Toward Regression-based Estimation of Localization Errors in Fingerprinting-based Localization

Filip Lemic*, Vlado Handziski†, Jeroen Famaey*
*Internet Technology and Data Science Lab (IDLab), Universiteit Antwerpen - imec, Belgium
†Telecommunication Networks Group (TKN), Technische Universität Berlin, Germany
Email: {filip.lemic, jeroen.famaey}@uantwerpen.be, handziski@tkn.tu-berlin.de

***Abstract*—Location information is a valuable source of context that can be utilized by end-user applications and wireless networks to optimize their performance and usability. When used, location information should ideally be considered jointly with the estimate of its accuracy. Most of the current approaches for estimating the accuracy rely on performing a static performance benchmark of a localization solution in a deployment environment, which fails to capture the dynamic nature of the environment. We address this problem for fingerprinting-based localization by grounding the estimation of localization errors on the low-level features, i.e. the RSS values from APs used in fingerprinting. We use these low-level features measured at different locations in an environment, as well as their respective localization errors, to train different regression models, allowing us to predict the localization errors at new locations, given new observed values of the low-level features at these locations. Our evaluation results show substantially better performance of the proposed regression-based estimation of localization errors compared to static performance benchmarks.**

I. Introduction

Location information of mobile devices is a valuable source of context information in wireless networks. As such, it has a potential to be used by the end-user applications for providing context-aware services, as well as by wireless networks for optimizing their performance. In practice, localization services feature a certain level of localization errors. These errors should be accounted for when leveraging location information, which is currently often not the case, as discussed in [1].

Nevertheless, there are some examples of leveraging practically obtainable (i.e. erroneous) location information. An obvious one is in end-user navigation systems (e.g. Google Maps), where, in addition to the location information of the user, a confidence interval around the location label is usually shown. Some examples also exist in the domain of context-aware communication. In [2], the authors propose a mechanism for location-based selection of mobile relays in wireless networks, where their selection algorithm takes into account the erroneous nature of location information. Moreover, in [3] the authors propose a location-based mechanism for Device-to-Device (D2D) link establishment. The mechanism, in addition to the devices' location information *per-se*, considers the quality of the provided information.

The current approaches for estimating localization errors rely on static performance benchmarks that are performed upon deploying a localization solution in a given environment [4]. These benchmarks typically provide some aggregate statistical metric (e.g. mean value, Cumulative Distribution Function (CDF)) for characterizing the localization accuracy for the environment [5]. Such spatially aggregated metrics do not account for the fact that the localization errors can vary substantially in different regions of the environment. For example, it has been shown in [6] that the localization errors for a number of solutions are considerably larger at the edges of an environment, in comparison to its center. This issue can be solved by generating a more rich static performance benchmark that captures localization errors for different regions of an environment. Even then, this approach does not account for dynamic changes in an environment (e.g. failures of anchor nodes, interference). Moreover, this approach requires a two-step process of generating location estimates and then assessing their quality by comparing them with ground truth information. In case the expected errors of the estimated location information are too high for a given use-case, the location information would be deemed useless and there would be no need for estimating it in the first place.

These issues have been sparsely discussed in the literature. One example is [7], where the authors hypothesize that the localization errors can be estimated based on the entropy of low-level features used for estimating location information, only to conclude that there is no significant correlation between the entropy and the localization errors.

To address these issues, there is a need for dynamic estimation of localization errors based only on the low-level features used for generating location information. If such an estimation would suggest acceptable error for a given use-case, the localization solution would be requested to generate and provide a location estimate. By leveraging the generated location estimate, the estimate of localization errors could then potentially be improved. Although we believe that a dynamic estimation of errors is needed for localization solutions in general, in this work we constrain ourselves to fingerprinting, one of the most promising and widely utilized localization solutions for office-like indoor environments. We train different off-the-shelf regression algorithms with Received Signal Strength (RSS) values from different Access Points (APs) used in fingerprinting, as well as with the observed localization errors in case location information is estimated using these RSS values. Using the trained models, we are able to predict the localization errors at new locations based solely on the observed RSS values at these locations. We also show consider the usage of the estimated location information as an input feature for regression.

By using WiFi as an example technology, we demonstrate the feasibility of regression-based estimation of localization errors. Specifically, our results show that regression-based estimation using only RSS values yields roughly 25% more accurate estimation of localization errors compared to static benchmarks characterized by the average localization error. Moreover, we demonstrate the consistency of our observations across a variety of environmental conditions and parameterizations of a representative fingerprinting solution. The observed improvement increases to roughly 40% in case the estimated location information is also used as an input for regression.

