Accurate WiFi-based Indoor Positioning with Continuous Location Sampling

J.E. van Engelen[1], J.J. van Lier[2], F.W. Takes[1], and H. Trautmann[3]

[1] Department of Computer Science (LIACS), Leiden University, The Netherlands
[2] Big Data & Analytics, KPMG Advisory N.V., The Netherlands
[3] Department of Information Systems, University of Münster, Germany

Abstract. The ubiquity of WiFi access points and the sharp increase in WiFi-enabled devices carried by humans have paved the way for WiFi-based indoor positioning and location analysis. Locating people in indoor environments has numerous applications in robotics, crowd control, indoor facility optimization, and automated environment mapping. However, existing WiFi-based positioning systems suffer from two major problems: (1) their accuracy and precision is limited due to inherent noise induced by indoor obstacles, and (2) they only occasionally provide location estimates, namely when a WiFi-equipped device emits a signal. To mitigate these two issues, we propose a novel Gaussian process (GP) model for WiFi signal strength measurements. It allows for simultaneous smoothing (increasing accuracy and precision of estimators) and interpolation (enabling continuous sampling of location estimates). Furthermore, simple and efficient smoothing methods for location estimates are introduced to improve localization performance in real-time settings. Experiments are conducted on two data sets from a large real-world commercial indoor retail environment. Results demonstrate that our approach provides significant improvements in terms of precision and accuracy with respect to unfiltered data. Ultimately, the GP model realizes continuous location sampling with consistently high quality location estimates.

Keywords: indoor positioning, Gaussian processes, crowd flow analysis, machine learning, WiFi

1 Introduction

The increasing popularity of wireless networks (WiFi) has greatly boosted both commercial and academic interest in indoor positioning systems. Requiring no specialized hardware, WiFi-based positioning systems utilize off-the-shelf wireless access points to determine the location of objects and people in indoor environments using wireless signals, emitted by a multitude of electronic devices. This location information finds a broad range of applications, including in robotics, indoor navigation systems, and facility management and planning in retail stores, universities, airports, public buildings, etc. [12, 13].

WiFi-based positioning systems use sensors to capture the signals transmitted by WiFi-equipped devices such as smartphones and laptops. Since the signals are

attenuated (i.e., reduced in strength) as they travel through physical space, a sensor close to a device emitting a signal will measure a higher signal strength than a sensor farther away from the device. By combining measurements of the same signal by different sensors, indoor positioning systems can approximate the signal's origin. Two broad groups of localization techniques exist: fingerprinting-based and model-based methods. The former can be regarded as a form of supervised learning, where a model is trained to predict a device's location based on grouped signal strength measurements by passing it pairs of measurement groups (sample features) and known locations (labels). Consequently, fingerprinting-based methods require *training data* [5, 24]. Model-based methods, on the other hand, require no training data and rely on physical models for the propagation of signals through space to determine the most likely location of the device [20]. They include lateration-based methods, which estimate the distance to multiple sensors based on the received signal strengths and use regression methods to determine the most likely position of the device.

The first concern with all of these approaches is the unpredictability of signal propagation through indoor environments. Static obstacles such as walls, ceilings, and furniture attenuate the transmitted signal and prohibit the construction of an accurate model for the received signal strength. Furthermore, the refraction of signals by these obstacles leads to multiple observations of the same signal by a single sensor, but at different signal strengths (the *multipath* phenomenon, see [16]). Dynamic obstacles such as people further complicate the process, as these cannot be modeled offline. The resulting observation noise results in significant variance of location estimates in all existing approaches [12]. The second problem encountered using these methods is that they can only estimate the location of a device when it transmits a signal. In practice, signals are transmitted infrequently and at irregular intervals, thus only allowing for intermittent positioning.

Both of these problems form major hurdles in the application of WiFi-based location analysis to real-world issues. With intermittent, inaccurate, and imprecise location estimates, any further data mining and analysis becomes exceedingly difficult. In this paper, we propose a novel method to mitigate these issues. We employ a nonparametric approach to approximate sensors' received signal strength distributions over time. For this, we use a Gaussian process (GP) model to obtain an a posterior estimate and associated variance information of signal strengths, which simultaneously allows us to resample from the distribution at any timestamp, and to sample only measurements with sufficiently low variance. The proposed approach can greatly enhance any model based on signal strengths. In addition to the Gaussian process model, we also compare several smoothing methods for location estimates. Ultimately, our main contribution is an accurate model that addresses each of the problems outlined above, reducing localization noise and providing accurate location estimates at arbitrary time intervals.

The rest of the paper is structured as follows. First, related work is outlined in Section 2. Section 3 gives a formal problem statement. Methods for smoothing and interpolation and the proposed Gaussian process model are discussed in Section 4 and 5. Experiments are covered in Section 6. Section 7 concludes.

2 Background and related work

Driven by the proliferation of WiFi-equipped devices, WiFi sensors have become a popular choice for use in indoor positioning systems. Requiring no specialized hardware, the accuracy of these approaches is generally in the range of 2 to 4 meters [12]. This makes them a good candidate for indoor localization purposes in many areas, including, among others, facility planning and indoor navigation. In robotics, localization of robots in indoor spaces forms a key challenge [22].

