Benchmarking of Localization Solutions: Guidelines for the Selection of Evaluation Points

Eli De Poorter, Tom Van Haute, Eric Laermans, Ingrid Moerman

Ghent University - imec - IDLab,
iGent Tower - Department of Information Technology,
Technologiepark 15, 9052 Gent (Zwijnaarde),
firstname.lastname@ugent.be

Abstract

Indoor localization solutions are key enablers for next-generation indoor navigation and track and tracing solutions. As a result, an increasing number of different localization algorithms have been proposed and evaluated in scientific literature. However, many of these publications do not accurately substantiate the used evaluation methods. In particular, many authors utilize a different number of evaluation points, but they do not (i) analyze if the number of used evaluation points is sufficient to accurately evaluate the performance of their solutions and (ii) report on the uncertainty of the published results. To remedy this, this paper evaluates the influence of the selection of evaluation points. Based on statistical parameters such as the standard error of the mean value, an estimator is defined that can be used to quantitatively analyze the impact of the number of used evaluation points on the confidence interval of the mean value of the obtained results. This estimator is used to estimate the uncertainty of the presented accuracy results, and can be used to identify if more evaluations are required. To validate the proposed estimator, two different localization algorithms are evaluated in different testbeds and using different types of technology, showing that the number of required evaluation points does indeed vary significantly depending on the evaluated solution.

Keywords: indoor localization, evaluation points, guidelines, benchmarking, standard error of mean value, confidence interval

1. Introduction

In recent years, location-based information has become indispensable in multiple application domains. As such, the amount of published research focusing on indoor localization has increased tremendously. The outcome of these publications are mostly promising, though, the presented results are often biased since a decent evaluation is not always applied. Solutions from scientific literature are often evaluated in different environments using different evaluation methodologies [1]. The chosen environment and the evaluation points can drastically influence the performance results of the solution. As a result, it is hard to compare the performance of different localization solutions from scientific literature [2], as well as to assess with which confidence the reported results should be interpreted.

This paper focuses on the latter aspect: identifying the confidence of reported results based on the number of used evaluation points. As such, a clear definition of 'an evaluation point' is appropriate. In this paper, an evaluation point is a physical location in a test environment whereof the exact coordinates are known by the system evaluator, but not by the localization system. At this physical location, measurement data is collected, and the localization algorithm has to process this data and try to estimate the coordinates of this physical location, without having knowledge of the ground truth. The euclidean distance between the estimated point and the evaluation point is defined as the 'error distance'. In general, localization systems aim to have a low average or median error distance.

Typically, using as few evaluation points as possible is preferred, since this reduces the efforts needed for evaluating the system. However, the error distances of localization systems often exhibit a large variance, with error distances ranging from several centimeters to tens of meters. As such, using too few evaluation points will not result in a reported average accuracy with a high confidence level. In contrast, using too many evaluation points complicates the evaluation process and might not always be necessary.

The main contributions of this paper are as follows. (i) Firstly, the paper investigates whether or not the number of evaluation points has an impact on the presented results. (ii) Secondly, the paper defines an estimator which can determine the number of evaluation points that is needed to obtain a reliable performance result. (iii) Thirdly, guidelines are defined that researchers can follow to improve the reliability of their presented performance results. (iv) Finally, the correctness of these guidelines are verified using multiple configurations of localization systems.

The remainder of this paper is structured as follows. Section 2 contains the related work. Section 3 describes in detail the experimental testbeds and indoor localization algorithms, as well as the data collection process. The behavior of the error distances is investigated in Section 4. The next section, Section 5 analyzes the behavior and properties of the mean value and standard deviation of the subsets. Based on these findings, guidelines for determining the amount of evaluation points and their corresponding confidence interval are provided in Section 6. Finally, the paper's conclusions can be found in Section 7.

2. Related work

This related work section discusses the number of evaluation points used in state of the art research papers. In addition, a number of recent evaluation benchmarks and best practices are discussed.

2.1. The number of evaluation points in scientific papers

In [3], the authors analyze and compare 21 research papers each presenting an Indoor Localization Sensing (ILS) solution. The paper summarizes their algorithm design, devices, test setup and the performance results. However, no information is given regarding the used methodology for selecting the amount and locations of the evaluation points for each solutions. Since it has been shown [2] that some locations (e.g. center of the room) typically have much smaller error distances than more challenging evaluation points (e.g. near the walls), it is difficult to objectively compare the proposed solutions without more information about the evaluation points.

To overcome this lack of comparability, some recent papers evaluate the performance of localization solutions in identical conditions. For example, Xiaowei Luo et al. present in [4] a comparative evaluation of Received Signal-Strength Index (RSSI) based indoor localization techniques for construction job sites. Four different algorithms (MinMax, Maximum Likelihood, Ring Overlapping Circle RSSI & k-nearest Neighbor) are each evaluated in two different test environments (Building & Job site). The authors selected 21 measurement points in the first testbed (44.87 m^2) and 18 points in the second testbed (32.26 m^2). Similarly, in [5] 8 different localization solutions are evaluated in the same hospital environment, using the same 73 evaluation points. Although in this case more evaluation points are used, it is not clear whether using such a dense evaluation point grid was either overkill (i.e. not necessary for a detailed analysis) or still insufficient. Although the comparisons in these papers are more objective than the previously mentioned survey due to the use of the same evaluation points, none of these papers indicates why they selected these specific amounts of evaluation points, nor whether the number of used evaluation points was sufficient for obtaining reliable, comparable results.

Since there are no clear indications on the required number of evaluation points, some authors prefer to oversample their environment. Examples include e.g. Gayathri Chandrasekaran et al., who present an empirical evaluation of the limits on localization using signal strength [6]. In their work, a trace-driven emulation is used to evaluate the performance of 12 different localization algorithms. They applied the "leave-one-out" approach to evaluate the algorithms using a fingerprint database. As such, a dataset of 400 evaluation points is split into 400 sets of 399 training points and one testing point. Similarly, the publicly available UJIIndoor-Loc dataset WiFi fingerprinting dataset [7] consists of 933 reference positions (some of which can be used as evaluation points). Although such a large dataset is very useful, obtaining such large qualities of evaluation points is not feasible for most researchers that want to evaluate their solutions.

As such, in summary, for most scientific papers the confidence levels of the reported accuracies are unknown and it is not clear how the number of evaluation points influences this confidence level. As a result, it is either impossible to objectively compare the performance of different scientific papers, and/or datasets need to be of such a magnitude that obtaining them is not feasible in most research environments.

2.2. Evaluation procedure in competitions

Besides the scientific papers mentioned before, the results discussed in this paper are also extremely relevant for indoor localization competitions. Some examples include the EvAAL competition, the Microsoft IPSN competition and the EVARILOS competition.

- EvAAL (Evaluation Ambient Assisted Living (AAL) through Competitive Benchmarking) was the first organization who started evaluating and comparing multiple solutions. Since their main focus is tracking, EvAAL benchmarks solutions use a path, not evaluation points. Currently, they have multiple tracks and the competition is held yearly at the Indoor Positioning & Indoor Navigation (IPIN) conference.

- Since 2014, Microsoft started organizing indoor localization competitions yearly at the Information Processing in Sensor Networks (IPSN) conference. In contrast to EvAAL, Microsoft uses predefined static evaluation points to determine the performance accuracy of the evaluated solutions. In 2014 and 2015, 20 evaluation points were used. However, to save time, in 2015 this number was reduced to only 15 [8]. Since the error distance for the different evaluation points showed a large variation [9], it is not clear if this number is sufficient to objectively compare the performance of all contestants.

- Finally, Evaluation of RF-based Indoor Localization Solutions (EVARILOS) organized a competition with a special focus on the benchmarking methodology. The main focuses were the evaluation procedure in combination with the performance metrics and its definitions. During the competition, two different environments were used with each a minimum of 20 evaluation points.

In these competitions, an (often monetary) award is given to the competitor that participated with the localization solution with lowest average error distance. Since these competitions are typically organized on a single day, during which multiple solutions need to be evaluated, they prefer to use as few evaluation points as possible. Often, several competitors have very similar performances, and it is not always clear if the winner actually performed best or just had the good fortune that the selected evaluation points were in his favor. Since the impact of winning such a competition is high (especially for participating companies), including information about the certainty of the outcomes in the evaluation procedure would be much more fair to the contestants.