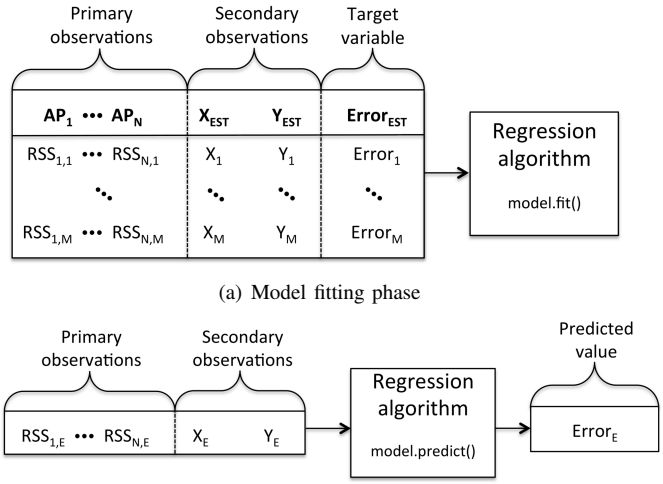
II. Regression-based Estimation of Localization Errors in Fingerprinting

Let us assume that N APs are used for fingerprinting in a deployment environment. Moreover, let us assume the availability of a static performance benchmark of a fingerprinting solution in the environment, where the benchmark provides a mapping between an RSS observation and the localization error of the solution for that observation. The localization error is specified as the Euclidean distance between the true location where the RSS observation is measured and the estimate of location information provided by the fingerprinting solution. Specifically, let us assume the availability of a set of M RSS observations from all APs, i.e. $[RSS_{1,i}, ..., RSS_{N,i}]$, $i = 1, ..., M$, where each set of observations maps to a certain localization error $Error_i$. In case of a missing RSS value in an observation, we substitute the missing value with the noise-floor figure, which increases the amount of information that can be used and, hence, improves the performance of regression-based estimation (as discussed in Section IV).

Regression is a predictive modeling technique based on a relationship between a target variable and independent variables (i.e. observations). The general idea of regression is to fit a curve to the data in a way that minimizes the differences between the distances of the data points and the curve. We call the vectors of RSS values the primary observations, while the resulting localization errors are considered as target variables for the fitting procedure for the regression algorithms, as depicted in Figure 1(a). Under the assumption that the estimates of location information are also available, one can also use this information as an observation in the fitting procedure of a regression algorithm. We call the estimated location information the secondary observation. We consider location estimates in a 2-dimensional (2D) plane and, therefore, we label them as (X_i, Y_i). Extending the problem to a 3-dimensional (3D) plane is straightforward.

The fitting procedure of a regression algorithm yields optimal parameterizations of the algorithm for the provided training data. The trained regression model can then be used for estimating localization errors (i.e. the predicted value) based on either only primary observations or on both primary and secondary observations, as depicted in Figure 1(b).

In this work, we consider a number of well-known regression algorithms, with details provided in e.g. in [8], [9]. In particular, we consider ordinary least squares ("OLS"), ridge ("Ridge"), lasso ("Lasso"), elasticNet ("Elastic"), polynomial ("Poly"), k-nearest neighbors ("kNN"), support vector ("SVR"), and random forest ("RF") regression algorithms.



(a) Model fitting phase

(b) Prediction phase

Figure 1: Regression-based estimation of expected localization errors

III. Evaluation Methodology

We approach the evaluation of the proposed procedure for the estimation of localization errors in fingerprinting through simulation and using WiFi as an example technology. The aim of the evaluation is twofold. First, we aim at demonstrating the feasibility of regression-based estimation of localization errors in fingerprinting. We do that by showing that regression-based estimation outperforms the estimation of localization errors based on static performance benchmarks. Second, for the two regression algorithms that perform best in the initial scenario, we aim at demonstrating their consistently better performance than the static performance benchmarks across a variety of fingerprinting-relevant parameterization scenarios.

The vector of RSS values observed from different WiFi APs in the simulation environment is selected as a fingerprint of a location, which is a well-known fingerprint creation procedure [10]. For calculating the similarity between a training and runtime fingerprint we use the Euclidean distance between RSS vectors, which is again a well-established and extensively used procedure [10]. In the post-processing procedure of fingerprinting, we use k-Nearest Neighbors (kNN) with parameter k set to 4, which has been shown to be optimal for the environment used in the simulation [10].

