LoRaWAN Fingerprinting with K-Means: the Relevance of Clusters Visual Inspection

Joaquín Torres-Sospedra[1], Michiel Aernouts[2], Adriano Moreira[1] and Rafael Berkvens[2]

[1] *ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal*

[2] *IDLab – Faculty of Applied Engineering, University of Antwerp – imec, Antwerp, Belgium*

Abstract

LoRaWAN-based positioning is emerging as an alternative positioning solution for battery-constrained IoT devices or GNSS-denied areas in urban environments. The data collected at the LoRaWAN Base Stations, such as the RSSI of received messages, can be merged to generate an RF fingerprint. Unsupervised crowdsourcing can be leveraged to build a large radio map covering a urban area at the expense of introducing noise of around tens of meters when labelling the reference data. As fingerprinting may have a low efficiency in a such a dense radio map, we propose to use K-Means clustering to make the position estimation faster. During our study, we found that clustering can also be used to detect large outliers in the radio map that can be subject to be removed. The rationale is to identify those samples within the cluster that are far from the geometric centroid of the cluster. This paper introduces the analysis of introducing K-Means clustering with outlier detection and the benefits it might bring. Although removing outliers have not had an outstanding increase in the positioning accuracy, the performed analysis has enabled a new metric that is moderately correlated with the positioning error. This correlation may be useful to detect unreliable position estimates and discard them. The results presented in this work, based on two LoRaWAN datasets, show that the average and median positioning error can be improved by 5 % to 10 % by discarding 4 % to 6 % of operational samples.

Keywords

Fingerprinting, Clustering, Scalability, LoRaWAN

1. Introduction

The Internet of Things (IoT) aims to interconnect a wide variety of objects, ranging from temperature sensors on mobile cooling containers to garbage bins in a city. In order to correctly interpret the measurements of such sensors, it is important to correlate them with location information. In many cases, IoT devices include a GNSS receiver for this purpose. However, this receiver only provides the device itself with location data, an Low Power Wide Area Network (LPWAN) such as LoRaWAN is often used to get sensor measurements and GNSS data to the user. This workflow is illustrated in Fig.1.

ICL-GNSS 2022 WiP, June 07–09, 2022, Tampere, Finland

✉ info@jtorr.es (Joaquín Torres-Sospedra); Michiel.Aernouts@uantwerpen.be (Michiel Aernouts); adriano.moreira@algoritmi.uminho.pt (Adriano Moreira); rafael.berkvens@uantwerpen.be (Rafael Berkvens)
0000-0003-4338-4334 (Joaquín Torres-Sospedra); 0000-0002-0527-3871 (Michiel Aernouts); 0000-0002-8967-118X (Adriano Moreira); 0000-0003-0064-5020 (Rafael Berkvens)

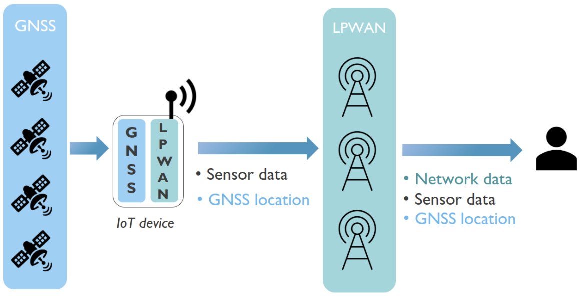

Figure 1: LPWANS are used to get sensor and location data to the user. Additionally, the user can access network metadata from the LPWAN.

An important constraint on IoT communication and localization technologies is that they must be as energy-efficient as possible, because IoT devices generally operate for multiple years using small batteries. This, and the fact that GNSS can normally only be used in outdoor environments, has motivated researchers to omit power-hungry GNSS receivers and instead leverage the existing LPWAN link and sensor data for localization purposes. For example, metadata such as the Received Signal Strength Indicator (RSSI), phase or timing information from multiple LPWAN receivers can be translated to distance estimations between each receiver and a transmitting IoT device. However, these methods strongly depend on the LPWAN network deployment and generally lead to high location estimation errors. A previous analysis on the choice between GNSS and LPWAN localization shows that the latter should only be favored over GNSS when a large location error is justifiable and when the energy budget of an IoT device is extremely limited [1]. In practice, this means that implementing GNSS receivers on low-power IoT devices is often feasible. That being said, LPWAN localization can certainly still prove its use, because not all applications require location data with GNSS-like accuracy. For example, a construction company might only want to know at which of its building sites its assets are located, which implies that an error of hundreds of meters can be accepted. LPWAN localization can also play an important role in multimodal localization, for example as a fallback solution when a tracking device is moving into GNSS-denied areas such as tunnels or indoor environments [2]. Moreover, it may act as a verification mechanism to detect GNSS spoofing.

In 2019, Aernouts et al. published an extended version of the LoRaWAN dataset described in [3]. Over a course of three months, 20 postal services cars carried LoRaWAN devices that periodically transmitted their latest GNSS location. As a result, the collected dataset contains 130430 entries with a ground truth location, the LoRa Spreading Factor (SF) used by the transmitter, timing data and Received Signal Strength (RSS) data for each receiving LoRaWAN gateway. It should be noted that the ground truth information was collected from GNSS receivers and, therefore, with potential errors of tens of meters. First, urban canyoning can decrease the GNSS accuracy, since the dataset is collected in a dense urban area. Second, the received GNSS coordinates of the transmitting device could differ from the actual device coordinates at receiving time because the total transmission time of a LoRa signal can take up to a few seconds, depending on the payload size and the SF. This effect becomes even more prominent when the transmitter travels at higher speeds.

