INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 6, Month: December, Year: 2021
Article Number: 4394, https://doi.org/10.15837/ijccc.2021.6.4394

CCC Publications
AGORA UNIVERSITY PRESS

Effect of Sample Sizes in Fingerprinting Database for Wi-Fi System

A.H.A. Sa'ahiry, A.H. Ismail, L.M. Kamaruddin, M.S.M. Hashim
M.S.M. Azmi, M.J.A. Safar, M. Toyoura

Ahmad Hakimi Ahmad Sa'ahiry*, Abdul Halim Ismail, Muhammad Juhairi Aziz Safar
Faculty of Electrical Engineering Technology
Universiti Malaysia Perlis, Malaysia
02600 Arau Perlis, Malaysia
*Corresponding author: a.hakimi@studentmail.unimap.edu.my
ihalim@unimap.edu.my, juhairi@unimap.edu.my

Latifah Munirah Kamaruddin
Faculty of Electronic Engineering Technology
Universiti Malaysia Perlis, Malaysia
02600 Arau Perlis, Malaysia
latifahmunirah@unimap.edu.my

Mohd Sani Mohamad Hashim, Muhamad Safwan Muhamad Azmi
Faculty of Mechanical Engineering Technology
Universiti Malaysia Perlis, Malaysia
02600 Arau Perlis, Malaysia
sanihashim@unimap.edu.my, safwanazmi@unimap.edu.my

Masahiro Toyoura
Department of Computer Science and Engineering
University of Yamanashi, Japan
4-3-11 Takeda, Kofu Yamanashi, 400-8511, Japan
mtoyoura@yamanashi.ac.jp

Abstract

Indoor positioning system has been an essential work to substitute the Global Positioning System (GPS). GPS utilizing Global Navigation Satellite Systems (GNSS) cannot provide an accurate positioning in the indoor due to the multipath effect and shadow fading. Fingerprinting method with Wi-Fi technology is a promising system to solve this issue. However, there are several problems with the fingerprinting method. The fingerprinting database collected has different sample sizes where the previous researcher does not indicate any standard for the sample size to be used. In this paper, the effect of the sample sizes in fingerprinting database for Wi-Fi technology has been discussed deeply. The statistical analyzation for different sample sizes has been analyzed. Furthermore, two methods which are K- Nearest Neighbor (KNN) and Deep Neural Network (DNN) are being used to examine the effect of the sample sizes in term of accuracy and distance error. The discussion in this paper will contribute to the better sample size selection depending on the method taken by the user. The result shows that sample sizes are an important metrics in developing the indoor positioning system as it effects the result of the location estimation.

Keywords: indoor positioning system, sample size, positioning accuracy, big data, fingerprinting, deep learning.

1 Introduction

The Global Positioning System (GPS) and Global Navigation Satellite Systems (GNSS) in general have been adopted as the primary positioning technology due to the highly accurate location information they provide on a global scale; However, this technology fails in certain environments, such as indoors or urban canyons. These GNSS failures are primarily due to the satellites' low received signal power due to degradation of signal as illustrated in Figure 1 and visibility in urban/indoor areas. As a result, non-GNSS navigation technologies are critical in these regions [11].

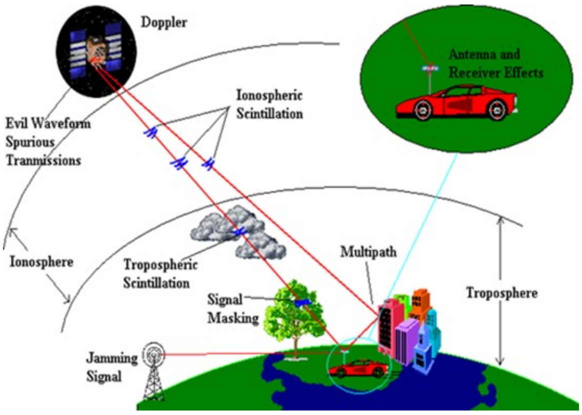

Doppler
Antenna and Receiver Effects
Ionospheric Scintillation
Evil Waveform Spurious Tranmissions
Ionosphere
Tropospheric Scintillation
Multipath
Troposphere
Signal Masking
Jamming Signal

Figure 1: Degradation signal of GPS and GNSS (Ma [10])

Numerous studies have been conducted over the last few years to address this issue. By utilizing a variety of technologies, including infrared [2], ultra- wide band [22], bluetooth [3], inertial [20], magnetic [8], and fusion of the technologies [20]. Wi-Fi is one of the most reliable technologies available. The reasons for this are that it is widely deployed within buildings because the world relies on Wi-Fi to connect to the internet for home networking, supporting the internet of things, and teaching. By utilizing Wi-Fi, no pre-deployment effort or infrastructure support is required. As a result, labor costs associated with installing new hardware to implement the system are reduced. The use of Wi-Fi technologies for indoor positioning has been discussed, as well as several methods and techniques.

Several methods have been used by the previous researcher such as TOA [19], AOA [19], and fingerprinting [5] method to predict user location. However, the most precise and provide a better accuracy is the fingerprinting method [9]. Fingerprinting methods works by taking the unique identification for each of the reference point for the database collection. Fingerprinting has two phases which are the offline and online phases as illustrated in Figure 2. Offline phases are a calibration or collecting the database while for online is predicting the location of the user. In the offline phases which is the data collection, previous researcher does not have a standard for choosing a sample data [7]. Additionally, the issue with the fingerprint database is the database must be taken by expert surveyor due to the collection of the database need a professional trained person. Otherwise, the database will not effectively be created, and this will produce a problem in future location estimation. Moreover, the expert surveyor will create a massive problem where the cost to hire the expert surveyor is excessively expensive. Other problem related to the expert surveyor is the time taken for the expert surveyor to collect the database is time consuming. This two problem of the professional surveyor responsible for compiling the database is expensive and time consuming [1] which need to be avoided.

Hence, crowdsourcing fingerprinting database was introduced. Crowdsourcing method is replacing the labor to stranger, where the stranger or anyone could contribute their signal into the fingerprinting database. However, the problem with the crowdsourcing is the sample size in the data base get from the stranger in each of the fingerprinting database reference point is different [18]. The sample size is the number of signal strength from the source collected. For example, the Wi-Fi signal strength sample size can be collected by taking the signal strength over time. In the crowdsourced database, the stranger does not know the system of the fingerprinting method. Thus, provide a different sample size for each of the stranger and will creates inaccuracy in predicting the user location. In this paper, the effect of the different sample sizes will be discussed to get a better understanding and the authors try to find the optimal number of sample sizes in term of fingerprint database for Wi-Fi signal.