In the simulation environment, we specify the locations and transmit powers of APs. RSS values obtained at each location are modeled using the COST 231 multi-wall model for indoor radio propagation [11]. The applicability of the model for WiFi fingerprinting has been demonstrated [12] and the model has been extensively used for simulating the behavior of fingerprinting solutions (e.g. [13], [14]). In the model, the first attenuation contribution is a one-slope term relating the RSS to the distance between an AP and the receiver. This term is characterized by the constant l_0, which is the path-loss at 1 m distance from the AP at the center frequency of 2.45 GHz, and

the path-loss exponent γ. The second attenuation contribution is a linear wall attenuation term, where the number of walls in the direct path between the AP and the receiver is counted and certain attenuation is assumed for each of them. The model outputs RSS values from the defined APs at a location of the receiver. A noise is then added to the RSS values, where the noise is modeled using a Gaussian distribution $\mathcal{N}(0, \sigma)$. Gaussian noise is frequently used to account for different variations caused by e.g. interference or quantization [14].

For the simulation environment, the TWIST testbed is selected [15], [16]. The TWIST testbed environment is an office building, with its outline as given in Figure 2. In the parameterization of the simulation model, measurements from the testbed were used in the least-square fitting procedure for minimizing the cost function between the measured RSS values and the modeled ones. The input parameters of the model are the constant l_c related to the least-square fitting procedure, the path-loss exponent γ, and the wall attenuation factor l_w. Additionally, a zero-mean Gaussian noise with standard deviation σ has been added to the modeled RSS values. If not explicitly stated otherwise, the parameters derived and used in the simulation are as $lc = 53.73$ dBm, $\gamma = 1.64$, $l_w = 4.51$ dBm, and $\sigma = 1$ dBm. The transmit power of each AP is set to 20 dBm. For most of our results, we defined a set of 4 APs, with their locations as indicated in Figure 2 (AP1, AP2, AP3, AP4). The receiver's true location has been selected randomly, followed by estimating its location using the selected fingerprinting solution. The procedure has been repeated 5000 times for generating the data points for the evaluation. The metric used for the evaluation is the "prediction error" $[m]$, defined as the absolute difference between the calculated and the estimated localization error. The results have been reported using regular box-plots.

We have divided our data points in a training and evaluation sets in the ratio of 80:20. For the hyper-parameter tuning we used a grid-search procedure on a training set. Based on cross-validation, this procedure yielded the close-to-optimal hyper-parameters for each regression algorithm. Intuitively, the optimal hyper-parameters will differ for other deployment environments. In addition, the main goal of our evaluation is to demonstrate the feasibility of regression-based estimation of localization errors, not necessarily its optimal performance. For these and for brevity reasons, we omit the hyper-parameters of the regression algorithms from the paper.

IV. Evaluation Results

The first box-plot in Figure 3 (and consequent figures) depicts a reference against which the proposed approach is compared. This box-plot shows the distribution of prediction errors in a static performance benchmark, where the prediction error for a given evaluation point (i.e. a data point from the evaluation set) is calculated as the absolute difference between the average localization error in the environment and the observed localization error for that particular point. The second box-plot in Figure 3 also depicts the distribution of prediction errors in a reference scenario. However, in this case, the prediction error for a given evaluation point is calculated by first estimating location information at that point, followed by calculating the localization error for that estimate. From a static performance benchmark, we then find the nearest evaluation point to the estimated location and take its localization error as the estimated localization error for that location estimate. The prediction error then equals the absolute difference between the calculated localization error and the estimated localization error for that evaluation point. As visible from the figure, the first and second box-plots are comparable, which indicates that in our static performance benchmark there is no strong spatial variability of the errors in different regions of the environment, i.e. the errors are equally distributed in the simulation environment. Intuitively, if there is a strong spatial variability of the errors, the prediction error depicted with the second box-plot would be considerably smaller than the one depicted with the first box-plot. Due to that, in the consequent figures we consider only the distribution of the absolute differences between the average localization error in the environment and the observed localization error for a particular evaluation point as a reference.