RSS data enables positioning with trilateration and fingerprinting. While the former requires knowing the location of the LoRaWAN Base Stations (BSs), the propagation model and the environment obstructions; the latter only requires a set of reference data at known positions, also known as the radio map. In this paper, we focus on passive fingerprinting, where a fingerprint is the set of RSSI measurements of a particular LoRaWAN message transmitted by a device and measured in the available LoRaWAN BSs in the operational area.

This technique requires two phases: the *offline* phase focuses on geo-referenced RSSI data collection (see radio map collection in [3]), whereas the *online* phase estimates the position of new fingerprints at unknown positions with, for instance, a k-Nearest Neighbour (k-NN)-based algorithm and the radio map.

However, fingerprinting is computationally demanding if the dataset contains thousands of samples, e.g. LoRaWAN datasets in [3]. In those datasets, every single operational fingerprint has to be compared with all the reference samples in the radio map, even if they significantly differ, to obtain the most similar ones and compute the final position estimate. Thus, clustering techniques have been applied to split the radio map into several smaller versions [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. In the operational stage, the identification of the most relevant cluster is done first (coarse search). Then, the position is estimated using the corresponding reduced radio map (fine-grained search). This two-step procedure is significantly faster that regular fingerprinting, specially in large datasets [15].

In this paper we propose a version of K-Means clustering with outlier detection where noisy fingerprints are removed. We hypothetise that the clusters generated with K-Means over the feature RSSI space can be de-noised by removing the reference samples which are significantly far for the cluster geometric centroid. It is worth noting that the proposed algorithmic solution is performed after generating the clusters with K-Means. K-Means clustering is an unsupervised model that groups similar data without, in this case, the location information (i.e., the labels). Thus, we consider that K-MEANS basic principles cannot be significantly re-formulated to make it more robust. The main contributions of this work include:

- Modification of K-Means to remove outliers from clusters according to the geometric information;
- Comprehensive comparison between applying K-Means without and with ourlier detection;
- A new metric which is correlated with the positioning error under some cases;
- A procedure to discard unreliable position estimations.

The remainder of this work is organised as follows. Section 2 introduces the related work on LoRaWAN, fingerprinting and clustering. Section 3 describes the materials and methods used in this work. Section 4 details the experimental setup and shows the empirical results. Section 5 provides the final discussion and conclusions about this work.

2. Related work

2.1. LoRaWAN and fingerprinting

LoRaWAN's relatively wide bandwidth of 125 kHz to 250 kHz makes it a suitable candidate for both RSS-based and time-based localization. Thanks to the widespread availability of LoRaWAN networks and datasets [3, 16, 17, 18, 19], many researchers have evaluated the performance of various localization methods. For instance, Pospisil et al. evaluated the performance of five Time Difference of Arrival (TDoA) algorithms through simulation and validated two of them with field measurements. They achieved a mean location error of 543 m in a test area of 4.58 km^2 [20].

The aforementioned LoRaWAN dataset by Aernouts et al. enabled many researchers to evaluate fingerprinting and machine learning approaches for localization. Pandangan et al. generated a hybrid dataset containing RSS and TDoA information based on the LoRaWAN dataset. Their hybrid dataset was then used to evaluate k-NN and Random Forest algorithms which resulted in a median error of 333 m and 194 m respectively [21]. This is a slight improvement compared to related research on Neural Network localization with the LoRaWAN dataset [22, 23]. Purohit et al. also used this dataset for their research on Neural Network localization. In their investigation of three different learning models, the Long Short-Term Memory (LSTM) model with 64 neurons came out on top with a mean error of 191 m [24]. Janssen et al. compared the location accuracy, R^2 score and evaluation time of ten Machine Learning algorithms using the LoRaWAN dataset. Their experiments show that the weighted k-NN and Random Forest algorithms result in the best accuracy and R^2 score, but Random Forest has a significantly faster computation time [25]. In a subsequent study, the authors extended their comparison with range-based localization using eight different path loss models and six weight functions. Their best path loss model - weight function combination yielded an estimation error of 700 m, which is significantly higher than the 340 m obtained with fingerprinting. Furthermore, this work provides a comprehensive overview of the trade-offs that must be made between range-based and fingerprinting-based localization, including accuracy, complexity, cost, etc. [26].

2.2. Clustering in fingerprinting

Clustering has been widely applied in Wi-Fi and BLE fingerprinting to reduce the computational cost and keep similar accuracy, being K-Means [4, 5], including K-Medoids [6, 7] and Fuzzy c-Means (FCM) [8, 9, 10] variants, the most popular. Other approaches, such as Affinity Propagation Clustering (APC) [11, 12] or Density-based spatial clustering of applications with noise (DBSCAN) [13, 14], have also been explored but their feasibility may depend on the dataset according to some preliminary experiments we performed.

Therefore, this work is focusing in K-Means clustering, trying to take benefit from the position information of the reference data to remove those reference samples that may poison the radio map. To enhance the performance of K-Means, we have used the Manhattan distance for distances computations in the feature (RSSI) space and the centroid initialization proposed in [27].

3. Materials and Methods

3.1. K-Means in fingerprinting

The core of the passive fingerprinting technique requires two phases: the *off-line* and *on-line* phases as explained before. In the *off-line phase*, reference fingerprints (s^t) are generated from a set of received LoRaWAN messages (that include their position from GPS) by the available LoRaWAN BS, generating thus a radio map ($\mathcal{T}$). In the *on-line phase*, the operational fingerprints (from unknown positions) are compared to the fingerprints stored in the radio map. Their position is estimated using the locations of the most similar fingerprints in the radio map, usually computing their centroid.