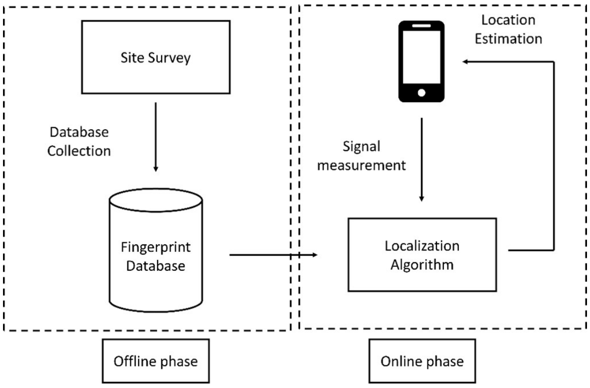


Figure 2: Fingerprinting workflows

2 Related Work

In the previous research work, it appears that researchers frequently use a variety of sample sizes to evaluate their research work, without explaining the rationale for the sample size selection. In [12], 40 observations were made for a set of 155 reference points that served as training data to eliminate human behavior's randomness. In [21], the authors collected one sample per second for five minutes (a total of 300 samples) in order to investigate wireless channel changes over time. Although it is acceptable, it does not indicate any reference for sample size and why 300 sample sizes were collected. Similarly with author In [17], a different sample sizes were used to analyze and compare various filtering strategies for real-world indoor 802.11 positioning systems. The authors determined the radio distribution at 250 uniformly spaced grid points over a 15 x 35 meters area. The authors of [4] proposed a technique called dynamic hybrid projection (DHP) for enhanced 802.11 localization. They collected 802.11 RSS data at 27 different reference locations on different days and with four different user orientations during their experiments. They selected 15 locations with a step of 1.5–2 meres from this sample to use as training data. The sample sizes were not declared, and the sample sizes were totally different for various researcher.

In the crowdsourced data, [15] has developed a human-computer interface for indicating location over intervals of varying duration, a client-server protocol for pre-fetching signature data for use in localization and location-estimation algorithm incorporating highly variable signature data. They describe an experimental deployment of their method in a nine-story building with more than 1,400 distinct spaces served by more than 200 wireless access points. The sample size for different location is diverse, this will be a problem for the online phases to locate the user location.

In summary, our review of the literature indicates that authors calibrate, train, test, and evaluate indoor positioning systems using a variety of sample sizes and patterns. The previous author does not specifically justify the reason of the number sample sizes collected. The authors in this paper hypothesis are the sample sizes is an important criterion to gain an accurate result. Hence, before making a prediction, the authors of this paper intend to investigate the effect of sample size on the accuracy and statistical properties of fingerprint database data.

Referring to [9], the author provides the following performance benchmarking for indoor wireless location system: accuracy, precision, complexity, scalability, robustness, and cost. However, in this paper the author will be using two performance metrics which are the accuracy and precision. Both metrics will be tested by using two methods. K-Nearest Neighbor (KNN) and Deep Neural Network (DNN). There are advance KNN that has been used, [6] has used WKNN which is Weighted K-Nearest Neighbor. In this work, KNN is chosen as a conventional method and to standardize the evaluation that been used by most of previous research. DNN is the most advance method using a neural network based on machine learning. There are many architecture in designing the deep learning model, one of it are [14], the author used a complex deep learning architecture by combining multiple machine learning algorithm. In this study, a basic architecture will be executed as in [13], the scope and aim are not to get the best prediction of the location. The DNN algorithm is enough to study the effect of the different sample size in term of the accuracy of the location prediction. Hence, using these two methods will verify the performance of the sample sizes in term of two of the verification metrics.

In this paper, the first section will be addressing the general problem which is the GPS and GNSS problem. The second section will be the related work where the previous researcher has done. The third section will be explaining on the data collection part where the authors will clarify how the data is collected, what hardware are being used and the configuration of the setup. For the fourth section, the data distribution based on its intensity will be analyzed to get a clear view of the data in the fingerprinting database. Then, the authors study the characteristic of the signal and the statistical analyzation in term of mean, mode, and standard deviation. The box plot also is plotted to know the stabilization of the sample sizes based on the median. For the performance metrics, the sample size in term of its accuracy and precision by using KNN and DNN method are evaluates. The fifth section is the conclusion and discussion from what have been discovered from the experiment and discussion based on the previous section.

3 Methodology

This section will explain the experimental setup of the data collection in collecting the fingerprinting database. The experimental area, hardware configuration, hardware used and all the setups will be explained in this section. The setup has been made by taking a consideration on the previous researcher work to get the optimize setyp and avoid any configuration error.

3.1 Overview

The flow to analyze of the effect of the sample sizes in fingerprinting database is presented in Figure 3. First the data is collected by using single access point. The data is then presented by two angle which are in two dimensional and three dimensional. Then, the data were analyzed with respect to the sample sizes critically by the signal characteristic and statistical approach. Afterward, the different sample sizes data were evaluated by using two methods which are the non-parametric techniques, K-Nearest Neighbor (KNN) and by using artificial intelligence method, Deep Neural Network (DNN).

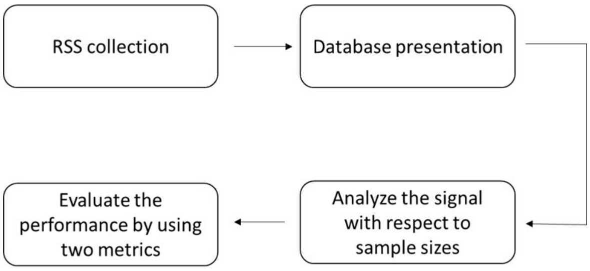

RSS collection
Database presentation
Analyze the signal with respect to sample sizes
Evaluate the performance by using two metrics

Figure 3: Works Overview

3.2 Data Collection

The data collection was made in Solid Mechanic and Acoustic lab in Universiti Malaysia Perlis as in Figure 4. It consists of 42 reference point, 6 reference point on X axis and 7 reference point on Y axis. The data taken is semi-controlled environment where all the configurations will be stay the same the only changes are the location of the mobile devices taken for each of the reference point.