Both physical models for signal propagation and machine learning methods are broadly applied. The latter include methods based on traditional machine learning techniques such as support vector machines and k-nearest neighbours [12]. Notably, a Gaussian process model for generating a likelihood model of signal strengths from location data is proposed in [5]. More recently, a neural network approach was suggested by Zou et al. [24]. Such fingerprinting-based models require training data to construct a prediction model, which has two major drawbacks [14]. First, training data can be expensive to obtain: to construct a proper localization model, training data consisting of signal strength measurements and known device coordinates is required from locations throughout the input space. Second, when extending a localization system to new environments or with additional sensors, the system needs to be re-trained with additional training data.

Approaches requiring no calibration data instead rely on physical models for the propagation of signals through space, only requiring the locations of the sensors to be known. They include approaches based on the measured angle of the incoming signal (*angle of arrival*) and approaches based on the time a signal takes to reach the sensor (*time of flight*) [8, 15]. Furthermore, distance-based models, localizing devices using a model of the propagation of signals through space, are often used [10, 23]. They exploit the fact that radio signals reduce in strength as they propagate through space, utilizing lateration-based techniques to find the most likely position of the device based on signal strength measurements at multiple sensors. Lastly, approaches exist that require neither training data nor knowledge about the environment (such as sensors locations). These approaches, known as *simultaneous localization and mapping* (SLAM [22]) methods, simultaneously infer environment information and device locations. Applications of this method using WiFi sensors include *WiFi SLAM* [4] and *distributed particle SLAM* [2]. For an extensive review of WiFi-based positioning systems and applications, we refer the reader to [3, 8, 12].

No single performance measure exists for evaluating localization quality, but consensus exists in literature that both accuracy and precision are important. In [12] and [3], different performance criteria are discussed; we will use a subset of these in our work. The main contribution of this paper is the proposition of a lateration-based model using Gaussian processes for signal strength measurements, addressing the most important performance criteria of accuracy, precision, and responsiveness simultaneously. Our approach enables high-quality location estimates at arbitrary timestamps. To the best of our knowledge, this is the first generic model for continuously estimating device locations.

3 Problem statement

Our aim is to construct a (1) precise, (2) accurate, and (3) responsive (i.e., able to provide continuous location estimates) positioning system for indoor environments where relatively few signal observations are present. Furthermore, the method should be robust and cost-effective.

Assume we are provided with a set of n signal strength measurements $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for a single device. Each measurement is generated by a device at some unknown position. The vector $\mathbf{x}_i$ consists of the observation timestamp, an identifier of the sensor receiving the signal, the signal strength, and information to uniquely identify the package transmitted by the device. Let $\mathbf{d}_t$ denote the position of the device at time t. Our objective, then, is to obtain a position estimate $\hat{\mathbf{d}}_t$ for the device at any timestamp t based on the measurements X. We wish to maximize the accuracy, i.e., minimize the distance between the expected estimated location and the actual location, $||\mathbb{E}[\hat{\mathbf{d}}_t] - \mathbf{d}_t||$, where $||\cdot||$ denotes the Euclidean norm. We also wish to maximize the precision, i.e., minimize the expected squared distance between the estimated location and the mean estimated location, $\mathbb{E}\left[||\hat{\mathbf{d}}_t - \mathbb{E}[\hat{\mathbf{d}}_t]||^2\right]$.

Of course, we cannot evaluate the expected accuracy and precision. Instead, we optimize the empirical accuracy and precision. Assume our calibration data consists of observations of a device at c different known locations $(\mathbf{p}_1, \ldots, \mathbf{p}_c)$. Each set C_j then consists of the estimated device positions when the device was at position $\mathbf{p}_j$. Let $\hat{\mathbf{p}}_j$ be the mean of the estimated positions for calibration point j, i.e., $\hat{\mathbf{p}}_j = \frac{1}{|C_j|}\sum_{\hat{\mathbf{d}} \in C_i} \hat{\mathbf{d}}$. The accuracy and precision are then calculated as an average over the accuracy and precision at all calibration points.

The empirical accuracy (Acc) is calculated as the mean localization error,

$$Acc = \frac{1}{c}\sum_{j=1}^{c} ||\hat{\mathbf{p}}_j - \mathbf{p}_j||. \tag{1}$$

The empirical precision ($Prec$) per calibration point is calculated as the mean squared distance between the estimated location and the mean estimated location. This yields, averaged over all calibration points,

$$Prec = \frac{1}{c}\sum_{j=1}^{c}\left(\frac{1}{|C_j|}\sum_{\hat{\mathbf{d}} \in C_j} ||\hat{\mathbf{d}} - \hat{\mathbf{p}}_j||^2\right). \tag{2}$$

4 Lateration-based positioning

WiFi-equipped devices emit radio signals when connected to an access point, and when scanning for a known access point. Sensors can be utilized to listen for these signals, recording the received signal strength. We henceforth refer to this received package and the associated signal strength as a *measurement*. Our

approach combines measurements of a signal by different sensors to determine its spatial origin. It requires no prior knowledge besides the sensor locations.

To localize a device, we need measurements of the same signal from multiple sensors. Each transmitted package contains information, including the package's sequence number and a device identifier, by which we can identify unique packages received by multiple sensors. We henceforth refer to these groups of measurements of the same package by different sensors as *co-occurrences*.