A summary of the number of typically used evaluation points in recent research papers and recent indoor localization competitions is shown in Table 1. Currently no rule of thumb exists regarding the number of used evaluation points, with numbers varying from 15 (probably not enough) to over 200 (probably overkill). Even very prestigious competitions sometimes limit themselves to using only 15 evaluation points, whereas many scientific papers seem to converge on using 20 evaluation points. Based on these strong variations in the number of used evaluation points, it is clear that no best practices are currently available regarding the number of used evaluation points.

Table 1: Overview of papers and competitions with the used test environment and evaluation points.

Summary	Test Environment	Area [m^2]	# Evaluation Points	Ref
RESEARCH PAPERS				
Indoor Localization Sensing solutions	Typical buildings	[12, 8 625]	n.a.	[3]
Comparative evaluation #1	Building	44.8	21	[4]
Comparative evaluation #2	40th floor under construction	32.13	18	[4]
ORBIT	ORBIT Indoor Testbed	334.45	400	[6]
Optimal landmark placement	3rd floor of CoRE Building	1 486	286	[10]
High performance	Outdoor, open field	100 , 144	36, 49	[11]
COMPETITIONS				
EvAAL 2011 - 2016	Home environment - Exhibition hall	n.a.	Tracking a path	[12]
Microsoft Indoor Localization 2014	Large office	300	20	[9]
Microsoft Indoor Localization 2015	Exhibition hall	2000	20	[8]
Microsoft Indoor Localization 2016	Large office	465	15	[13]
EVARILOS 2014	Office	450	20	[14]

2.3. Benchmarking approaches for localization solutions

Finally, the need for better evaluation methods has also been recognized in the scientific community. Recent papers that proposed improved evaluation methods include e.g. the definition of more meaningful evaluation metrics [1, 15] and the definition of standardized evaluation scenarios [16]. However, although these efforts are extremely important for improving comparability between the performance of localization solutions, these works do not propose any metrics to quantify the trustworthiness that can be assigned to the reported performance results. In particular, none of these benchmarking proposals include suggestions on how many evaluation points should be used.

As such, to the best of our knowledge, our work is the first to (i) theoretically and experimentally analyze the impact of the number of evaluation points on the reliability of the reported accuracies and (ii) propose a methodology to easily identify the number of required evaluation points during the evaluation process.

3. Test set-up

The results discussed in this paper have been established using two test environments, two localization algorithms and multiple technologies (Zigbee, WiFi and Bluetooth). This section describes the evaluation environments and algorithms, as well as the process of data gathering.

3.1. Test environments

3.1.1. w-iLab.t II

As a first evaluation environment, the w-iLab.t II test facility was used. It is located in Zwijnaarde (Ghent, Belgium) above a cleanroom. Due to its open nature and the presence of many metal obstacles (metal pipes), the environment can be considered representative of a challenging industrial environment. The total area measures 66m by 21m and is shown in Figure 1. Since the test environment is surrounded by metal walls, it is shielded from outside wireless interference. Only weak signals can be observed sporadically at the borders of the test area.

To collect ground truth information, robots can drive through the testbed and collect measurement data at predefined evaluation points. Since we want to evaluate the impact of the number of evaluation points, a very fine 2m x 2m grid of 203 evaluation points was defined where data was collected

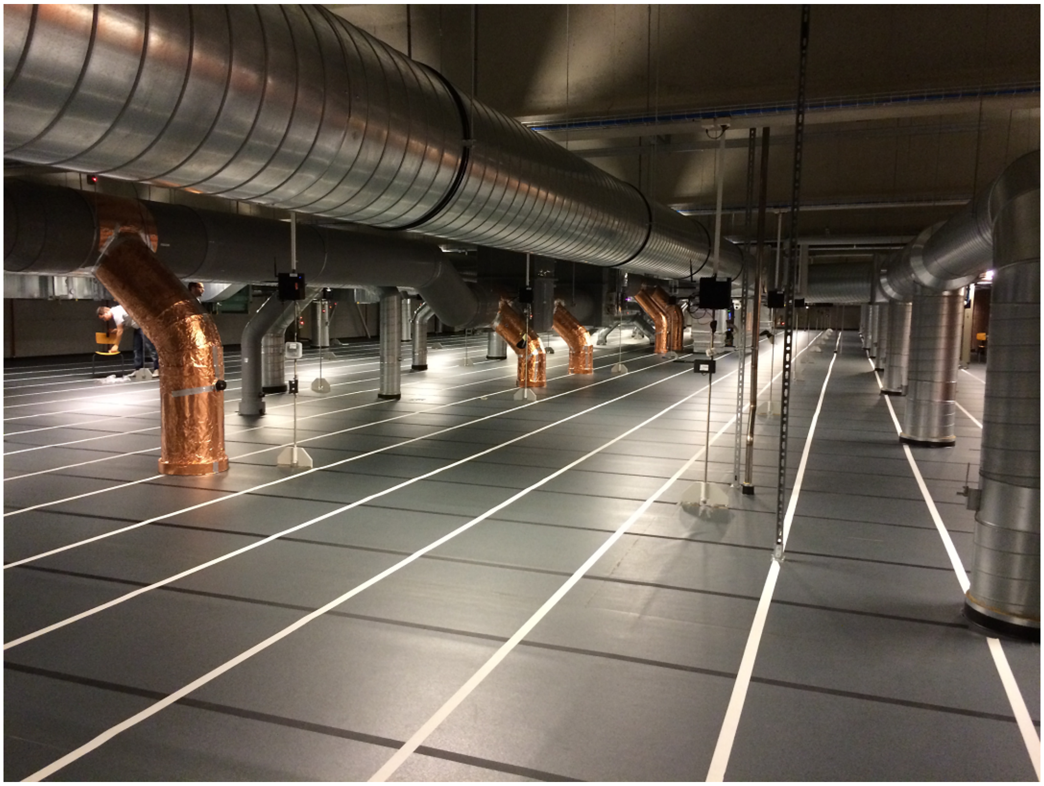


Figure 1: The w-iLab.t II testbed environment located in Zwijnaarde, near Ghent. The testbed is an open space containing metal pipes and obstacles causing multipath fading. Therefore, it can be considered as a challenging environment.

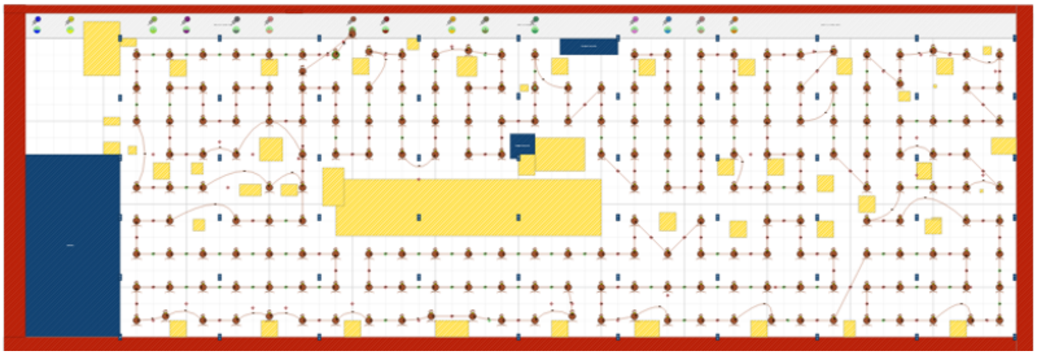


Figure 2: Indication of the 203 evaluation points that were used in the w-iLab.t II test environment in Zwijnaarde. Yellow areas are not available due to pipes and other constructions. The blue areas indicate the technical equipment (e.g. dataracks) of the testbed. At these marked locations, it is not feasible to drive a robot.

(Figure 2). A few points are missing in the grid, mainly in the middle of the test environment, due to the metal obstacles and pipes. Nevertheless, this number of evaluation points is far larger than the number of used evaluation points in most scientific literature, especially in terms of number of evaluation points per square meter.

3.1.2. Hospital environment

A second measurement campaign was performed in an actively used hospital environment (the Sint-Jozefskliniek hospital in Izegem, Belgium). The measurements were performed in the "surgical day hospital" ward, located in a new building on the first floor. The end of the corridor was available for the experiments, while the rest of the ward was in "normal operation", meaning patients and nurses were present and were walking around. The floor plan of the ward is depicted in Figure 3. Patient rooms 9, 10 and 11 were used for the evaluation. A dense evaluation grid of 1 m by 1 m was marked on the floor resulting in 73 evaluation locations. Note that the grid was positioned in such a way that grid lines are 10 cm away from the wall. During the data collection, all doors were open.