The subsequent groups of box-plots in Figure 3 depict the errors achieved by different regression algorithms. The first box-plot in each group (a group characterized by box-plots of the same color) depicts errors achieved when only primary observations (i.e. RSS values) are used for the fitting of a particular model. The second box-plot in a group depicts the observed prediction errors in case the secondary observations (i.e. estimated location information) are also included in the fitting procedure. This depiction is followed in the subsequent figures. As visible from the figure, some regression algorithms achieve substantially better estimation of localization errors than the reference. In particular, polynomial and kNN algorithms yield respectively 15% and 25% better results than the reference estimation in case primary observations are used in the fitting of the model. If also the secondary observations are used in the model fitting, the improvement is roughly 25% and 40% in comparison to the reference for polynomial and kNN algorithms, respectively. We believe the reason for the best performances achieved by the polynomial and kNN algorithms are related to low dimensionality of independent variables and relatively large number of data points.

In the second step, we evaluate the consistency of the estimation of localization errors across various fingerprinting-relevant parameters. We do that for the two regression algorithms that achieved the best performance in the initial evaluation scenario - polynomial and kNN.

First, we evaluate the influence of the number and spatial distribution of training points in fingerprinting on the performance of estimation algorithms. For the environment depicted in Figure 2, we define 40, 105, and 420 training points, which translates roughly to a regular 2D grid with the cell sizes of 3, 2, and 1 m, respectively. In addition to the regular 2D grid that is usually used in the generation of training sets in fingerprinting, we also evaluate the influence of a hexagonal training grid, which is a more optimal spatial

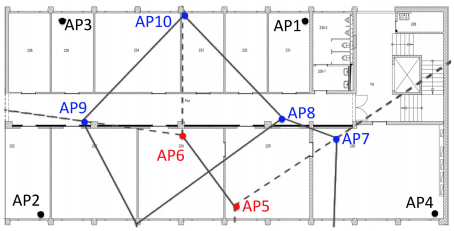


Figure 2: Floor-plan of TWIST

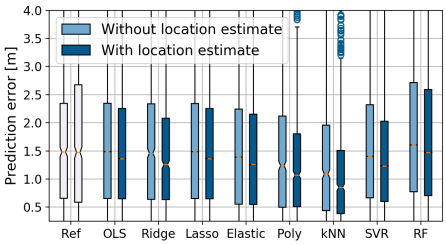


Figure 3: Regression algorithms

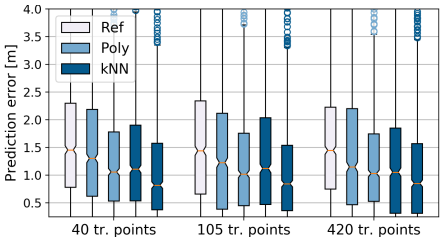


Figure 4: Number of training points

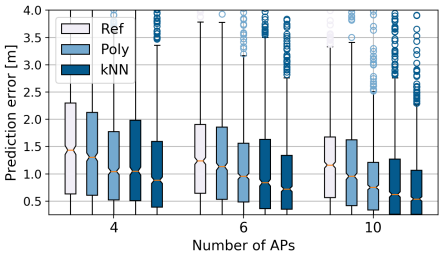


Figure 5: Number of APs

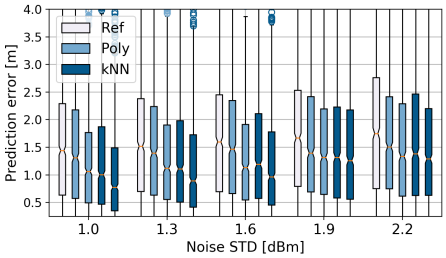


Figure 6: Noise levels

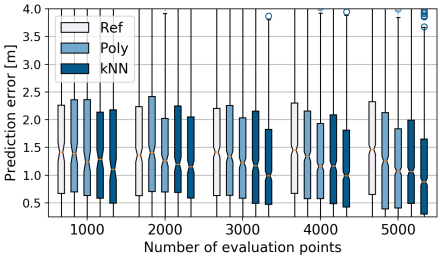


Figure 7: Number of data points

distribution of training points in fingerprinting [14], as well as random placement (i.e. no grid), which is the usual spatial distribution of training points in case the training set is generated by crowd-sourcing [17]. The results are depicted in Figure 4 for the regular 2D grid only, since we have not observed a significant influence of spatial distributions on the prediction errors, in case the same number of training points is used across spatial distributions. In other words, although different spatial distributions of training points influence the absolute values of localization errors in fingerprinting [14], these influences are too small to have an observable effect on the prediction error because the same amount of observations is used for model fitting for different spatial distributions. Furthermore, the prediction error slightly improves (i.e. 2-5%) with the increase in the number of training points, which is consistent across spatial distributions. This is because the increase in the number of training points increases the amount of information used for the model fitting.