4.1 Localization model

Having identified a co-occurence, we wish to translate this set of the package's signal strength measurements into an estimate of the transmitting device's location. We do so by modeling the propagation of the signal through space, and finding the device coordinates and parameters that minimize the difference between the modeled signal strengths and the observed signal strengths.

To model the propagation of a signal in space from a transmitting antenna to a receiving antenna, we make use of the Friis transmission equation [6]. It defines the relation between the transmitted signal strength P_t (from an antenna with gain G_t), the received signal strength P_r (at an antenna with gain G_r), the wavelength λ of the signal, and the distance R between the antennas. Antenna gain is a measure for the efficiency of the antenna, and is constant for each antenna. Using the dBm unit for the signal strengths, we calculate P_r by

$$P_r = P_t + G_t + G_r + 20 \cdot \log\left(\frac{\lambda}{4\pi R}\right), \tag{3}$$

where $\log(\cdot)$ denotes the logarithm with base 10. Equation 3 is premised on the assumption that the path between the two antennas is unobstructed (the *free space* assumption). This assumption is captured in the last term of the equation, which is the logarithm of the inverse of what is known as the free-space path loss $(\frac{4\pi R}{\lambda})^2$ (also known as the free-space loss factor [1]). However, in most real-world indoor environments, we cannot assume free space. To combat this problem, we make use of an empirically derived formula for modeling propagation loss from [9]. It introduces the *path loss exponent* n as the exponent in the fraction of the path loss equation. For free-space environments, $n = 2$, yielding the original transmission equation. For environments with obstructions, generally $n > 2$ [18]. We introduce a method for estimating n in Section 4.2. We note that, by using a physical model for estimating device locations based on measurements, our lateration-based method does not require an expensive training phase. Further motivation for using lateration-based methods over fingerprinting-based methods can be found in the fact that, unlike in most machine learning problems, the true model generating our observations is known (up to corrections for the multipath phenomenon). This valuable prior knowledge is discarded in most fingerprinting-based methods.

Given the signal strength model, we formulate the localization problem as a regression problem, minimizing the sum of squared differences between the

measured signal strength, and the signal strength obtained when calculating P_r from the Friis transmission equation using our estimated location parameters. Assuming i.i.d. measurements with additive Gaussian noise, this corresponds to the maximum likelihood estimator. We define the model for the transmission strength based on the Friis transmission equation from Equation 3 and incorporate sensor i's path loss exponent n_i, rewriting it for a single sensor as:

$$\begin{aligned} P_r &= P_t + G_t + G_r + n_i \cdot 10 \cdot \log\left(\frac{\lambda}{4\pi R_i}\right) \\ &= P_t + G_t + G_r + n_i \cdot 10 \cdot \log\left(\frac{\lambda}{4\pi}\right) - n_i \cdot 10 \cdot \log R_i \\ &= \rho - n_i \cdot 10 \cdot \log R_i, \end{aligned}$$

where $\rho = P_t + G_t + G_r + n_i \cdot 10 \cdot \log\left(\frac{\lambda}{4\pi}\right)$ is the bias term to be estimated, and the path loss exponent n_i is assumed known and constant. The model corresponds to the transmission strength model defined in [9]. The resulting system of equations is underdetermined in the general case where G_r is dependent on the sensor. Thus, we assume that G_r is constant across all sensors, making the system overdetermined. In our case, as all WiFi sensors are of the same type, this is reasonable. Expressing R_i in terms of the sensor location vector $\mathbf{s}^{(i)} = (s_x^{(i)}, s_y^{(i)})^T$ and the device location estimate vector $\mathbf{d} = (d_x, d_y)^T$, we obtain our model:

$$f_i(\theta) \equiv f_i(\theta | \mathbf{s}^{(i)}, n_i) = \rho - n_i \cdot 10 \cdot \log ||\mathbf{s}^{(i)} - \mathbf{d}||, \tag{4}$$

where $\theta \equiv (\rho, d_x, d_y)^T$ are the parameters to be estimated. We are now in place to define our loss function J:

$$J(\theta) \equiv J(\theta | \mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}, \mathbf{n}) = \sum_{i=1}^{N} (P_r^{(i)} - f_i(\theta))^2, \tag{5}$$

where $P_r^{(i)}$ is the measured signal strength at sensor i. We wish to minimize this function, i.e., we want to find $\hat{\theta} \in \arg\min_\theta J(\theta)$. Since $f_i(\theta)$ is nonlinear in its parameters θ, we cannot find the solution analytically. Therefore, we make use of Newton's method, which iteratively minimizes the loss function using the update rule $\theta_{t+1} = \theta_t - \frac{f'(\theta_t)}{f''(\theta_t)}$ for determining the set of parameters θ at iteration $(t+1)$. The initial state can be chosen in several ways, e.g., by taking the weighted mean position of all sensors that received the signal.

The positioning procedure described thus far localizes each co-occurrence independently, without taking into account previous or future estimated device positions. In other words, it assumes the locations over time for a single device are independent. Furthermore, it is premised on the assumption that the signal propagates through free space, which does not generally hold. In the remainder of this section, we propose several methods to overcome these shortcomings and improve the quality of fits in the sense of the outlined performance criteria.