After generating the radio map $\mathcal{T}$, similar fingerprints in the feature RSSI space are grouped by K-Means clustering algorithm. It is expected that fingerprints within a cluster would be also close in the geometrical space. The output of K-Means provides the K cluster centroids, $\mathcal{C}_i, \forall i \in [1, \ldots, K]$, and the reduced radio map for every cluster $\mathcal{T}_i, \forall i \in [1, \ldots, K]$. The centroids and reference fingerprints are both vectors representing the feature RSSI space, thus having as many values as LoRaWAN BSs.

As an illustrative example, a few clusters over the LoRaWAN 2017/18 dataset are shown in Fig. 2. The gray dots represent the reference fingerprints in the radio map, whereas the coloured ones represent the samples in the cluster. The number of reference fingerprints and their dispersion in the geometric space depends on the cluster.

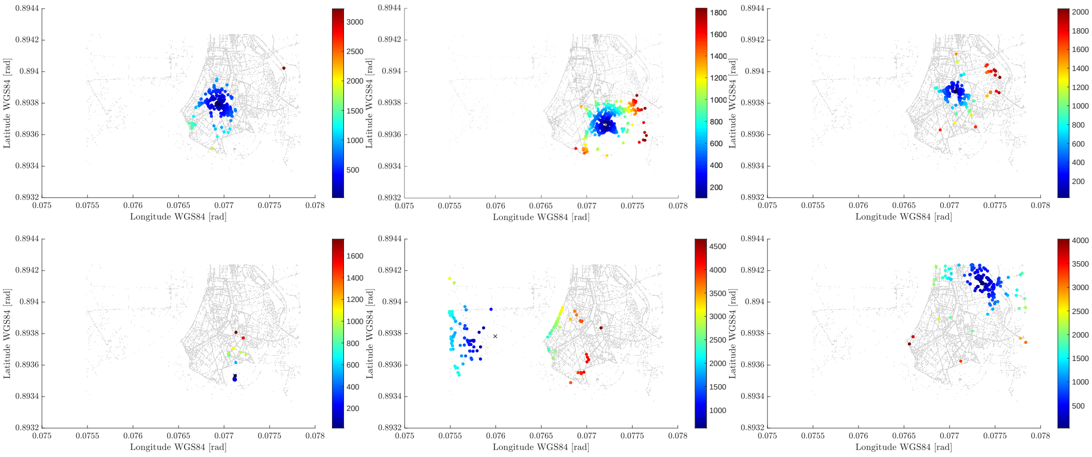

Figure 2: Example of six illustrative clusters generated by K-Means in LoRaWAN 2017/18 dataset. Color indicates distance [m] to geometric centroid.

In the operational phase, the search of most similar reference fingerprints is done in a two-step process. First, the operational fingerprint is compared to all the cluster centroids (RSSI space) to retrieve the one reporting the lowest Euclidean distance. Second, the search of most similar reference fingerprints is done over the corresponding reduced radio map, $\mathcal{T}_i$.

3.2. Analysis of clustering with K-Means

Previous results in the literature show that K-Means in fingerprinting reduces the computational cost at the expense of a slightly higher positioning error. This reduction on time is specially relevant in large datasets [28].

In this paper, the same clustering model has been applied to both LoRaWAN datasets, being K the squared root of the samples in the radio map as suggested in [28]. These results, which are shown in Section 4.2, were in phase with the results reported in the literature.

However, to avoid the adoption of a black box approach while using k-Means, an additional overall analysis on the clusters was performed. In particular, the location (longitude and latitude in WGS84 format) of the reference samples in the reduced radio map was visually inspected for each cluster, showing a relevant output in many clusters.

Fig. 2 shows six illustrative examples of the clusters generated with K-Means. Despite their size and dispersion depend on the cluster, most of them report cases where the fingerprints are very far (reddish points in the figure) from the current geometric centroid and close to others geometric centroids. Those outliers share similar RSSI values with respect other reference fingerprints in the cluster, but thet geographically far from them. Among other factors, this effect may be caused by the positioning errors introduced by the GNSS receivers.

3.3. Removing noisy samples from clusters

The idea to remove noisy samples from the radio map is simple. Given the samples (fingerprints) of a cluster, their geometric centroid (in the WGS84 space) is calculated. All samples whose distance to the geometric centroid is higher than twice the median value are removed. This is only applied to those clusters where the maximum distance is higher than 5 times the median value. i.e., it is only applied to those clusters having significant outliers. The proposed model is described in Algorithm 1, which has 3 stages: clusters generation (line 2), clusters cleaning (ln. 3–12) and position estimation (ln.15–22). First and second stage can be performed once per dataset, so their timing can be neglected when providing the computational costs of providing a position estimate in the online phase.

$\mathcal{T}$ is the radio map, $\mathcal{V}$ is the set with the test/evaluation samples, k is the number of nearest neighbors for k-NN. A sample (fingerprint) is represented with $\mathbf{s}$ and has N_{bs} elements (one for each LoRaWAN BS), whereas its position is represented and its position (longitude and latitude in WGS84) with $\mathbf{pos}$. For K-Means, K is the number of clusters ($K = \sqrt{|\mathcal{T}|}$ as suggested in [15]), $\mathcal{C}$ represents the clusters RSSI centroids and $\mathcal{G}$ represents the clusters geometric lat/lon centroids. $\dot{\mathcal{T}}_c$ stands for the clean reduced radio map for cluster c.

The other parameters for the outlier detection were set based on the researchers experience. e.g., the threshold used to remove the noisy samples, 2 times the median distance, has been selected as the distances to the geometric centroid usually increase gradually.

