Position error and entropy of probabilistic Wi-Fi fingerprinting in the UJIIndoorLoc dataset

2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 4-7 October 2016, Alcalá de Henares, Spain

Position Error and Entropy of Probabilistic Wi-Fi Fingerprinting in the UJIIndoorLoc Dataset

Rafael Berkvens and Maarten Weyn
MOSAIC, Faculty of Applied Engineering
University of Antwerp – iMinds
Antwerp, Belgium
rafael.berkvens@uantwerpen.be

Herbert Peremans
ENM, Faculty of Applied Economics
University of Antwerp
Antwerp, Belgium

***Abstract*—The accuracy of a positioning system is usually expressed as its average position error in an experiment. However, when the ground truth is no longer available, it would still be useful to know the reliability of a position estimate based on a single measurement. To obtain a reliability metric, we hypothesize that there is a relation between the uncertainty in a position's posterior probability distribution, expressed as its conditional entropy, and the position error of the position that is derived from this distribution. In this paper, we present the correlation between these two metrics as calculated for the UJIIndoorLoc Wi-Fi fingerprinting dataset, using a new probabilistic sensor model. We found that there is no significant correlation between the conditional entropy and the position error. However, we learned that our sensor model is usually very certain in the dataset, and saw that the suggestion of a correlation improves when we increase the uncertainty by selecting a fixed, larger variance. Interestingly, the position error results improve as well.**

***Index Terms*—Wi-Fi, fingerprinting, evaluation, conditional entropy, reliability**

I. Introduction

The reliability of the position estimate by a positioning system contains large amounts of information, especially during the deployment of the system, *i.e.*, when no ground truth is available. With the ground truth absent, the error of the position estimate cannot be calculated. You can rely on a previously calculated median error; however, there remains a fifty percent chance that the current estimate is worse than the median error. If the system has multiple sensor modalities to estimate a position, it can change its reliability estimate depending on the availability of these modalities [1]. The level of agreement between the sensing modalities could also be used to create a dynamic reliability estimate for the individual modalities [2].

In a single sensor modality situation, we hypothesize that we can use the uncertainty of the posterior probability distribution of the position estimate to indicate a reliability of that estimate. It seems intuitive that a distribution with high certainty, such as a Gaussian distribution with low variance, will have a good position estimate, while a distribution with low certainty, such as a Gaussian distribution with high variance, can have anything from a good to a bad position estimate. This was inspired by our work with the mean mutual information between measurements of a specific position in the environment and all positions in that environment [3]. To test our hypothesis, we will use the conditional entropy of the posterior probability distribution of the position, given a single measurement. The conditional entropy expresses the uncertainty of a distribution. If our hypothesis holds, then there must be a significant correlation between the position's error and the conditional entropy of a single measurement.

As a testing environment, we chose the UJIIndoorLoc dataset [4]. This is a large dataset of Wi-Fi fingerprints, covering three buildings with up to five floors. The dataset is open access and includes both a training and validation dataset. We use a new probabilistic sensor model based on a Gaussian kernel, which incorporates both access points that are seen in a fingerprint, and access points that are not seen and thus have no Received Signal Strength indicator (RSSI) value for that fingerprint. In previous research, we tend to choose a fixed variance in the sensor model, while we see that in other literature the variance is trained and thus different for each access point at each reference position [5], [6]. We will take this opportunity to compare these two approaches.

This paper continues as follows. First, in Section II, we briefly discuss the dataset that is used to calculate our results. We also explain the probabilistic sensor model, which is used with two different approaches in standard deviation selection. In this section, we end with the formulas for the conditional entropy calculation. Then, in Section III, we provide three sets of results. The first is the cumulative distribution of the position error, with a summarizing table. The second is also a cumulative distribution and summarizing table, but then of the conditional entropy. The third is the correlation between these two metrics. Finally, in Section IV we provide our conclusions.

II. Methods

The large open access UJIIndoorLoc dataset [4] consists of two collections of Wi-Fi fingerprints, acquired in three university buildings of up to five floors. A Wi-Fi fingerprint is the set of RSSI values of each Wi-Fi access point in the environment, as measured at specific reference positions. In other words, if the environment contains A APs, then the measurement w would be $w_1, w_2, \ldots, w_A$ where each w_a can be any value, or $\emptyset$ if the AP is not visible at that position. The two collections in the UJIIndoorLoc dataset are a training dataset and a validation dataset. The training dataset

contains 19,937 measurements, distributed over 933 distinct positions. The validation dataset contains 1111 measurements, distributed over 1074 distinct positions. In the environment, there are 520 access points. The sensor model can be trained using the training database, and results can be obtained using the separate validation database.