3.2. Localization algorithms

To analyze the impact of the number of evaluation points on the reported accuracy, two different localization approaches where selected: a localization algorithm based on signal strength that does not use prior information for its

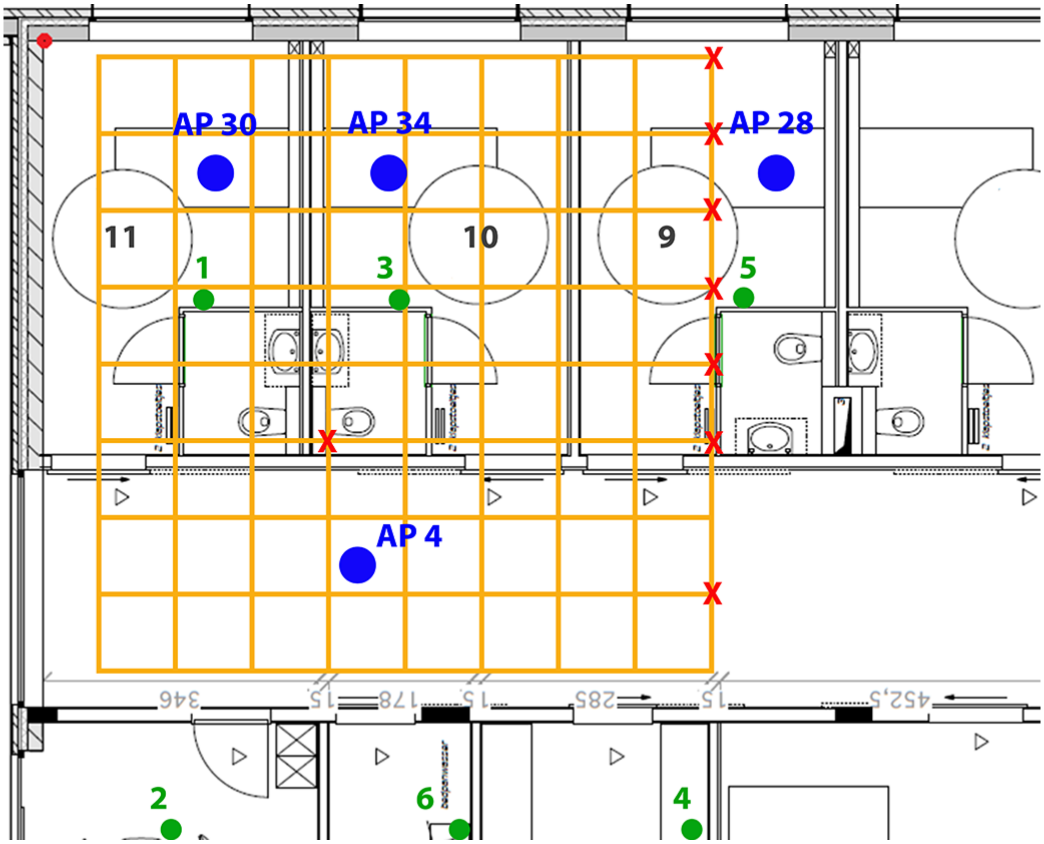


Figure 3: The floor plan of the hospital environment. The yellow grid represents the available evaluation points. The grid points with a red cross do not have measurement data, this results in 73 available evaluation points. Blue dots refer to the location of the WiFi AP, the green ones represent the location of the Zigbee and BLE nodes.

position estimates, and a tracking algorithm based on Viterbi that keeps track of previous position estimates to improve the accuracy of position estimates.

3.2.1. Multilateration algorithm

As a first localization algorithm, a multilateration approach based on received RSSI-values is selected. This specific algorithm first averages all the RSSI values received per reachable anchor point. Next, the algorithm applies multilateration on all possible combinations with three different anchor nodes. Finally, it averages all the x and y coordinates separately in order to return the final x and y coordinate of the estimated location. Similar approaches are very common in scientific literature, and many scientific papers that describe various optimizations of this basic principle are available. As such, the algorithm is representative for a wide range of existing solutions.

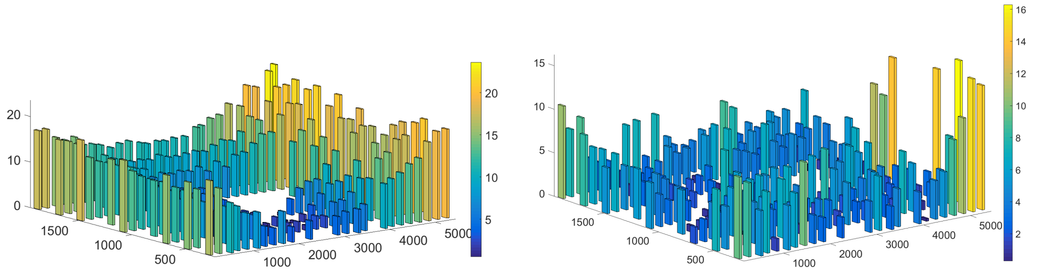
3.2.2. Viterbi based algorithm

Besides point based localization, the inclusion of tracking information is a common approach to increase the accuracy of location estimates. For this paper, a tracking based solution as described in [17] based on the Viterbi algorithm was used. This algorithm uses semantic data to process and estimate the path that needs to be tracked. As expected, the accuracy of this solution is higher than the first algorithm (see Section 4), but it can only be used for continuous tracking of persons or goods.

3.3. Data collection

Indoor localization solutions typically measure environmental data, such as RSSI values, and transform these measurements into position estimates. The authors organized two measurement campaigns to capture a raw dataset of WiFi, Zigbee and BLE RSSI traces at the mentioned evaluation points mentioned in Section 3, once in the w.iLab.t II testbed [1] and once in the hospital [5]. To eliminate the influence of environmental randomness (e.g. interference, RSSI fluctuations, human activities, etc.), multiple indoor localization algorithms can use this dataset as input for position estimations, allowing comparable results between the different algorithm runs.

For the data collection, RSSI and Time of Arrival (ToA) values were collected during 90 seconds at each evaluation point using an STM-32W as a mobile device. During these 90 seconds, the mobile device sequentially sends unicast message to each available anchor node (in a loop). The anchors receiving such a message will reply with an Acknowledgement (ACK),



(a) Multilateration algorithm

(b) Solution based on Viterbi

Figure 4: Visual representation of the error distances in a 3D bar plot. The X and Y axes represent the coordinates of the testbed (in cm) whilst the Z axis shows the error distance (in m). Generally, error distances are higher at the edges of the test environment.

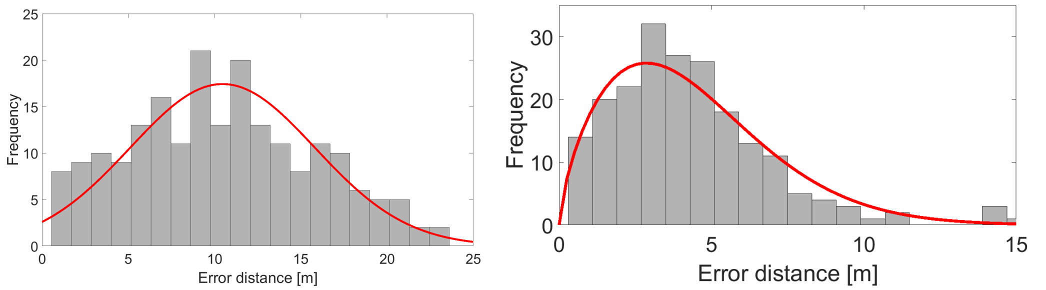
of which the time of arrival and RSSI is stored. In the w-iLab.t II testbed an automated robot based on a Roomba vacuum cleaner was used to collect the data, whereas in the hospital data collection at each evaluation point was done manually after calibrating all evaluation points through laser measurements.

4. Evaluation of the behavior of the error distances

Before investigating the impact of different numbers of evaluation points, this section first discusses the behavior of the complete set of error distances obtained at all evaluation points. Since the conclusions of this section are similar for both evaluation environments, we will for now limit ourselves to discussing the behavior of the localization systems in w-iLab.t II.