Algorithm 1 k-NN for positioning with K-Means and outlier detection

```
input $\mathcal{T}$, $\mathcal{V}$, $k$, $K$
$\mathcal{C}_i$,$\mathcal{T}_i \leftarrow$ Apply $K$-Means to $\mathcal{T}$
for $i = 1$ to $K$ do
    $\mathcal{G}_i \leftarrow$ Compute geometric centroid of samples in $\mathcal{T}_i$
    $G \leftarrow \left\{geoDistance\left(\mathbf{pos}_j, \mathcal{G}_i\right), \forall j \in \mathcal{T}_i\right\}$
    if $max\,(G) > (5 \cdot median\,(G))$ then
        // Remove samples far from geometric centroid
        $\dot{\mathcal{T}}_i \leftarrow \left\{\mathbf{s}_j^{\mathcal{T}_i} \in \mathcal{T}_i : G_j \leq (2 \cdot median\,(G))\right\}$
    else
        $\dot{\mathcal{T}}_i \leftarrow \mathcal{T}_j$ // No cleaning for cluster $i$
    end if
end for
for $i = 1$ to $|\mathcal{V}|$ do
    Identify most relevant cluster, $c$
    Set the reduced radio map $\dot{\mathcal{T}}_c$
    for $j = 1$ to $|\dot{\mathcal{T}}_c|$ do
        Compute distance between $\mathbf{s}_i^{\mathcal{V}}$ and $\mathbf{s}_j^{\dot{\mathcal{T}}_c}$
    end for
    Sort distances in RSS space
    Select the $k$ closest candidates
    Estimate position lat/lon
end for
Return: Estimated positions for all samples in $\mathcal{V}$
```

4. Experiments and Results

4.1. Experimental Setup

In order to assess the proposed clustering model with outlier detection, we have compared the results between the plain k-NN, the optimization rule proposed by Moreira [29], K-Means without outlier detection and K-Means with outlier detection. To estimate the final position, we have used the simple 1-NN algorithm using the *Euclidean distance*. The models have been run 10 times to minimise the random initialization of K-Means.

For the experiments, two datasets collected in the city of Antwerp between end of 2017 and beginning of 2019 [3, 30] have been used, namely LoRaWAN 2017/18 and LoRaWAN 2018/19. Both datasets were collected to evaluate fingerprint localization algorithms in large outdoor environments and, according to the database authors, the RSSI of the LoRaWAN messages could hold an additional GPS error. This feature makes them appropriate for assessing the proposed algorithm to remove noise from clusters. For both datasets, the samples have been sorted by timestamp and then split for training and testing, the first $\approx 80\%$ of samples have been used for training and the last $\approx 20\%$ of samples have been used for evaluation. This division has been performed to avoid having data from the same device and day on both subsets, $\mathcal{T}$ and $\mathcal{V}$.

The evaluation metrics include the Averaged Positioning Error (APE), $\bar{\epsilon}$; the Median Positioning Error (MPE), $\tilde{\epsilon}$; and the Averaged Operational Time (AOT), $\bar{\tau}_{fp}$, and consider all the 10 execution runs. The APE and MPE are included in the ISO18305 standard, whereas the AOT refers to the average time required to process an operational fingerprint and provide the position estimate. In contrast to the plain k-NN algorithm, where all fingerprints hold similar operational time, the operational time may significantly vary depending on the cluster. i.e., K-Means clustering does not guarantee that all clusters are equally distributed, so the time required to perform the fine-grained search will depend on the selected cluster. Therefore, the standard deviation is also reported for the operational time.

4.2. Results

This subsection is devoted to show the empirical results. First, a comparison with traditional fingerprint models is introduced. Then, a comprehensive analysis about the consequences of removing noise from the radio map is performed. Finally, the possible benefits of the proposed model are described.

4.2.1. Comparative analysis

Table 1 introduces the main results for the comparative analysis. It includes the plain k-NN algorithm ($k = 1$), the optimization rule based on common strongest anchor proposed in Moreira et al. [29], and K-Means clustering without and with the outlier detection (OD). Fig. 3 introduces the Empirical Cumulative Distribution Function (ECDF) plots of the positioning error and operational time of the four methods for both datasets.

Table 1
Main results: APE, MPE and AOT

	Lorawan 2017/18			Lorawan 2018/19		
Method	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]
plain k-NN	558.3	374.1	2464.0 (13.7)	375.6	169.3	2518.2 (11.0)
Moreira [29]	563.4	377.0	202.9 (149.7)	375.5	167.7	306.4 (223.9)
K-Means	566.1	379.2	17.1 (7.3)	379.7	174.7	28.2 (16.0)
K-Means OD	559.3	369.6	16.4 (6.8)	378.8	168.0	26.8 (15.5)

In general, the four models provide similar results in terms of positioning error being the main difference their computational cost. The two solutions based on K-Means report the lowest computational cost with an averaged execution time below 20 ms and 30 ms respectively.

Removing the outliers not only made K-Means slightly more accurate but also slightly more efficient in the operational stage as the proposed approach removed 8.6% and 9.5% of reference fingerprints on each dataset respectively. However, the improvements may be marginal.