The main objective is to test our hypothesis on the correlation between conditional entropy, or uncertainty, and the error of the position estimate. The conditional entropy, a metric from information theory, is only applicable on a posterior probability distribution. Therefore, we must use a probabilistic sensor model of the form $P(pos \mid w)$, where pos is an element of the set of positions Pos in the environment and which reads as the probability of a position given a Wi-Fi fingerprint measurement. The distribution is written as $P(Pos \mid w)$. Using Bayes' rule, we derive:

$$P(pos \mid w) = \frac{P(w \mid pos)P(pos)}{\sum_{pos \in Pos} P(w \mid pos)P(pos)}, \quad (1)$$

where $P(w \mid pos)$ is the likelihood of the measurement given a position; and $P(Pos)$ is the prior probability distribution over the positions, which we assume to be uniform, since we are doing localization using only a single measurement. A common practice in probabilistic sensor models is to assume that the likelihood $P(w \mid pos)$ is independent for each access point in the fingerprint w [5], [7], or:

$$P(w \mid pos) = \prod_{a=1}^{A} P(w_a \mid pos). \quad (2)$$

The likelihood of a RSSI value for a single access point given the position can then be calculated using a Gaussian kernel:

$$P(w_a \mid pos) = \frac{1}{\sqrt{2\pi}\sigma_a} \exp - \frac{(w_a - \mu_a)^2}{2\sigma_a^2}, \quad (3)$$

where $P(w_a \mid pos)$ is the likelihood value for a specific position pos, μ_a is the mean RSSI value of the access point at that position, and σ_a is its standard deviation. The value of μ_a is obtained from the training dataset. The value of σ_a is obtained in two separate ways, effectively creating two separate approaches. Once it is obtained from the training dataset; we call this approach the trained σ. The other approach is by selecting a single standard deviation for all access points at all reference positions; we call this approach the fixed σ. The value of the fixed σ is the median standard deviation of the set of standard deviations derived from the training set. We compare these two approaches in our results.

To calculate the parameters μ_a and σ_a, we must take into account that an access point may not be seen at a reference position. In fact, this is usually the case in the UJIIndoorLoc training dataset, where on average only 18 of the 520 access points have a value in a fingerprint. We have created a probabilistic sensor model based on the Gaussian kernel in Equation (3) that considers these so called 'missing' values, see Figure 1. The idea of this sensor model is that there are two possible situations: there is a RSSI value w_a associated with the access point a, or there is no value associated to it. If there is a value associated to the access point, we can use the Gaussian kernel in Eq. (3). If there is no value associated to the access point, we assume that its RSSI value must have been too low, *i.e.*, under the minimum RSSI value in the training dataset w_{th}. The probability of this situation can be found as the integrated probability of having any value under w_{th}:

$$P(w_a = \varnothing \mid pos) = \int_{-\infty}^{w_{th}} \frac{1}{\sqrt{2\pi}\sigma_a} \exp - \frac{(w - \mu_a)^2}{2\sigma_a^2} dw. \quad (4)$$

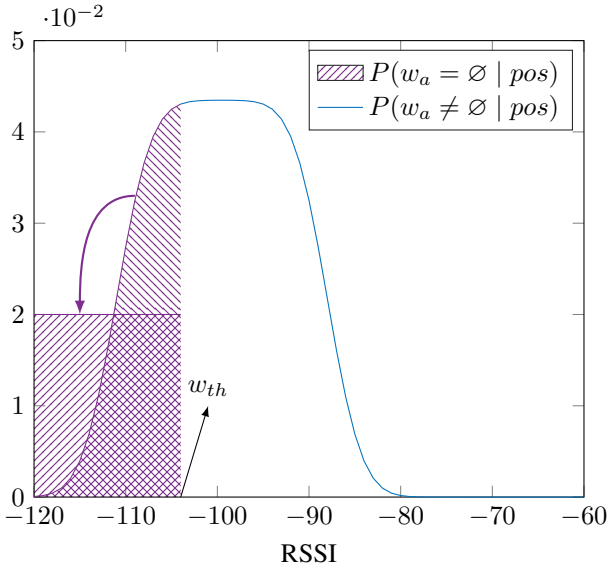


Fig. 1. A graphical representation of our sensor model for a specific access point a at a specific reference position. When the access point has a RSSI value in a fingerprint, the likelihood of being at the reference position can be calculated with a normal distribution. The likelihood of not having an RSSI value in a fingerprint at the reference position can be found by integrating the normal distribution until the threshold RSSI value w_{th}.