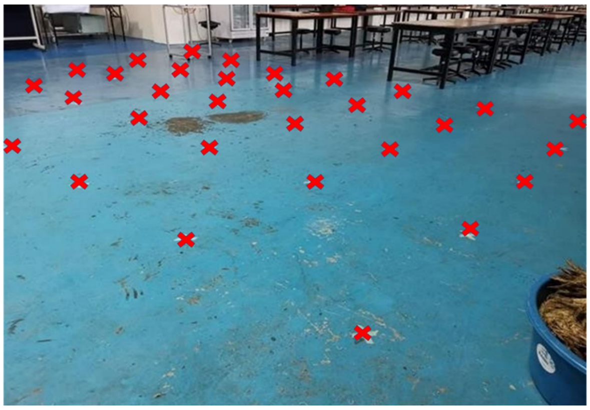


Figure 4: Experimental Test Bed

To get the situation as exact as in real environment, people who are walking in the fingerprinting database area was considered. Hence, this will give an exact real environment to get a precise result in discussing the effect of sample size. The data collection procedure is the modelled similarly to paper the approach delineated in [16].

The equipment used are TP-link (TD-W8961N) with nominal frequency 2.4 GHz and android phone are used which is Mi A2 Lite mobile phone. TP-link is used as the access point and Mi A2 Lite is used as the devices for collecting the fingerprinting database. Mobile phone is chosen because in crowdsourced most of the user who will contribute into the fingerprinting database will be using a mobile phone. Thus, this experiment database will be collected by using a mobile phone. The time taken for each of the reference point taken are 20 minutes to get 1000 samples for each of the reference point. Then, in each of the reference point, the sample were divided into 8 different sample sizes which are 10, 20 ,30 ,40, 50, 100, 500 and 1000 sample sizes. 10 until 50 is consider as the small sample size while the medium is 100 and 500. 1000 is the large sample size. This number will be discussed in the further analyzation to know the perfect number in choosing the number of sample size. This is to test the effect of the sample size in fingerprinting database for Wi-Fi system.

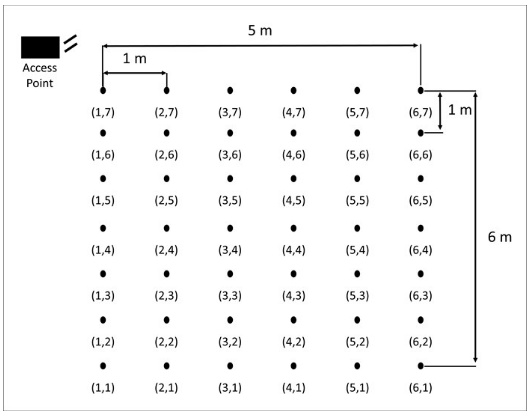


Figure 5: Experimental Area

The reference point of the grid system of the fingerprint database is illustrated in Figure 5. To make sure only the sample size is the main variable, all the other variables are constant, while only the sample size of the Received Signal Strengh (RSS) of Wi-Fi is different. The fingerprinting database has 42 reference point, where the gap location between each reference point is 1 meter. The fingerprinting database for the vertical axis is 6 meter while in the horizontal axis is 5 meters. One access point has been used in this data collection process which is labelled as AP1.

In this work, we consider an access point by using one router with 2 fixed antennae at a nominal 2.4 GHz frequency which is TP-link (TD-W8961N). The axis of the reference point is based on the

X and Y axis as indicated in Figure 5. The database is represented as R_{xy}, where R is the received signal strength (RSS) of Wi-Fi for 42 reference point. x is the X axis while y is the Y axis for each of the reference point in the experimental test bed. This reference point is labelled so it is easier for the database organization for the used in estimating the user location.

3.3 K-Nearest Neighbor (KNN)

The sample size are tested using two methods in predicting the user location. Two method which were implemented are KNN and DNN. KNN is a conventional method that has been used by most of the researcher to locate the user data by taking the nearest neighbors. The advanced method is the DNN method. It based on artificial intelligent where it trains the data before it started to predict using the trained model.

KNN is a supervised machine learning algorithm. KNN does not need a specialized training phase as it calculates through distance from its neighbors. Hence, it is also does not need a gaussian data to get an accurate value. The step for predicting the user location will be first to label the training data. In this test, the label data is the access point 1. The variation of sample will be the manipulated variable for the analyzing purposes. Next, is by choosing the K of the algorithm which is how many neighbors for the algorithm to classify into how many group. In this case, the K is chosen in 3 variation which are 1, 3, and 5 because the K is important in KNN algorithm. Hence, to analyze a better effect of the sample sizes in the fingerprinting database, the K will be varying. By using Euclidean distance in equation below, It will calculate and find the nearest neighbor which the user is located. If it finds the nearest neighbors are in their group, then it will classify them as their group.

$$d = \sqrt{(x_1 - x_2)^2_+(y_1 - y_2)^2},$$

Where d is distance, x_1 and y_1 is X and Y axis of one of the reference points in the database. x_2 and y_2 is second X and Y axis in another reference point.

3.4 Deep Neural Network (DNN)

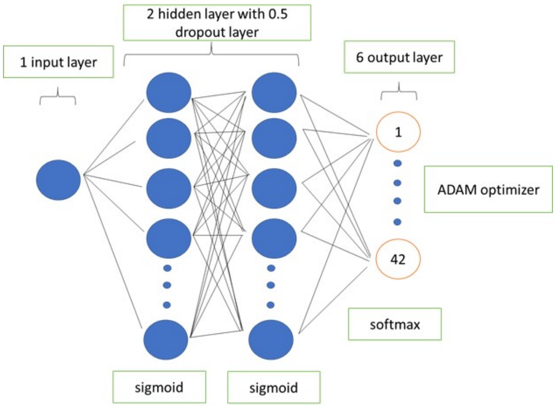


Figure 6: Deep Neural Network (DNN) Model

In the DNN method, the classifier used are the same as the KNN which is supervised machine learning method. DNN will try to predict the location by adjusting the model of the training data. There are two phases for DNN in estimating the location. The first one is to train the data and after getting the solid model, the model is then used to predict the location.

The first step in using a DNN classifier is to label the data. By using the Access Point 1 (AP1) as the feature, the data is labelled. Then the data need to be trained. In training, the hyperparameter of the model will be chosen heuristically. In this experiment the variable that has been considered is the sample sizes, hence the hyperparameter will be fix through all the sample sizes. Two hidden layers has been used in this experiment with 250 nodes in the first layer and 20 nodes in the second layer. 0.2 dropout were used to avoid overfitting. The activation function for both hidden layers are by using sigmoid as it can handle negative value. For the output activation function, SoftMax will be used in the model because SoftMax is an output's classifier as it takes the highest value in the output nodes to predict the user location. The model is illustrated in Figure 6.

Finally, the optimizer for the model was chose by using the Adam optimizer. It is back propagation where the algorithm learnt from the output and try to optimize the model by adjusting the error in output. The validation of the model used are by using accuracy metric as in subsection 3.5.