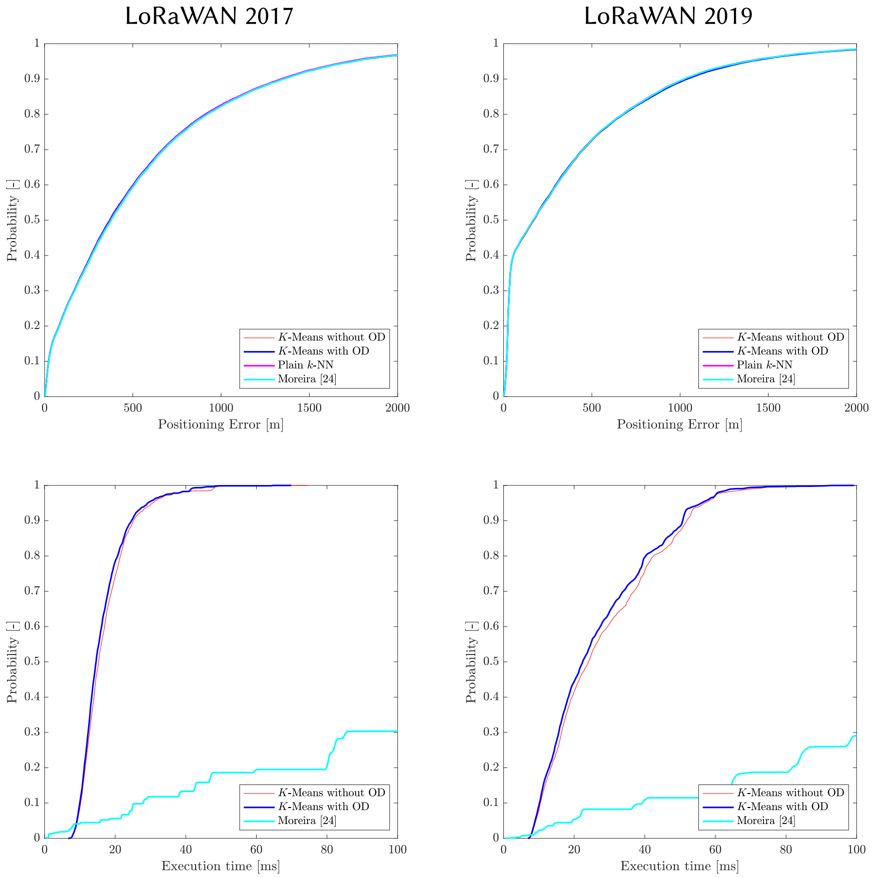

Figure 3: ECDF of positioning error and execution time for both datasets

4.2.2. A comprehensive analysis of removing noise

Despite K-Means without and with outliers detection having similar performance according to the previous results, positioning can be based on two sources, namely a *noisy* radio map and a *clean* radio map. This enables to exploit some additional information at the operational stage as there are samples where both approaches based on K-Means (without and with outliers detection) do not agree in estimating the position. This happens in 12% and 32% of samples on each dataset, respectively.

Thus, the evaluation set can be split into two subsets, one where both estimators agree and provide the same position estimation ("*same*" in table and figure), and the other where they disagree ("*diff*"). Table 2 and Fig. 4 show the corresponding results and ECDFs.

According to Table 2 and the ECDFs plots from Fig. 4, the subset "*same*" is generally better than the subset "*diff*" in both metrics, positioning error and execution time, specially in the first dataset (LoRaWAN 2017/18). i.e. when the two estimators –without and with outlier detection– agree, the positioning results are better that when they disagree. If both estimators disagree, the positioning error provided with K-Means with outlier detection is better.

Table 2
Results of K-Means without and with outlier detection

		Lorawan 2017/18			Lorawan 2018/19		
K-Means	Subset	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]
without OD	Same	545.2	363.9	16.9 (7.1)	373.4	169.0	24.3 (14.7)
without OD	Diff	734.5	530.8	18.1 (8.6)	393.3	190.3	36.6 (15.6)
with OD	Same	545.2	363.9	16.3 (6.7)	373.4	169.0	23.7 (14.4)
with OD	Diff	673.1	419.8	16.4 (7.6)	390.3	164.8	33.7 (15.4)

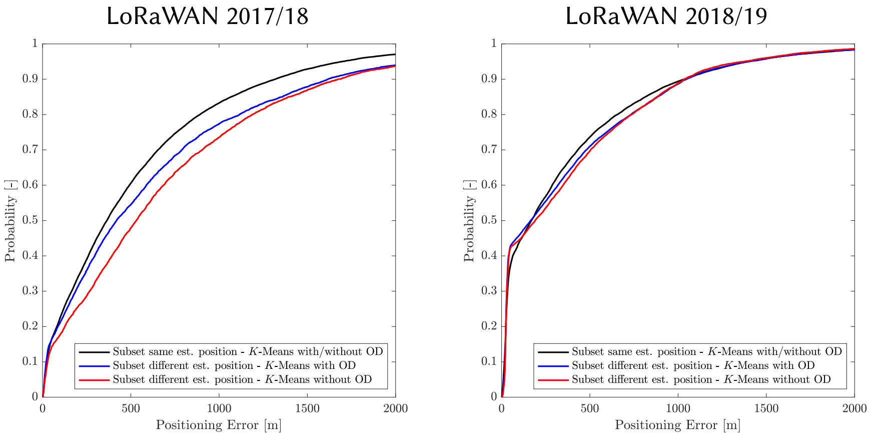

Figure 4: ECDF of the models based on K-Means

We hypothesise that a divergence between the position estimate between both K-Means models may indicate the quality of the position estimate provided by the proposed model. In particular, we explore the correlation between the distance between position estimations when they disagree and the positioning error using K-Means with outlier detection. First, that relation is shown as a scatter plot in Fig. 5 (top) and as a density heat map (color representing the amount of samples for a particular range in both dimensions) in Fig. 5 (bottom). In addition, we computed the Pearson correlation, which provided a correlation factor of 0.64 and 0.52 for the two datasets respectively. Thus, it seems that when both models based on K-Means disagree, the distance between the two estimators may indicate the positioning error.

The scatter plots are dense as the number of test samples is large and the experiments have been run 10 times. Fig. 6 shows the boxplot of the positioning errors for different distances between estimates. The correlation trends between the distance between estimates and the positioning error using K-Means with the proposed outlier detection can be seen more clearly in the figure. However, it can also be seen that in distances above around 2000 m seems to be less reliable as the error and its variability are both high. i.e., the the lowest variability is provided in range $[0, \ldots, 500[$ and it is increasing as the distance between estimates also increases and the number of cases is significant. For the ranges including the largest distances between estimations, there are only a few cases in both datasets.