Lastly, we note that there is a range in antenna gain that influences the RSSI measurements. This is true for Wi-Fi measurements in general, but specifically for the UJIIndoorLoc dataset, which was collected by different people, using different devices and without any information on their orientation when they collected the Wi-Fi fingerprints. We convolve a uniform distribution with a width equal to the antenna gain range over the sensor model to take the uncertainty about the actual gain into account. This results in the flat top of the distribution in Figure 1. If we would use a normal distribution as probability density function for the antenna gain, rather than a uniform distribution, the result would again be a normal distribution with a mean and variance as the sum of the original means and variances [8]. In the distribution, we indicate the probability of a RSSI value under the threshold, *i.e.*, the access point is not seen, as a uniform distribution, since we have no information on what the correct RSSI value would be if we had been able to measure it.

We then create a search space over a likely range of values for μ_a and σ_a. Within this search space, calculate the

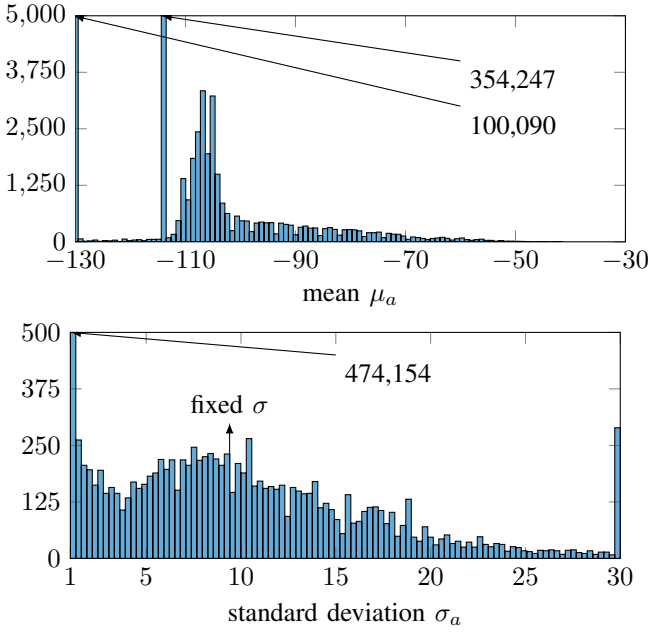


Fig. 2. Histogram distribution of the parameters found at each reference position for each access point after training the sensor model. The arrow indicates the fixed standard deviation σ.

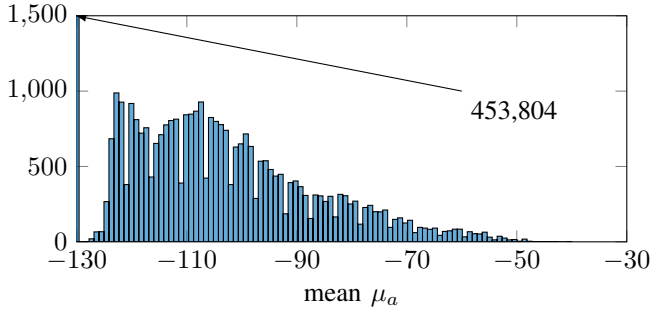


Fig. 3. Histogram distribution of the parameter μ found at each reference position for each access point after training the sensor model with the fixed σ.

likelihood of the values w_a. Finally, we select the parameters μ_a and σ_a that maximize this likelihood. The distributions of the parameters that we found this way for the trained σ approach are shown in Figure 2. The distribution of μ_a has two large peaks: these occur when an access point is never seen at a reference position. The difference between the peaks is due to numerical rounding. When an access point is never seen, the likelihood of the parameters is nearly equal for a certain set of small μ_a and σ_a values. When more than 27 samples are taken at the reference position, and the access point is not seen in any of those 27 samples, the likelihood in this set equalizes, due to rounding. This is when the real smallest μ_a and σ_a is selected; the difference in performance afterwards is negligible. The same peaks would have been visible in the distribution of σ_a, if the step size of the histogram was smaller.