Deep learning learns from previous data which the user feeds into it. The main objective of classification are by using DNN to take the highest probability of the output which has been optimized by the model. The process to choose the highest probability undergoes a certain set of formula. First, by entering the input node layer by multiplying with x_i and the weigh w_i to get the result of the next node in the hidden layer.

$$u = \sum_{i=1}^{n} x_i w_i$$

Where x_i is

$$x_i = R_{xy}$$

Where R_{xy} is the received signal strength (RSS) of Wi-Fi in one of the access point, the unit is in dBm.

On the hidden layer, u is then inserted into non-linear activation function which is the sigmoid to get the output value between 0 to 1,

$$S(u) = \frac{1}{1 + e^{-u}}$$

Where u is the input and S is the output.

After iterating process complete, the result of the second hidden layer will be inserted into the SoftMax equation. This will give the highest probability in continuous digit,

$$F(\overrightarrow{\mathrm{S}})_i = \frac{e^{S_i}}{\sum_{j=1}^{k} e^{S_j}}$$

Where F is the output, $\overrightarrow{\mathrm{S}}$ is the input vector gain from previous outcome on sigmoid activation equation, k is the number of classes in the output as in this example is 42 classes, S_i is the standard exponential function for input vector and S_j is the standard exponential function for output vector

3.5 Validation Accuracy and Distance Error

The accuracy is determined by evaluating KNN and DNN approaches. In the population of the data, 20% of the data is used to evaluate the performance of the different sample sizes. The data is evaluated by calculating the true reference point over the estimated position. The result will be converted in term of percentage , to make the value easily digestible by the reader, the ratio is simply multiplied by 100. Distance error is one of the evaluations where the distance is calculated by using Euclidean distance. Between two of the reference points, the distance error was gain and the result is plotted in cumulative distribution frequency (CDF) graph. The frequency of distance errors is explored through an examination of the average and maximum errors.

To get the accuracy, the ratio of the correct prediction and the total number of the prediction sample is applied,

$$A = \frac{CP}{TP} \times \mathit{100}$$

where A is accuracy, CP is the number of correct prediction and TP is the total number of the prediction sample. The accuracy gain in term of percentage as in table 2.

4 Result and Discussion

In this section, the fingerprinting database will be analayzed and shown in two form by using heatmap (2D) and distribution map (3D). RSS signal characteristic of different sample sizes will be shown and discussed in this section. Then, statistical analysis in term of mean, mode and standard deviation will be addressed to know the effect of different sample sizes. Box plot is plotted for the different sample sizes at 4 reference point to view the median stabilization where the median is important in fingerprinting method.

4.1 RSS Intensity Heatmap

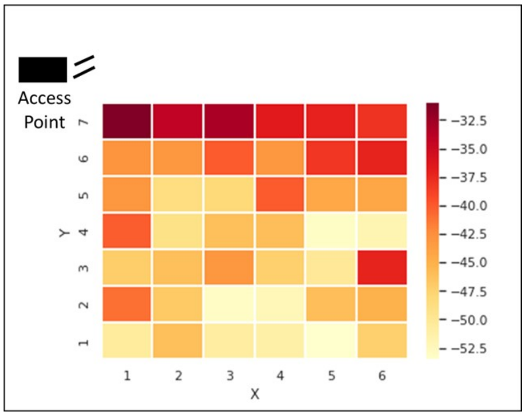


Figure 7: RSS Intensity Heatmap

In getting the clear view of intensity signal strength measuremen,t the heatmap is created as illustrated in Figure 7. In the heatmap, the darker area presenting that the signal strength is strong as in (1,7) coordinate which is the strongest position. As the device is further away from the access point or router the signal strength becomes weaker as in coordinate (5,6). The heatmap will give the clear presentation of the intensity of the RSS data to make a rough analyze for the fingerprinting database.

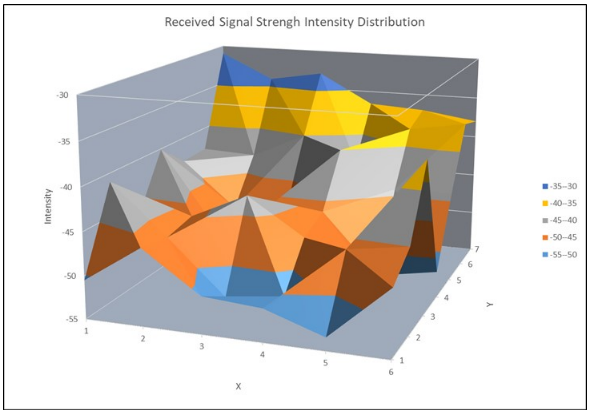


Figure 8: RSS Intensity Distribution

The RSS intensity distribution is illustrated in Figure 8 to get the clear presentation of the signal strength of fingerprinting database collected. The dark blue and yellow color indicated that the it is the strongest signal where the range is between -40 dBm to -30 dBm where it is the nearest point to access point. On average, the signal is between – 50 dBm to -45 dBm, where the signal is in the majority of the distribution. The weakest signal strength are between -55 dBm to -50 dBm. It is the furtheest from the access point in position (6,1). The signal strength is fluctuated over the location.

Some of the location acquire the the lowest signal strengh even in the middle position like in (6,3) coordinate. This is expected as the RSS is not linearly proportional over the distance.

4.2 RSS characteristic

Received signal strength (RSS) from the router is taken to get the properties of the signal. Variety of the sample sizes is taken to examine the effect of the different sample sizes. In Figure 9, the different sample sizes for RSS characteristic is shown. The RSS from Wi-Fi does not produce a smooth line, this is expected as Wi-Fi signal is not stable over time.

