Comprehensive Analysis of Distance and Similarity Measures for Wi-Fi Fingerprinting Indoor Positioning Systems

Joaquín Torres-Sospedra, Raúl Montoliu, Sergio Trilles, Óscar Belmonte, Joaquín Huerta

Institute of New Imaging Technologies, Universitat Jaume I, Avda. Vicente Sos Baynat S/N, 12071, Castellón, Spain.
[jtorres,montoliu,strilles,belfern,huerta]@uji.es

Abstract

Recent advances in Indoor Positioning Systems led to a business interest in those applications and services where a precise localization is crucial. Wi-Fi fingerprinting based on Machine Learning and Expert Systems are commonly used in the literature. They compare a current fingerprint to a database of fingerprints, and then return the most similar one/ones according to: 1) a distance function, 2) a data representation method for Received Signal Strength values, and 3) a thresholding strategy. However, most of the previous works simply use the *Euclidean distance* with the raw unprocessed data. There is not any previous work that studies which is the best distance function, which is the best way of representing the data and which is the effect of applying thresholding. In this paper, we present a comprehensive study using 51 distance metrics, 4 alternatives to represent the raw data (2 of them proposed by us), a thresholding based on the RSS values and the public *UJIIndoorLoc* database. The results shown in this paper demonstrates that researchers and developers should take into account the conclusions arisen in this work in order to improve the accuracy of their systems. The IPSs based on k-NN are improved by just selecting the appropriate configuration (mainly distance function and data representation). In the best case, 13-NN with *Sørensen distance* and the *powed* data representation, the error in determining the place (building and floor)

has been reduced in more than a 50% and the positioning accuracy has been increased in 1.7 meters with respect to the 1-NN with *Euclidean distance* and raw data commonly used in the literature. Moreover, our experiments also demonstrates that thresholding should not be applied in multi-building and multi-floor environments

Keywords: Indoor Localization, Distance measures, Similarity measures, k-NN, Wi-Fi fingerprint

1. Introduction

Automatic user localization is a hot research topic nowadays with an expected $\$2.60$ billion market in 2018 (Markets & Markets, 2014). Context-aware applications based on user's location need to know the precise localization to provide location-based services (Estevez & Carlsson, 2014; Neves et al., 2014; Torres-Sospedra et al., 2015), monitor people (Calderoni et al., 2015), and track *Internet-of-Things*' objects (Le et al., 2014), among others. Although outdoor localization is already solved due to the inclusion of Assisted Global Positioning System (A-GPS) sensors in smartphones, indoor positioning is still an open problem due to the low GPS coverage inside buildings and the lack of floor identification in GPS.

There is a large number of technologies to develop Indoor Positioning Systems (IPSs): Radio-Frequency Identification (RFID) (Jin et al., 2006; Montaser & Moselhi, 2014; Calderoni et al., 2015), Bluetooth (Feldmann et al., 2003; Li, 2014), Wireless Local Area Network (WLAN or Wi-Fi) (Bahl & Padmanabhan, 2000; Lau & Chung, 2007; del Corte-Valiente et al., 2009; Gansemer et al., 2010a; Segou et al., 2010; Machaj et al., 2011; Marques et al., 2012; Chen et al., 2013; Lan & Shih, 2013; Le et al., 2014), ZigBee (Martí et al., 2012), Ultrasound (Ijaz et al., 2013), Magnetic field variations (Chung et al., 2011; Guo et al., 2014), and even LED light (Kuo et al., 2014), among others. A combination of technologies has also been used (Martí & Marín, 2011; Baniukevic et al., 2013; Li et al., 2015). Wi-Fi is a good choice for indoor positioning technology due to the ubiquity of Wireless Access Points (WAPs), embedded

Wi-Fi connectivity in modern mobile phones, and the use of pre-existing Wi-Fi network infrastructure(s)

The IPSs based on Wi-Fi fingerprinting are preferred to those based on the *Propagation Model*, *Angle Of Arrival*, *Time Of Arrival* and *Time Difference Of Arrival* because they do not require any very specialized hardware, line-of-sight to the emitter and knowing emitter's location to operate (Zhou et al., 2014a; Yu et al., 2014). Figure 1 shows a example of fingerprinting where a smartphone detects 6 WAPs. However radio-wave propagation through indoor environments is harsh (Karimi, 2013), so there is no guarantee that the Received Signal Strength Indicator (RSSI or, simply, RSS) values from a given WAP collected at different location inside a building would reflect the architectural aspects such as similarity among Wi-Fi fingerprints taken at the same floor (Campos et al., 2014). Although there exists a general model to represent the RF signal propagation loss, it is not possible to create an appropriate model for a particular scenario since it would be altered by many factors (Ward et al., 1997; Yim, 2008). Moreover, this model requires to know the position of emitter, which may be unknown

Wi-Fi fingerprinting is a complex subject which can profit by well-established Expert Systems techniques by implementing *Machine Learning* techniques (k-NN, Neural Networks, Support Vector Machines, Decision Trees, among others). Since RADAR Indoor Positioning System publication (Bahl & Padmanabhan, 2000), many IPS use k-NN algorithm (Cover & Hart, 1967) and the *Euclidean distance* as base metric, even in recent contributions (Yu et al., 2014; Hu et al., 2015). In this paper, we mainly concentrate on the Wi-Fi Indoor Positioning Systems based on fingerprinting and k-NN

After analyzing the already developed techniques based on Wi-Fi and the k-NN based Indoor Positioning Systems, we realized that there are three main issues related to fingerprinting and k-NN algorithm

- The first issue is the lack of a deep comparison of distance/similarity measurements for the k-NN algorithm and Wi-Fi fingerprinting. Even in recent con-

tributions (e.g., Haque & Assi (2015); Yang et al. (2015)), the *Euclidean distance* is the measurement considered for calculating the distance or similarity between two signals provided by the same WAP in most of cases. However, there are many more alternative functions to be used with k-NN as it can be seen in the comprehensive survey on distance/similarity measurements presented by Cha (2007). Moreover, most of observations contains missing values and the *Euclidean distance* may not be appropriate Calderoni et al. (2015) to deal with them. According to our experiments, some WAPs can appear and disappear in a fingerprint with respect to a prior fingerprint taken at the same place less than a second before. This behaviour tends to occur with weak signals (very low RSS values)

- The second issue is that the logarithmic nature of the Wi-Fi fingerprints is not usually considered and the differences between fingerprints are linearly computed. For instance, the difference between a received signal strength of $-100dBm$ and $-97dBm$ has the same weight in fingerprinting than the difference between $-60dBm$ to $-57dBm$ when the they are linearly computed. The difference is only $3dBm$ in both cases, but the power differences in Watts (W) are $0.1pW$ and $1nW$ ($1000pW$) respectively. Although this difference is significantly different in both cases, it is not usually considered in existing techniques

- The third issue is that fingerprints contain information provided by far detected WAPs. Some previous works only considers the strongest signals to provide indoor locations, so they have introduced thresholding techniques to remove, a priori, irrelevant information from fingerprints. They assume that this information should be removed because it injects noise in the expert system used for providing indoor location. In contrast, some other works do not apply thresholding. It is not clear whether applying thresholding benefits or worsen the accuracy of IPSs.

To deal with these three interesting issues we introduce a deep comparative study where:

1. We have tested the performance of more than 50 different measures in the real problem of indoor positioning
2. We have performed performing an analysis of 4 alternative data representations, two of them consider the logarithmic nature of Wi-Fi signals propagation and have been proposed by us in this paper
3. We have studied if the presence of very far detected WAPs degrades the accuracy of an indoor localization system by applying thresholding based on the RSS values

As far as we know, there is not any prior comprehensive study for Wi-Fi fingerprinting in the realm of expert systems and machine learning for indoor positioning, such as the one we introduce in this paper

The remainder of this work is organized as follows. Section 2 shows the related work about distance/similarity measures, data representation and thresholding. Section 3 introduces the material and methods used in this contribution. Section 4 describes the experiments carried out and shows the results. Section 5 discusses about the results and compared the new alternatives to the traditional 1-NN based on the *Euclidean distance*. Section 6 brings the conclusions

2. Related Work

Although the global navigation satellite systems (GPS, GLONASS or GALLILEO) work extremely well in most of outdoor scenarios, their accuracy drastically decreases in indoor scenarios due to multi-path propagation, signal distortion, refraction, and absorption, among other factors. Thus the necessity of developing intelligent and accurate Indoor Positioning Systems that should operate in "any" indoor scenario. Although many indoor localization technologies have been proposed or adopted to attempt it, we

focus on Wi-Fi fingerprinting due to the ubiquity of wireless networks and the proliferation of such kind of low-cost techniques

Wi-Fi fingerprinting is based on pattern-matching algorithms such as the distance-based technique *k-Nearest Neighbor* (Cover & Hart, 1967) (k-NN), which is detailed in Section 3.2. Since the development of RADAR, the *pioneer* IPS based on k-NN (Bahl & Padmanabhan, 2000), many similar systems have been proposed. Although other advanced techniques have also been used to elaborate IPSs based on signal fingerprinting (*Bayesian Inference* (Zhou et al., 2014b), *Neural Networks* (Kuo et al., 2013; Campos et al., 2014), *Decision Trees* (Yim, 2008), *Random Forest* (Calderoni et al., 2015), among others), we concentrate our work on k-NN based techniques in this paper due to its presence in the literature

2.1. Measurements for k-NN

Only a few well-known distance metrics appear in most of the Indoor Positioning Systems based on k-NN. Moreover, the study of which distance/similarity function could perform better in the developed methods is not usually included. Furthermore, a basic k-NN algorithm is sometimes implemented for comparison purposes when advanced or complex methods are proposed. Many of the developed or proposed k-NN based algorithms use just one metric which usually is the *Euclidean distance*, the *Manhattan distance* or, in a few cases, the *Mahalanobis distance*.

Yim (2008) introduced decision tree-based IPS to increase the efficiency of IPSs such as k-NN. In the experimental setup, the reference fingerprint database contains the average of measures taken at the same reference point instead of having the full discrete measures set. Although the computational efficiency of their decisions tree-based IPS was superior to the simple 1-NN, the accuracy of 1-NN was better or similar in most of the cases they presented. However, the distance metric used for 1-NN was not explicitly described

Yu et al. (2014) introduces a Cluster-based version of the k-NN technique to reduce signal unsteadiness, reduce interferences and improve positioning. They only considered the *Manhattan distance* as *performance metric* since it was slightly better than *Euclidean distance* (according to prior works they referenced) and their workspace was a rectangular rectangle

Calderoni et al. (2015) introduces Random Forest in a positioning system based on RSS values of RFID beacons. For clustering purposes, the *Partial distance* is used to compare the reference fingerprints to the cluster centroids (in the RSS space). This distance is based on the *Euclidean distance* and it is used to minimize the missing values in the observations (fingerprints). Moreover, another interesting step is data normalization to represent data in a easier-to-manage format

However, many works based on k-NN only consider the *Euclidean distance* (or a equivalent one) as base metric to compare two Wi-Fi fingerprints. Campos et al. (2014) proposes a technique that combines natural (only RSS space) and architectural data (architectonic restrictions) to provide indoor location. Zhuang et al. (2014) introduces a matching weight coefficient based on RSS values and order to provide Indoor Location. In case of multiple candidates, the proposed system uses the 1-NN algorithm to assign the location. Chapre et al. (2014) also uses *Channel State Information* and exploits the frequency & spatial diversity using *Multiple Input Multiple Output*. Li et al. (2015) combines Wi-Fi fingerprinting (as in (Cheng et al., 2014)) and *magnetic matching* to enhance the accuracy of the positioning systems. k-NN and *Euclidean distance* are the basis of many modern IPS, even though RADAR was introduced in 2000