When using the fixed standard deviation, we still train the mean parameter for each access point in the same manner. The distribution of the mean trained with the fixed standard deviation is shown in Figure 3. For comparison, the fixed standard deviation has been indicated on the distribution of the trained standard deviation. The distribution of μ_a for the fixed σ approach has only a single peak, as opposed to the two peaks in the distribution of the trained σ. This is because the same issue with the nearly identical likelihoods does not occur for the selected standard deviation.

During the actual localization phase, we apply the same sensor model: if there is a value for access point a, we use the part of the distribution in Figure 1 above the threshold value w_{th}; if there is no value for access point a, we use the probability of the area below the curve. The likelihood of the complete fingerprint given a reference position, *i.e.*, for all access points at the reference position, can then be found by multiplying the likelihoods of the individual access points, see Eq. (2). The posterior probability distribution can subsequently be found using Bayes' rule, see Eq. (1), where we assume a uniform prior distribution for all positions.

We select the position with the highest posterior probability as the position estimate, which is also known as the Maximum A Posteriori (MAP) estimate. This facilitates in selecting a building and floor as well, since these are set to those of the selected position.

Using the position estimate, we calculate the position error. We chose to calculate a three dimensional Euclidean distance, based on the latitude, longitude, and floor values in the UJIIndoorLoc dataset. During last year's competition [3] it was indicated that the latitude and longitude values are actually expressed in meters. Additionally, the average floor height was indicated as four meters. Thus, by multiplying the floor identification value by four, we obtained a metric value for our third dimension. Finally, we did not take the building identification value into account, since the positions are already spatially separated by the latitude and longitude coordinates; selecting the wrong building will inevitably increase the position error.

Since we use the MAP estimate, there is a minimum position error that can be achieved—we will always select a position from the training dataset as our position estimate, which is unlikely to be exactly a position from the validation dataset. On average, there is about 1.67 m between the position of a validation sample and the nearest training position. To illustrate the performance of the positioning system, we applied our model to the private dataset that was used at last year's competition [9]. The organizers of the competition evaluated our results with the sensor model, so that we could compare with our own and the winning team's result.

From the position's posterior probability distribution given a measurement w in Eq. (1) we can calculate the conditional entropy of this distribution. The general formula of the entropy is, see [10]:

$$H(X) = \sum_{x \in X} P(x) \log P(x), \tag{5}$$

where X is a random variable and $p(x)$ is the posterior probability value of an outcome x. This is the mean value of the Shannon's information content of each probability value in the probability distribution. The conditional entropy is the entropy of a conditional probability distribution:

$$H(X \mid y) = \sum_{x \in X} P(x \mid y) \log P(x \mid y), \tag{6}$$

which is more applicable to our situation. In the notation of our problem, the conditional entropy becomes:

$$H(Pos \mid w) = \sum_{pos \in Pos} P(pos \mid w) \log P(pos \mid w), \tag{7}$$

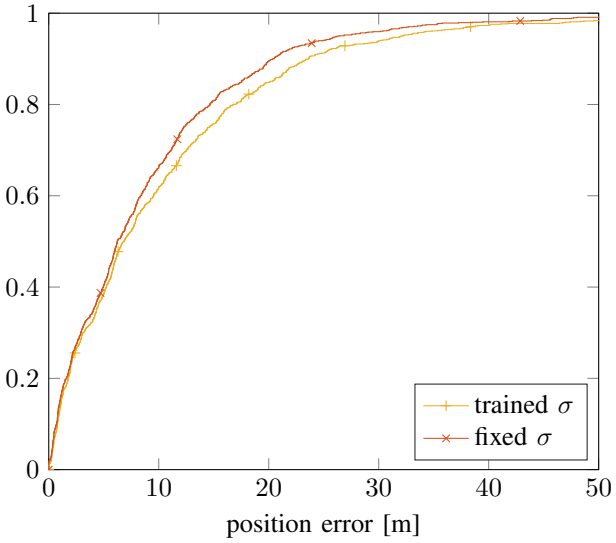
where the logarithm is a base two logarithm, so the result is expressed in bits. The conditional entropy, or uncertainty, of a distribution, is maximal for a uniform distribution. This value can be derived by $\log N_{pos}$, where N_{pos} is the number of positions in the training dataset, in our case $N_{pos} = 933$. The minimal conditional entropy is 0 bit, which is achieved when the posterior probability value at one position is 1 and 0 for all other positions. Our hypothesis is that the conditional entropy of a distribution indicates the reliability of a position estimate, *i.e.*, there is correlation between the conditional entropy and position error.