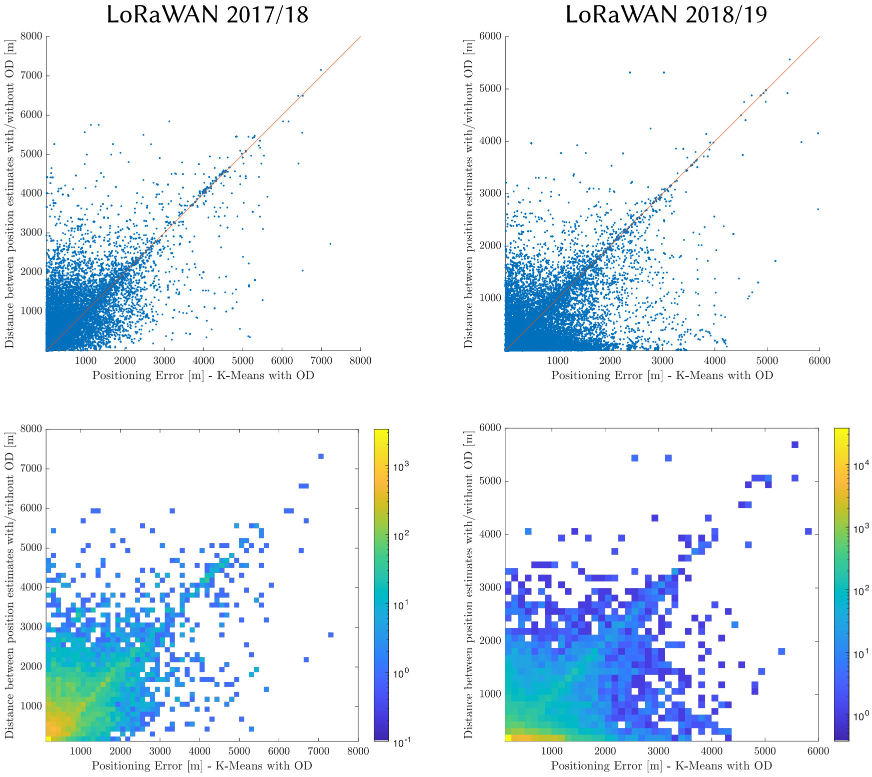

Figure 5: Scatter plots and heat maps relating distance between estimates and positioning error using the proposed method

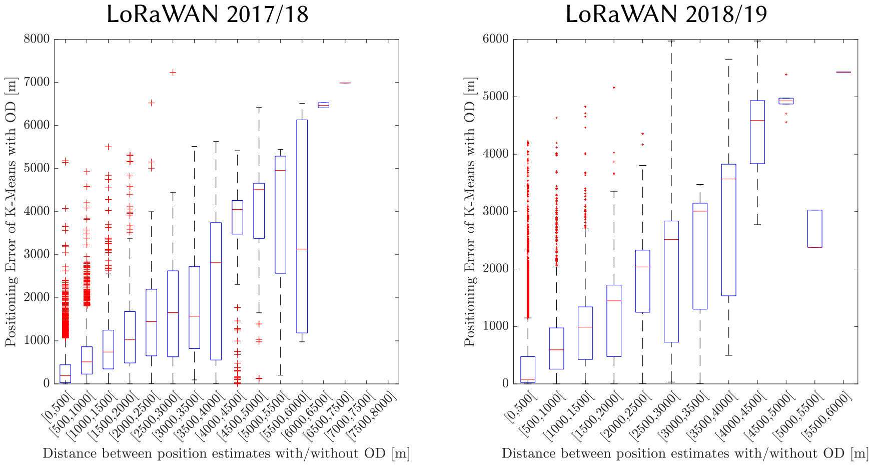

Figure 6: Relation between distances among estimations and positioning error

4.2.3. Possible benefits of combining noisy and cleaned data sets

Positioning using K-Means has shown to be very efficient in terms of computational time. Computing the position estimate with fingerprints from the original reduced radio maps and the cleaned (without outliers) reduced radio maps is feasible. For any operational fingerprint, if the two position estimates differ, their distance could be used as an indicator of reliability (see Figs. 5-6). For instance, if this distance is higher than a predefined threshold, the position estimate could be discarded.

We consider that positioning can take benefit of discarding unreliable samples. In general, these operational fingerprints may have a large positioning error attached. Therefore, the positioning error of the remaining fingerprints (the ones that are reliable) should be better. The only requirement is to set a threshold on the distance between the two position estimates. Table 3 and Fig. 7 show the results using K-Means with outlier detection and different thresholds, where rs stands for reliable samples.

Table 3
Results of combining K-Means without and with outlier detection removing unreliable operational samples

	Lorawan 2017/18				Lorawan 2018/19			
Threshold	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]	rs[%]	$\bar{\epsilon}$[m]	$\tilde{\epsilon}$[m]	$\bar{\tau}$[ms]	rs[%]
–	559.3	369.6	16.4 (6.8)	100.0	378.8	168.0	26.8 (15.5)	100.0
2000 m	547.3	365.8	16.4 (6.8)	99.1	373.9	165.9	26.9 (15.5)	99.7
1000 m	536.1	356.8	16.5 (6.8)	96.5	363.1	159.2	27.0 (15.5)	98.1
500 m	533.7	352.0	16.5 (6.9)	93.6	355.8	150.2	27.2 (15.6)	96.1
250 m	535.6	353.0	16.5 (6.9)	91.6	351.0	141.9	27.4 (15.6)	94.0
125 m	537.0	354.1	16.5 (6.9)	90.7	344.9	129.7	27.5 (15.7)	91.4