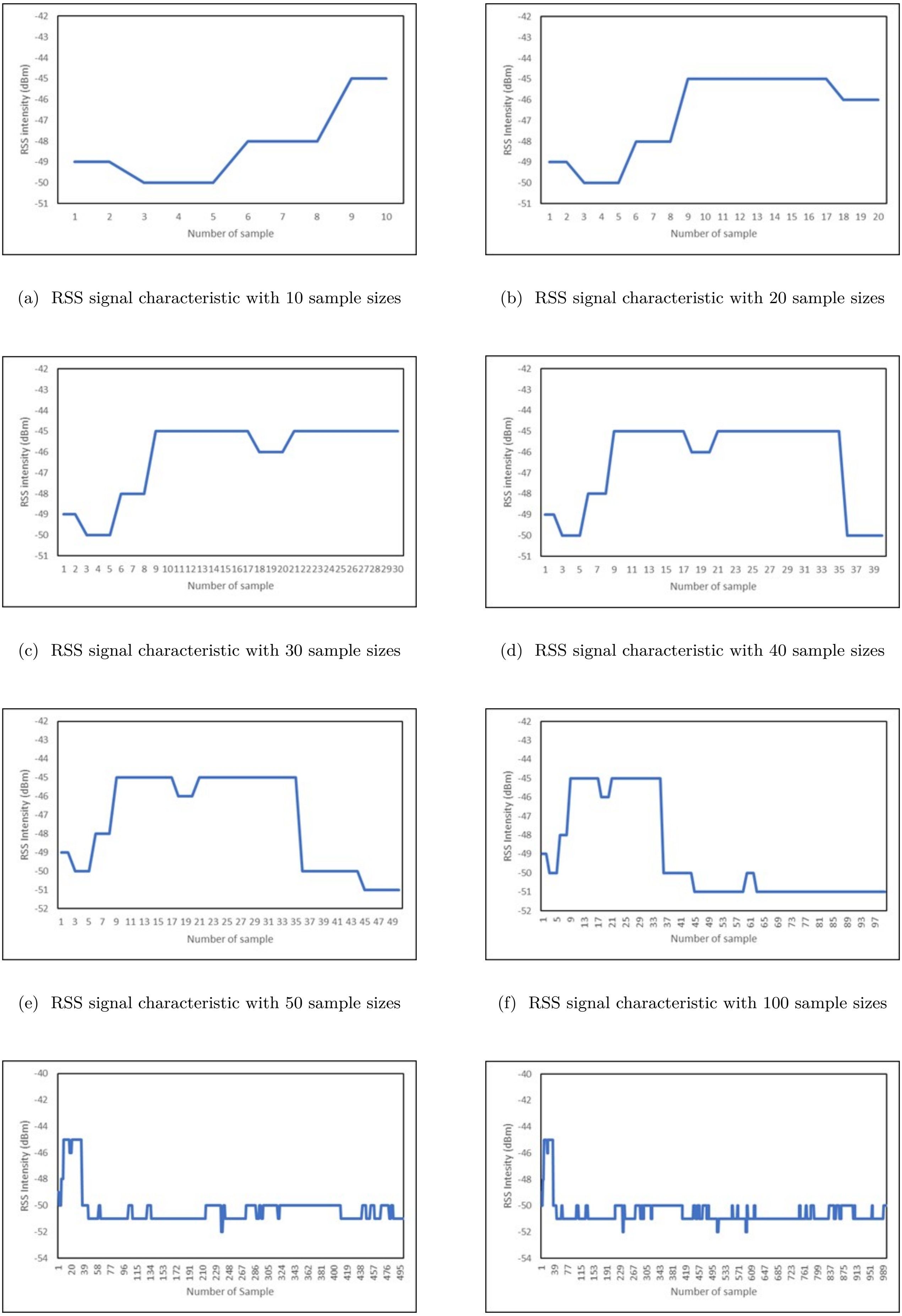
In Table 1, one of the reference point which is (0,1) coordinate has been taken to analyzed the statistical properties of the Wi-Fi signal strengh. The mean, mode, and standard deviation for 8 of the variation of samples are compared. The average or mean of the sample size started with -48.2 dBm and reduces to -46.75 dBm. This is due to the outlier at the starting of the sample collected. As the sample increases, the mean started to get stable result. At 500 sample sizes, the different between 1000 and 500 sample sizes mean are just only -0.28 dBm. This show that 500 sample is a promising result to take as the threshold for fingerpriningt database sample size. The mode or the maximum dBm is -45 and it is the same for all number of sample sizes.

Standard deviation shows that for the first 10, 20 and 30 number of sample size are below 2. At 40 sample sizes the standard deviation started to increase. This means that the dispersion is getting bigger. At 500 sample sizes the standard deviation starts to decrease significantly and follow with 1000 sample sizes. This show that the respectable dispersion of the number of sample sizes start at 500 and the best is at 1000 which get 1.04 in standard deviation.

Table 1: Statistic of different number of sample sizes

Number of samples	Mean (dBm)	Mode (dBm)	Standard Deviation (dBm)
10	-48.20	-45	1.87
20	-46.75	-45	2.00
30	-46.17	-45	1.82
40	-46.50	-45	2.10
50	-47.32	-45	2.51
100	-49.30	-45	2.54
500	-50.25	-45	1.34
1000	-50.53	-45	1.03

In Figure 9, RSS characteristic of one reference point in coordinate (0,1) with variation of sample 10, 20 ,30 ,40, 50, 100, 500, and 1000 is shown. At smaller sample size as in Figure 9a, 9b and 9c the range is between in -50 dBm and -45 dBm. The range started increase by 1 in the larger sample size as in Figure 9d, 9e and 9f. The largest sample sizes are 500 and 1000 as in Figure 9g and 9h giving a range between -52 dBm and -45 db. In Figure 9a, with 10 sample sizes the range could not be seen as it has only 10 sample sizes. On the bigger sample size as in Figure 9g and 9h the signal characteristic shows that it has outlier in the beginning. This is one of the information that can be used, as it shows early of signal may be not a proper way to be choose as it can be an outlier.