III. Results

There are three sets of results: the position error, the conditional entropy, and the correlation of those two.

The position error in the validation dataset is summarized in Figure 4. The mean position error is 10.49 m for the approach that trains a standard deviation for each access point at each reference position, and a mean error of 9.20 m for the approach that uses a fixed standard deviation for all access points. The median position error is 6.85 m for the trained σ approach, and 6.23 m for the fixed σ approach. The median position error is a lot smaller than its mean, which indicates that larger errors also occur. This is supported by the 95th percentile of the position error, which is as large as 32.10 m for the trained σ approach and 26.38 m for the fixed σ approach. These errors can be explained by the rather large floor fail ratio, which indicates how often the approaches select a wrong floor. This floor fail ratio is 13.86 % for the trained σ approach, and 9.90 % for the fixed σ approach.

The difference between the position error of the trained σ approach and the fixed σ approach is small. The mean differs only about one meter, the median just more than half a meter. However, this difference is significant according to the two-sample t-test, with a p-value of 0.0064. The difference can be explained as the trained σ approach being much less likely to select a reference position with a set of RSSI values that differ from the set of trained values, even if this position is actually the true position. It will rather select a position that better fits its trained parameters. The fixed σ approach will more easily select such positions. Both systems, however, select positions near the true position, which indicates that these positions reflect the values in the fingerprint of the validation sample



	trained σ	fixed σ
mean [m]	10.49	9.20
median [m]	6.85	6.23
95th percentile [m]	32.10	26.38
floor fail [%]	13.86	9.90
building fail [%]	0.09	0.09

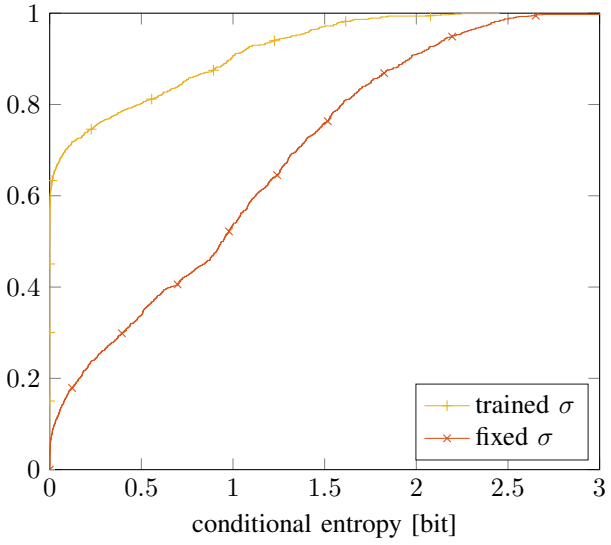
Fig. 4. Cumulative distribution function with summarizing table of the position error when using the UJIIndoorLoc validation dataset with $N = 1111$.

TABLE I
Comparison with 2015 competition, track 3, results.

$N = 5179$	MOSAIC our system	RTLS@UM winning team	trained σ 2016	fixed σ 2016
mean [m]	11.64	6.20	10.34	9.01
median [m]	6.72	4.57	8.64	6.34
floor fail [%]	6.14	6.26	19.00	12.01
building fail [%]	1.35	0.00	0.00	0.00

better. We think that this is caused by the rather sparsely sampled environment.