The error distances of both localization algorithms in the w-iLab.t II evaluation environment are visualized in Figure 4. As expected, the Viterbi solution on average has lower error distances due to the inclusion of historical data in the algorithm. Furthermore, Figure 4 shows that generally the error distances are larger at the edges of the area compared to the center of the environment, a conclusion which is similar to [2].

To gain better insight in the behavior of the errors, a histogram of the error distances of the algorithms is shown in Figure 5. By evaluating multiple possible distributions, it was found that the error distances of the multilateration algorithm from Figure 5a matches best with a normal distribution ($X \sim N(\mu, \sigma)$ with $\mu = 10.46$m and $\sigma = 5.36$m). This is verified using



(a) Multilateration algorithm: normal distribution

(b) Solution based on Viterbi: Weibull distribution

Figure 5: Histogram of both the error distances in all available evaluation points with corresponding matching distributions (5% significance level): the normal distribution for the multilateration algorithm (a) and the Weibull distributions of the Viterbi algorithm (b).

a χ^2-test (chi-square test), validating the null hypothesis that the error distances are distributed normally at a 5% significance level. However, the error distances obtained by the Viterbi algorithm are lower, resulting in a shift of the histogram bins to the left side, see Figure 5b. As a consequence, the resulting distribution of the error distances does not fit a normal distribution (the χ^2-test rejects the null hypotheses at a significance level of 5 %). In this case, the best matching distribution is the `Weibull` distribution where the probability density function is described as:

$$f(x;\lambda;k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0 \end{cases} \qquad (1)$$

whereby $k > 0$ is the *shape* parameter and $\lambda > 0$ is the *scale* parameter. Since the Weibull distribution is more general than the normal distribution, it can also be used to represent the normally distributed datasets from Figure 5a. However, in general normal distributions are preferred since they are easier to work with due to their lower complexity. As such, the simplification towards the normal distribution is only possible when the algorithm has relatively large error distances.

In the remainder of this paper, the mean average error distance obtained from this set of 203 evaluation points will be considered as the correct mean error distance value of the evaluated solution. In other words, we assume that *the mean value will not change if additional evaluation points are included. This is a safe assumption since this amount of evaluation points is ten times higher than typically used numbers of evaluation points (cf. 203 instead of a typical value of 20 evaluation points).*

Takeaway:

- The distribution of the error distances for localization algorithms can be represented by the Weibull distribution.
- The distribution of error distances for algorithms with larger error distances can be simplified towards the normal distribution ($X \sim N(\mu, \sigma)$).

5. Impact of the number of evaluation points on the mean value and the standard deviation

This section analyzes the impact of the number of evaluation points on two statistical values: the mean and standard deviation of the reported error distances. To illustrate the impact of the evaluation point selection procedure, Figure 6 plots the Cumulative Distribution Function (CDF) of the average error for different subsets of 20 evaluation points. Depending on the selected evaluation points subset, the reported median error distance varies from around 5m up to almost 15m. This large variance of the possible CDF plots shows that the amount of evaluation points can have a significant effect on the behavior of the reported error distances. Preferably, sufficient evaluation points should be used so that adding more evaluation points, or selecting a different subset of evaluation points, does no longer significantly influence the behavior of the CDF.

In the following analysis, the collection of evaluation points is defined by the variable X, whereas the corresponding error distances of this collection of evaluation points is represented as $e = f(X)$. Figure 7 shows a flowchart of the steps executed to obtain the presented results. The process is divided in two steps. First, Section 4 analyzed the statistics of the entire dataset using all 203 evaluation points, resulting in an average error distance and variance that is considered absolutely correct. Next, Section 5 calculates the statistics of multiple smaller subsets of n evaluation points. For each n ($n = [2, 202]$), 200 randomly selected subsets are used to calculate the reported mean error distance and the variance of the error distance. This results in a matrix of data for each statistical parameter, as shown in Figure 8.

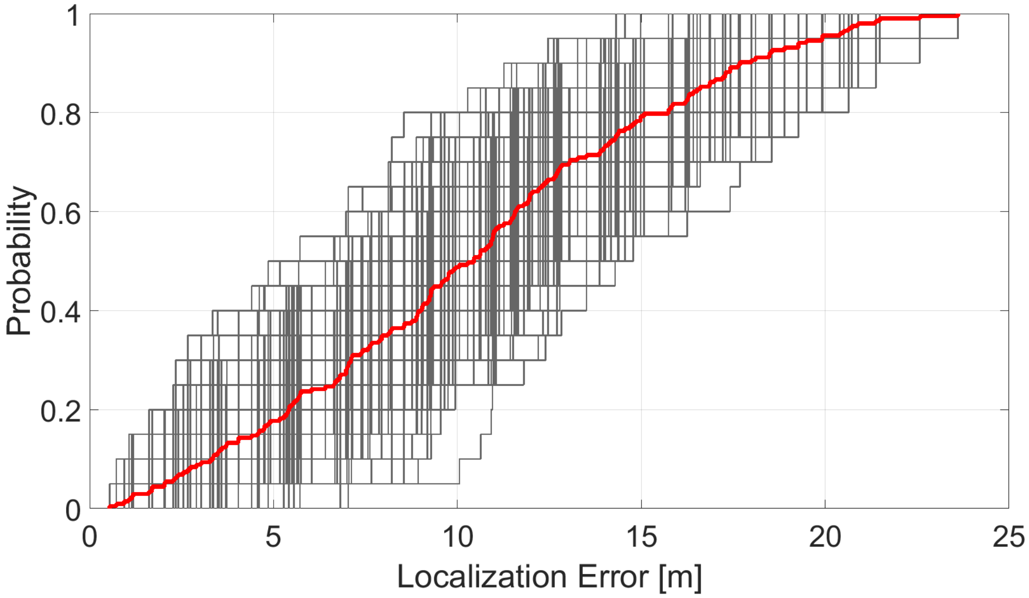


Figure 6: Multiple CDF plots for a subset of 20 evaluation points. Grey lines demonstrate the CDF when using different subsets of evaluation points whilst the red line is the final CDF when all 203 evaluation points are used.

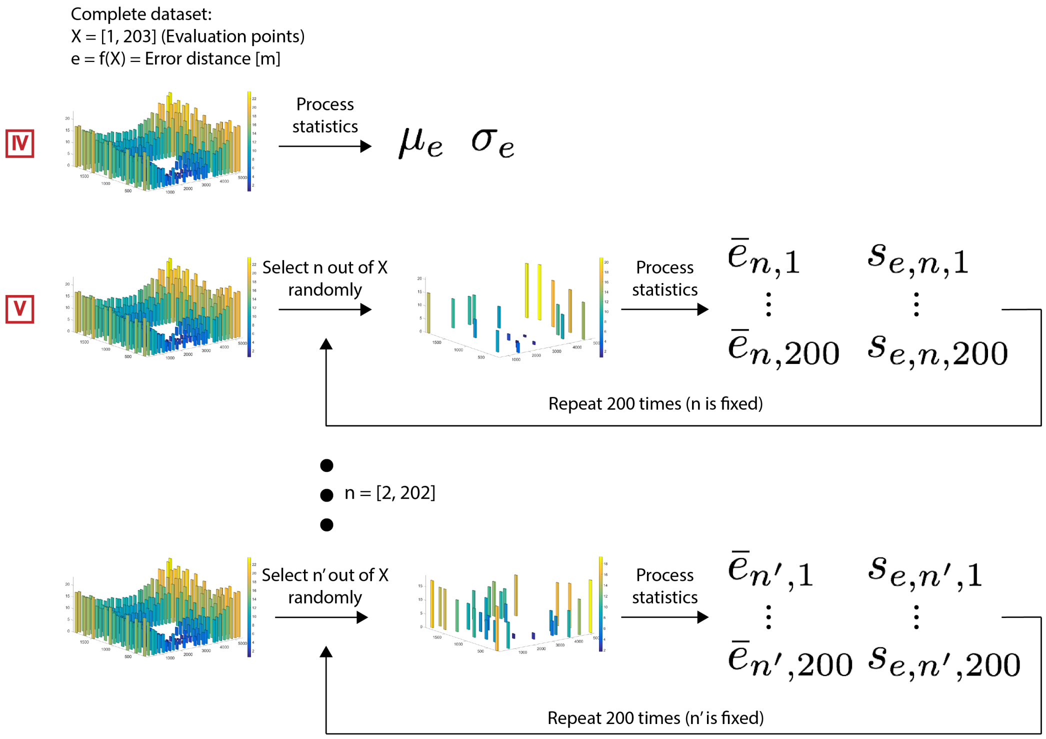

Complete dataset:
X = [1, 203] (Evaluation points)
e = f(X) = Error distance [m]
IV
Process statistics
$\mu_e \quad \sigma_e$
V
Select n out of X randomly
Process statistics
$\bar{e}_{n,1} \quad s_{e,n,1}$
$\vdots \quad \vdots$
$\bar{e}_{n,200} \quad s_{e,n,200}$
Repeat 200 times (n is fixed)
n = [2, 202]
Select n' out of X randomly
Process statistics
$\bar{e}_{n',1} \quad s_{e,n',1}$
$\vdots \quad \vdots$
$\bar{e}_{n',200} \quad s_{e,n',200}$
Repeat 200 times (n' is fixed)

Figure 7: The flowchart of the data-processing. Section 4 discusses the error statistics obtained by using all 203 evaluation points. Next, Section 5 presents the reported error distance statistics when using random subsets of n evaluation points. The process of selecting random evaluation points is repeated 200 times for each value of n.

