	3.02	46.25	1	1	2.48	42.79	0.82	0.93
LIB 2	4.18	46.17	1	1	2.27	139.69	0.54	3.03
MAN 1	2.82	156.01	1	1	2.06	156.4	0.73	1
MAN 2	2.47	14.37	1	1	1.86	22.3	0.75	1.55
SIM	3.24	254.25	1	1	2.41	232.13	0.74	0.91
TUT 1	9.59	18.88	1	1	4.45	58.84	0.46	3.12
TUT 2	14.37	2.76	1	1	8.09	3.21	0.56	1.16
TUT 3	9.59	79.5	1	1	8.55	93.89	0.89	1.18
TUT 4	6.36	79.87	1	1	5.4	293.25	0.85	3.67
TUT 5	6.92	11.98	1	1	5.26	39.07	0.76	3.26
TUT 6	1.94	624.81	1	1	1.91	728.08	0.98	1.17
TUT 7	2.69	511.79	1	1	2.24	599.18	0.83	1.17
UJI 1	10.81	599.87	1	1	6.56	697.81	0.61	1.16
UJI 2	8.05	2938.38	1	1	6.09	4678.64	0.76	1.59
Average	-	-	1	1	-	-	0.739	1.814

B. Results

Table II shows the results for four models of Wi-Fi fingerprinting based on k-NN: (1) plain k-NN, without any optimization; (2) Moreira, which applies the heuristic proposed by Moreira *et al.* [12], (3) Gallagher, optimized as proposed by Gallagher *et al.* [11], and (4) k-means. For the later model, we considered 3 values of k for k-means: 25, $rfp1 = \sqrt{n}$ and $rfp2 = \frac{n}{25}$, where n is the number of reference samples.

TABLE II
GENERAL NORMALIZED RESULTS FOR TRADITIONAL METHODS.

	Simple Conf.		Best Conf.	
Method	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$
plain k-NN	1.000	1.000	0.739	1.814
Moreira	1.154	0.068	1.003	0.097
Gallagher	0.977	0.357	0.751	0.599
k-means $k = 25$	1.029	0.100	0.871	0.185
k-means $k = rfp1$	1.048	0.073	0.890	0.127
k-means $k = rfp2$	1.059	0.076	0.919	0.122

For all models, the *Best Configuration* is providing significantly better accuracy than the *Simple Configuration* at the expense of a significantly higher computational cost. The best configuration includes computationally expensive distance metrics, such as Log-Gaussian Distance [18], in some datasets.

As expected, k-NN model reports the largest computational times. The Moreira model provides the lowest general computational cost in the two configuration cases. However, it provides the highest mean positioning error. In contrast, the Gallagher model has an accuracy similar to the plain k-NN model but the time cost is just reduced to a third at best.

The solutions based on the k-means model provide a good trade-off between the accuracy and time cost dimensions. Although their mean accuracies are slightly worse than those obtained for the other models, their mean computational cost is reduced more than ten times. Figures 2–4 introduce additional analyses on the clusters generated by k-means, considering all evaluated operational fingerprints.

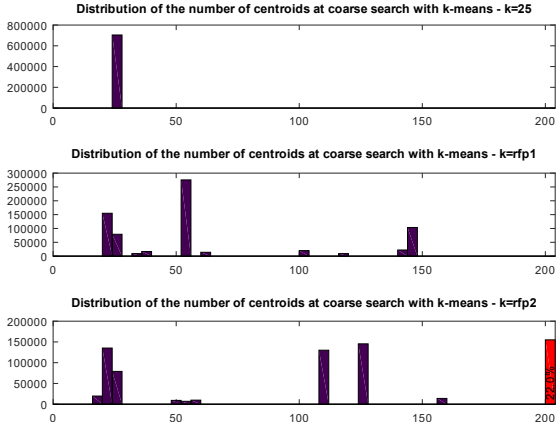


Fig. 2. Histograms of the number of cluster's centroid comparisons (coarse search in k-means) for each operational sample (> 200 in red)

Figure 2 shows the clusters involved in the coarse search, which can be fixed using the same k in all datasets. However, the number of clusters varies when they depend on a heuristic. For the case of $k = rfp1$, the majority of coarse searches involve more than 50 clusters, reaching almost 150 in some cases. A similar behavior is obtained in $k = rfp2$, where the coarse search involves more than 200 clusters in 22% of cases.