Second, we evaluate the influence of the number of APs used for fingerprinting on the performance of regression algorithms. In order to do that, we introduce additional APs in the deployment environment in locations as depicted in Figure 2. We introduce new APs based on Voronoi diagrams, which is shown to be the optimal approach in placing new APs for fingerprinting purposes [14]. The limitation of the method is that it requires a placement of an AP in each Voronoi vertice, hence it is not always possible to introduce a single new AP, but a number of them. For this particular environment, we first introduce 2 new APs (AP5, AP6), followed by introducing 4 more (AP7, AP8, AP9, AP10), as depicted in Figure 2. The results are depicted in Figure 5. As visible in the figure, the increase in the number of APs substantially improves the prediction error for both the reference and regression-based estimations. For the reference this is because the absolute localization errors are also reduced with the introduction of new APs. For the regression-based estimation the reduction of prediction errors is partially also caused by the increase in the number of observations used for model fitting.

Third, we evaluate the influence of different noise levels on the performance of regression-based estimation. To do that, we use $\sigma = 1dBm$ in the model fitting phase, and increase σ as depicted in Figure 6 in the prediction phase. As visible in the figure, the prediction errors for both reference and regression-based estimations increase with the increase in the noise levels. For both methods, this is because the absolute localization errors increase with the increase of the noise level. The results also demonstrate consistently better performance of regression-based estimation than the reference, across different noise levels. This can serve as an indicator that the regression-based estimation can perform well under varying interference conditions in an environment.

Finally, we evaluate the influence of the number of data points on the performance of the regression algorithms. The results are given in Figure 7. As we increase the number of data points from 1000 to 5000, we observe a decrease of the prediction errors for regression-based estimation. However, the error distributions are statistically unchanged for the reference estimation. These results demonstrate the main weakness of the regression-based estimation. The regression methods require a relatively large amount of data for accurately estimating localization errors, while the amount of data necessary for the reference evaluation can be lower.

V. Exploratory Data Analysis

In this section, we present the results of a set of standard exploratory data analysis techniques for regression. The indications we provide can be used in future work for improving the performance of the regression-based estimation of localization errors in fingerprinting. We present the indications for the kNN algorithm. Similar indications have been observed for polynomial regression and these are therefore omitted.

Figure 8 depicts the Quantile-Quantile (QQ) plot of prediction errors. As visible in the figure, the prediction errors are fairly normally distributed with an exception of outliers quantiles, which suggests that a linear transformation of any variables would bring little improvement to the model. The

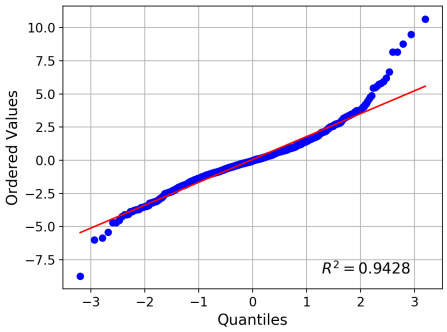


Figure 8: QQ-plot

same indication is given by a relatively high R^2 score of roughly 94%. However, detection and removal of the outliers from the data could benefit the performance of the model.

The studentized residuals are plotted against the predicted values in Figure 9. There is a visible pattern in the depicted data (i.e. a funnel shape), which is an indicator of both the non-linearities in the data used for model fitting and in the heteroskedasticity in the model. There are two well-established approaches in tackling both the non-linearities and heteroskedasticity issues. The first would involve transforming the predicted variable. However, the prediction errors are currently normally distributed, as indicated in Figure 8 and as desired. The transformation would also influence the distribution of prediction errors, which would negatively influence the performance of the model. The second approach, which we believe is more suitable for this scenario, is to try to tune the regression algorithms (e.g. changing the distance metric for the kNN regression) and modify the observation variables (e.g. rescaling the data, reducing dimensionality). Furthermore, as a rule of thumb, studentized residuals with values bigger than 2 can be considered as outliers. As visible in Figure 9, there is a number of outliers in the data. The performance of the model could potentially benefit if these outliers are removed.