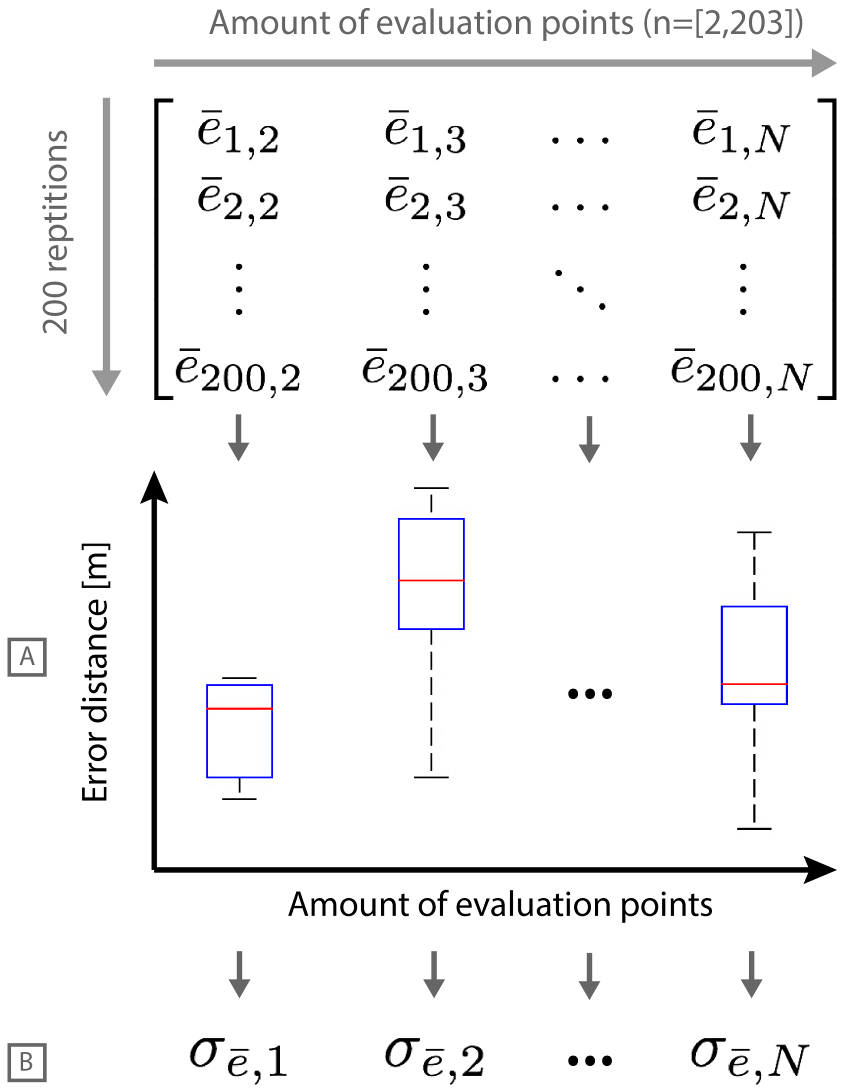

Amount of evaluation points (n=[2,203])
200 reptitions
$\bar{e}_{1,2}$ $\bar{e}_{1,3}$ $\cdots$ $\bar{e}_{1,N}$
$\bar{e}_{2,2}$ $\bar{e}_{2,3}$ $\cdots$ $\bar{e}_{2,N}$
$\bar{e}_{200,2}$ $\bar{e}_{200,3}$ $\cdots$ $\bar{e}_{200,N}$
A
Error distance [m]
Amount of evaluation points
B
$\sigma_{\bar{e},1}$ $\sigma_{\bar{e},2}$ ••• $\sigma_{\bar{e},N}$

Figure 8: The process described in Figure 7 results in a matrix of average error distances $\bar{e}$, with data entries of 200 random selections of each time n evaluation points. This allows calculating the spread ($\boxed{\text{A}}$) and variance ($\boxed{\text{B}}$) that is present within the subsets of n evaluation points, as shown later in respectively Figure 9 and Figure 10.

5.1. Multilateration algorithm in w-iLab.t II

Since most scientific papers report the average point error as their main evaluation criteria, this section formulates an answer to the question “How many evaluation points are required in order to obtain a reported mean value which is close enough to the actual mean error distance?”. To answer this question, it is important to evaluate the spread of the reported mean error for different subsets of evaluation points. To this end, Figure 9 shows the spread of the reported error as a function of the number of used evaluation points. This graph is achieved by processing the data using Trace V as depicted in Figure 7. This results in a matrix that consists of mean values $\bar{e}_{i,j}$. The i-th column represents the repetition ID whilst the j-th column represents the number of used evaluation points. Figure 9 represents a boxplot of the mean values off the vertical matrix entries ($[\bar{e}_{1,n}, \bar{e}_{200,n}]$). As expected, the spread of the mean error distance is larger when the number of evaluation points is low since more subset combinations are possible. A histogram of the spread is also shown for $n = 10$, 30, 60, 100 and 190, as represented by the red lines. χ^2-tests confirm that the spread of these histograms is normally distributed as well.

Takeaway: The average reported error from random subsets of evaluation points (with fixed n) can be represented using a normal distribution.

Now that we know the spread and distribution of the reported average error distance, it is possible to calculate the likelihood that an atypical subset of n evaluation points is selected. To this end, we calculate the standard deviation σ of the reported mean error for different subsets of n evaluation points. This standard deviation represents how many reported error distances are close to the mean value of the entire dataset. With e defined as the error distance and N the number of used evaluation points, the standard deviation of the complete dataset can be calculated using the Equation 2.

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |e_i - \mu_e|^2 \text{ with } \mu_e = \frac{1}{N}\sum_{i=1}^{N} e_i} \quad (2)$$

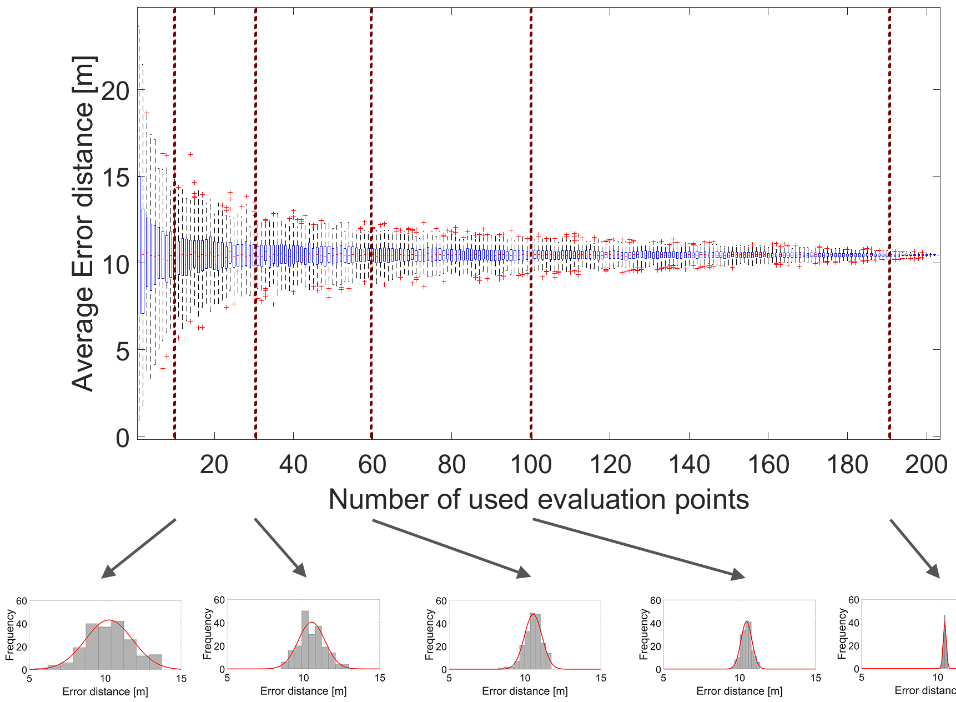


Figure 9: Boxplots representing the spread of the mean error distances with different subsets of evaluation points. Below the boxplots, histograms visualize the spread of respectively $n = 10$, 30, 60, 100 and 190 evaluation points.