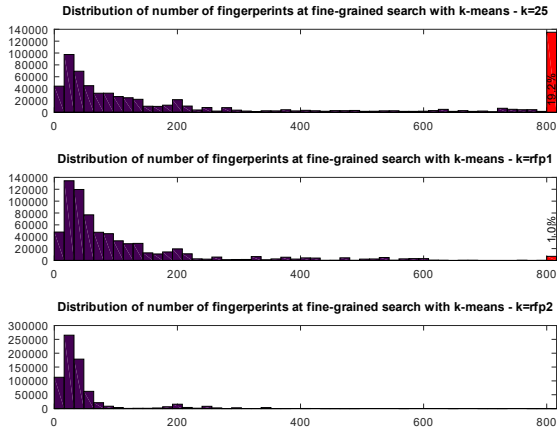


Fig. 3. Histogram of the number of fingerprint comparisons (fine-grained search in k-means) for each operational sample (> 800 in red)

Figure 3 shows that the number of fingerprints in the coarse search is usually low, less than 200 in the vast majority of cases. In $k = 25$, the fine-grained search involve more than 800 reference samples in 19.2% of cases. Having a heavy fine-grained search might happen when the dataset is large and k is too low, but also when the clusters are not equally distributed.

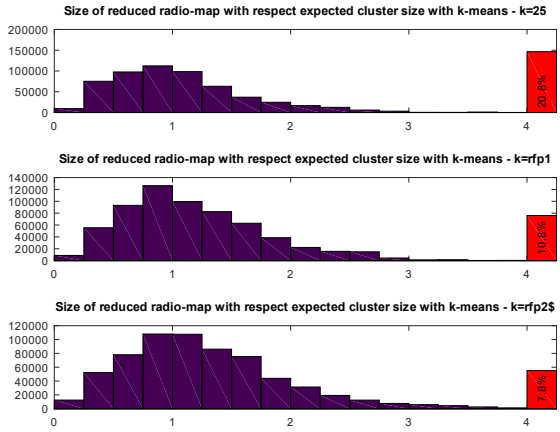
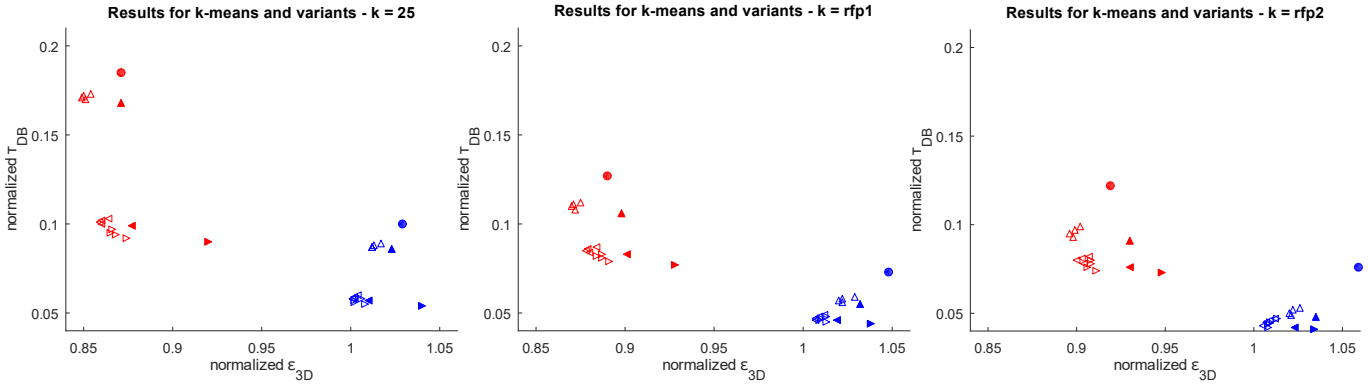


Fig. 4. Histogram of the ratio $\frac{a}{e}$. a is the number of fingerprint comparisons (fine-grained search in k-means) and e is number of comparisons to be performed if the clusters had the same size for a radio map (> 4 in red)