(a) RSS signal characteristic with 10 sample sizes

(b) RSS signal characteristic with 20 sample sizes

(c) RSS signal characteristic with 30 sample sizes

(d) RSS signal characteristic with 40 sample sizes

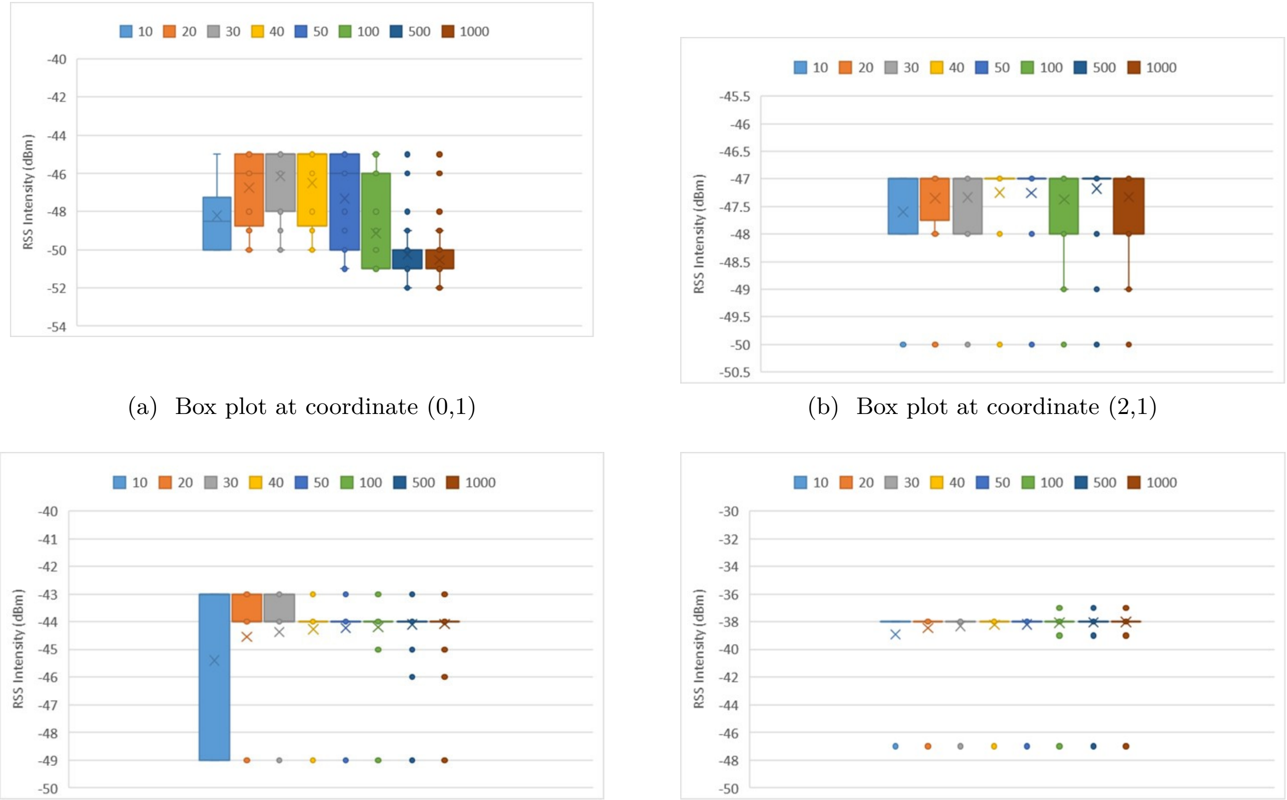
(e) RSS signal characteristic with 50 sample sizes

(f) RSS signal characteristic with 100 sample sizes

(g) RSS signal characteristic with 500 sample sizes

(h) RSS signal characteristic with 1000 sample sizes

Figure 9: RSS in one reference point (0,1) with varies of sample sizes.



(a) Box plot at coordinate (0,1)

(b) Box plot at coordinate (2,1)

(c) Box plot at coordinate (7,2)

(d) Box plot at coordinate (7,6)

Figure 10: Box plot of different number of sample sizes for 4 reference point.

The median distribution of the sample sizes is shown in Figure 10. To avoid outlier interference, the box plot has been plotted for 4 reference point. The reference point taken are in Figure 10a at coordinate (0,1), Figure 10b at coordinate (2,1), Figure 10c at coordinate (7,2) and Figure 10d at coordinate (7,6). In Figure 10a, the median of the data started stabilize when its reach 500 sample sizes. For reference point at coordinate (2,1) in Figure 10b, it is the same with Figure 10a, where it acquire stablity when the sample sizes are 500. In Figure 10c at coordinate (7,2), the median reaches its stability on 50 sample sizes. In Figure 10d at coordinate (7,6), even with 20 sample sizes the median has reach the stability. However, to get accuracy stability 50 sample sizes is considered a good threshold as it shows no difference at 100 and above sample sizes in term of its median.

4.3 Accuracy prediction using Deep Neural Network (DNN) and K-Nearest Neighbors (KNN)

The accuracy of DNN and KNN is shown in Table 2 in term of its percentage. For 10 sample sizes, the accuracy is 7% and by increase the sample size to 20 the accuracy starts to increase to 14%. The accuracy for DNN method keeps increasing as the sample size increase. The accuracy increases near to 4% over the sample size population. The highest accuracy is in 1000 sample sizes at 41% gain, nearly to 50%. This indicates by using DNN method, increase in sample sizes will increase the accuracy. Thus, provide a better elocation estimation.

At K = 1 the KNN accuracy shows that in small sample sizes the accuracy starts to increase from 21% to 35% at 10 sample sizes to 30 sample sizes. However, the accuracy fluctuates where it drops back to 27% and started increase back until 39%. The maximum accuracy is 39% at 500 sample sizes. At K=3, the accuracy is unstable as the accuracy at 17% in 10 sample sizes. It fluctuate in small area until it reaches the maximum accuracy which is 30% in 1000 sample sizes. For K = 5, the result shows almost similar as K= 3 where the accuracy fluctuate in a small area. The highest accuracy is in 500 sample sizes which gets 37%. This shows that KNN method does not rely on the sample size as the KNN method depend on number K, the user applies. For example, in this test the K number is

5 which it will take the nearest 5 neighbors to classify the location for the majority neighbors. Hence, KNN does not solely depends on the sample size unlike the DNN method as the sample size increase, the accuracy also will increase.

Table 2: Accuracy of two different algorithm

Sample Size	CNN(%)	KNN(%)		
		K = 1	K = 3	K = 5
10	7	21	17	17
20	14	28	24	13
30	22	35	22	14
40	24	27	22	15
50	26	28	23	23
100	30	34	23	21
500	39	39	28	37
1000	41	38	30	24

4.4 Distance Error

In this subsection, Deep Neural Network (DNN) distance error graph is discussed. The main objective of this subsection is to know the relationship between the sample sizes and distance error of the fingerprinting database. Likewise, the graph shows the maximum and the average distance error of the different sample sizes and different algorithm that has been used.

The accuracy metric equation does not same with the distance error. To compute the distance error between each of the sample sizes. The cumulative distribution frequency (CDF) was plotted. The distance error gain by comparing the predicted location with the true location. The distance error is calculated by using Euclidean distance as in subsection 3.3 . The graph of the CDF distance error for DNN and KNN is then plotted as in Figure 11 and Figure 12.