However, this formula is only valid when collecting a "complete set" of 203 evaluation points, which is not the case when selecting smaller subsets of n evaluation points. Therefore, the standard deviation σ_e is estimated by examining a random sample taken from the complete dataset and computing a statistic of the sample. This statistic is called an estimator, namely the `sample standard deviation` s_e. In order to calculate the value of this estimator, Bessel's correction [18] needs to be applied, resulting in Equation 3, with N representing the number of used evaluation points and $\bar{e}$ being the sample mean.

$$\sigma_e \approx s_e = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|e_i - \bar{e}|^2 \text{ with } \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i} \tag{3}$$

The behavior of the standard deviation ($\sigma_{\bar{e}}$, see Figure 8 B) of the reported mean error $\bar{e}$ is depicted in Figure 10. The standard deviation decreases exponentially when the number of used evaluation points N increases. In other words: the sample mean values are more stable and have less variation if the amount of evaluation points increases. This demonstrates that adding more evaluation points indeed lowers the risk of reporting an incorrect average error distance. However, the benefits of adding evaluation points decreases over time. After a certain number of evaluation points, the "cost" of adding additional evaluation points will not compensate the "profit" of a lower standard deviation. This is also known as `the law of diminishing returns`.

Based on the central limit theorem, the true standard deviation of the mean values can be calculated as follows:

$$\sigma_{\bar{e}} = \frac{\sigma_e}{\sqrt{N}} \tag{4}$$

In this formula, σ_e is the true standard deviation of the entire dataset. Again, when using only a subset of N evaluation points, an estimator can be used for $\sigma_{\bar{e}}$ which is also known as the `standard error of the mean (SEM)`:

$$\sigma_{\bar{e}} \approx SEM = \frac{s_e}{\sqrt{N}} \tag{5}$$

In conclusion, it is possible to estimate the behavior of the standard deviation of the mean values for each value of n ($n = [0, N]$) by only using the

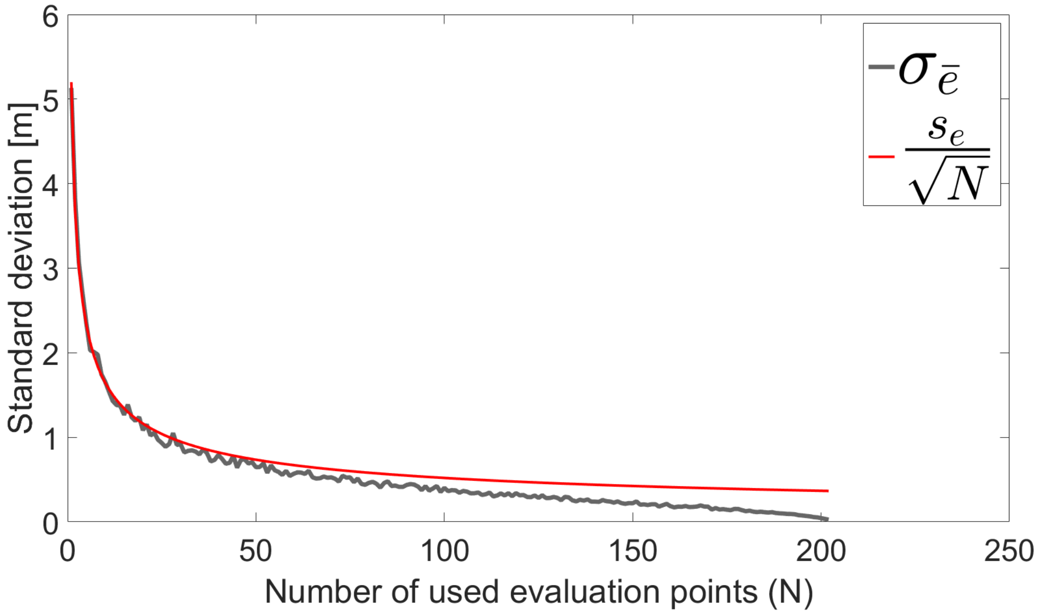


Figure 10: Standard deviation of the mean values of the error distances $\sigma_{\bar{e}}$ are represented by the grey curve. The estimator SEM is visualized by the red curve and is an efficient estimation of the standard deviation values when the amount of evaluation points (N) is lower.

sample standard deviation of one single sample. This estimation is shown in Figure 10 by the red curve, using estimator s_e as the estimated standard deviation based on a subset of 30 evaluation points. To be valid as an estimator, the central limit theorem requires a "large" number of items for the subset, typically 30 or more. Lower number of evaluation points can also be used, but the resulting estimator can have higher shifts [19].

The estimator is sufficient if the complete set is considered as an infinite set of evaluation points or if the complete set is unknown. However, as seen in Figure 10, for large numbers of evaluation points the estimator no longer corresponds to the standard deviation. This is because the estimator assumes that the sample size is much smaller than the size of the complete dataset. As such the estimator is no longer correct if the sample size is too large, e.g. 100 or more in Figure 10. To obtain a better estimation also for larger number of evaluation points, the estimator must be corrected by multiplying a `finite population correction (FPC)`[20].

$$\sigma_{\bar{e}} \approx SEM = \frac{s_e}{\sqrt{N}} * \sqrt{\frac{N-n}{N-1}} \tag{6}$$

Figure 11 demonstrates the effect of the finite population correction. The estimator SEM approximate the true standard deviations using only two parameters from the subset of evaluation points: the standard deviation of the sample and the total amount of evaluation points. In the cases whereby the total amount is unknown, only one single parameter will be sufficient.

Takeaway:

- A higher standard deviation raises the possibility of retrieving an atypical subset.
- The standard deviation of the mean values can be estimated for each value of n using the standard deviation of the error distances s_e of one single subset.

5.2. Verification

We now can estimate the expected standard deviation of the mean values using only a small subset of evaluation points using Equation 3. These formulas were verified using the multilateration localization solution in w.iLab.t II. This section verifies if the formulas above also can be applied for other