Figure 4 shows that the relative cluster size with respect the expected size –i.e. equally distributed partition with $\frac{n}{c}$ samples per cluster– is usually around 1. However, it is 4 times higher than expected in 20.8% (k=25), 10.8% (k=*rfp1*) and 7.8% (k=*rfp2*) of cases. k-means provides unbalanced subsets of the radio map, specially in complex datasets with multiple devices and a non-regular spatial distribution of reference points

TABLE III
GENERAL NORMALIZED RESULTS FOR THE THREE PROPOSED VARIANTS UNDER DIFFERENT PARAMETRIZATION CONDITIONS.

	k-means - Variant 1				k-means - Variant 2				k-means - Variant 3			
	Simple Conf.		Best Conf.		Simple Conf.		Best Conf.		Simple Conf.		Best Conf.	
	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$	$\tilde{\epsilon}_{3D}$	$\tilde{\tau}_{DB}$
k=25 ρ=00	1.023	0.086	0.871	0.168	1.010	0.057	0.877	0.099	1.040	0.054	0.920	0.090
k=25 ρ=03	1.012	0.087	0.851	0.170	1.001	0.058	0.860	0.100	1.008	0.055	0.874	0.092
k=25 ρ=06	1.012	0.087	0.849	0.171	1.001	0.058	0.859	0.101	1.002	0.056	0.868	0.094
k=25 ρ=09	1.013	0.088	0.850	0.172	1.002	0.059	0.860	0.102	1.002	0.057	0.865	0.095
k=25 ρ=12	1.017	0.089	0.854	0.173	1.004	0.060	0.864	0.103	1.006	0.058	0.866	0.097
$k = rfp1$ ρ=00	1.032	0.055	0.898	0.106	1.019	0.046	0.901	0.083	1.038	0.044	0.928	0.077
$k = rfp1$ ρ=03	1.022	0.056	0.872	0.108	1.007	0.046	0.880	0.084	1.013	0.045	0.891	0.079
$k = rfp1$ ρ=06	1.020	0.057	0.870	0.110	1.007	0.047	0.878	0.085	1.009	0.046	0.887	0.081
$k = rfp1$ ρ=09	1.022	0.058	0.871	0.111	1.009	0.048	0.879	0.086	1.010	0.047	0.884	0.082
$k = rfp1$ ρ=12	1.029	0.059	0.875	0.112	1.012	0.049	0.884	0.087	1.013	0.048	0.887	0.083
$k = rfp2$ ρ=00	1.035	0.048	0.930	0.091	1.023	0.042	0.930	0.076	1.034	0.041	0.948	0.073
$k = rfp2$ ρ=03	1.021	0.049	0.898	0.093	1.005	0.043	0.904	0.078	1.008	0.042	0.911	0.074
$k = rfp2$ ρ=06	1.020	0.050	0.896	0.095	1.007	0.045	0.900	0.080	1.008	0.044	0.906	0.076
$k = rfp2$ ρ=09	1.022	0.052	0.899	0.097	1.009	0.046	0.903	0.081	1.009	0.045	0.908	0.078
$k = rfp2$ ρ=12	1.026	0.053	0.902	0.099	1.012	0.047	0.907	0.082	1.013	0.047	0.908	0.080



(a) (b) (c)

Fig. 5. Visual representation of the results. ● stands for the k-means algorithm, △ stands for variant 1 of the k-means algorithm, ◁ stands for variant 2 of the k-means algorithm, ▷ stands for variant 3 of the k-means algorithm. The filled triangles stand for the case where $\rho = 0$ in the proposed variants. The symbols printed in blue are for the single configuration, whereas the symbols printed in red are for the best configuration.

Table III and Figure 5 show the general results for the three proposed variants under different parametrization conditions, namely k for k-means and ρ for the relevance calculation.

According to the general results shown in the table, the Variant 3 is always providing the lowest general computational cost. This makes sense as it applies the improved coarse search introduced in Variant 1 and a more restrictive filtering in fine-grained search than Variant 2. However, Variant 1 is reporting the best general results for the IPS with the *Best Configuration*, whereas Variant 2 is better for the *Simple configuration*. For the three values of k, the variants improve the original k-means in both dimensions, as shown in Figure 5.