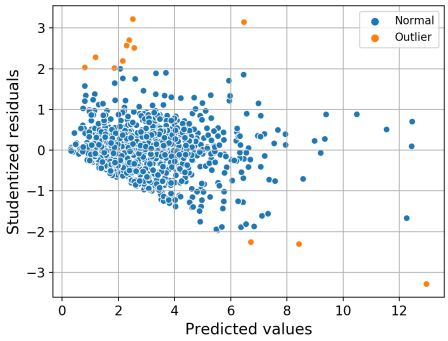


Figure 9: Studentized residuals vs. predicted values

VI. Conclusion

In this paper, we demonstrated the feasibility of regression for estimating localization errors in fingerprinting localization. In particular, we have shown that polynomial and kNN regression algorithms yield better performance than the reference estimation based on static performance benchmarks. These improvements are consistent across a number of environmental and fingerprinting-related parameters. Moreover, we have indicated potential directions for improvement of regression-based estimation of localization errors, which include modifying the regression algorithms and removing outliers from the data. Our future work will be oriented toward further exploration of these insights. We will also investigate different Machine Learning (ML) methods for estimating localization errors, with primary focus on deep learning. Furthermore, we will evaluate the possibility of using ML for estimating localization errors for other types of solutions. Finally, to strengthen our findings we will test the developed ML methods on different experimental datasets with potentially non-uniform distributions of localization errors in different environmental regions.

Acknowledgments

This research received funding from the ICON project MuSCLe-IoT. MuSCLe-IoT is realized in collaboration with imec, with project support from VLAIO (Flanders Innovation and Entrepreneurship). Project partners are imec, Flash Private Mobile Networks, Engie M2M, Sensolus, and Aertssen.

References

[1] R. Di Taranto *et al.*, "Location-aware communications for 5g networks," *IEEE Signal Processing*, vol. 31, no. 6, pp. 102–112, 2014.

[2] J. J. Nielsen *et al.*, "Location-quality-aware policy optimisation for relay selection in mobile networks," *Wireless Networks*, vol. 22, no. 2, pp. 599–618, 2016.

[3] F. Lemic, A. Behboodi, V. Handziski, *et al.*, "Location-based decision-making mechanism for device-to-device link establishment," in *Vehicular Technology Conference (VTC-Fall)*, IEEE, 2017, pp. 1–6.

[4] T. Haute *et al.*, "Performance analysis of multiple indoor positioning systems in a healthcare environment," *International journal of health geographics*, vol. 15, no. 1, p. 7, 2016.

[5] D. Lymberopoulos *et al.*, "A realistic evaluation and comparison of indoor location technologies: Experiences and lessons learned," in *Information processing in sensor networks*, ACM, 2015, pp. 178–189.

[6] F. Lemic, V. Handziski, *et al.*, "Experimental evaluation of rf-based indoor localization algorithms under rf interference," in *Localization and GNSS (ICL-GNSS)*, IEEE, 2015, pp. 1–8.

[7] R. Berkvens, H. Peremans, and M. Weyn, "Conditional entropy and location error in indoor localization using probabilistic wi-fi fingerprinting," *Sensors*, vol. 16, no. 10, p. 1636, 2016.

[8] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[9] R. S. Michalski *et al.*, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[10] F. Lemic *et al.*, "Experimental Decomposition of the Performance of Fingerprinting-based Localization Algorithms," in *Indoor Positioning and Indoor Navigation (IPIN'14)*, 2014.

[11] A. Borrelli *et al.*, "Channel models for ieee 802.11 b indoor system design," in *International Conference on Communications (ICC)*, IEEE, vol. 6, 2004, pp. 3701–3705.

[12] G. Caso and L. De Nardis, "On the applicability of multi-wall multi-floor propagation models to wifi fingerprinting indoor positioning," in *Future Access Enablers of Ubiquitous and Intelligent Infrastructures*, Springer, 2015, pp. 166–172.

[13] F. Lemic *et al.*, "Toward extrapolation of wifi fingerprinting performance across environments," in *17th International Workshop on Mobile Computing Systems and Applications*, ACM, 2016, pp. 69–74.

[14] A. Behboodi *et al.*, "Hypothesis testing based model for fingerprinting localization algorithms," in *Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th*, IEEE, 2017, pp. 1–6.

[15] F. Lemic *et al.*, "Infrastructure for benchmarking rf-based indoor localization under controlled interference," in *Ubiquitous Positioning, Navigation and LBS (UPINLBS), 2014*, IEEE, 2014, pp. 26–35.

[16] ——, "Demo abstract: Testbed infrastructure for benchmarking rf-based indoor localization solutions under controlled interference," in *European Wireless Sensor Networks (EWSN'14)*, 2014, pp. 1–5.

[17] Q. Jiang, Y. Ma, K. Liu, and Z. Dou, "A probabilistic radio map construction scheme for crowdsourcing-based fingerprinting localization," *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3764–3774, 2016.