4.2 Estimating the path loss exponent

The general Friis transmission equation is premised on the assumption that the radio signal propagates through free space, which is generally not the case in indoor environments. To combat this, [9] proposes an empirical adjustment to the model, introducing path loss exponent n in the Friis transmission equation, where n grows as the free-space assumption is relaxed.

Considering Equation 4, we see that the received signal strength can be rewritten as $P_r = \rho - nx$, where $x = 10 \cdot \log(R)$. As all parameters in ρ are assumed to be constant with respect to R, P_r is linear in $\log(R)$. Now, using calibration data for which R, the distance between device and sensor, and P_r, the received signal strength at the sensor, are known, we can apply linear regression to estimate ρ and n. Having estimated path loss exponent n, we can use it in our model to account for attenuation effects induced by the surroundings.

4.3 Smoothing and filtering fits

Another improvement on fit quality, in the sense of in particular precision, but also accuracy, can be achieved by exploiting the knowledge that, during short periods of time, the position of a device is not expected to change significantly. This opens up the possibility of simply smoothing the estimated locations through time. Here, we outline the most common smoothing and filtering methods, which can be applied to the estimated x- and y-coordinates individually.

First, we consider using the *exponential moving average* (EMA) [7]. Due to its $O(1)$ complexity in smoothing a single data point, and because it only depends on previous observations, it is a good candidate for systems requiring large-scale and real-time localization. In its simplest form, the EMA assumes evenly spaced events, and calculates the smoothed value x'_i at step i as a weighted average of x'_{i-1}, the smoothed value at step $i-1$, and x_i, the unsmoothed input value. Using α to control the amount of smoothing, we obtain $x'_i = x_i \cdot \alpha + x'_{i-1} \cdot (1-\alpha)$.

Second, we consider *Gaussian smoothing* [21], which can be seen as the convolution of the signal with a Gaussian kernel. Like EMA, the filter generally assumes evenly spaced observations. However, by adjusting the weighting of each sample based on its observation timestamp, we can apply it to non-evenly-spaced observations as well. In discrete space, we calculate the smoothed value x'_i as the weighted average of all observations, where the weight of x_j is dependent on the time difference between observations j and i. Denoting n as the number of observations and t_i as the observation timestamp of sample i, we write

$$x'_i = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j} \quad \text{, where} \quad w_j = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_j - t_i)^2\right). \tag{6}$$

Thus, the filter simply smooths the locations' x- and y-coordinates. Theoretically, the filter has time complexity $O(n)$ for a single observation, but this can be reduced to $O(1)$ for observations sorted by timestamp. As a large part of a Gaussian's density is concentrated close to its mean, we can simply use a small central section of the Gaussian without expecting significant changes in results.

Third, we consider a more sophisticated smoothing approach: the *Savitzky-Golay* filter [19], which smooths values by fitting a low-degree polynomial (centered around the observation to be smoothed) to the observed values in their vicinity. It then predicts the smoothed value at time t by evaluating the fitted polynomial function at its center. This corresponds to the bias term in the fitted polynomial. Each observation is smoothed independently, and makes use of the observations within some pre-specified window around the observation to be smoothed. For evenly spaced observations, an analytical solution exists; numerical methods are required for our non-evenly-spaced observations.

5 Gaussian processes for measurement resampling

Smoothing and filtering approaches address the first of the two most significant problems of WiFi-based indoor localization: they improve accuracy and precision. However, they do not tackle the second problem, concerning the scarcity of measurements. We introduce a method to address both of the issues simultaneously, allowing arbitrary resampling of measurements while limiting variance.

Our method generates a model for the signal strengths measured by the sensors for a single device, and then resamples signals from this model. The model is constructed by means of a Gaussian process, making use of the fact that the signal of a device as measured by a sensor is expected to vary only slightly over small time intervals. Resampling facilitates the construction of new measurements at arbitrary timestamps, and reduces variance in signal strengths at the same time, by interpolating between signals received from the device around the requested timestamp. Before continuing to the implementation of this method, we provide a brief explanation of Gaussian processes; for details on prediction with Gaussian processes, we refer the reader to [17].

5.1 Gaussian processes

Assume that we have a data set consisting of n observations, and that each observation is in $\mathbb{R}^d$. We denote $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ denote the feature vector and the target value, respectively, for the ith data point. The observations are drawn i.i.d. from some unknown distribution specified by

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{7}$$

where ϵ_i is a Gaussian distributed noise variable with 0 mean and variance σ_i^2.

Since we want to predict a target value for previously unseen inputs, our objective is to find a function $\hat{f}$ that models f. A Gaussian process estimates the posterior distribution over f based on the data. A model is sought that finds a compromise between fitting the input data well, and adhering to some prior preference about the shape of the model. This prior preference touches on a fundamental concept in Gaussian processes: they are based on the assumption that some similarity measure between two inputs exists that defines the correlation

between their target values based on their input values. More formally, it requires a kernel function $k(\mathbf{x}, \mathbf{x}')$ to be specified that defines the correlation between any pair of inputs $\mathbf{x}$ and $\mathbf{x}'$. Furthermore, a Gaussian process requires a prior mean μ (usually 0) and variance σ^2 (chosen to reflect the uncertainty in the observations).