There are some exceptions where several distance/similarity functions have been considered. del Corte-Valiente et al. (2009) introduced a comparison with 5 different measures including: *Euclidean*, *Manhattan*, *Chi-Squared*, *Bray-Curtis* and *Mahalanobis*. That last measure, *Mahalanobis*, provided the lowest error in positioning and this particular measure has also been widely used or reviewed in other works (Duvallet

& Tews, 2008; Biswas & Veloso, 2010; Beder & Klepal, 2012)

Machaj et al. (2011) proposed an elaborate rank-based fingerprinting algorithm. To compare two rank-based fingerprints, they tested the following rank distance measures: *Spearman distance*, *Spearmans footrule*, *Jaccard coefficient*, *Hamming distance* and *Canberra distance*. Their algorithm provided the best results when the *Spearmans footrule* similarity measure was used

Marques et al. (2012) performed a study of the impact of several similarity functions on the accuracy in their WI-FI fingerprint-based positioning system. They studied the *Euclidean*, *Manhattan* and *Tanimoto* distances, and the experimental results proved that the *Manhattan's* one was the best choice

Farshad et al. (2013) investigated different definitions for Wi-Fi fingerprinting. They used the deterministic k-NN technique with three different distance measures: *Euclidean distance*, *Manhattan distance* and *Mahalanobis distance*. According to their experiments, the *Manhattan distance* seems to be slightly better than the other two metrics in an office environment. However, in a shopping centre environment, the *Mahalanobis distance* was better choice

Hu et al. (2015) introduces a new metric which combines the *Minkowsky distance* for computing the average RSS distance and the *Jaccard similarity* to get WAP sets similarity. This new distance is compared to traditional methods including 1) selecting the shared WAPs with strong RSS values and 2) filling non-observed WAPs with a weak RSS value. However, traditional distances and similarity metrics are not explicitly included in this comparison. Moreover, a semi-supervised affinity propagation version of the Weighted k-NN algorithm is successfully proposed

Niu et al. (2015) introduces ZIL: a Wi-Fi fingerprinting system that uses ZigBee radio. A weighted fingerprint matching algorithm is applied to align a pair of fingerprints effectively. Then, the k-NN algorithm is implemented with three different distances the *weighted Euclidean distance*, the *weighted Manhattan distance* and *relative*

entropy (also known as *Kullback-Leibler divergence*). They found that the *weighted Manhattan distance* provided the best performance

However, there are many more alternative functions to be used for this particular problem as it can be seen in the comprehensive survey on distance/similarity measurements presented by Cha (2007). In particular, 45 different measures were reviewed and categorized into 8 families depending on their similitude. Some of them were well-known measures such as the *Euclidean distance* or the *Manhattan distance* (*City Block* - L_1 in that paper). Cha performed a study to assess the similarities among the 45 well-known measures with cluster analysis and randomly generated data; and a new family was introduced with 6 new measures based on the extracted syntactic relationships. One of the objectives of this paper is to assess the efficiency and suitability of these 51 $(45 + 6)$ measures on the real *Indoor Location* problem

2.2. Logarithmic strength values and nature of signal propagation

Fingerprinting techniques relies on the Received Signal Strength values. A RSS value indicates the power of a particular received Wi-Fi (IEEE 802.11) radio signal and it is expressed in dBm. In this particular case, it indicates the ration between the received signal's power and a reference power of one milliwatt (mW) according to Eq.1

$$RSS = 10 \cdot \log_{10}\left(\frac{Power_{mW}}{1mW}\right) \tag{1}$$

Figure 2 shows the relation between dBm and mW for the $[-104, \ldots, 0]$ and $[-75, \ldots, -45]$ intervals according to Eq.1. It can be noticed that the same difference in dBm have different weight depending on the signal strength values themselves. A difference between $0dBm$ & $-10dBm$ corresponds to a difference of $0.9mW$ $(9 \cdot 10^{-1}mW)$, the difference between $-50dBm$ & $-60dBm$ is $0.000009mW$ $(9 \cdot 10^{-6}mW)$, and the difference between $-90dBm$ & $-100dBm$ is $0.0000000009mW$ $(9{\cdot}10^{-10}mW)$. However, the computed difference in existing fingerprinting algorithms

tends to be the same for the three cases, $10dBm$

Existing Wi-Fi fingerprinting techniques does not consider the relation between the signal strength and power as shown in (Gansemer et al., 2010a; Segou et al., 2010; Chen et al., 2013) and most of the papers reviewed in Section 2.1. In this paper we propose two new alternative ways to represent RSS values. One based on the exponential function and the other based on the pow function. Our aim is to improve the integration of dBm logarithmic scale with k-NN methodologies for Wi-Fi fingerprinting

2.3. Thresholding to remove WAPs

Thresholding consists in removing the RSS values from a fingerprint that may provide irrelevant information or introduce noise. This technique is, a priori, not directly related to feature selection or LDA-PCA. Some technical forums and scientific works suggested that a static threshold values should be introduced. In fact, constant-based thresholding is applied in the experimental evaluation carried out in Yim (2008) and Stella et al. (2014), whereas Niu et al. (2015), Hu et al. (2015) and Machaj & Brida (2015) use RSS-value based thresholding

Yim (2008) only processes the information provided by 5 different WAPs to provide indoor location inside the Micro Lab as suggested in Kaemarungsi & Krishnamurthy (2004). Stella et al. (2014) only selects the 3 strongest detected WAPs to generate the fingerprints. Although this particular kind of thresholding was useful in those works, it would not be appropriate for general use IPSs. One main disadvantage of these techniques is that a single Advanced Wireless Access Points (AWAPs) can emit multiple Virtual Wireless Access Points (VWAPs) through a single antenna (Farshad et al., 2013). When we select the n strongest signals, they could be provided by the same hardware emitter and, therefore, positioning accuracy may drastically decrease

Martin et al. (2010) stated that values below $-85dBm$ were too inconsistent to be leveraged as reference, and values above $-80dBm$ were desirable. Lin & Hung (2014) introduced a threshold value of $-70dBm$ for indoor localization, whereby RSS values

lower than $-70dBm$ were not eligible for indoor localization. ZIL (Niu et al., 2015) introduces a Wi-Fi fingerprinting using ZigBee radio where values below $-90dBm$ are omitted. This is also the case of Machaj & Brida (2015), where the threshold value for Wi-Fi RSS values is also $-90dBm$. As above-mentioned, Hu et al. (2015) compares their proposed metric to one that only considers the shared strongest WAPs with a RSS value higher $-85dBm$. In most of the previous thresholding examples, the experiments were carried at controlled single-building single-floor scenarios in university buildings

Slightly different thresholding strategies have also been applied. Gansemer et al. (2010b) proposed a positioning algorithm with four threshold parameters, three of them directly related to RSS values in dBm. Kai et al. (2013) introduced an adjustable WAP filter based on the number of samples containing the WAP identifier

Thresholding has been used to filter the weakest signals from Wi-Fi fingerprints. Due to the variability of threshold values and strategies found in the literature, we also include an experiment to determine whether the presence of very far detected WAPs degrades the accuracy of an indoor localization system. In particular, we have concentrated in RSS-value based thresholding for the experiments

3. Material and methods

This section describes materials and methods used in the experiments. First, the *UJIIndoorLoc database* is briefly described. Then, we describe the *k-Nearest Neighbor* (k-NN) classifier and how it has been adapted to perform indoor localization. Moreover this section shows the distance/similarity measures that have been used to perform the experiments and the four alternatives we have used to represent the fingerprint vector values. The measures have been grouped according to their syntactic similarities as in Cha (2007). Finally, the procedure to study whether it is necessary to filter low-intensity values from Wi-Fi fingerprints is explained.

3.1. The UJIIndoorLoc Database

The *UJIIndoorLoc* (Torres-Sospedra et al., 2014) is a multi-building & multi-floor database based on Wi-Fi fingerprinting, that covers a real scenario with three heterogeneous buildings. The database contains 21048 Wi-Fi fingerprints, each fingerprint is represented as a 520-element vector. Each vector contains the original intensity value for the detected WAPs and the default value $+100dBm$ for those WAPs that were not detected. Moreover, the database is split into two well differentiated sets: the training set (reference fingerprints) with 19937 fingerprints and the validation set (operational fingerprints) with 1111 records. The former one contains the reference database, and the later is used to test the positioning system based on k-NN.

With this database, the results of the Indoor Positioning System are more realistic since: 1) It was generated by means of more than 20 users and with 25 different devices; 2) Some samples for validation where taken by users and devices that not participated in generating the training set; and 3) The validation fingerprints were taken four months later than the reference (training) ones. These three features negatively affect the precision of an IPS. For instance, according to prior experiments we performed, testing an IPS with the same device and user reports very low positioning error when the evaluation is done a few days after the reference fingerprints were taken. *UJIIndoorLoc* the obtained error in positioning is more realistic. The aim of this paper is not to introduce a high-precision Indoor Positioning System, but introducing a realistic comprehensive study whose conclusions would be used by researchers to improve their already developed or new systems

We have selected UJIIndoorLoc to perform the experiments since it covers a multi-building multi-floor scenario, the training set and the validation set are well-differentiated, and the diversity of samples has been guaranteed by collecting data with different users and devices. Moreover, researchers and developers can also use this database to evaluate their already existing solutions or their new methodologies, distance/similarity

metrics and, even, data representations for indoor positioning. So existing and further advances could be directly compared to the alternatives we have implemented and proposed in this paper

More information about this database can be found in Torres-Sospedra et al. (2014). This database has been donated to the UCI repository of Machine Learning (Bache & Lichman, 2013) and it is publicly available[1] for research purposes. So, the results shown in this paper can be reproduced or fairly compared to other IPS systems based on Wi-Fi fingerprinting.

3.2. k-NN for Indoor localization

The *k-Nearest Neighbor* rule (k-NN) is a distance-based classifier which compares a current sample to all the labelled samples from a database (Cover & Hart, 1967). This classifier requires generating a database for the comparisons (commonly known as a training set) where all the samples are properly labelled. In the case of indoor positioning, the samples are Wi-Fi fingerprints (vectors with the WAP intensities) and the labels are the numerical values related to the real-world coordinates (longitude, latitude, altitude/floor and building).

There are two steps to estimate the position of a current fingerprint with k-NN:

1. The distance (or similarity) with respect to the current fingerprint is calculated for all the training fingerprints.
2. The k nearest (or most similar fingerprints) in the feature space are used to obtain the estimated position. In the simplest scenario, $k = 1$, the k-NN algorithm calculates the distance of a current fingerprint with respect to all the training fingerprints. The current position corresponds to the position (as longitude, latitude and altitude) of the training fingerprint which reported the lowest distance (or highest similarity).

[1] `https://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc`

However, the concept *classification* is complex in indoor positioning since a *class label* includes: *longitude*, *latitude*, *altitude/floor* and *building*. So, the procedure to obtain the final position should be more elaborated for $k > 1$. Particularly, we propose that the final position can be estimated as follows:

1. The distance/similarity between the current fingerprint and all the fingerprints recorded at the reference database is calculated.
2. The k "*nearest*" reference fingerprints (to the current test one) are extracted according to the measure employed (lowest distance or highest similarity).
3. A simplest voting procedure is used to estimate the building from the k-nearest fingerprints. Each extracted fingerprint provides a single vote to the building in which the reference fingerprint was taken. The most voted building is assigned as the estimated building.
4. The simplest voting procedure is also applied to estimate the floor inside the previously estimated building. In this case, those fingerprints that belong to the estimated building are only considered. The most voted floor inside the estimated building is assigned as the estimated floor.
5. The centroid (longitude and latitude) is calculated using only the fingerprints that belong to the estimated building and estimated floor.