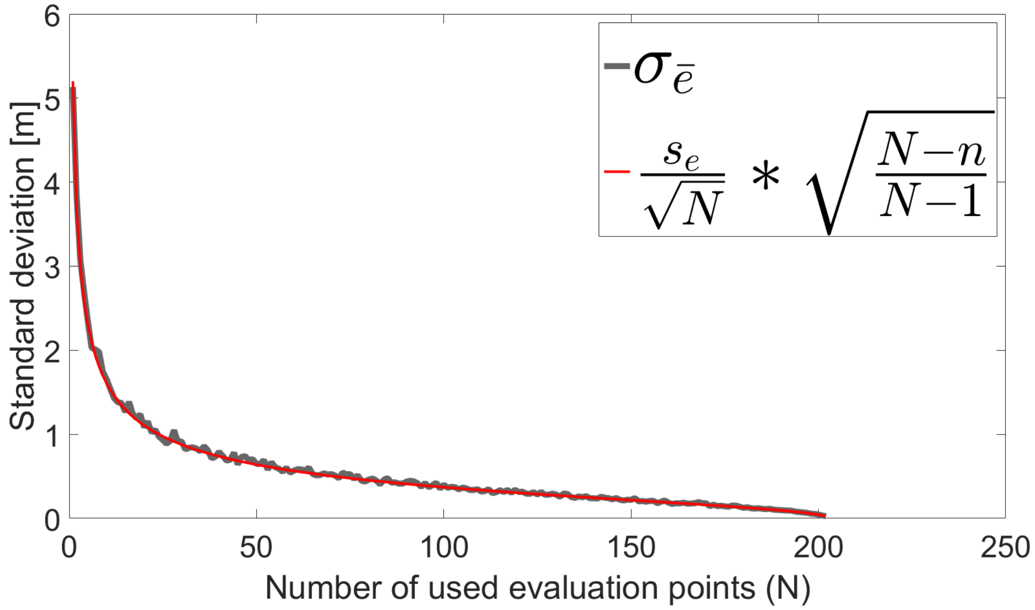


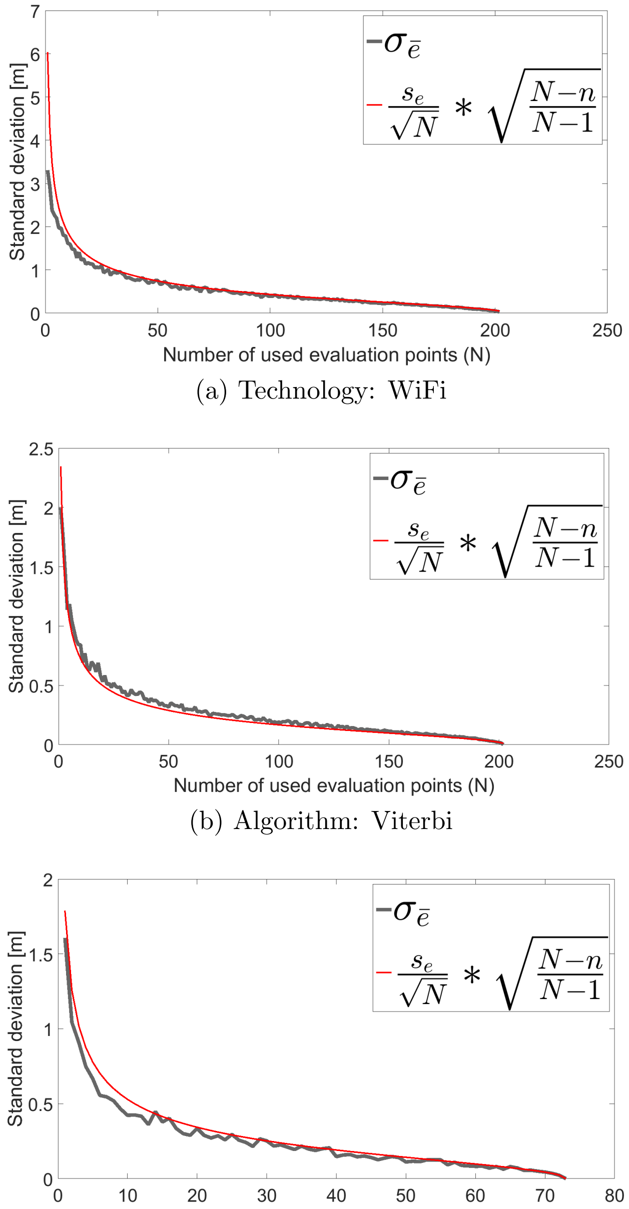
Figure 11: Standard deviation of the mean values of the error distances $\sigma_{\bar{e}}$ in combination with the estimator SEM whereby the finite population correction is applied.

localization solutions and environments. To this end, the same analysis procedure was applied for the more advanced localization algorithm (Viterbi), a different environment (hospital) and a different technology (WiFi). The different localization solution combinations are explained in Section 3. For completeness, all the standard deviations and mean values of the complete dataset are listed in Table 2.

The relation between $\sigma_{\bar{e}}$ and $\frac{s_e}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$ for each new configuration is shown in Figure 12. These results show that a single experiment using 30 evaluation points can predict the true standard deviation in function of the used amount of evaluation points. As such, this information can be used to decide if adding additional evaluation points is necessary.

6. General guidelines

The main conclusion of Section 5 is the fact that the standard error can be predicted based on the standard deviation of the mean values for subsets with a different amount of evaluation points. This approximation can be used to define general guidelines which can help an experimenter to (i) decide if



(a) Technology: WiFi

(b) Algorithm: Viterbi

(c) Environment: Hospital

Figure 12: Visualization of the calculated standard deviation of the mean values with the standard error function. Also for different localization solutions and environments the estimator performs well. In all cases, s_e is the standard deviation of 30 error distances, selected randomly in the environment.

Table 2: Summary of the statistical results. All available mean μ and standard deviations σ are summarized in this table. These numbers are calculated using the entire dataset of 203 evaluation points (w.iLab.t II) or 73 evaluation points (hospital).

Algorithm	Technology	w-iLab.t II [m]		Hospital [m]	
		μ_e	σ_e	μ_e	σ_e
Multilateration	Wifi	10.77	5.79	2.68	1.37
	Zigbee	10.46	5.36	3.08	1.64
	BLE	n.a.	n.a.	3.12	1.51
Viterbi	Zigbee	4.35	2.76	n.a.	n.a.

additional evaluation points should be considered and (ii) derive confidence levels for the reported error results.

Guidelines for determining the necessary number of evaluation points:

Step 1 Select as randomly and independently as possible $N = 20$ evaluation points in the chosen test environment.

Step 2 Calculate the mean value $\bar{e}_N$ and the standard deviation $s_{e,N}$ of the dataset.

Step 3 Calculate the current confidence interval based on these statistics:

$$\left[\bar{e}_N - 1.96 * \frac{s_{e,N}}{\sqrt{N}}, \bar{e}_N + 1.96 * \frac{s_{e,N}}{\sqrt{N}}\right] \tag{7}$$

Step 4 Define the desired confidence bounds CB ($\bar{e}_N \pm CB$) and calculate the amount of evaluation points N' needed:

$$N' = \left(\frac{1.96 * s_{e,N}}{CB}\right)^2 \tag{8}$$

Step 5 Go back to **Step 1** with $N = N'$ and verify that the confidence bounds are changed.

The guidelines consist of 5 steps. Rather than enforcing a minimum number of evaluation points, the guidelines assume that the experimenter wants at least a confidence bound on their reported results, expressed as the maximum uncertainty in meters. Based on this confidence level, the guidelines will predict how many evaluation points are required to achieve this confidence level. The actual confidence level can be selected by the experimenter depending on available experimentation time or required overall accuracy of the results.

To illustrate the guidelines, the process has been applied to all algorithms, environments and technology combinations discussed in Section 3. The results for each configuration are presented in Table 3 using a confidence bound of $1m$. To achieve this confidence level, the number of required evaluation points varies strongly, ranging from 20 up to 109 required evaluation points. These outcomes demonstrate that the minimum number of required evaluation points depends strongly on the actual localization solution. However, by following the guidelines, every localization solution combination achieved a confidence interval whereby the "estimated mean value" μ_e finds itself within the given confidence bounds.

Takeaway:

- Guidelines are given to determine the number of evaluation points required to report mean errors within a desired confidence interval.
- The number of evaluation points required depends on the specifics of the localization solution (i.e. the algorithm, technology, etc.) and the test environment. However, the guidelines can be applied to any combination of technologies, algorithms and environments.

7. Conclusions and future work

In current scientific literature, evaluation procedures for analyzing the performance of localization solutions are often still lacking. Since evaluating a localization solution is time consuming, in many cases the reported accuracy results are obtained using a limited number of evaluation points. Based on current evaluation procedures, it is not possible to identify if the resulting accuracy reports are trustworthy or not. Although some results are probably reliable, the lack of analysis on the confidence levels of these reported results makes it very difficult to identify which of these reports can be trusted and to compare these objectively. To remedy these shortcomings, this paper studied the behavior of reported error distances, as well as the influence of the number of evaluation points on the reported accuracy.

First, a study on the behavior of the error distances shows that the histogram of error distances can be represented using a Weibull distribution. If

Table 3: Applying the guidelines on the multiple datasets of error distances. The amount of necessary evaluation points depends on the combination of the selected algorithm, technology and environment. It also depends on the desired confidence bounds of the mean value. In this example, a confidence bound of 1 m was chosen.