Our new probabilistic sensor model has a better performance than the one we used on the Wi-Fi fingerprinting competition of last year, see Table I. Conversely, the ratio of selecting a wrong floor has increased, which may be caused by explicitly incorporating the antenna gain ranges. The winners of last year's competition include some information that we do not yet model in our sensor model. For example, they shift the range of RSSI values of one device to the range of RSSI values of another device at the same reference position [11]. In our sensor model, we avoid using hardware specific calibration. Additionally, they included both the training and validation dataset as training data before calculating the results on the competition dataset. It is difficult to further analyze these results, as the detailed results are private to the dataset owners and competition organizers. We conclude that our sensor models provide viable, yet not competitive positioning results;



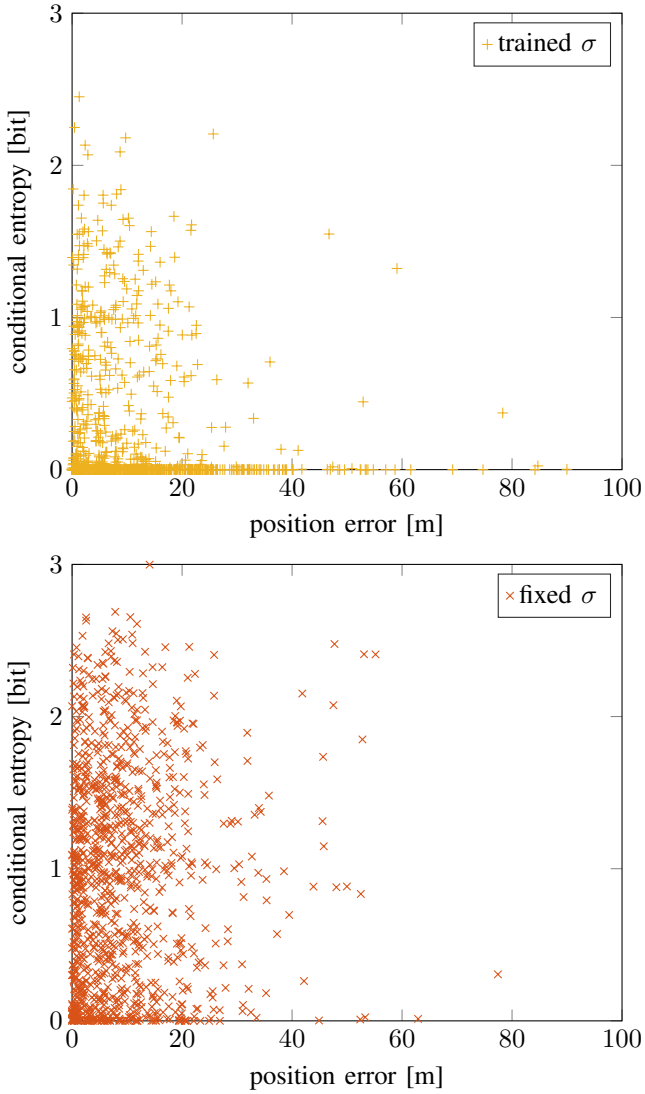
	trained σ	fixed σ
mean [bit]	0.24	0.95
median [bit]	1.26×10^{-5}	0.95
95th percentile [bit]	1.32	2.20

Fig. 5. Cumulative distribution function with summarizing table of the conditional entropy of the position's posterior probability distribution when using the UJIIndoorLoc validation dataset, where the maximum entropy is 9.87 bit.

they might be enough in some situations, and can certainly be used to test our hypothesis.

The conditional entropy in the validation dataset is summarized in Figure 5. The mean conditional entropy of the trained σ approach is 0.24 bit. This approach has a median conditional entropy that is almost zero. The mean conditional entropy of the fixed σ approach is 0.95 bit. Its median conditional entropy is equal to its mean. The 95th percentile of the conditional entropy is 1.32 bit for the trained σ approach and 2.20 bit for the fixed σ approach. The distributions generated by our sensor model are thus very certain, with nonzero probabilities at only few reference positions. On the other hand, the environment is sampled such that there are sometimes less than ten reference positions in a single corridor. This limited amount of samples might not capture the variations in the Wi-Fi signal enough to be reflected in the posterior distributions.