The k-NN distance-based classifier has only one parameter (apart from the distance/similarity measure used to rank the training instances) which has to be properly set to obtain optimal results: the value of k. This values represents the number of samples from the database (fingerprints whose position is well-known) which are used to estimate the position of a device from a Wi-Fi fingerprint. Setting a low value of k, such as 1, may be not adequate since only a single sample is considered to estimate the final position, whereas a high k value may consider points which are very far from the current position and, therefore, degrade the IPS accuracy Bahl & Padmanabhan (2000).

Note that providing an optimal Indoor Localization algorithm is out of this contri-

bution's scope. Therefore, the simple k-NN algorithm, which has high computational cost at operating stage, is used to perform the comprehensive study

3.3. Distance & Similarity Measures for k-NN

As previously mentioned, Cha (2007) introduced a comprehensive study of 51 distance and similarity measures using cluster analysis with random data. One of the main contributions of the present paper is to extend the study of Cha (which was done with synthetic data) by using data from a real problem. Therefore, their relationship and suitability can be obtained with real data. The current subsection briefly reviews the 9 categories introduced by Cha to group all the measures.

We will use the nomenclature used in Cha (2007) herein after. There are well-known measures that seem to have been omitted but they appear under another name. For instance, the *Manhattan distance* used in Marques et al. (2012) appeared as *City Block* L_1 in Cha (2007). In addition, very different measures can have similar name. This is the case of *Jaccard distance* and *Jaccard coefficient* in Machaj et al. (2011).

The distance/similarity equations of this paper are based on the equations shown by Cha (2007). As an illustrative example, Eq.(2) shows the equation for the *Euclidean distance*. Where P and Q refer to the two vectors, the distance between them is being calculated; and d refers to the vectors length (number of features of the Wi-Fi fingerprinting problem).

$$distance_{euclidean}(P,Q) = \sqrt[2]{\sum_{i=1}^{d} |P_i - Q_i|^2} \tag{2}$$

3.3.1. Family 1: The Minkowski family

The *Minkowski family* measures, L_p, includes: the *Euclidean distance* (L_2); the *City Block distance* (also known as *Manhattan* or *Taxicab* distances) (L_1); the *Minkowski distance* (L_p); and the *Chebyshev distance* (L_∞). The *Minkowski distance*, see Eq.(3),

is the generalized formula for this family.

$$distance_{L_p}(P,Q) = \sqrt[p]{\sum_{i=1}^{d} |P_i - Q_i|^p}, \forall p \in N^+ \tag{3}$$

Note that not all equations are shown in this paper because they are already fully detailed in Cha (2007). Only one is shown to see the differences between families.

3.3.2. Family 2: The L_1 family

The L_1 *family* measures are based on the *City Block distance* (L_1) and this family includes: *Sørensen distance*, *Gower distance*, *Soergel distance*, *Kulczynski distance*, *Canberra distance* and *Lorentzian distance*. In distance-based methods, such as k-NN, *Gower distance* (see Eq.(4)) and the original *City Block* (see Eq.3 with $p = 1$) distances are equivalent since the distance provided by *Gower* is the *City Block* value divided by a constant (the number of features).

$$distance_{gower}(P,Q) = \frac{1}{d} \cdot \sum_{i=1}^{d} |P_i - Q_i| = \frac{1}{d} \cdot distance_{L_1}(P,Q) \tag{4}$$

Although the distance value provided by the other L_1 *family* measures are not proportional to the values provided by the original L_1 measure, most of them include the following term: $\sum_{i=1}^{d} |P_i - Q_i|$. For instance, Eq.(5) shows the *Sørensen distance*.

$$distance_{sorensen}(P,Q) = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} (P_i + Q_i)} \tag{5}$$

3.3.3. Family 3: The Intersection family

The *Intersection family* contains: *Intersection distance*, *Wave Hedges distance*, *Czekanowski distance*, *Motyka distance*, *Kulczynski similarity*, *Ruzicka similarity* and *Tanimoto distance*. In distance-based methods, such as k-NN, some distances are equivalent. This is the case of *Soergel*, *Tanimoto* and *Ruzicka*. Moreover, both *Kulczynski* measures are inversely proportional.

It is worth mentioning that some *family 3* measures resemble the L_1 *family* since they include the $|P_i - Q_i|$ term. In fact, *Czekanowski* and *Sørensen* are the same mea-

sure (note that *Sørensen distance* was catalogued as L_1 family) see Eq.(6). This is also the case of *intersection distance* (see Eq.(7)) which is proportional to *Gower distance* as denoted with $\propto$ symbol.

$$\begin{aligned} distance_{czekanowski}(P,Q) = & \quad 1 - similarity_{czekanowski}(P,Q) \\ = & \quad \frac{\sum_{i=1}^{d}|P_i - Q_i|}{\sum_{i=1}^{d}(P_i + Q_i)} \\ = & \quad distance_{sorensen}(P,Q) \end{aligned} \tag{6}$$

$$distance_{intersection}(P,Q) = \frac{1}{2} \cdot \sum_{i=1}^{d} |P_i - Q_i| \propto distance_{gower}(P,Q) \tag{7}$$

3.3.4. Family 4: The Squared L_2 family

The *Squared L_2 family* or χ^2 *family* is based on the *Euclidean distance* and includes the following distance measures: *Squared Euclidean*; *Pearson* χ^2; *Neyman* χ^2; *Squared* χ^2; *Probabilistic Symmetric* χ^2; *Divergence*; *Clark*; and *Additive Symmetric* χ^2. All of these contain the squared of *Euclidean distance* term, $(P_i - Q_i)^2$, weighted by different factors. For instance, the *Neyman* χ^2 *distance* is shown in Eq.(8):

$$distance_{neyman}(P,Q) = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i} \tag{8}$$

3.3.5. Family 5: The Inner Product family

The *Inner Product family* radically differs from the previous families and introduces the scalar product of two vectors. This product provides a scalar value and, according to Cha (2007), it corresponds to the number of matches if it is used for binary vectors. As stated by Cha, most of this family measures are frequently used in information retrieval and biological taxonomy for the binary feature vector comparison. In this family, the measures are not proportional among them and they are: *Inner Product similarity* in Eq.(9); *Harmonic mean similarity*; *Cosine similarity*; *Kumar-Hassebrook similarity*;

Jaccard distance; and *Dice distance*. Eq.(9) shows the *Inner product distance* as an example of this family.

$$similarity_{innerproduct}(P,Q) = \mathbf{P}\cdot\mathbf{Q} = \sum_{i=1}^{n}(P_i \cdot Q_i) \tag{9}$$

3.3.6. Family 6: The Fidelity family

The sixth family is the *Fidelity family* or *Squared-chord family* and it includes: *Fidelity similarity* (see Eq.(10)), *Bhattacharyya distance*, *Hellinger distance*, *Matusita distance* and *Squared-chord distance*. This family resembles the measures introduced in the *Inner Product family*, but the square root is applied to the vector values.

$$similarity_{fidelity}(P,Q) = \sum_{i=1}^{n}\sqrt{P_i \cdot Q_i} \tag{10}$$

3.3.7. Family 7: The Shannon's Entropy family

Shannon's Entropy family contains those distance measures based on the *Shannon's* concept of probabilistic uncertainty: *Kullback Leibler* (see Eq.11), *Jeffreys*, *K divergence*, *Topsøe*, *Jensen-Shannon*, and *Jensen difference*. *Jensen-Shannon distance* corresponds to *Topsøe* divided by 2.

$$distance_{kullback-leibler}(P,Q) = \sum_{i=1}^{n}\left(P_i \cdot \log\frac{P_i}{Q_i}\right) \tag{11}$$

3.3.8. Family 8: The Combinations family

Combinations family contains all those distance measures which combine different approaches: *Taneja*; *Kumar-Johnson*; and $Avg(L_1, L_\infty)$. In fact, $Avg(L_1, L_\infty)$ see Eq.(12), corresponds to the mean value provided by *City Block* (L_1) and *Chebyshev* (L_∞) distances.

$$distance_{Avg(L_1,L_\infty)}(P,Q) = \frac{distance_{L_1}(P,Q) + distance_{L_\infty}(P,Q)}{2} \tag{12}$$

3.3.9. Family 9: The Vicissitude family

Additionally, Cha (2007) included six distances which were not in the literature. They were grouped into the *Vicissitude family*: *Vicis-Wave Hedges*; *Vicis-Symmetric* χ^2

(with three different variants); *min-Symmetric* χ^2; and *max-Symmetric* χ^2. As stated by Cha, a large number of new distance/similarity measures can be relayed by studying the syntactic relations and may be useful in some applications. For instance, the equation for the third version of the *Vicis-Symmetric* is given by Eq.(13):

$$distance_{viccissymmetric3}(P,Q) = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{max(P_i, Q_i)} \tag{13}$$

3.4. Fingerprint data representation

Each fingerprint contains a list with the intensity values of the detected WAPs. Fingerprints can be represented as fixed-size vectors (such as in the *UJIIndoorLoc* database) where each index corresponds to a WAP registered by the IPS. So, the vector representation contains the original intensity values and a default significant value is used to denote those WAPs which were not detected.

Some distance and similarity measures do not allow the use of negative values. Several measures apply the square root to the vector values, and some other measures are based on logarithmic values. Four alternatives were initially considered for representing the RSS levels with positive values: *positive values* Eq.(14); *zero-to-one normalized values* Eq.(15); *exponential values* Eq.(16); and *powed values* Eq.(17). Where the two last ones, *exponential* and *powed*, have been introduced by us in this contribution.

Positive values data representation, Eq.(14), simply subtracts the minimum possible value. So, new low values stand for low signal, whereas higher values indicate that the signal is stronger. In this representation, the lowest possible value is 0 and it is used to show that the WAP has not been detected.

$$Positive_i(x) = \begin{cases} (RSS_i - min) & \textit{If } WAP_i \textit{ is present in the fingerprint } x \\ & \textit{and } RSS_i \geq \tau \\ 0 & \textit{otherwise} \end{cases} \tag{14}$$

where i is the WAP identifier, and RSS is the actual intensity level provided by i-th WAP. min is the lowest RSS value minus 1 considering all the of fingerprints and WAPs of the database. τ is the threshold value, these intensities lower than the threshold are considered as not-detected WAPs, and the lowest possible value is assigned in the new representation. In some indoor systems, τ is statically set to a fixed *default* value (Gansemer et al., 2010b; Kai et al., 2013; Martin et al., 2010; Lin & Hung, 2014), e.g. $-75dBm$ or $-85dBm$. To avoid thresholding, τ should be set to a value lower than the minimum possible RSS value, i.e lower than $-104dBm$ for the dataset used in this paper.

Similarly, we introduced the *zero-to-one normalized* representation, Eq.15, which corresponds to the *positive values* representation but the intensity values are normalized in the positive $[0 \dots 1]$ range.

$$ZeroToOneNormalized_i(x) = \frac{Positive_i(x)}{-min} \tag{15}$$

These two last representations (*positive* and *zero-to-one normalized*) maintain the linearity of the original values. Note that τ is introduced in Eq.(14) for thresholding purposes. Although some previous works such as Gansemer et al. (2010b) ,Kai et al. (2013), Martin et al. (2010), and Lin & Hung (2014) applied threshold values, we have initially considered setting min as the threshold value to avoid removing any intensity. Instead, we have introduced a thresholding study in Section 2.3 to assess the suitability

of applying threshold values to Wi-Fi fingerprints.