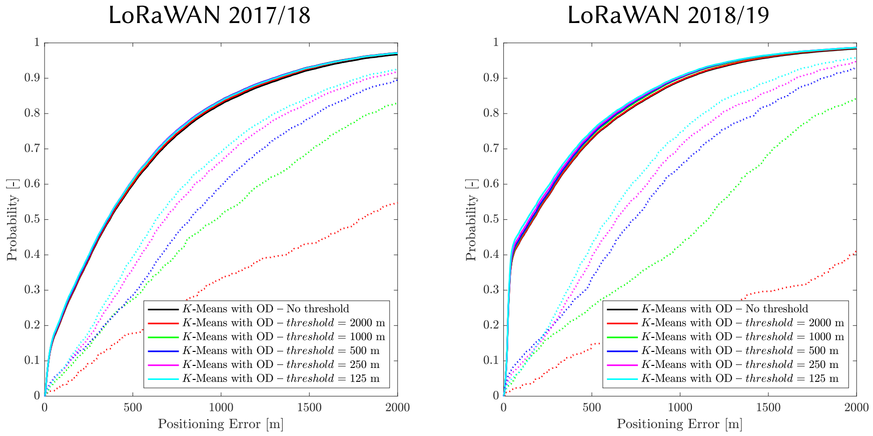

Figure 7: ECDF of K-Means with outlier detection and samples removal

The ECDF is shown for both sets, reliable samples (solid) and unreliable samples (dashed). In general, as the threshold decreases, the more samples are considered unreliable and the lower the positioning error of the reliable samples. However, the presence of low positioning errors in the set of unreliable samples increases. i.e., the lower the threshold (e.g., 125 m), the better the results of the reliable samples, but also the higher the probability of discarding a good position estimate.

According to the results presented in Table 3 and Fig. 7, the threshold depends on the dataset. For the two LoRaWAN datasets we have used, the threshold is 500 m (LoRaWAN 2017/18) and 125 m (LoRaWAN 2018/19), as they provide good results in terms of positioning error of the reliable samples in their respective datasets. On the other hand, the lower the threshold the more samples (including good estimations) are removed.

5. Discussion and Conclusions

K-Means is often applied to fingerprinting as a black box to obtain a similar average positioning error with a significantly lower computational cost. In this paper, we have applied it to two large datasets, getting results in phase to what has been reported in state-of-the art works about Wi-Fi and BLE fingerprinting.

Visual inspection on the generated clusters has shown that they might contain noisy fingerprints which are close to the cluster centroid in the RSSI space but on different locations. Thus, as the reference data provides the fingerprints (RSSI vectors) and their locations, we have proposed a simple rule to remove the noisy samples from clusters.

Although the results are not outstanding, having two ways to estimate the position has enabled a new metric based on the distance between the two position estimates. For samples where both estimators diverge, this metric has shown to be moderately correlated to the positioning error provided by the proposed K-Means clustering with outlier detection.

Being able to detect unreliable position estimates at the operational stage is an important step as a better accuracy can be ensured for the reliable ones. In this case, the average and median positioning error can be improved by 5 % to 10 % by discarding the 4 % to 6 % of operational samples.

In this paper, we propose a model to clean the clusters. It is of utmost importance to not blindly trust on Machine Learning models if they were used as black boxes. Visual inspection allowed to detect noisy samples and get a new metric correlated to the positioning error. Further efforts will be devoted to improve noise removal with different strategies.

Acknowledgments

A. Moreira gratefully acknowledge funding from FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

[1] M. Aernouts, T. Janssen, R. Berkvens, M. Weyn, Lora localization: With gnss or without?, IEEE IoT Magazine *(Submitted)* (2022).

[2] M. Aernouts, F. Lemic, B. Moons, J. Famaey, J. Hoebeke, M. Weyn, R. Berkvens, A Multimodal Localization Framework Design for IoT Applications, Sensors 20 (2020) 4622. doi:10.3390/s20164622.

[3] M. Aernouts, R. Berkvens, K. Van Vlaenderen, M. Weyn, Sigfox and lorawan datasets for fingerprint localization in large urban and rural areas, Data 3 (2018). URL: https://www.mdpi.com/2306-5729/3/2/13. doi:10.3390/data3020013.

[4] A. Anuwatkun, J. Sangthong, S. Sang-Ngern, A diff-based indoor positioning system using fingerprinting technique and k-means clustering algorithm, in: 16th International Joint Conference on Computer Science and Software Engineering, 2019, pp. 148–151.

[5] S. G. Lee, C. Lee, Developing an improved fingerprint positioning radio map using the k-means clustering algorithm, in: Int. Conf. on Information Networking, 2020, pp. 761–765.

[6] J. Cheng, Y. Cai, Q. Zhang, J. Cheng, C. Yan, A new three-dimensional indoor positioning mechanism based on wireless lan, Mathematical Problems in Engineering 2014 (2014).

[7] H. Lin, L. Chen, An optimized fingerprint positioning algorithm for underground garage environment, in: Int. Conf. on Information Networking, 2016, pp. 291–296. URL: https://doi.ieeecomputersociety.org/10.1109/ICOIN.2016.7427079. doi:10.1109/ICOIN.2016.7427079.

[8] H. Zhou, N. Van, Indoor fingerprint localization based on fuzzy c-means clustering, 2014, pp. 337–340. doi:10.1109/ICMTMA.2014.83.

[9] D. J. Suroso, P. Cherntanomwong, P. Sooraksa, J. Takada, Fingerprint-based technique for indoor localization in wireless sensor networks using fuzzy c-means clustering algorithm, in: International Symposium on Intelligent Signal Processing and Communications Systems, 2011. doi:10.1109/ISPACS.2011.6146167.