A linear relationship between the position error and the conditional entropy cannot be observed in our data, see Figure 6. While the Pearson's correlation coefficient ρ is nonzero with a p-value under 5 %, the graphical representation shows that it is likely to be caused by outliers. A true linear relationship is not what we expect; rather, we expect that the conditional entropy will increase if the position error increases, but not necessarily the reverse. This is certainly not the case for the trained σ approach. Most of its conditional entropy is zero or near zero, while having position errors of more than 80 m. When the conditional entropy is larger, the position error tends to be under 20 m, which is quite the opposite of what we



	trained σ	fixed σ
Pearson's ρ	−0.12	0.08
p-value	9.29×10^{-5}	0.01

Fig. 6. Relation between position error and conditional entropy. The table shows the Pearson's ρ correlation coefficient of the position error and the conditional entropy. The p-value indicates the probability of the null hypothesis, namely that there is no correlation.

expect. However, this behavior is improved when using the fixed σ approach. While a few outliers exist with a position error of around 60 m, most of the samples with a large position error now have an increased conditional entropy. On the other hand, the conditional entropy has increased for most samples, including those with small position error.

IV. Conclusion

Our hypothesis is that the uncertainty or conditional entropy in a posterior probability distribution of the positions in an environment given a Wi-Fi fingerprint measurement relates to the distance between the true position and the position selected

by the maximum a posteriori approach. This hypothesis is based on the intuition that a very certain distribution should yield a better result than an uncertain distribution; that there is some relationship between uncertainty and position error. This relationship is not visible when using our new probabilistic sensor model in the UJIIndoorLoc Wi-Fi fingerprinting dataset.

We learned that our sensor model usually exhibits a low conditional entropy, which indicates a high certainty in the posterior distribution, while not necessarily indicating a position close to the true position. We see two possible reasons for this discrepancy. First, the training data has been sampled on reference positions that are a few meters apart. The distance between these positions introduces a considerable variation in the Wi-Fi signal. This variation can also be observed in the validation samples, in which it must not necessarily correspond to the variation of the signal at the nearest reference position. We want to better incorporate this spatial variation in our next training procedure. Second, the hypothesis may be incorrect. This must be validated by performing the same experiment in different environments.

Additionally, we studied the difference between using a standard deviation that is trained specifically for each access point at each reference position and using a single, fixed standard deviation for each of those. The difference in position error is relatively small, yet significant according to the t-test. The difference in conditional entropy is considerably larger, which is to be expected because the fixed standard deviation was usually larger than the trained standard deviation. The conditional entropy for both approaches is still small compared to the maximum amount of conditional entropy that is possible in the environment. Neither of the approaches provides much evidence for the hypothesis, as discussed before.

References

[1] U. Schatzberg, Y. Amizur, L. Banin, and J. Segev, "Systems and methods for providing variable position precision," U.S. Patent Application 13/536,398, 2014.

[2] A. Jacobson, Z. Chen, and M. Milford, "Autonomous Multisensor Calibration and Closed-loop Fusion for SLAM," *Journal of Field Robotics*, vol. 32, no. 1, pp. 85–122, jan 2015.

[3] R. Berkvens, M. Weyn, and H. Peremans, "Localization Performance Quantification by Conditional Entropy," in *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*. Banff, AB, Canada: IEEE, 2015, pp. 1–7.

[4] J. Torres-Sospedra, R. Montoliu, A. Martnez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Indoor Positioning and Indoor Navigation (IPIN), 2014 International Conference on*. Busan, Korea: IEEE, 2014.

[5] I. Bisio, F. Lavagetto, M. Marchese, and A. Sciarrone, "Smart probabilistic fingerprinting for WiFi-based indoor positioning with mobile devices," *Pervasive and Mobile Computing*, vol. in press, feb 2016.

[6] Y. Luo, O. Hoeber, and Y. Chen, "Enhancing Wi-Fi fingerprinting for indoor positioning using human-centric collaborative feedback," *Human-centric Computing and Information Sciences*, vol. 3, no. 1, pp. 1–23, 2013.

[7] S. He and S. H. G. Chan, "Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 466–490, 2016.

[8] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill Europe, 2002.

[9] F. Potorti, P. Barsocchi, M. Girolami, J. Torres-Sospedra, and R. Montoliu, "Evaluating indoor localization solutions in large environments through competitive benchmarking: The EvAAL-ETRI competition," in *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*. Banff, AB, Canada: IEEE, oct 2015, pp. 1–10.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY: John Wiley & Sons, 1991.

[11] A. Moreira, M. J. Nicolau, F. Meneses, and A. Costa, "Wi-Fi fingerprinting in the real world - RTLS@UM at the EvAAL competition," in *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*. Banff, AB, Canada: IEEE, 2015, pp. 1–10.