Although in the majority of works the intensity levels are used as lineal values as provided by the devices, the RSS values are provided in decibels dBm which has a logarithmic scale. We introduce the *exponential representation*, Eq.16, and *powed representation*, Eq.17, to break linearity of the original intensity values provided by the device. In both equations, we have introduced α and β to represent the case-based parameters for *exponential* and *powed* representations respectively.

$$Exponential_i(x) = \frac{\exp\left(\frac{Positive_i(x)}{\alpha}\right)}{\exp\left(\frac{-min}{\alpha}\right)} \tag{16}$$

$$Powed_i(x) = \frac{(Positive_i(x))^{\beta}}{(-min)^{\beta}} \tag{17}$$

According to prior experiments we carried out, signal fluctuations are generally more likely to occur when the transmitter is far. If the distance with respect to the transmitter (WAP) is medium/high, then the probability of interferences is higher. E.g., there can be more people and moving objects between the transmitter and the device which is capturing the signal and the signal can be blocked by them. If the device is near to a WAP, then the values received have less fluctuations across time. So, the two last equations (Eq.16 and Eq.17) tend to represent the RSS values as they really are. Moreover, they also tend to penalize more those fluctuations related to good signal values, and the differences of WAPs reporting weak/low signal have less importance. E.g., the new values for $-104dDm$ and $-94dBm$ (very low signal) are practically identical to these two new representations but the new normalized values for $-11dBm$ and $-1dBm$ (very strong signal) are very different.

Figure 3 introduces the relationship between the original captured values and the new values provided by the four representations. The relation provided by *positive values* and *zero-to-one normalized* is equivalent, and it corresponds to the lineal relation shown in the graphic. It can be seen that *positive* and *zero-to-one normalized*

representations maintain the linearity of the values. Moreover, Figure 3 also shows that *exponential* and *powed* representations penalize differences when the signal strength is high.

Finally, the *exponential* and *powed* representations required to setup a few parameters: the denominator constant (α) in the former was set to 24, and the exponent (β) in the later was set to the mathematical constant e. Although these parameters were selected after performing some prior experiments, this optimization is out of the paper's scope.

4. Experiments and results

This section presents the experimental results and the discussion which include: evaluation of the measures with different data representations, thresholding study and setting the value of k.

4.1. Experiment 1: Distance/Similarity measures evaluation using four different data representations

Four different vector-based representations have been proposed to represent the original Wi-Fi fingerprint values (see Section 3.4). Table 1 introduces the results of this first experiment where the $success$ and $error$ are shown using the fingerprints from the validation set. $Success$ corresponds to the percentage of testing fingerprints whose building and floor is correctly predicted, whereas $error$ corresponds to the real world distance between the actual position (stored in the validation set) and the predicted position. This error only considers the fingerprints whose building and floor was correctly predicted. So, the best positioning alternative is the one which provides the highest $success$ rate and the lowest $error$. In Table 1, the best case (best data representation) for each measure is highlighted in bold print. Moreover, 53 different measures have been tested with 1NN (k-NN with $k = 1$), 1NN has been used to assess the suitability of the measures with a simple estimator, note that there are three implementations of the *Minkowski* L_n measure ($n = 3$; $n = 4$; and $n = 5$). The measures

are introduced and grouped in the table according to the families reviewed in Section 3.3.

For clarification purposes, the analysis of results are shown in three different parts.

4.1.1. Relationship/Equivalences among measures

According to the results shown in Table 1, some measures are proportional and/or equivalent when they are used in k-NN algorithms. In fact, there are only a total of 36 different non-proportional alternatives to use as base measure for algorithms based on k-NN from the 53 measures tested (there were 3 versions of the *Minkowski L_3 distance*). An example of this fact is the case of *City Block L_1*, *Gower*, and *Intersection* (Group 1 of equivalent measures) which provide the same *success* and *error* for the four data representations. However, these three measures are different and they are catalogued into different families. Although these three measures may provide different distance/similarity values, the values they provide are proportional to $\sum |p - q|$. So, they are equivalent when they are used in distance-based ranking algorithms such as k-NN. This behavior also occurs with higher values of k, but complete results have not been provided here because optimizing the value of k due to the table size. Note that the term *Group* and *Family* are not directly related. In fact, *Group 1 of equivalent measures* contains measures from three different families. The three mentioned equivalent measures corresponds to an extract of equivalent measures, the equivalent measures are:

Group 1: *City Block L_1*, *Gower*, and *Intersection*

Group 2: *Euclidean L_2* and *Squared Euclidean*

Group 3: *Sørensen*, *Soergel*, *Kulczynski distance*, *Czekanowski*, *Motyka*, *Kulczynski similarity*, *Ruzicka* and *Tanimoto*

Group 4: *Squared χ^2* and *Probabilistic Symmetric χ^2*

Group 5: *Kumar-Hassenbrook*, *Jaccard distance* and *Dice*

Group 6: *Fidelity* and *Bhattacharyya*

Group 7: *Hellinger*, *Matusita* and *Squared-Chord*

Group 8: *Topsøe* and *Jensen difference*

4.1.2. General analysis of data representation

We second analyze the four data representations from Table 1. In general, the best result for each measurement is provided by the *exponential representation* or the *powed representation* considering the *success* and the *error*. The lineal representations (*positive representation* and the *zero-to-one normalized representation*) are equivalent in almost all of the measures, except for *Lorentzian distance* and *Additive Symmetric* χ^2. Moreover, both lineal representations provide the best result only for the *Inner Product*. Thus, in the case of Wi-Fi fingerprint, selecting the most appropriate data representation is a step which should be seriously considered, because the lineal representations can be outperformed by using an alternative (with the *exponential* or *powed*) representation.

When the results for *Euclidean* L_2 *distance* are analyzed, there is not a clear winner since *exponential* and *powed* representations are both equally good according to the *error* and *success*. The former representation provided lower *error* in meters but it also provided worse *success*. Although the latter representation provided a higher *error*, the *success* was the best for the *Euclidean distance*. Anyway, both representations outperform the results obtained with the *positive* and *zero-to-one normalized* representations. The *success* increased by more than 2% and the error was reduced by one meter (*exponential representation*).

As was expected, the use of non-lineal data representations is more suitable for the indoor location problem. The *exponential* and *powed* data representations tend to represent the RSS values as they really are, and they also tend to highly penalize fluctuations related to good signal intensities. Therefore, the *success rate* and error are improved with respect to the lineal data representations.

4.1.3. General analysis of measures

The best result for each representation (*positive&normalized*, *exponential* and *powed*) are respectively: *Sørensen* (and group 3 measures), *Neyman* χ^2 and, again, *Sørensen* (and group 3 measures). The *Euclidean distance* combined with the *positive representation* (and the *zero-to-one normalized*) has a success rate of 89.92% and an error of 7.90 meters. However, this positioning accuracy was improved by *Sørensen* (and group 3 measures) with a success rate of 92.17% and an error of 7.33 meters. The success rate was thus improved with 2.25% and the error was reduced by 0.57 meters when using the same data representation of the RSS values. *Neyman* χ^2 provided a success rate of 93.79% and an error of 6.99 meters with the *exponential* representation, and *Sørensen* (and group 3 measures) provided a success rate of 94.78% and an error of 6.86 meters with the *powed* representation. In those two last cases, the difference with respect to the *Euclidean distance* and *positive representation* is remarkably high.

Sørensen (and group 3 measures) and some χ^2-based measures provide good results mainly due to the normalization realized in their expressions as stated in del Corte-Valiente et al. (2009). In fact, the normalization used in *Sørensen distance* (also known as *BrayCurtis coefficient*) provides a final distance value in $[0 \dots 1]$ range which can be considered a degree or percentage of dissimilarity that does not depend on absolute RSS values.

Finally, an indoor positioning system can be improved by selecting the appropriate measure and RSS values representation. Some already developed indoor positioning systems rely on metrics based on *Euclidean distance* or the *Manhattan's distance* combined with the positive data representation. The results presented in this section show that the *success rate* can be improved by almost a 5% and the *error in positioning* can be reduced by more than 1 meter by using an alternative measure and data representation, instead of the common measures combined with positive data representation.

4.2. Experiment 2: Thresholding analysis

Prior to analyzing the complete results provided after applying the threshold values to the RSS intensities, we performed a first test in order to obtain the effect of thresholding on the fingerprints taken in places with low Wi-Fi coverage.

Figure 4 shows the percentage of void training and validation fingerprints after applying the threshold value. The plots for training and validation samples resembles the sigmoidal function. Void fingerprints are those that do not detect any WAP, so positioning can not be done with them. As was expected, this percentage increases as the intensity threshold increases. In fact, all the fingerprints that are taken in areas with low Wi-Fi coverage, detect only a few WAPs with low intensity value. So, thresholding can remove all the detected WAPs in a fingerprint depending on the threshold value. Moreover, Figure 5 shows the mean number of detected WAPs by a device after applying threshold values. As was expected, this number also decreases as the threshold value increases.

Figures 4 and 5 show that high threshold values should not be used due to the high rate of void fingerprints. E.g. the number of void training and validation fingerprints was 345 (1.74%) and 8 (0.72%) respectively for $-85dBm$ as threshold value, and these values increased to 1361 (6.85%) void training fingerprints and 40 (3.6%) void validation fingerprints using $-75dBm$ as the threshold value. Moreover, the number of detected WAPs per fingerprint decreases as the threshold value increases. If the threshold value is $-74dBm$, then fingerprints contains an average of 6 WAPs; but it decreases to 3 WAPs when the threshold value is $-63dBm$.

Although it seems that thresholding can be disadvantageous, we decided to apply a fine-grained analysis of thresholding values using the $[-104dBm, \ldots, -75dBm]$ interval as previously did in other works. Figures 6 and 7 graphically show the $success$ and the $error$ results according to the selected threshold values. The former figure shows the results using the *Euclidean distance* with the four data representations,

whereas the later shows the results using the best measure for each representation according to the experiments introduced in Section 4.1 (*Sørensen* -and group 3 measures- for *positive*, *zero-to-one normalized* and *powed* representations, and *Neyman* χ^2 for *exponential representation*). In both plots, lineal stands for *positive* and *zero-to-one normalized* representations since both are totally equivalent for the measures herein used.

On the one hand, the $error$ tends to increase as the threshold value increases according to the plots provided for *Euclidean distance* (Figures 6 and 7); although the $error$ decreases when the threshold is around $-91dBm$ for *powed representation*, the $success$ rate also decreases for this value. On the other hand, $success$ tends to decrease as the threshold value increases. $Success$ is only improved by applying a threshold value around $-88dBm$ for *positive*, *zero-to-one* and *exponential* representations, but the corresponding $error$ is highly increased.

Similarly, the same behavior is provided for the plots shown in Figure 7. Only the threshold value of $-97dBm$ provided slightly better results for the *exponential* representation: it reported a $success$ of 93.88% and an error of $6.97m$. For the *powed* representation, the threshold value $-90dBm$ also provided slightly better results. Anyway, these few improvements were not significant since they provided an improvement in $error$ around 1 or 2 cm and an increase of $success$ lower than 0.1%.

In both cases, *Euclidean distance* (Figure 6) and best distance (Figure 7), the results do not vary by applying threshold values from $N.T.$ (where no threshold is applied) to $-100dBm$. There are only a few cases in which a WAP reported a value lower than $-100dBm$; in particular, 85 training fingerprints (0.42%) and 6 validation fingerprints (0.54%) contained intensity values lower than $-100dBm$. Therefore, the best general results (considering $error$ and $success$) are provided when the original data without thresholding is used. The threshold value may be set to $-100dBm$ but the results will not be altered.

4.3. Experiment 3: Setting the value of k for k-NN

Here we present an experiment to find the evolution of classification results by using k values inside the [1,3,5,7,9,11,13,15,17,19,21,23] range. In particular, the evolution of $error$ and $success$ rate is shown according to the value of k for the three representations. We first show this for the *Euclidean distance* as reference in Figure 8, and for the best measure for each representation in Figure 9.