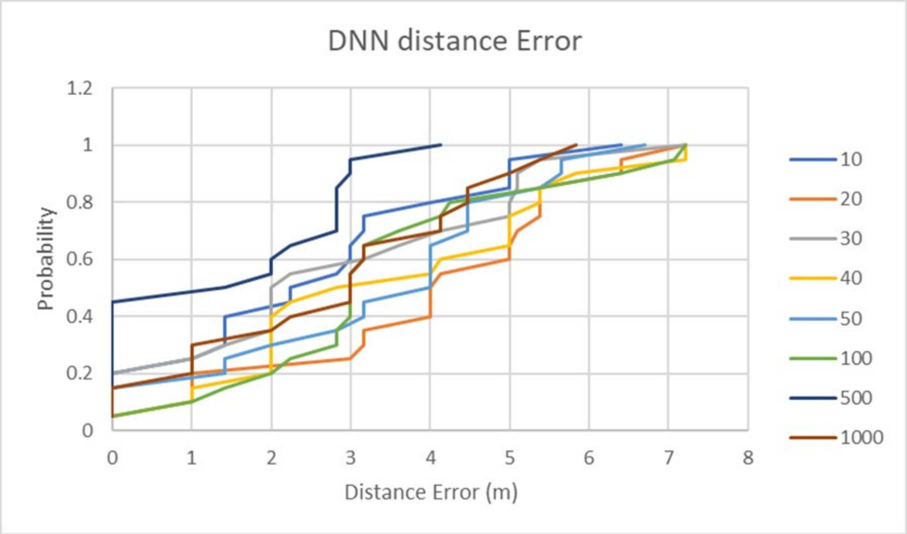


Figure 11: DNN Distance Error

4.5 Deep Neural Network (DNN) Distance Error

In Figure 11, 500 sample sizes show the most optimize result which is estimately 1.5 meter. This is followed by 30 sample sizes, then 10, 40, 1000, 50 and 20 sample sizes. This indicates that the distance error does not have any relationship with the sample sizes. In the accuracy test, the DNN method is affected by the sample size. As the sample size increase the accuracy also will be increase. However, for the distance error result, it shows that the different distance error does not have any relationship between the sample size.

Distance maximum error acquire the same result but have a small significant different in term of the result in its arrangement. However, it still does not show a linear relationship between sample size and distance error. In 500 sample sizes, the maximum distance error is 4.2 meter which is the most optimize distance error. While the others, ranging around 5 meters to 8 meters. The worst are 20, 30, 40 and 100 which is at 7.2-meter distance error. For 1000, sample sizes, the maximum distance error is 5.8 meter. Hence, this shows that the distance error for DNN method does not provide a linear relationship between the sample size and distance error.

4.6 K-Nearest Neighbor (KNN) Distance Error

This subsection used a K-Nearest Neighbor to estimate the location. By calculating the distance error, the maximum and average distance error was calculated. The aim is correlated to the DNN distance error. The only difference is the distance error equation.

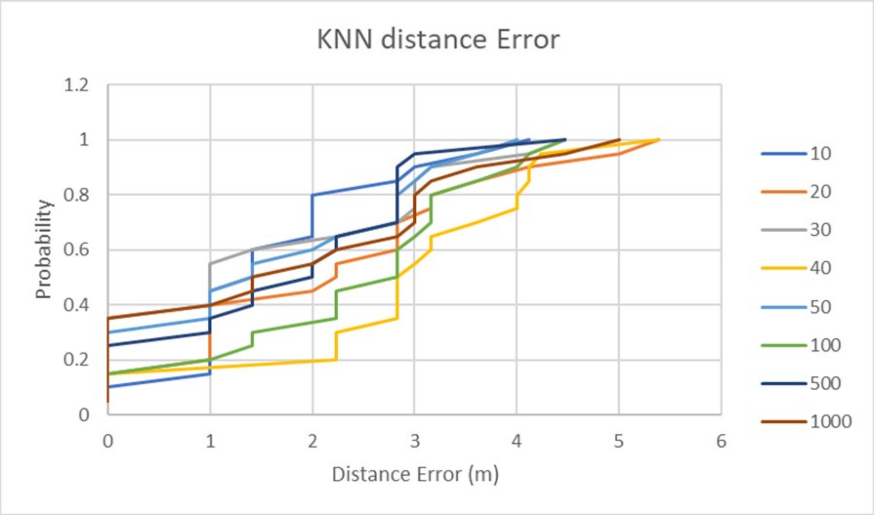


Figure 12: KNN Distance Error

In the KNN prediction method as shown in Figure 12, the distance error for different sample sizes show the same as its accuracy validation. The distance error is unpredictable in term of the sample sizes. On average, the most optimize sample sizes is 30 which give the distance error 1 meter. This followed by 10, 50, 1000, 500, 20, 100 and 40 sample sizes. This sequence shows the distance error is not linearly proportional to the distance error. On the maximum distance error. The result ranging from 4 meter to 6 meter. The highest distance error is in 20 and 40 sample sizes which giving 5.2-meter distance error. The most least distance error for the least maximum error is in 50 sample sizew which gain 4 meters. The distance error in KNN method is unpredictable as it does not give any correlation between the sample size and the distance error.

5 Conclusion

This paper discussed the effect of sample sizes in fingerprinting database for Wi-Fi system. Through intensive data analysis the different sample sizes for fingerprinting database is discussed. First, the data was collected in 5 x 6-meter area. Then, the intensity and distribution of the RSS is then presented. Analyzation of the data through a statistical measure is then discussed. Lastly, two methods which are the conventional method, KNN and the most advance method, DNN is applied to examine the effect of the sample size in term of accuracy and distance error.

From the characteristic of the signal from different sample sizes, the outlier could be the major problem where the outlier in the beginning of the sample could affect the accuracy atrociously as the standard deviation will be increase. If the database has only 10 sample sizes, the outlier could affect the accuracy of the prediction in future. In statistical analysis, the mean, mode, and standard deviation were discussed. The mean was affected to the outlier as in the 10 sample sizes the mean is higher compared to 20 sample sizes. This is due to the outlier in the smaller sample sizes. For the standard deviation, 10 sample sizes giving 1.87 and it start increasing until 100 sample sizes. Then, it drops down at 500 sample sizes and decrease until 1.04 standard deviation in 1000 sample sizes. This shows that at 500 sample sizes, the dispersion of the signal is optimum and more sample giving less dispersion or error of the signal. Afterward, the box plot is plotted to show the different sample sizes in term of the median. From 4 reference point analyzed, different location give a different stabilization in term of the median. However, the most stabilize median is when it reaches 500 sample sizes as it can be seen in 4 of the reference points discussed.

In the performance metrics to test the accuracy and distance error from the different sample sizes, two methods are applied which are KNN and DNN. For DNN, the accuracy is gradually increase over the sample size population. However, for KNN the accuracy is not linearly proportional to the sample size as even the sample size increases the accuracy does not give a better accuracy. In distance error, the DNN and KNN method both giving an unpredictable result. Both methods do not show any relationship between the sample size and the distance error. Nevertheless, the sample size still effects the distance error as showed in Figure 11 and Figure 12 in the result section. In short, the sample size effect the accuracy and distance error of the fingerprint system. However, the sample size does not show a linear relationship in accuracy and distance error except in DNN accuracy test.