5.2 A Gaussian process model for signal strength measurements

We propose a method for modeling the signal strength over time for a single device and a single sensor. We assume that each measurement consists of at least a timestamp, the measured signal strength, and information to uniquely identify the co-occurence (see Section 4), the device, and the sensor. In our model, we instantiate a single Gaussian process with one-dimensional feature vectors for each sensor to model the development of the signal over time. We will later elaborate on possibilities to extend this model to include additional information.

For our GP models, we propose using the exponential kernel or a similar kernel that exploits the fact that signal strengths vary relatively little over time. The measurements variance σ^2 should be chosen based on prior knowledge from the calibration data, and should reflect the observed variance in signal strength measurements. An important aspect of the GP is that it includes approximations of the variance on the posterior distribution. Consequently, we can reason about the certainty of our prediction at a specific timestamp. We exploit this knowledge in sampling from our signal strength distributions, discarding timestamps where the variance exceeds a predetermined upper bound. By sampling measurements for multiple sensors at the same timestamp, we can generate new co-occurences at arbitrary timestamps, provided that the variance is below our prespecified threshold. The latter reveals the true power of the GP model: resampling means that we are much less affected by the intermittence in the transmittal of signals by devices. Furthermore, we have estimates of the quality of the measurements, and we can vary in the trade-off between the number of measurements to generate, and their quality. We note that this capability is not natively present in existing techniques, even in those that are based on Gaussian processes: their models output device locations independently based on given measurements.

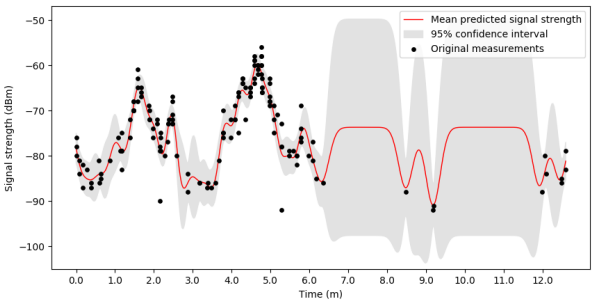


Fig. 1: Gaussian process model applied to a single sensor's signal measurements.

Figure 1 displays an example, where we apply the Gaussian process model to measurements of a device from a single sensor. We use data of a person carrying a device along a predefined path in a store over a time period of 15 minutes. We fit a Gaussian process model with an exponential kernel to these measurements, optimizing the hyperparameters ℓ and σ^2 using the Limited-memory BFGS algorithm [11]. The resulting mean estimate and 95% confidence interval are plotted in Figure 1. As can be seen from this image, the posterior variance decreases in regions where many observations are present. This makes sense, considering we assume additive Gaussian noise: when multiple measurements are made in a small time range, we can more accurately estimate the target value for that time. Resampling the mean, i.e., maximum-likelihood prediction of the obtained distribution, we can generate arbitrarily many new measurements. Using a fixed upper bound on the variance at which to resample points allows us to quantify the quality of the obtained measurements.

An additional advantage of Gaussian processes is that we can incorporate additional information about the measurements in our kernel function k. For instance, we can amend the feature vectors to include the sensors' x- and y-coordinates, and make use of the fact that nearby sensors should obtain similar signal strength measurements for the same device at nearby timestamps.

6 Experiments and results

We are now ready to evaluate the performance of our localization algorithm and the various suggested improvements. Our two experimental setups, data sets, and methods of comparison to baseline approaches are described in Section 6.1. We first optimize the path loss exponent hyperparameter in Section 6.2. Based on these experiments, we evaluate the smoothing approaches and our proposed Gaussian process model in Section 6.3 and Section 6.4, respectively.

6.1 Experimental setup

Our experiments were conducted in an 84×110 meter indoor retail environment. A total of 85 sensors were positioned throughout a single floor of the store at irregular locations, at around 3 meters above ground level. The sensors, off-the-shelf WiFi access points, report integer dBm signal strengths. A multitude of obstacles were present, including furniture and walls.

The Gaussian process model generates new measurements by sampling from the modeled signal strength distributions for all sensors at regular intervals, discarding measurements with variance exceeding a certain threshold. We set this threshold such that it is just below the observation noise passed to the model. This way, we essentially require a significant contribution from multiple nearby measurements to generate a measurement at a specific timestamp: if only a single measurement is near the input timestamp, the estimated variance will be close to the observation noise, and thereby above our variance threshold. However, it is possible to adjust this threshold to balance measurement quality and quantity.

Table 1: Properties of fixed calibration and visitor path data sets.

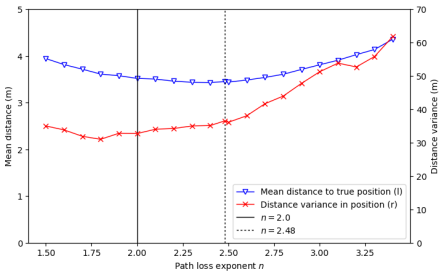
Data set	# Measurements	# Fits	Time span
Fixed	107 751	3170	90.3 min.
Visitor Path	202 528	9004	248.8 min.