For the *Euclidean distance*, in Figure 8, the best values for k are shown inside range of 5 to 9 in general. The results for this range reach: 7.16 meters of error and a success of 90.10% for the *positive* and *normalized* representations (both represented by *lineal* in the figures); 6.44 meters of error and a success of 92.53% for the *exponential* representation; and 6.98 meters of error and a success of 93.16% for *powed* representation. Note that *exponential* and *powed* representations provide the best results for this distance; the former provided the lowest error in meters and the best success rate was obtained with the latter. In general, the error reduces (approx. 70 cm for the lineal representation and 50 for the other representations) and the success rate slightly improves (less than 1%) for increasing values of k.

Figure 9 provides the best measure of the different data representations, the worst results are provided when k-NN only considers the nearest neighbor ($k = 1$). According to the $success$ rate, the interval 3 to 23 is good for: *positive*, *normalized* and *powed* representations. However, this interval is narrower for the *exponential* representation: k from 3 to 11. According to the $error$ rate, the positioning error decreases while the k value increases until k with values close to 13. From this value, $k > 13$, the $error$ increases as the value of k increases.

Regarding setting the k value, it seems that low values (e.g. $k = 1$) are not suitable because the system relies only on a single candidate or on a few similar candidates. High values, such as 23, are also not suggested because the error is raised and the success rate decreased. It seems that using an intermediate value balances the diversity

on samples and better positioning is obtained. In fact, the best overall configuration is $k = 13$ for *powed* representation since it reports an error lower than 6.2 meters and a success rate higher than 95%. It is worth mentioning that the difference of this best case with respect to the traditional alternative (*Euclidean distance* and *Positive* representation) is notably high; almost 2 meters in $error$ and more than 5% in the $success$ rate.

5. Discussion

The results obtained in our experiments provide us with new insights about how to choose a suitable distance metrics in Wi-Fi fingerprinting based applications. This new insights can be summarised as follows

First, according to the experiments performed, selecting the most appropriate data representation is a crucial step. All the measures except the *Inner Product* provide better positioning results when an alternative data representation, *exponential* or *powed*, is used instead of a traditional lineal representation. We strongly recommend to use the alternative data representations that break the data linearity in order to obtain a representation that resembles the original nature of Wi-Fi signals and penalizes differences in strong signals (close WAPs)

Second, the most suitable measure for each representation (*positive&normalized*, *exponential* and *powed*) are respectively: *Sørensen* (and group 3 measures); *Neyman* χ^2; and *Sørensen* (and group 3 measures). Moreover, other measures such as *Cosine distance* and *Kumar-Hassebrook* (group 5 of equivalent measures) also provide good results. These results justify the application of a study in depth on distance/similarity measures because the default metrics, *Euclidean distance* or *City Block*, are not always the most appropriate ones

Third, applying a thresholding method to remove those Wi-Fi signals with very low intensity is not needed because the accuracy in locating them did not significantly improve. In fact, the performance of the localization system decreases if the thresh-

old value is set to a high default value such as $85dBm$ or $75dBm$. We suggest to avoid using thresholding on fingerprinting techniques, since the accuracy decreases as threshold value increases

Fourth, selecting the most appropriate value of k for k-NN is also important to obtain more accurate results. In fact, the $success$ has reached a very high rate, 95.2%, and the positioning error has reached the lowest overall value, $6.19m$, for $k = 13$ and *Sørensen distance* (and group 3 measures) combined with the *powed* data representation. In contrast to the best configuration for k and *Euclidean distance* which provided approximately a success of 90% and an error of $7.2m$

Finally, it is worth mentioning that we have achieved a success rate of 89.92% and an error in positioning of $7.90m$ with the traditional 1-NN algorithm using the *Euclidean distance* with the UJIIndoorLoc database, which is a very challenging realistic multi-building multi-floor database. For the *Manhattan distance* (*City Block*) the error in positioning is better but the success rate is worse ($7.60m$ and 88.03% respectively). With the alternative metrics and the data representations we propose, we can achieve a high success rate of 94.78% and an error rate of $6.86m$ (*Sørensen* and *powed representation*) with 1-NN. The error in positioning has been reduced in $1m$ (which corresponds to a relative reduction of 15% approx.) and the success rate has been increased in almost 5% (which corresponds to a relative reduction of 48.21% wrong-building & wrong-floor errors) compared to the traditional approach commonly used in the literature (e.g., Marques et al. (2012); Farshad et al. (2013); Campos et al. (2014); Yu et al. (2014); Zhuang et al. (2014); Li et al. (2015), among many others). As stated in Campos et al. (2014), a medium in-floor error is tolerable, but wrong-floor errors may not be acceptable at all because it is much easier to move within the same floor than among floors. By adopting an alternative metric and data representation, we have not only decreased the in-floor error by $1m$, but also reduced the wrong-building and wrong-floor errors by a half. We consider that other existing techniques can take

benefit from the increase in success and accuracy supported by alternative metrics and data representations shown in this paper

6. Conclusions

It is hard to apply signal propagation algorithms to Wi-Fi fingerprinting positioning due to harsh nature of signal propagation, architectural conditions, diversity of emitters, so solutions based on machine learning techniques and expert systems are good candidates to tackle this problem. Most of the presented solutions are based on the k-NN distance based classifier which in turn mainly use the *Euclidean distance* as main metrics. However, it could be not the better alternative by default due to the nature of Wi-Fi fingerprints values. This paper has introduced a comprehensive comparative study of distance/similarity measures, data representation and thresholding using the challenging realistic multi-building multi-floor *UJIIndoorLoc* database

This work demonstrates that selecting the best configuration for k-NN (metric, data representation and k value) is crucial for indoor positioning. The alternative to traditional k-NN and *Euclidean distance* we propose for indoor positioning is based on the *Sørensen distance*, *powed* representation and $k = 13$. This alternative not only decreased the in-floor error by $1.7m$, but also reduced the severe wrong-place (wrong building and/or wrong floor) errors by more than a half with respect to the traditional 1-NN based on raw data and *Euclidean distance*. Moreover, this work also demonstrates that applying a thresholding method to remove weak Wi-Fi signals is not needed and should be avoided in multi-building multi-floor scenarios. It seems that the presence of distant WAPs in fingerprints can capture the temporal variability inherent to signal propagation and, therefore, provide robustness to the indoor positioning algorithm. We consider that researchers can use the herein introduced study to improve their positioning systems by applying an alternative measure and/or data representation, specially when they are based on k-NN and *Euclidean distance*

As future work we are considering to perform a comprehensive study of cluster-

ing techniques using the 51 distance/similarity measures for indoor positioning. The complexity of the reference database used in k-NN could be reduced and the computational costs at operational stage could be also highly decreased. Moreover, we consider that the application of the proposed alternatives to represent RSS data could be useful in other expert systems based on *Neural Networks*, *Support Vector Machines*, and *Bayesian Inference*. In general, those systems do not have explicit knowledge about the logarithmic nature of RSS values. Another interesting work would focus on increasing the $success$ rate as maximum as possible, which is a slightly different indoor positioning problem. In some scenarios, it is crucial that the expert system provides the correct 'place' rather than the exact coordinates

Acknowledgements

This work was supported by Ministerio de Economía y Competitividad under the project "Smart Ways" (Convocatoría Retos-Colaboración, RTC-2014-1466-4). Sergio Trilles had a grant by Generalitat Valenciana (FPIACIF/2012/112) that partially supported this work

We would like to thank all the current and past members of the *Geospatial Technologies Research Group* and *UBIK Geospatial Solutions S.L.* who contributed on creating and enriching the *UJIIndoorLoc* database

We also thank Javier Fernandez, Ángel Ramos, Álvaro Arranz and Guillermo Amat for their collaboration and comments, as members of project "Percepción" (Ministerio de Industria, Energía y Comercio, Programa Avanza2, TSI-0206012012-50)

References

Bache, K., & Lichman, M. (2013). UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.

Bahl, P., & Padmanabhan, V. (2000). 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. (pp. 775–784). volume 2.

Baniukevic, A., Jensen, C., & Lu, H. (2013). Hybrid indoor positioning with wi-fi and bluetooth: Architecture and performance. In *Proceedings of the 14th IEEE International Conference on Mobile Data Management (MDM'2013)* (pp. 207–216). volume 1.

Beder, C., & Klepal, M. (2012). Fingerprinting based localisation revisited a rigorous approach for comparing rssi measurements coping with missed access points and differing antenna attenuations. In *Proceedings of the third International Conference on Indoor Positioning and Indoor Navigation* (pp. 1–7).

Biswas, J., & Veloso, M. M. (2010). Wifi localization and navigation for autonomous indoor mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation* (pp. 4379–4384).

Calderoni, L., Ferrara, M., Franco, A., & Maio, D. (2015). Indoor localization in a hospital environment using random forest classifiers. *Expert Systems with Applications*, *42*, 125 – 134.

Campos, R. S., Lovisolo, L., & de Campos, M. L. R. (2014). Wi-fi multi-floor indoor positioning considering architectural aspects and controlled computational complexity. *Expert Systems with Applications*, *41*, 6211 – 6223.

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*, 300–307.

Chapre, Y., Ignjatovic, A., Seneviratne, A., & Jha, S. (2014). Csi-mimo: Indoor wi-fi fingerprinting system. In *Local Computer Networks (LCN), 2014 IEEE 39th Conference on* (pp. 202–209).

Chen, Y., Lymberopoulos, D., Liu, J., & Priyantha, B. (2013). Indoor localization using fm signals. *IEEE Transactions on Mobile Computing*, *12*, 1502–1517.

Cheng, J., Yang, L., Li, Y., & Zhang, W. (2014). Seamless outdoor/indoor navigation with wifi/gps aided low cost inertial navigation system. *Physical Communication*, *13, Part A*, 31 – 43. Indoor Navigation and Tracking.

Chung, J., Donahoe, M., Schmandt, C., Kim, I. J., Razavai, P., & Wiseman, M. (2011). Indoor location sensing using geo-magnetism. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services* MobiSys '11 (pp. 141–154). New York, NY, USA: ACM.

del Corte-Valiente, A., Gómez-Pulido, J. M., & Gutiérrez-Blanco, O. (2009). Efficient techniques and algorithms for improving indoor localization precision on wlan networks applications. *International Journal Communications, Networks and System Sciences*, *2*, 645–651.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, *13*, 21–27.

Duvallet, F., & Tews, A. D. (2008). Wifi position estimation in industrial environments using gaussian processes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2216–2221).

Estevez, A. G., & Carlsson, N. (2014). Geo-location-aware emulations for performance evaluation of mobile applications. In *Proceedings of the IEEE/IFIP Conference on Wireless On-demand Network Systems and Services*.

Farshad, A., Li, J., Marina, M., & Garcia, F. (2013). A microscopic look at wifi fingerprinting for indoor mobile phone localization in diverse environments. In *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on* (pp. 1–10).

Feldmann, S., Kyamakya, K., Zapater, A., & Lue, Z. (2003). An indoor bluetooth-based positioning system: Concept, implementation and experimental evaluation. In *Proceedings of the International Conference on Wireless Networks* (pp. 109–113). CSREA Press.

Gansemer, S., Großmann, U., & Hakobyan, S. (2010a). Rssi-based euclidean distance algorithm for indoor positioning adapted for the use in dynamically changing wlan environments and multi-level buildings. In *Proceedings of the 1st the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2010)*.

Gansemer, S., Pueschel, S., Frackowiak, R., Hakobyan, S., & Großmann, U. (2010b). Improved rssi-based euclidean distance positioning algorithm for large and dynamic wlan environments. *International Journal of Computing*, *9*, 37–44.