Algorithm	Multilateration	Multilateration	Viterbi	Viterbi
Technology	Zigbee technology	Wi-Fi technology	Zigbee technology	Zigbee technology
Environment	w-iLab.t II	w-iLab.t II	w-iLab.t II	Hospital
Step 1	$N = 20$	$N = 20$	$N = 20$	$N = 20$
Step 2	$\bar{e}_{20} = 9.88\ m$	$\bar{e}_{20} = 10.81\ m$	$\bar{e}_{20} = 3.89\ m$	$\bar{e}_{20} = 2.89\ m$
	$s_{e,20} = 4.63\ m$	$s_{e,20} = 5.32\ m$	$s_{e,20} = 2.28\ m$	$s_{e,20} = 1.79\ m$
Step 3	$[9.88 \pm 2.03]\ m$	$[10.81 \pm 2.33]\ m$	$[3.89 \pm 1.00]\ m$	$[2.89 \pm 0.78]\ m$
	$[7.85, 11.91]\ m$	$[8.48, 13.14]\ m$	$[2.89, 4.89]\ m$	$[2.11, 3.67]\ m$
Step 4	Desired $CB = 1\ m$	Desired $CB = 1\ m$	Desired $CB = 1\ m$	Desired $CB = 1\ m$
	$\Rightarrow \boldsymbol{N' = 82}$	$\Rightarrow \boldsymbol{N' = 109}$	$\Rightarrow \boldsymbol{N' = 20}$	$\Rightarrow \boldsymbol{N' = 20}$
Step 5	$\bar{e}_{82} = 11.17\ m$	$\bar{e}_{109} = 10.45\ m$	$\bar{e}_{20} = 3.89\ m$	$\bar{e}_{20} = 2.89\ m$
	$s_{e,82} = 5.40\ m$	$s_{e,109} = 5.63\ m$	$s_{e,20} = 2.28\ m$	$s_{e,20} = 1.79\ m$
	$[11.17 \pm \mathbf{1.16}]\ m$	$[10.45 \pm \mathbf{1.05}]\ m$	$[3.89 \pm \mathbf{1.00}]\ m$	$[2.89 \pm \mathbf{0.78}]\ m$
	$\boldsymbol{[10.01, 12.33]\ m}$	$\boldsymbol{[9.40, 11.50]\ m}$	$\boldsymbol{[2.89, 4.89]\ m}$	$\boldsymbol{[2.11, 3.67]\ m}$
Mean $\bar{e}$	$10.46\ m$	$10.77\ m$	$4.35\ m$	$3.08\ m$

the algorithm has larger error distances, this distribution can be simplified towards a normal distribution. Next, the impact of using subsets of evaluation points was investigated. It was shown that, within a subset of n evaluation points, the reported mean distance can vary strongly, with reported mean errors for some algorithms varying between 5m and 15m, even when using 20 evaluation points. It was also shown that the reported accuracy behaves as a normal distribution with decreasing variance for larger numbers of evaluation points. Not only do we prove mathematically that the confidence of the reported means increases when using additional evaluation points, we also quantify by how much. More specifically, the paper shows it is possible to estimate the standard deviation of the mean value for a certain amount of evaluation points using the standard error of the mean value. Finally, easy to follow guidelines are given to calculate the confidence one has in the reported mean error, as well as to calculate how many additional evaluation points are required to meet predefined confidence levels.

The work presented in this paper will allow researchers to more accurately calculate the confidence they have in their reported work, and will allow them to make informed choices regarding the number of required evaluation points. As such, it signifies a large step forward towards more standardized evaluation methods for localization solutions.

Biography

[1] T. Van Haute, E. De Poorter, F. Lemic, V. Handziski, N. Wirstrom, T. Voigt, A. Wolisz, I. Moerman, Platform for benchmarking of rf-based indoor localization solutions, Communications Magazine, IEEE 53 (9) (2015) 126–133.

[2] T. Van Haute, E. De Poorter, I. Moerman, F. Lemic, V. Handziski, A. Wolisz, N. Wirström, T. Voigt, Comparability of rf-based indoor localisation solutions in heterogeneous environments: An experimental study, Int. J. Ad Hoc Ubiquitous Comput. 23 (1/2) (2016) 92–114. `doi:10.1504/IJAHUC.2016.078483`.
URL `http://dx.doi.org/10.1504/IJAHUC.2016.078483`

[3] N. Li, B. Becerik-Gerber, Performance-based evaluation of rfid-based indoor location sensing solutions for the built environment, Advanced Engineering Informatics 25 (3) (2011) 535–546.

[4] X. Luo, W. J. OBrien, C. L. Julien, Comparative evaluation of received signal-strength index (rssi) based indoor localization techniques for construction jobsites, Advanced Engineering Informatics 25 (2) (2011) 355–363.

[5] T. Van Haute, E. De Poorter, P. Crombez, F. Lemic, V. Handziski, N. Wirström, A. Wolisz, T. Voigt, I. Moerman, Performance analysis of multiple indoor positioning systems in a healthcare environment, International Journal of Health Geographics 15 (1) (2016) 7. doi:10.1186/s12942-016-0034-z.
URL http://dx.doi.org/10.1186/s12942-016-0034-z

[6] G. Chandrasekaran, M. A. Ergin, J. Yang, S. Liu, Y. Chen, M. Gruteser, R. P. Martin, Empirical evaluation of the limits on localization using signal strength, in: Sensor, Mesh and Ad Hoc Communications and Networks, 2009. SECON'09. 6th Annual IEEE Communications Society Conference on, IEEE, 2009, pp. 1–9.

[7] J. Torres-Sospedra, R. Montoliu, A. Martnez-Us, J. Avariento, T. Arnau, J. Benedito-Bordonau, J. Huerta, Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems., IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN), Busan, Korea.

[8] Microsoft, Microsoft indoor localization competition 2015, http://research.microsoft.com/en-us/events/indoorloccompetition2015/, [Online; accessed 18-April-2016] (2015).

[9] D. Lymberopoulos, J. Liu, X. Yang, R. R. Choudhury, S. Sen, V. Handziski, Microsoft indoor localization competition: Experiences and lessons learned, GetMobile: Mobile Computing and Communications 18 (4) (2015) 24–31.

[10] Y. Chen, J.-A. Francisco, W. Trappe, R. P. Martin, A practical approach to landmark deployment for indoor localization, in: Sensor and Ad Hoc Communications and Networks, 2006. SECON'06. 2006 3rd Annual IEEE Communications Society on, Vol. 1, IEEE, 2006, pp. 365–373.

[11] M. Marks, E. Niewiadomska-Szynkiewicz, J. Kolodziej, High performance wireless sensor network localization system, International Journal of Ad Hoc and Ubiquitous Computing 17(2/3):122.

[12] J. A. Álvarez-García, P. Barsocchi, S. Chessa, D. Salvi, Evaluation of localization and activity recognition systems for ambient assisted living: The experience of the 2012 evaal competition, Journal of Ambient Intelligence and Smart Environments 5 (1) (2013) 119–132.

[13] Microsoft , Microsoft indoor localization competition 2016, `http://research.microsoft.com/en-us/events/msindoorloccompetition2016/`, [Online; accessed 18-April-2016] (2016).

[14] F. Lemic, V. Handziski, A. Wolisz, T. Constambeys, C. Laoudias, S. Adler, S. Schmitt, Y. Yang, Experimental evaluation of rf-based indoor localization algorithms under rf interference, in: Localization and GNSS (ICL-GNSS), 2015 International Conference on, IEEE, 2015, pp. 1–8.

[15] H. Liu, H. Darabi, P. Banerjee, J. Liu, Survey of wireless indoor positioning techniques and systems, Trans. Sys. Man Cyber Part C 37 (6) (2007) 1067–1080. `doi:10.1109/TSMCC.2007.905750`.
URL `http://dx.doi.org/10.1109/TSMCC.2007.905750`

[16] Iso/iec 18305 standard test and evaluation of localization and tracking systems, http://www.iso.org/iso/home/standards.htm.

[17] J. Trogh, D. Plets, L. Martens, W. Joseph, Advanced real-time indoor tracking based on the viterbi algorithm and semantic data, International Journal of Distributed Sensor Networks 2015.

[18] W. J. Reichmann, Use and abuse of statistics, Oxford University Press, 1964.

[19] S. S. Kar, A. Ramalingam, Is 30 the magic number? in sample size estimation, National Journal of Community Medicine 4 (1).

[20] L. Isserlis, On the value of a mean as calculated from a sample, Journal of the Royal Statistical Society 81 (1) (1918) 75–81.