Two data sets, listed in Table 1, were used. The *fixed calibration data set* was constructed by positioning a device at several fixed, known locations throughout the store, and emitting packages for a known amount of time. The path travelled was approximately 500 meters long. A person carrying the device moved at roughly $1.4\,\mathrm{m\,s^{-1}}$, and traversed the path a total of 22 times. Close to 6,000 packages were received from this device during the evaluated time period. The positioning system's accuracy and precision can be evaluated by comparing the estimated locations to the real locations. This corresponds with most well-known experimental setups from literature [12]. The second *visitor path data set* was constructed by moving a device along the same known path through the environment several times in a row. The path travelled by the device is unique and known, but the associated timestamps (where the device was at what time) are not. Thus, we cannot evaluate accuracy and precision in the same way as with fixed calibration data. Instead, we rely on the shortest distance to the path, i.e., the length of the shortest line segment from the estimated location to the known path, referred to as *distance of closest approach*, or DOCA.

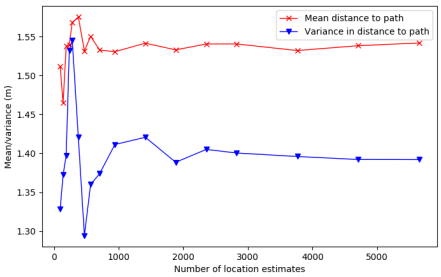
As data sets used in indoor localization and implementations of indoor localization methods are generally not open-source, we compare our methods to a baseline, namely the lateration-based model applied to the raw co-occurrences of measurements. The performance of this baseline corresponds with empirical results from other studies using similar methods [3].

6.2 Results: Path loss exponent

We estimate the path loss exponent (see Section 4.2) using the fixed calibration data. The true distance between device and sensor is combined with the received signal strength in a linear regression model. We apply the approach to estimate the path loss exponent globally, using 65 000 measurements with 81 unique sensors. The regression yields a path loss exponent of approximately 2.48 with an R^2 coefficient of 0.38. We validate this result by comparing predictor accuracy and precision using different path loss exponents. Figure 2a shows the mean of the distances to the calibration points, and the distance variance, for different path loss exponent values. Path loss exponent 2.0 (solid, black line) corresponds to the original localization model (which assumes free space), whereas the dashed black line corresponds to the path loss exponent calculated based on the linear regression model. The figure shows that the accuracy of the calculated exponent (2.48) is very close to that of the empirical optimum; the variance, however, is slightly higher. Precision is relatively constant for $n \in [2.0, 2.5]$, and accuracy only marginally improves with an adjusted path loss exponent.



(a) Results of path loss exponent experiments.



(b) DOCA and DOCA variance for varying sample counts in GP model.

Fig. 2: Result diagrams for experiments in Section 6.2 and Section 6.4.

6.3 Results: Smoothing and filtering

Our smoothing and filtering approaches, introduced in Section 4.3, attempt to improve localization accuracy and precision by averaging the estimated x- and y-coordinates over small time periods. All experiments were conducted using the path loss exponent estimated in Section 6.2. Based on empirical experiments with multiple hyperparameter combinations, hyperparameters of the smoothing and filtering algorithms were set:

- **Exponential Moving Average**: a constant smoothing factor α was used, with $\alpha = 0.5$. Experiments were conducted with $\alpha \in [0.1, 0.9]$.
- **Gaussian smoothing**: a Gaussian with standard deviation $\sigma = 5000$ milliseconds was used. We experimented with $\sigma \in [1000, 20000]$.
- **Savitzky-Golay filtering**: a third-degree polynomial was fitted to a window of 50 seconds centered around the data point to be smoothed. Experiments with polynomials of degree 2 and 4 demonstrated inferior performance.

Results are shown in Table 2, demonstrating how, on the fixed position calibration data, all smoothing methods vastly outperformed the baseline. The average distance between the mean estimated location and the true location was reduced by 37%. Gaussian smoothing outperformed both EMA and Savitzky-Golay filtering, but the differences in accuracy were minimal. All smoothing methods show a significant improvement in precision (variance was reduced by 75% for Gaussian smoothing), which is expected as they generally shift estimates towards the mean.

For the visitor path data set, we note from Table 2 that the DOCA (see Section 6.1) forms a lower bound on the distance between the estimated location and the actual location. This results from the fact that each actual location is on the path, but it is unclear where on the path the device was located at a given timestamp. Especially for methods with a higher deviation from the actual

Table 2: Smoothing and filtering results on fixed and visitor path data set. Results are in meters.

Method	**Fixed data**		**Visitor Path data**	
	Accuracy	Precision	DOCA	Variance
Unfiltered	3.44	6.01	2.08	5.65
EMA	2.26	1.75	1.90	1.81
Savitzky-Golay	2.29	2.12	1.66	1.44
Gaussian smoothing	2.17	1.49	1.42	1.18

location, this improves observed performance, as a significantly misestimated location can still have a small DOCA when it is close to another path segment.