Guo, Y., Chen, Y., & Liu, J. (2014). Indoor location estimation based on local magnetic field via hybrid learning. In F. Sun, K.-A. Toh, M. G. Romay, & K. Mao (Eds.), *Extreme Learning Machines 2013: Algorithms and Applications* (pp. 189–207). Springer International Publishing volume 16 of *Adaptation, Learning, and Optimization*.

Haque, I., & Assi, C. (2015). Profiling-based indoor localization schemes. *Systems Journal, IEEE*, *9*, 76–85.

Hu, X., Shang, L., Gu, F., , & Han, Q. (2015). Improving wi-fi indoor positioning via ap sets similarity and semi-supervised affinity propagation clustering. *International Journal of Distributed Sensor Networks*, .

Ijaz, F., Yang, H. K., Ahmad, A., & Lee, C. (2013). Indoor positioning: A review of indoor ultrasonic positioning systems. In *Proceedings of the Advanced Communication Technology* (pp. 1146–1150).

Jin, G. Y., Lu, X. Y., & Park, M. S. (2006). An indoor localization mechanism using

active rfid tag. In *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing* (pp. 4 pp.–). volume 1.

Kaemarungsi, K., & Krishnamurthy, P. (2004). Modeling of indoor positioning systems based on location fingerprinting. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies* (pp. 1012–1022 vol.2). volume 2.

Kai, Z., Binghao, L., Dempster, A., & Lina, C. (2013). A comparison of algorithms adopted in fingerprinting indoor positioning systems. In *Proceedings of the International Global Navigation Satellite Systems Society Symposium*.

Karimi, H. A. (2013). *Advanced Location-Based Technologies and Services*. CRC Press.

Kuo, R., Shieh, M., Zhang, J., & Chen, K. (2013). The application of an artificial immune system-based back-propagation neural network with feature selection to an RFID positioning system. *Robotics and Computer-Integrated Manufacturing*, *29*, 431 – 438.

Kuo, Y. S., Pannuto, P., Hsiao, K. J., & Dutta, P. (2014). Luxapose: Indoor positioning with mobile phones and visible light. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking* MobiCom '14 (pp. 447–458). New York, NY, USA: ACM.

Lan, K. C., & Shih, W. Y. (2013). On calibrating the sensor errors of a pdr-based indoor localization system. *Sensors*, *13*, 4781–4810.

Lau, E. E. L., & Chung, W. Y. (2007). Enhanced rssi-based real-time user location tracking system for indoor and outdoor environments. In *Proceedings of the 2007 International Conference on Convergence Information Technology* (pp. 1213–1218). Washington, DC, USA: IEEE Computer Society.

Le, W., Wang, Z., Wang, J., Zhao, G., & Miao, H. (2014). A novel wifi indoor positioning method based on genetic algorithm and twin support vector regression. In *Control and Decision Conference (2014 CCDC), The 26th Chinese* (pp. 4859–4862).

Li, H. (2014). Low-cost 3d bluetooth indoor positioning with least square. *Wireless Personal Communications*, (pp. 1–14).

Li, Y., Zhuang, Y., Lan, H., Zhang, P., Niu, X., & El-Sheimy, N. (2015). Wifi-aided magnetic matching for indoor navigation with consumer portable devices. *Micromachines*, *6*, 747.

Lin, C. Y., & Hung, M. T. (2014). A location-based personal task reminder for mobile users. *Personal and Ubiquitous Computing*, *18*, 303–314.

Machaj, J., & Brida, P. (2015). Wireless positioning as a cloud based service. In N. T. Nguyen, B. Trawiski, & R. Kosala (Eds.), *Intelligent Information and Database Systems* (pp. 430–439). Springer International Publishing volume 9012 of *Lecture Notes in Computer Science*.

Machaj, J., Brida, P., & Piche, R. (2011). Rank based fingerprinting algorithm for indoor positioning. In *Proceedings of the 2nd the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2011)* (pp. 1–6).

Markets, & Markets (2014). *Indoor Location Market by Positioning Systems, Maps and Navigation, Location based analytics, Location based services, Monitoring and emergency services - Worldwide Market Forecasts and Analysis (2014 - 2019)*. Technical Report.

Marques, N., Meneses, F., & Moreira, A. (2012). Combining similarity functions and majority rules for multi-building, multi-floor, wifi positioning. In *Proceedings of the 3th the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2012)*.

Martí, J. V., & Marín, R. (2011). Ariel: Advanced radiofrequency indoor environment localization: Smoke conditions positioning. *Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems*, *0*, 1–8.

Martí, J. V., Sales, J., Marín, R., & Jiménez-Ruiz, E. (2012). Localization of mobile sensors and actuators for intervention in low-visibility conditions: The zigbee fingerprinting approach. *International Journal of Distributed Sensor Networks*, (p. 10 pages).

Martin, E., Vinyals, O., Friedland, G., & Bajcsy, R. (2010). Precise indoor localization using smart phones. In *Proceedings of the International Conference on Multimedia* MM '10 (pp. 787–790). New York, NY, USA: ACM.

Montaser, A., & Moselhi, O. (2014). RFID indoor location identification for construction projects. *Automation in Construction*, *39*, 167 – 179.

Neves, A. R. d. M., Carvalho, A. M. G., & Ralha, C. G. (2014). Agent-based architecture for context-aware and personalized event recommendation. *Expert Systems with Applications*, *41*, 563 – 573.

Niu, J., Wang, B., Shu, L., Duong, T., & Chen, Y. (2015). Zil: An energy-efficient indoor localization system using zigbee radio to detect wifi fingerprints. *Selected Areas in Communications, IEEE Journal on*, *33*, 1431–1442.

Segou, O., Mitilineos, S., & Thomopoulos, S. (2010). Dale: A range-free, adaptive indoor localization method enhanced by limited fingerprinting. In *Proceedings of the 1st the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2010)* (pp. 1–8).

Stella, M., Russo, M., & Begui, D. (2014). Fingerprinting based localization in heterogeneous wireless networks. *Expert Systems with Applications*, *41*, 6738 – 6747.

Torres-Sospedra, J., Avariento, J., Rambla, D., Montoliu, R., Casteleyn, S., Benedito-Bordonau, M., Gould, M., & Huerta, J. (2015). Enhancing integrated indoor/outdoor mobility in a smart campus. *International Journal of Geographical Information Science*, .

Torres-Sospedra, J., Montoliu, R., Martínez-Usó, A., Avariento, J. P., Arnau, T. J., Benedito-Bordonau, M., & Huerta, J. (2014). 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*.

Ward, A., Jones, A., & Hopper, A. (1997). A new location technique for the active office. *Personal Communications, IEEE*, *4*, 42–47.

Yang, L., Chen, H., Cui, Q., Fu, X., & Zhang, Y. (2015). Probabilistic-knn: A novel algorithm for passive indoor-localization scenario. In *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st* (pp. 1–5).

Yim, J. (2008). Introducing a decision tree-based indoor positioning technique. *Expert Systems with Applications*, *34*, 1296 – 1302.

Yu, F., Jiang, M., Liang, J., Qin, X., Hu, M., Peng, T., & Hu, X. (2014). 5g wifi signal-based indoor localization system using cluster k-nearest neighbor algorithm. *International Journal of Distributed Sensor Networks*, .

Zhou, M., Tian, Z., Xu, K., Yu, X., Hong, X., & Wu, H. (2014a). SCaNME: Location tracking system in large-scale campus wi-fi environment using unlabeled mobility map. *Expert Systems with Applications*, *41*, 3429 – 3443.

Zhou, M., Zhang, Q., Tian, Z., Qiu, F., & Wu, Q. (2014b). Integrated location fingerprinting and physical neighborhood for wlan probabilistic localization. In *Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on* (pp. 1–5).

Zhuang, J., Zhang, J., Zhou, D., Pang, H., & Huang, W. (2014). An improved wi-fi indoor positioning method via signal strength order invariance. In *Computer and Information Technology (CIT), 2014 IEEE International Conference on* (pp. 3–6).

List of Tables

Table 1: Results of experiment 1: *Error in positioning* and *Success Rate* of 53 distance/similarity measurements and 4 alternative data representations.

	positive		0-1 norm		exponential		powed	
Measure	success	error	success	error	success	error	success	error
City Block L_1	88.03%	7.60	88.03%	7.60	**90.73%**	**7.06**	90.46%	7.13
Euclidean L_2	89.92%	7.90	89.92%	7.90	**92.35%**	**6.90**	**92.71%**	**7.40**
Minkowski L_3	90.37%	8.58	90.37%	8.58	**92.98%**	**7.23**	92.71%	7.53
Minkowski L_4	89.74%	8.91	89.74%	8.91	92.80%	7.54	**92.89%**	**7.69**
Minkowski L_5	89.47%	8.90	89.47%	8.90	**92.98%**	**7.56**	92.98%	7.90
Chebyshev L_∞	86.05%	9.53	86.05%	9.53	**91.81%**	**7.81**	91.81%	8.09
Sørensen	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Gower	88.03%	7.60	88.03%	7.60	**90.73%**	**7.06**	90.46%	7.13
Soergel	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Kulczynski d	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Canberra	82.81%	9.48	82.81%	9.48	**87.58%**	**7.51**	84.16%	8.55
Lorentzian	81.10%	8.11	87.76%	7.56	**90.73%**	**7.07**	90.55%	7.17
Intersection	88.03%	7.60	88.03%	7.60	**90.73%**	**7.06**	90.46%	7.13
Wave Hedges	83.98%	8.90	83.98%	8.90	**86.14%**	**7.86**	81.91%	8.24
Czekanowski	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Motyka	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Kulczynski s	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Ruzicka	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Tanimoto	92.17%	7.33	92.17%	7.33	90.73%	7.05	**94.78%**	**6.86**
Squared Euclidean	89.92%	7.90	89.92%	7.90	92.35%	6.90	**92.71%**	**7.40**
Pearson χ^2	48.42%	14.47	48.42%	14.47	**88.30%**	**8.08**	48.51%	14.74
Neyman χ^2	77.05%	11.33	77.05%	11.33	**93.79%**	**6.99**	81.91%	10.77
Squared χ^2	86.05%	8.73	86.05%	8.73	91.09%	7.28	**91.45%**	**6.92**
Probabilistic Symmetric χ^2	86.05%	8.73	86.05%	8.73	91.09%	7.28	**91.45%**	**6.92**
Divergence	80.02%	10.23	80.02%	10.23	**89.74%**	**7.81**	84.97%	8.63
Clark	55.63%	16.37	55.63%	16.37	**80.74%**	**9.56**	39.69%	24.63
Additive Symmetric χ^2	78.85%	11.39	78.85%	11.20	2.70%	60.54	**79.48%**	**11.55**
Inner Product	**79.48%**	**11.23**	**79.48%**	**11.23**	24.30%	20.03	77.05%	12.93
Harmonic mean	81.73%	10.69	81.73%	10.69	55.99%	12.12	**90.19%**	**9.81**
Cosine	92.53%	7.85	92.53%	7.85	92.98%	6.88	**94.69%**	**7.59**
Kumar-Hassebrook	92.53%	7.57	92.53%	7.57	92.80%	7.01	**94.33%**	**7.00**
Jaccard distance	92.53%	7.57	92.53%	7.57	92.80%	7.01	**94.33%**	**7.00**
Dice	92.53%	7.57	92.53%	7.57	92.80%	7.01	**94.33%**	**7.00**
Fidelity	76.06%	11.70	76.06%	11.70	2.88%	60.22	**79.66%**	**11.45**
Bhattacharyya	76.06%	11.70	76.06%	11.70	2.88%	60.22	**79.66%**	**11.45**
Hellinger	84.16%	9.39	84.16%	9.39	**91.09%**	**7.33**	**91.00%**	**7.27**
Matusita	84.16%	9.39	84.16%	9.39	**91.09%**	**7.33**	**91.00%**	**7.27**
Squared-Chord	84.16%	9.39	84.16%	9.39	**91.09%**	**7.33**	**91.00%**	**7.27**
Kullback Leiber	44.82%	18.80	44.82%	18.80	25.74%	24.03	**61.66%**	**15.28**
Jeffreys	80.74%	10.56	80.74%	10.56	**91.09%**	**7.33**	90.10%	8.23
K divergence	60.94%	17.71	60.94%	17.71	15.84%	30.07	**79.39%**	**13.95**
Topsøe	84.88%	9.02	84.88%	9.02	91.09%	7.30	**91.45%**	**7.14**
Jensen-Shannon	88.48%	8.76	88.48%	8.76	91.90%	8.89	**94.69%**	**7.44**
Jensen difference	84.88%	9.02	84.88%	9.02	91.09%	7.30	**91.45%**	**7.14**
Taneja	79.30%	10.88	79.30%	10.88	**91.36%**	**7.34**	86.68%	9.34
Kumar-Johnson	78.76%	11.95	78.76%	11.95	**91.27%**	**7.41**	80.38%	11.38
Avg(L_1, L_∞)	88.39%	7.82	88.39%	7.82	**91.18%**	**7.09**	91.09%	7.11
Vicis-Wave Hedges	79.21%	11.00	79.21%	11.00	**89.11%**	**7.76**	79.21%	11.70
Vicis-Symmetric χ^2 *1*	78.94%	11.41	78.94%	11.41	**90.10%**	**8.54**	79.93%	11.64
Vicis-Symmetric χ^2 *2*	78.94%	11.38	78.94%	11.38	**91.09%**	**7.41**	78.94%	11.70
Vicis-Symmetric χ^2 *3*	87.40%	8.28	87.40%	8.28	91.18%	7.26	**91.27%**	**6.95**
min-Symmetric χ^2	80.11%	10.87	80.11%	10.87	**91.90%**	**7.63**	79.93%	11.09
max-Symmetric χ^2	60.40%	11.97	60.40%	11.97	**91.09%**	**7.40**	62.65%	11.99