Regarding the value of k for k-means, there is still a trade-off between the value of k and the results. However, the improved coarse and fine-grained searches make the differences between $k = rfp1$ and $k = rfp2$ insignificant in terms of positioning accuracy for the Simple Configuration. In general, the lowest computational load is provided when $k = rfp2$.

The threshold value ρ of the proposed variants has a significant impact on the results. The time cost increases as ρ increases. The ρ value indicates how restrictive or permissive the relevance function is for the coarse-search filtering. Furthermore, large and low ρ values are not suitable. The lowest threshold ($\rho = 0$, solid triangles in Figure 5) is too restrictive and relevant fingerprints are discarded for the fine-grained search, whereas the highest threshold ($\rho = 12$) is too permissive so that outliers are included in the position computation.

If we balance the results of all the proposed alternatives, including the different parameters and base IPS configurations, it seems that the proposed Variant 2 with $\rho = 3$ is a good choice. This particular variant with that threshold value significantly improves the traditional k-means in both dimensions (positioning error and computational time) independently of the value of k (for k-means).

V. Conclusions

This paper introduced three new variants to improve the coarse and fine-grained search in Wi-Fi fingerprinting when k-means clustering is used to partition the full radio map. The proposed Variant 2, with an improved coarse search and a soft-filtered fine-grained search, seems to be a good choice in terms of positioning accuracy and computational costs.

The optimization of the coarse grained search makes it more computationally efficient, especially when the number of clusters is large. As a side effect, removing non-relevant clusters reduces the presence of outlier centroids and, therefore, the position accuracy is slightly improved. The proposed filtering at coarse search based on relevant clusters works when it is neither so restrictive nor so permissive (i.e. $\rho = 3$).

The generated clusters may significantly differ in size. The time cost of the fine-grained search depends on the cluster where the operational fingerprint falls into. Some clustering benefits might be lost if the cluster is oversized. Variants 2 and 3 successfully deal with this issue, reducing the computational cost of the traditional k-means to almost a half.

Finally, we consider that this work is just the first step to improve the accuracy of k-means in Wi-Fi fingerprinting problems. The machine learning models, such as k-means and k-NN, were designed for general-purpose problems and, therefore, might not totally fit Wi-Fi fingerprinting. The indoor positioning community should try to have a better understanding of the machine learning models in order to introduce some specific knowledge about, for instance, the signal propagation. Including this knowledge about the strongest AP has improved the accuracy of k-means in both dimensions in our work. As future work, we envision the definition of more refined variants, a comprehensive dataset-wise analysis and the inclusion of other well-known clustering models.

References

[1] H. Huang and G. Gartner, "Current trends and challenges in location-based services," *ISPRS International Journal of Geo-Information*, vol. 7, no. 6, 2018.

[2] N. Klepeis, W. Nelson, W. Ott, *et al.*, "The national human activity pattern survey (nhaps): A resource for assessing exposure to environmental pollutants," *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 01, pp. 231–252, 2001.

[3] G. Shtar, B. Shapira, and L. Rokach, "Clustering wi-fi fingerprints for indoor–outdoor detection," *Wireless Networks*, vol. 25, no. 3, pp. 1341–1359, Apr. 2019.

[4] P. Bahl and V. N. Padmanabhan, "Radar: An in-building RF-based user location and tracking system," in *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings.*, vol. 2, 2000, pp. 775–784.

[5] H. Shin and H. Cha, "Wi-fi fingerprint-based topological map building for indoor user tracking," in *International Conference on Embedded and Real-Time Computing Systems and Applications*, 2010.

[6] A. Razavi, M. Valkama, and E.-S. Lohan, "K-means fingerprint clustering for low-complexity floor estimation in indoor mobile localization," in *IEEE Globecom Workshops (GC Wkshps)*, 2015.

[7] Y. Chen, D. Lymberopoulos, J. Liu, *et al.*, "Indoor localization using FM signals," *IEEE Transactions on Mobile Computing*, vol. 12, no. 8, pp. 1502–1517, 2013.

[8] G. Caso, L. De Nardis, and M.-G. Di Benedetto, "A mixed approach to similarity metric selection in affinity propagation-based wifi fingerprinting indoor positioning," *Sensors*, vol. 15, no. 11, pp. 27 692–27 720, 2015.