In general, the visitor path test results are in conformity with the fixed calibration test results. Every smoothing and filtering approach substantially improves on the baseline accuracy, and Gaussian smoothing outperforms the other two smoothing approaches. A decrease of 32% in mean DOCA relative to the baseline method was attained, and variance was significantly reduced. To visualize the effect smoothing has on our original location estimates, we include a visualization of the location estimates of a single traversal of the path. The actual path, the originally estimated path, and the smoothed path (using Gaussian smoothing) are depicted in Figure 3a. This visualization shows the noisy nature of the original estimates, and the extent to which smoothing ensures that the estimated path corresponds with the actual path followed.

6.4 Results: Gaussian process measurement resampling

Lastly, we consider the Gaussian process measurement resampling model introduced in Section 5, operating on the raw measurements obtained by the sensors. Because it outputs location estimates based on newly generated measurements sampled from this model, both the number of location estimates and their timestamps differ between the Gaussian process model and the smoothing approaches. To still allow for a comparison between the results of the other approaches and the measurement resampling model, we choose the sampling interval (the time between potential measurements to be sampled) such that the number of resulting location estimates roughly corresponds to the number of location estimates for the raw measurements. The results of the Gaussian process model on the fixed and visitor path data sets are listed in comparison to the original model and the Gaussian smoothing model in Table 3. The number of location estimates is listed alongside the performance metrics.

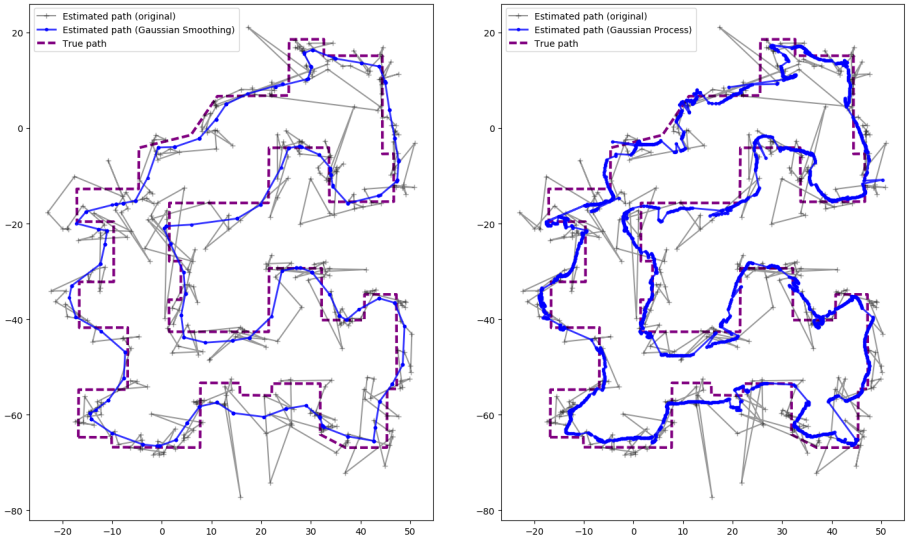
The model is able to greatly improve the quality of location estimates when compared to the unfiltered estimates: on the fixed calibration data, accuracy improved from $3.44m$ to $2.16m$, and precision improved from $6.01m$ to $1.76m$. On the visitor path data, DOCA decreased from $2.08m$ to $1.54m$, and the variance in DOCA decreased from $5.65m$ to $1.50m$. As such, the observed performance of

Table 3: Gaussian process measurement resampling results on fixed and visitor path data set. Results are in meters.

Method	**Fixed data**			**Path data**		
	# Fits	Accuracy	Precision	# Fits	DOCA	Variance
Unfiltered	3170	3.44	6.01	9004	2.08	5.65
Gaussian smoothing	3170	2.17	1.49	9004	1.42	1.18
Gaussian process	3463	2.16	1.76	9095	1.54	1.50

the Gaussian process measurements resampling model is generally similar to the performance measured using the top-performing smoothing method (Gaussian smoothing). The similarity in performance of these two different approaches is remarkable, as the smoothing model operates on the location estimates, whereas the Gaussian process model operates on the signal strength measurements. As can be expected, the Gaussian process model achieves less variance reduction than the Gaussian smoothing approach. This can be explained by the fact that the Gaussian smoothing approach operates on the location estimates, thereby directly impacting the precision performance criterion.

In general, our results show that the Gaussian process model is able to significantly improve localization *accuracy* and *precision*. However, its true advantage surfaces when considering the *responsiveness* performance criterion: the Gaussian process model allows for arbitrary resampling, meaning that we can generate



(a) Gaussian smoothing. (b) Gaussian process.

Fig. 3: Result diagrams for experiments in Section 6.

arbitrarily many measurements. We investigate this property further by evaluating the effect the number of potential sampling points has on the accuracy and precision of the location estimates. To this end, we sample at different numbers of equally spaced intervals over a time period of approximately 45 minutes, during which the path was traversed 5 times. In Figure 2b, the tradeoff between the number of resulting location estimates and accuracy and precision is depicted. The accuracy remains constant, and the precision converges as the number of samples increases, substantiating the claim that the number of location estimates can be increased arbitrarily, without significantly impacting performance.

A visualization of the path estimated by the GP model when using an exceedingly large number of location estimates is provided in Figure 3b. Here, the baseline model from Figure 3a is compared with a Gaussian process model where approximately 3 times as many points were sampled. The figure highlights the ability of the Gaussian process to provide location estimates at points in time where, previously, no location estimates were possible.