In the future, multiple devices will be further examining in a crowdsource data, the user contribute to the crowdsource fingerprinting database will have multiple hardware. Hence, this will create a bigger problem due to the diversity signal from the multiple devices. The experiment will be statistically discussed to improve the fingerprinting system in Wi-Fi technologies in predicting the user location. This will help to improve the previous researcher work to contribute in indoor positioning system.

Funding

The authors would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2018/ICT05/UNIMAP/02/4 from the Ministry of Education Malaysia.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Bolliger, P. (2008). Redpin - adaptive, zero-configuration indoor localization through user collaboration, Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-Less Environments - MELT ’08, 55, 2008.

[2] Chunhan Lee, Yushin Chang, Gunhong Park, Jaeheon Ryu, Seung-Gweon Jeong, Seokhyun Park, Jae Whe Park, Hee Chang Lee, Keum-shik Hong, & Man Hyung Lee. (2004). Indoor positioning system based on incident angles of infrared emitters, 30th Annual Conference of IEEE Industrial Electronics Society, 3, 2218–2222, 2004.

[3] Dinh, T.-M. T., Duong, N.-S., & Sandrasegaran, K. (2020). Smartphone-Based Indoor Positioning Using BLE iBeacon and Reliable Lightweight Fingerprint Map. IEEE Sensors Journal, 20(17), 10283–10294, 2020.

[4] Fang, S.-H., & Wang, C.-H. (2011). A Dynamic Hybrid Projection Approach for Improved Wi-Fi Location Fingerprinting, IEEE Transactions on Vehicular Technology, 60(3), 1037–1044, 2011.

[5] Ismail, Abdul Halim, Mizushiri, Y., Tasaki, R., Kitagawa, H., Miyoshi, T., & Terashima, K. (2017). A Novel Automated Construction Method of Signal Fingerprint Database for Mobile Robot Wireless Positioning System, International Journal of Automation Technology, 11(3), 459–471, 2017.

[6] Ismail, A. H., & Terashima, K. (2018). Prediction of WiFi signal using kalman filter for fingerprinting-based mobile robot wireless positioning system, Journal of Telecommunication, Electronic and Computer Engineering, 10(1–15), 17–21, 2018.

[7] Kanaris, L., Kokkinis, A., Fortino, G., Liotta, A., & Stavrou, S. (2016). Sample Size Determination Algorithm for fingerprint-based indoor localization systems, Computer Networks, 101, 169–177, 2016.

[8] Kim, B., & Kong, S.-H. (2016). A Novel Indoor Positioning Technique Using Magnetic Fingerprint Difference, IEEE Transactions on Instrumentation and Measurement, 65(9), 2035–2045, 2016.

[9] Liu, H., Darabi, H., Banerjee, P., & Liu, J. (2007). Survey of Wireless Indoor Positioning Techniques and Systems, IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 37(6), 1067–1080, 2007.

[10] Ma, C., Jee, G.-I., Macgougan, G., Lachapelle, G., Bloebaum, S., Cox, G., Garin, L., & Shewfelt, J. (2001). GPS Signal Degradation Modeling, Proceedings of International Technical Meeting of the Satellite Division of the Institute of Navigation, 1–12, 2001.

[11] Raquet, J., & Martin, R. K. (2008). WiFi-based indoor positioning, IEEE Communications Magazine, 53(3), 150–157, 2008

[12] Roos, T., Myllymäki, P., Tirri, H., Misikangas, P., & Sievänen, J. (2002). A Probabilistic Approach to WLAN User Location Estimation, International Journal of Wireless Information Networks, 9(3), 155–164, 2002.

[13] Sa'ahiry, A. H. A., Ismail, A. H., Kamarudin, L. M., Zakaria, A., & Nishizaki, H. (2021). An Experimental Study of Deep Learning Approach for Indoor Positioning System Using WI-FI System Proceedings of SympoSIMM 2020, 113–124, 2021.

[14] Tarekegn, G. B., Juang, R. T., Lin, H. P., Adege, A. B., & Munaye, Y. Y. (2021). DFOPS: Deep-Learning-Based Fingerprinting Outdoor Positioning Scheme in Hybrid Networks, IEEE Internet of Things Journal, 8(5), 3717–3729, 2021.

[15] Teller, S., Ryan, R., Battat, J., Charrow, B., Ledlie, J., Curtis, D., & Hicks, J. (2008). Organic Indoor Location Discovery, Mit-Csail-Tr-2008-075.

[16] Thewan, T., Ismail, A. H., Panya, M., & Terashima, K. (2016). Assessment of WiFi RSS using design of experiment for mobile robot wireless positioning system, FUSION 2016 - 19th International Conference on Information Fusion, Proceedings, July, 855–860, 2016.

[17] Wang, H., Szabo, A., Bamberger, J., Brunn, D., & Hanebeck, U. D. (2008). Performance comparison of nonlinear filters for indoor WLAN positioning, Proceedings of the 11th International Conference on Information Fusion, FUSION 2008, May 2014.

[18] Yang, S., Dessai, P., Verma, M., & Gerla, M. (2013). FreeLoc: Calibration-free crowdsourced indoor localization, 2013 Proceedings IEEE INFOCOM 2481–2489, 2013.

[19] Yang, C., & Shao, H. (2015). Non-GNSS radio frequency navigation, IEEE International Conference on Acoustics, Speech and Signal Processing, 5308–5311, 2015.

[20] Ye, F., Chen, R., Guo, G., Peng, X., Liu, Z., & Huang, L. (2019). A Low-Cost Single-Anchor Solution for Indoor Positioning Using BLE and Inertial Sensor Data, IEEE Access, 7, 162439–162453, 2019.

[21] Youssef, M., & Agrawala, A. (2005). The Horus WLAN location determination system, Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services - MobiSys '05, 205, 2005.

[22] Zheng-dong, L., Xing-jie, C., Xiu-ling, L., Juan, C., Yan, H., & Hai-mei, X. (2020). Design of Ultra-Wideband Localization System Based on Optimized Time Difference of Arrival Algorithm, IEEJ Transactions on Electrical and Electronic Engineering, 15(8), 1176–1182, 2020.

Cite this paper as:

Sa'ahiry, A. H. A., Ismail, A. H., Kamaruddin, L. M., Hashim, M. S. M., Azmi, M. S. M., Safar, M. J. A., & Toyoura, M. (2021). Effect of Sample Sizes in Fingerprinting Database for Wi-Fi System, *International Journal of Computers Communications & Control*, 16(6), 4394, 2021.
https://doi.org/10.15837/ijccc.2021.6.4394