[9] T. King, T. Butter, M. Brantner, *et al.*, "Distribution of fingerprints for 802.11-based positioning systems," in *2007 International Conference on Mobile Data Management*, May 2007, pp. 224–226.

[10] B. Li, I. Quader, and A. G. Dempster, "On outdoor positioning with wi-fi," *Journal of Global Positioning Systems*, vol. 7, 2008.

[11] T. J. Gallagher, B. Li, A. G. Dempster, *et al.*, "A sector-based campus-wide indoor positioning system," in *2010 International Conference on Indoor Positioning and Indoor Navigation*, 2010.

[12] A. Moreira, M. J. Nicolau, F. Meneses, *et al.*, "Wi-fi fingerprinting in the real world - RTLS@UM at the EvAAL competition," in *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, Oct. 2015.

[13] J. Luo and L. Fu, "A smartphone indoor localization algorithm based on wlan location fingerprinting with feature extraction and clustering," *Sensors*, vol. 17, no. 6, 2017.

[14] A. Arya, P. Godlewski, and P. Melle, "A hierarchical clustering technique for radio map compression in location fingerprinting systems," in *71st IEEE Vehicular Technology Conference*, 2010.

[15] A. Kushki, K. N. Plataniotis, and A. N. Venetsanopoulos, "Kernel-based positioning in wireless local area networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 689–705, 2007.

[16] N. Marques, F. Meneses, and A. Moreira, "Combining similarity functions and majority rules for multi-building, multi-floor, WiFi positioning," in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, IEEE, Nov. 2012.

[17] O. Abdullah, I. Abdel-Qader, and B. Bazuin, "K-means-jensen-shannon divergence for a wlan indoor positioning system," in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, Oct. 2016, pp. 1–5.

[18] A. Cramariuc, H. Huttunen, and E. S. Lohan, "Clustering benefits in mobile-centric WiFi positioning in multi-floor buildings," in *2016 International Conference on Localization and GNSS*, 2016.

[19] C. Park and S. H. Rhee, "Indoor positioning using wi-fi fingerprint with signal clustering," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2017, pp. 820–822.

[20] A. Anuwatkun, J. Sangthong, and S. Sang-Ngern, "A diff-based indoor positioning system using fingerprinting technique and k-means clustering algorithm," in *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2019.

[21] B. Wang, X. Liu, B. Yu, *et al.*, "An improved wifi positioning method based on fingerprint clustering and signal weighted euclidean distance," *Sensors*, vol. 19, no. 10, 2019.

[22] B. Altintas and T. Serif, "Improving rss-based indoor positioning algorithm via k-means clustering," in *17th European Wireless 2011 - Sustainable Wireless Technologies*, Apr. 2011, pp. 1–5.

[23] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[24] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.

[25] J. Torres-Sospedra, R. Montoliu, S. Trilles, *et al.*, "Comprehensive analysis of distance and similarity measures for Wi-Fi fingerprinting indoor positioning systems," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9263–9278, 2015.

[26] E.-S. Lohan, J. Torres-Sospedra, H. Leppäkoski, *et al.*, "Wi-fi crowdsourced fingerprinting dataset for indoor positioning," *MDPI Data*, vol. 2, no. 4, Oct. 2017.

[27] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proceedings of the Fifth Conference on Indoor Positioning and Indoor Navigation*, 2014, pp. 261–270.

[28] G. M. Mendoza-Silva, P. Richter, J. Torres-Sospedra, *et al.*, "Long-term wifi fingerprinting dataset for research on robust indoor positioning," *Data*, vol. 3, no. 1, 2018.

[29] T. King, T. Haenselmann, and W. Effelsberg, "On-demand fingerprint selection for 802.11-based positioning systems," in *2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks*, Jun. 2008, pp. 1–8.

[30] J. Torres-Sospedra, D. Quezada-Gaibor, G. M. Mendoza-Silva, *et al.* (Jun. 2020). Supplementary Materials for "New Cluster Selection and Fine-grained Search for k-Means Clustering and Wi-Fi Fingerprinting", [Online]. Available: https://zenodo.org/record/3751042.