List of Figures

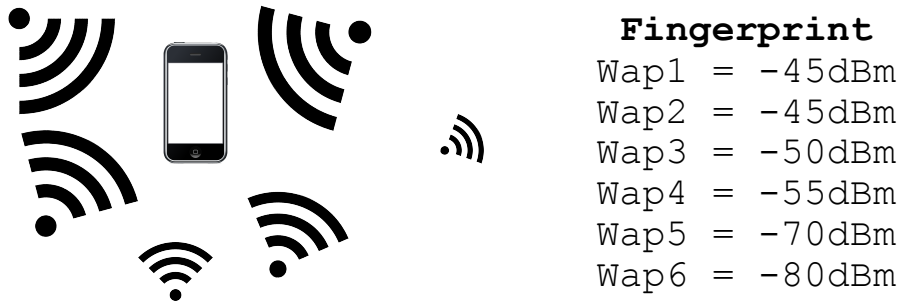

Figure 1: Basic example of fingerprint where the mobile phone has detected 6 different WAPs with different intensities. Note that WAPs have not to be directly oriented to the device in order to be detected.

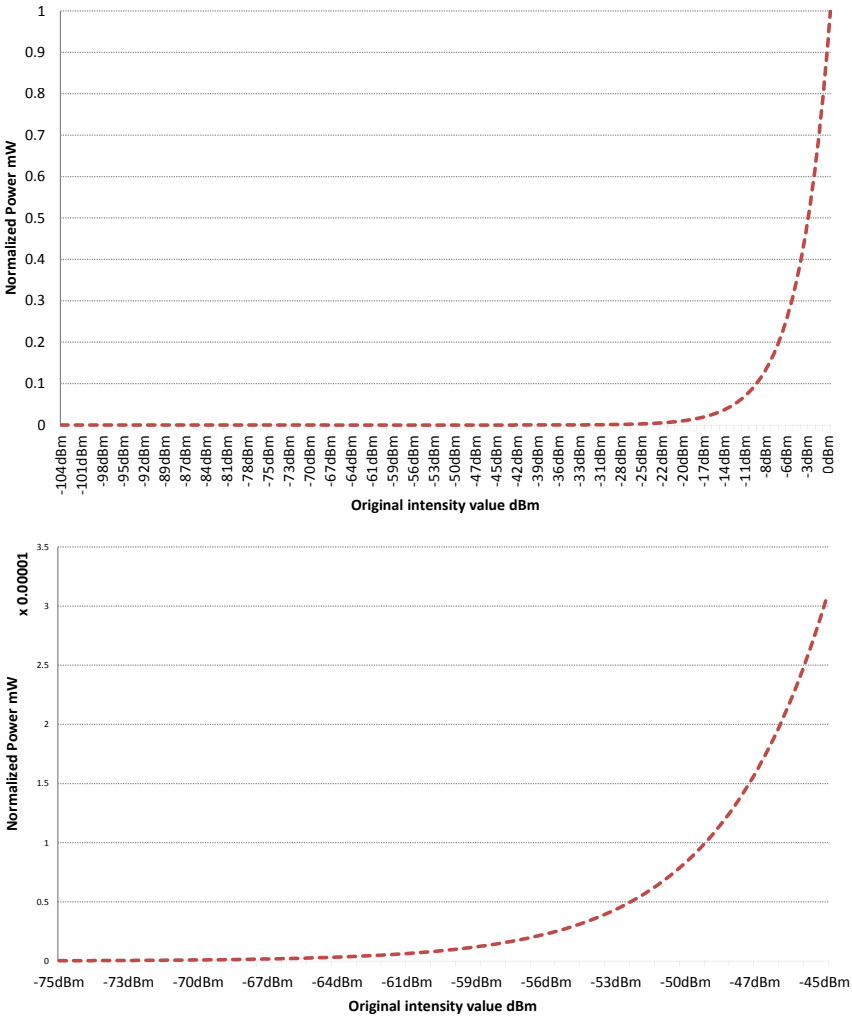

Figure 2: Relation between RSS values (dBm) and Signal Power (mW)

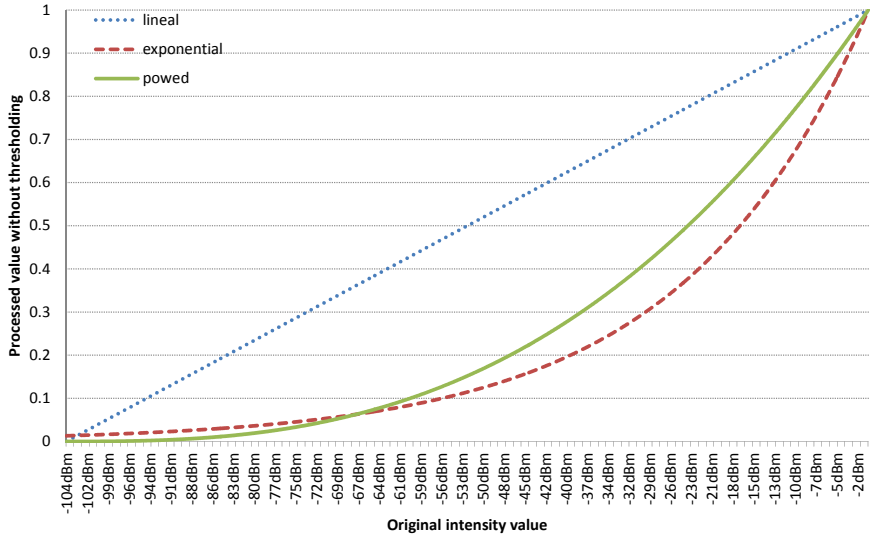

Figure 3: Visual meaning of the different representations used. *Lineal* stands for *positive* and *normalized* since both are proportional. The image shows the representation without applying thresholding.