7 Conclusion and future work

In this paper, we have presented a novel method for continuously estimating device positions in indoor environments based on intermittent WiFi signals. Using a Gaussian process model for signal strength measurements at WiFi sensors, we realized continuous location estimation, while also significantly improving estimation accuracy and precision. Moreover, we have investigated several smoothing approaches for indoor positioning systems, which improve accuracy and precision to a similar degree, bypassing the computational costs of Gaussian processes. On our validation set of known, fixed device positions, our algorithms improved localization accuracy from $3.44m$ to $2.17m$ and precision from $6.01m$ to $1.49m$. Performance on the visitor path data set of movements along a known path also significantly improved: the mean distance to the true path was reduced from $2.08m$ to $1.42m$, and the distance variance was reduced from $5.65m$ to $1.18m$. These results accomplish the goals set out in Section 1: an accurate, precise location estimator that is able to sample locations at arbitrary timestamps.

Further research opportunities and novel applications are manifold. The results pave the way for significant improvements in dwell-time analysis, visitor tracking, and other major application areas. In future work we aim to derive movement patterns of visitors through time. Such approaches could also be used to automatically infer layouts of indoor environments, identifying obstacles, paths, and open space based on the shape of the estimated movement distributions.

Acknowledgements Authors acknowledge support from the European Research Center for Information Systems (ERCIS). The third author was supported by the European Research Council (ERC), EU Horizon 2020 grant agreement number 638946. Authors thank people who volunteered to generate calibration data sets.

References

1. Balanis, C.A.: Antenna theory: analysis and design. John Wiley & Sons (2016)
2. Faragher, R., Sarno, C., Newman, M.: Opportunistic radio SLAM for indoor navigation using smartphone sensors. In: IEEE PLANS. pp. 120–128 (2012)
3. Farid, Z., Nordin, R., Ismail, M.: Recent advances in wireless indoor localization techniques and system. JCNC **13**, 1–12 (2013)
4. Ferris, B., Fox, D., Lawrence, N.D.: WiFi-SLAM using Gaussian process Latent Variable Models. In: IJCAI. pp. 2480–2485 (2007)
5. Ferris, B., Hähnel, D., Fox, D.: Gaussian processes for signal strength-based location estimation. In: Robotics: Science and Systems. vol. 2, pp. 303–310 (2006)
6. Friis, H.T.: A note on a simple transmission formula. Proceedings of the IRE **34**(5), 254–256 (1946)
7. Gardner, E.S.: Exponential smoothing: The state of the artpart ii. International Journal of Forecasting **22**(4), 637–666 (2006)
8. Gu, Y., Lo, A., Niemegeers, I.: A survey of indoor positioning systems for wireless personal networks. IEEE Communications surveys & tutorials **11**(1), 13–32 (2009)
9. Hata, M.: Empirical formula for propagation loss in land mobile radio services. IEEE Transactions on Vehicular Technology **29**(3), 317–325 (1980)
10. Langendoen, K., Reijers, N.: Distributed localization in wireless sensor networks: a quantitative comparison. Computer Networks **43**(4), 499–518 (2003)
11. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical programming **45**(1), 503–528 (1989)
12. Liu, H., Darabi, H., Banerjee, P., Liu, J.: 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)
13. Lymberopoulos, D., Liu, J., Yang, X., Choudhury, R.R., Sen, S., Handziski, V.: Microsoft indoor localization competition: Experiences and lessons learned. GetMobile: Mobile Computing and Communications **18**(4), 24–31 (2015)
14. Madigan, D., Einahrawy, E., Martin, R.P., Ju, W.H., Krishnan, P., Krishnakumar, A.: Bayesian indoor positioning systems. In: IEEE INFOCOM. vol. 2, pp. 1217–1227 (2005)
15. Niculescu, D., Nath, B.: Ad hoc positioning system (APS) using AOA. In: IEEE INFOCOM. vol. 3, pp. 1734–1743 (2003)
16. Pahlavan, K., Li, X., Makela, J.P.: Indoor geolocation science and technology. IEEE Communications Magazine **40**(2), 112–118 (2002)
17. Rasmussen, C.E.: Gaussian processes for machine learning. The MIT Press (2006)
18. Sarkar, T.K., Ji, Z., Kim, K., Medouri, A., Salazar-Palma, M.: A survey of various propagation models for mobile communication. IEEE Antennas and propagation Magazine **45**(3), 51–82 (2003)
19. Savitzky, A., Golay, M.J.: Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry **36**(8), 1627–1639 (1964)
20. Seco, F., Jiménez, A.R., Prieto, C., Roa, J., Koutsou, K.: A survey of mathematical methods for indoor localization. In: IEEE WISP. pp. 9–14. IEEE (2009)
21. Smith, S.W.: The Scientist and Engineer's Guide to Digital Signal Processing. California Technical Publishing (1997)
22. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics. The MIT Press (2005)
23. Yang, J., Chen, Y.: Indoor localization using improved rss-based lateration methods. In: IEEE GLOBECOM. pp. 1–6 (2009)
24. Zou, H., Jiang, H., Lu, X., Xie, L.: An online sequential extreme learning machine approach to wifi based indoor positioning. In: IEEE WF-IoT. pp. 111–116 (2014)