[10] C. Zhang, N. Qin, Y. Xue, L. Yang, Received signal strength-based indoor localization using hierarchical classification, Sensors 20 (2020). doi:10.3390/s20041067.

[11] P. A. Karegar, Wireless fingerprinting indoor positioning using affinity propagation clustering methods, Wireless Networks 24 (2018) 2825–2833. URL: https://doi.org/10.1007/s11276-017-1507-0. doi:10.1007/s11276-017-1507-0.

[12] G. Caso, L. De Nardis, M.-G. Di Benedetto, A mixed approach to similarity metric selection in affinity propagation-based wifi fingerprinting indoor positioning, Sensors 15 (2015). doi:10.3390/s151127692.

[13] M. Zhou, Y. Wei, Z. Tian, X. Yang, L. Li, Achieving cost-efficient indoor fingerprint localization on wlan platform: A hypothetical test approach, IEEE Access 5 (2017) 15865–15874. doi:10.1109/ACCESS.2017.2737651.

[14] B. Wang, X. Liu, B. Yu, R. Jia, X. Gan, An Improved WiFi Positioning Method Based on Fingerprint Clustering and Signal Weighted Euclidean Distance, Sensors 19 (2019). URL: https://pubmed.ncbi.nlm.nih.gov/31109054https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6567165/. doi:10.3390/s19102300.

[15] J. Torres-Sospedra, P. Richter, A. Moreira, G. M. Mendoza-Silva, E. S. Lohan, S. Trilles, M. Matey-Sanz, J. Huerta, A comprehensive and reproducible comparison of clustering

and optimization rules in wi-fi fingerprinting, IEEE Transactions on Mobile Computing 21 (2022) 769–782. doi:10.1109/TMC.2020.3017176.

[16] P. Masek, M. Stusek, E. Svertoka, J. Pospisil, R. Burget, E. S. Lohan, I. Marghescu, J. Hosek, A. Ometov, Measurements of LoRaWAN Technology in Urban Scenarios: A Data Descriptor, Data 6 (2021). URL: https://www.mdpi.com/2306-5729/6/6/62. doi:10.3390/data6060062.

[17] K. Mikhaylov, M. Stusek, P. Masek, R. Fujdiak, R. Mozny, S. Andreev, J. Hosek, On the performance of multi-gateway lorawan deployments: An experimental study, in: 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1–6. doi:10.1109/WCNC45663.2020.9120655.

[18] L. Bhatia, M. Breza, R. Marfievici, J. A. McCann, Loed: The lorawan at the edge dataset: Dataset, in: Proceedings of the Third Workshop on Data: Acquisition To Analysis, DATA '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 7–8. URL: https://doi.org/10.1145/3419016.3431491. doi:10.1145/3419016.3431491.

[19] R. Cardell-Oliver, C. Hübner, M. Leopold, J. Beringer, Dataset: Lora underground farm sensor network, in: Proceedings of the 2nd Workshop on Data Acquisition To Analysis, DATA'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 26–28. URL: https://doi.org/10.1145/3359427.3361912. doi:10.1145/3359427.3361912.

[20] J. Pospisil, R. Fujdiak, K. Mikhaylov, Investigation of the performance of tdoa-based localization over lorawan in theory and practice, Sensors (Switzerland) 20 (2020) 1–22. doi:10.3390/s20195464.

[21] Z. A. Pandangan, M. C. R. Talampas, Hybrid LoRaWAN Localization using Ensemble Learning, in: 2020 Global Internet of Things Summit (GIoTS), IEEE, 2020, pp. 1–6. doi:10.1109/GIOTS49054.2020.9119520.

[22] G. G. Anagnostopoulos, A. Kalousis, A Reproducible Comparison of RSSI Fingerprinting Localization Methods Using LoRaWAN, in: 16th Workshop on Positioning, Navigation and Communications, 2019.

[23] I. Daramouskas, V. Kapoulas, M. Paraskevas, Using Neural Networks for RSSI Location Estimation in LoRa Networks, in: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), IEEE, 2019, pp. 1–7. doi:10.1109/IISA.2019.8900742.

[24] J. Purohit, X. Wang, S. Mao, X. Sun, C. Yang, Fingerprinting-based Indoor and Outdoor Localization with LoRa and Deep Learning, in: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, IEEE, 2020, pp. 1–6. doi:10.1109/GLOBECOM42002.2020.9322261.

[25] T. Janssen, R. Berkvens, M. Weyn, Comparing Machine Learning Algorithms for RSS-Based Localization in LPWAN, in: Lecture Notes in Networks and Systems, volume 96, 2020, pp. 726–735. doi:10.1007/978-3-030-33509-0_68.

[26] T. Janssen, R. Berkvens, M. Weyn, Benchmarking RSS-based localization algorithms with LoRaWAN, Internet of Things 11 (2020) 100235. doi:10.1016/j.iot.2020.100235.

[27] D. Arthur, 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.

[28] J. Torres-Sospedra, D. Quezada-Gaibor, G. M. Mendoza-Silva, J. Nurmi, Y. Koucheryavy,

J. Huerta, New cluster selection and fine-grained search for k-means clustering and wi-fi fingerprinting, in: 2020 Int. Conf. on Localization and GNSS (ICL-GNSS), 2020.

[29] A. Moreira, M. J. Nicolau, F. Meneses, A. Costa, 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, ????

[30] M. Aernouts, R. Berkvens, K. Van Vlaenderen, M. Weyn, Sigfox and LoRaWAN Datasets for Fingerprint Localization in Large Urban and Rural Areas, 2019. doi:10.5281/zenodo.3904158, https://doi.org/10.5281/zenodo.3904158.