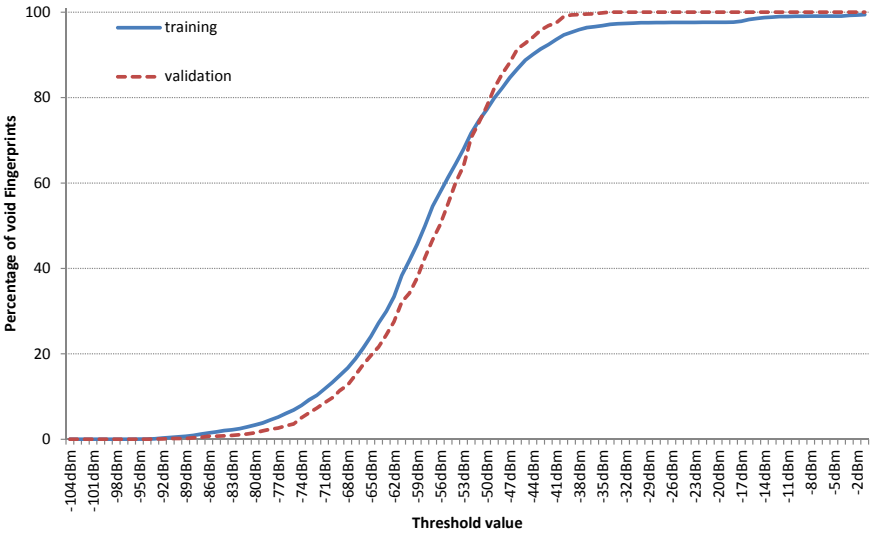

Figure 4: Percentage of void fingerprints after applying thresholding

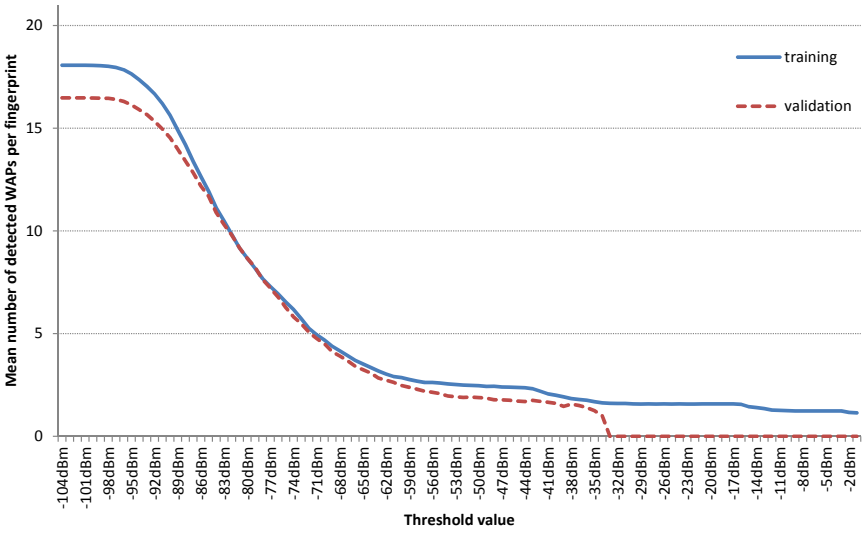

Figure 5: Mean number of detected WAPs after thresholding

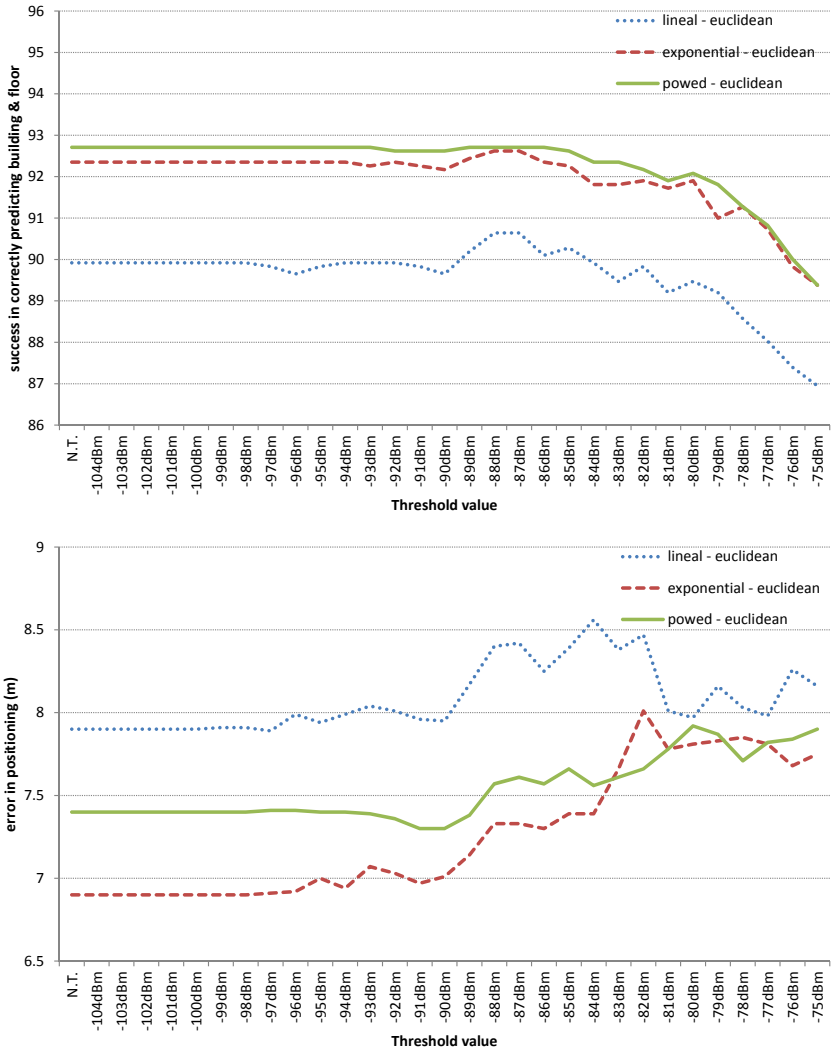

Figure 6: Evolution of success (top) and error (bottom) with respect to the threshold values for the *Euclidean distance* and the different representations. *Lineal* stands for *positive* and *zero-to-one normalized* representations since both are equivalent on the corresponding measure used. *N.T.* stands for the case in which threshold was not applied.

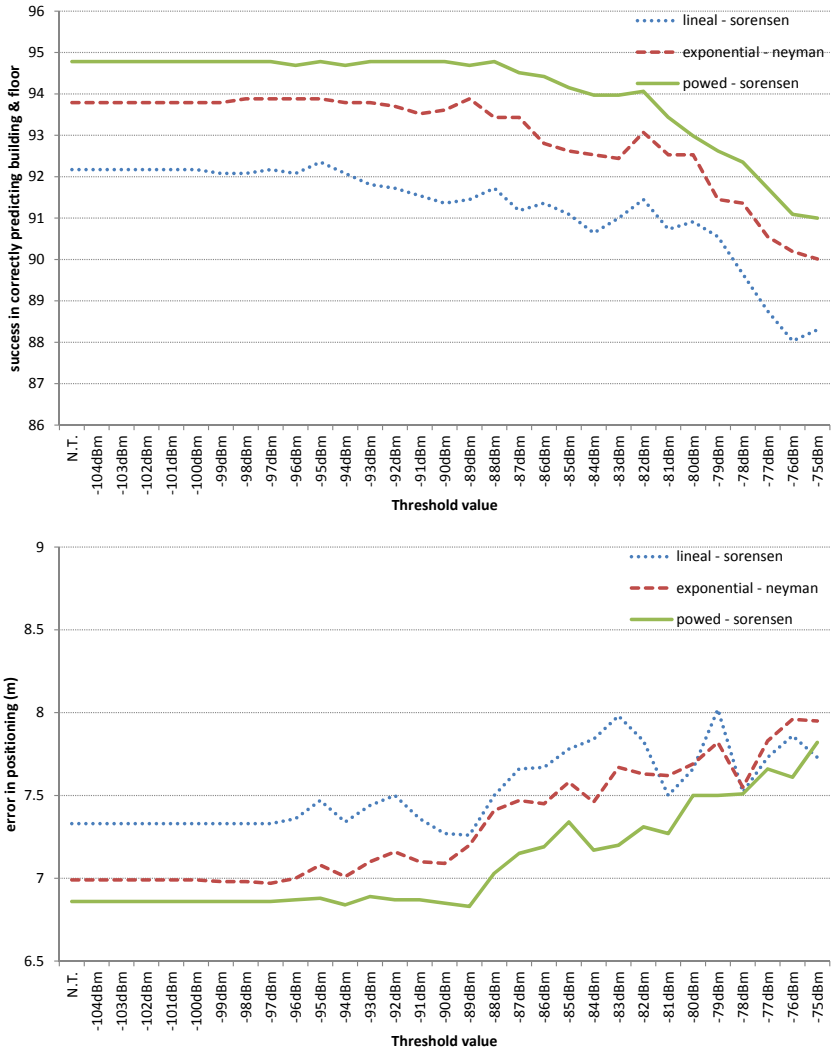

Figure 7: Evolution of success (top) and error (bottom) with respect to threshold values in the best measure for the different representations. *Lineal* stands for *positive* and *zero-to-one normalized* representations since both are equivalent on the corresponding measure used. *N.T.* stands for the case in which threshold was not applied.

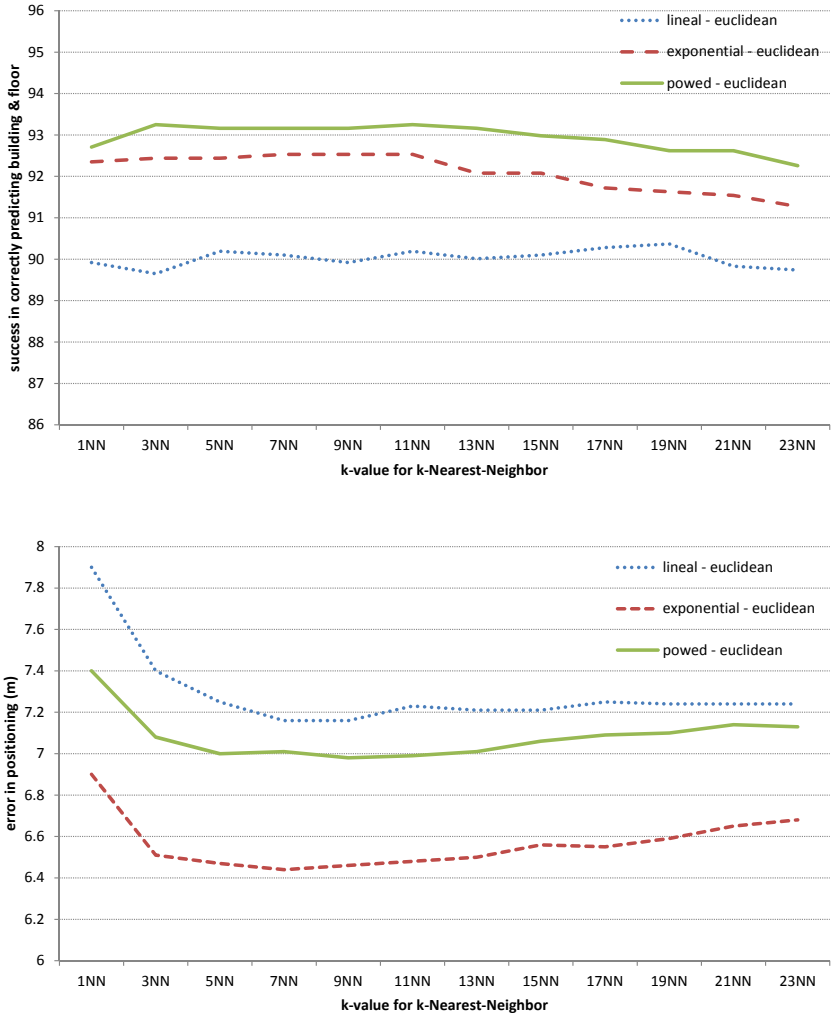

Figure 8: Evolution of success (top) and error (bottom) with respect to the k value for the *Euclidean distance* and the different representations. *Lineal* stands for *positive* and *zero-to-one normalized* representations since both are equivalent on the corresponding measure used.

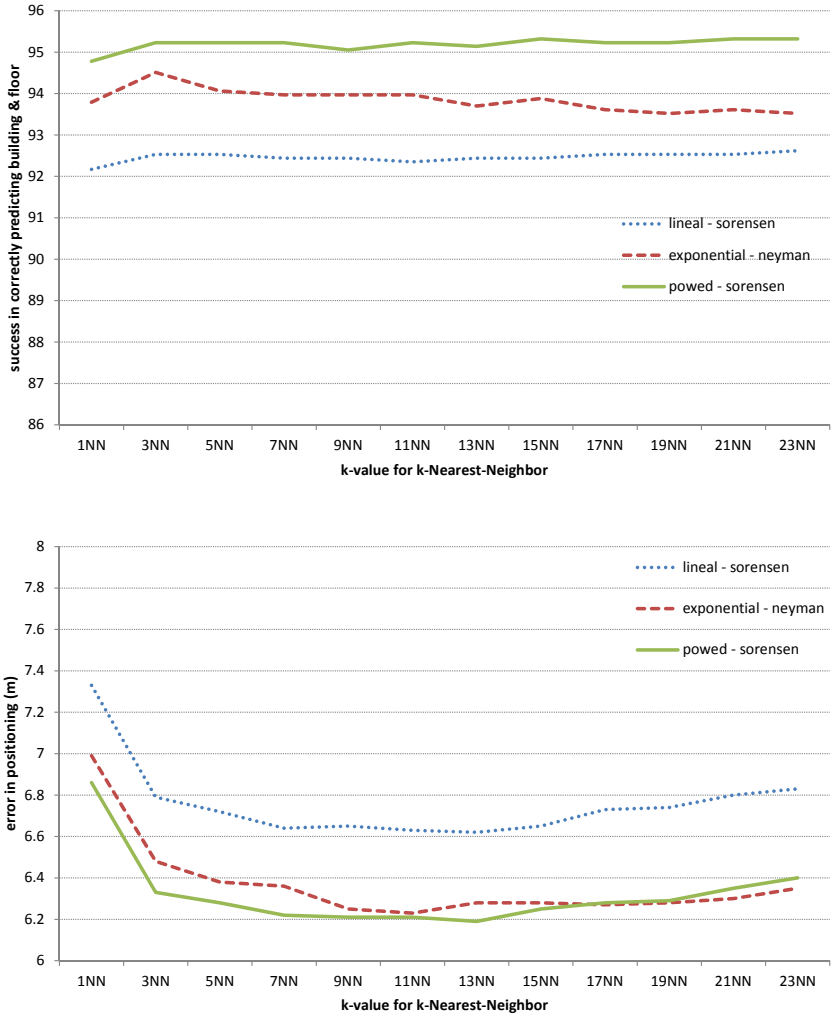

Figure 9: Evolution of success (top) and error (bottom) with respect to k-values in the best measure for the different representations. *Lineal* stands for *positive* and *normalized* representations since both are equivalent on the corresponding measure used.