CROWDSOURCE-BASED SIGNAL STRENGTH FIELD ESTIMATION BY GAUSSIAN PROCESSES

Irene Santos and Petar M. Djurić†*

* ETSI, Universidad de Sevilla
Dept. Teoría de la Señal y Comunicaciones
41092 Sevilla, Spain.
E-mail: irenesantos@us.es

† Stony Brook University
Dept. Electrical and Computer Engineering
Stony Brook, NY 11794 USA.
E-mail: petar.djuric@stonybrook.edu

ABSTRACT

We address the problem of estimating a spatial field of signal strength from measurements of low accuracy. The measurements are obtained by users whose locations are inaccurately estimated. The spatial field is defined on a grid of nodes with known locations. The users report their locations and received signal strength to a central unit where all the measurements are processed. After the processing of the measurements, the estimated spatial field of signal strength is updated. We use a propagation model of the signal that includes an unknown path loss exponent. Furthermore, our model takes into account the inaccurate locations of the reporting users. In this paper, we employ a Bayesian approach for crowdsourcing that is based on Gaussian Processes. Unlike methods that provide only point estimates, with this approach we get the complete joint distribution of the spatial field. We demonstrate the performance of our method and compare it with the performance of some other methods by computer simulations. The results show that our approach outperforms the other approaches.

***Index Terms*—** Sensor networks, Bayesian estimation, regression, spectrum sensing, Gaussian processes.

1. INTRODUCTION

Due to the rapid growth of wireless communication technologies, radio frequency (RF) spectrum monitoring has gained significant research interest. Spectrum monitoring amounts to detecting intruders in a spectrum band of interest and finding vacant channels that have no interference from other users [8, 14]. Current RF spectrum monitoring approaches suffer from one main drawback: they do not scale well and their coverage area cannot be easily extended due to cost problems. One appealing solution is to use measurements of users with low-cost but also low-accuracy sensors in large-scale geographical areas, and to exploit crowdsourcing for spectrum monitoring [13]. One can expect that when the density of users becomes high enough, a method based on measurements from large number of not very accurate devices will produce better spectral maps than based on measurements from a small number of accurate (and expensive) instruments.

In spectrum monitoring, we typically have access to received signal strength (RSS) measurements with errors made at a set of known locations, and the goal is to estimate as accurately as possible the RSS at any location in an area of interest. The estimation of RSS values is not just used for spectrum sensing; it has also been successfully exploited for localization [6, 12], tracking [5], distance estimation [9] and distributed asynchronous regression [7].

Existing approaches for RSS estimation apply spatial interpolation techniques to the data, such as Ordinary Kriging (OK) and Inverse Distance Weighting (IDW) [10, 11]. One class of methods estimate the path loss exponent from the measurements and treat this estimate as the true value. Furthermore, many of these methods produce point estimates of signal strength without providing the uncertainties of the estimates. Another class of methods exploits probability theory to compute soft estimates of the RSS. For example, in [8], the authors model the RSS as a multivariate Gaussian distribution with an exponential correlation model, where the path loss exponent and the locations of users are perfectly known. Similarly, in [3, 6] the authors apply a Gaussian Process (GP) to model the RSS, and for estimation they use measurements from known locations. All these approaches assume a log-normal path loss model and perfect knowledge of the user locations.

In this paper, we extend the approach in [6] to propose a more complex GP implementation. Specifically, we take into account that the locations of the users that provide RSS measurements are not accurate and that the path loss exponent is unknown. The location errors of the users are assumed Gaussian, and the path loss exponent, too, is modeled as a Gaussian random variable. The parameters of the path loss

The work of I. Santos was partially funded by Spanish government (Ministerio de Economía y Competitividad TEC2016-78434-C3-2-R) and by the European Union (FEDER). P. M. Djurić was supported by the National Science Foundation under Award CNS 1642965.

ISBN 978-0-9928626-7-1

Gaussian are estimated according to the empirical Bayesian approach. In our paper, we fix the locations where we estimate the RSS field. Thus, at any time instant, we can summarize all the information from previous measurements and can easily update the estimate with new measurements. Thereby, we make the approach scalable. We note that, if necessary, we can readily inject time variation of the field (not addressed in the paper).

The paper is organized as follows. We first describe the notation and our model in Section 2. In Section 3, we explain the modeling of the path loss exponent with a Gaussian distribution and how we obtain its hyper-parameters. We describe how we apply GPs to our system in Section 4. Simulation results are presented in Section 5. They show how our approach outperforms interpolation and graph signal-based techniques and how the estimation error reduces when the density of users increases. We provide conclusions in Section 6.

2. SYSTEM MODEL

We consider an area with many users who have low-cost sensing devices of RSS measurements. These users measure the RSS of one transmitter with a known location $x_0 \in \mathbb{R}^2$. The locations of the users are not perfectly known, that is, we only have the estimates of their locations. We denote the estimate of the location of the ith user by $\widehat{x}_i$, and we model it according to

$$\widehat{x}_i \sim \mathcal{N}\left(\widehat{x}_i; x_i, \sigma_x^2 I_2\right), \tag{1}$$

where $x_i \in \mathbb{R}^2$ is the exact location of the ith user, and I_2 is the 2×2 identity matrix.

The measurement (expressed in dB) of the ith user is modeled by

$$z_i = P - 10\alpha \log_{10}(d_i) + v_i + w_i, \tag{2}$$

where P is the transmitter power measured at a distance of 1 m, α is the path loss exponent, d_i is the distance (in meters) between user i and the transmitter, v_i is attenuation due to shadowing effects, and w_i is some unrelated additive noise. Recall that d_i is not known, but instead its estimate $\widehat{d}_i$ is available. Therefore, we modify (2) to

$$z_i = P - 10\alpha \log_{10}(\widehat{d}_i) + u_i + v_i + w_i, \tag{3}$$

where $\widehat{d}_i$ is now used instead of $\widehat{d}_i$, and u_i is error that reflects the imprecisely known location of the users. We adopt the following:

$$u_i \sim \mathcal{N}\left(u_i; 0, \frac{\sigma_u^2}{\widehat{d}_i^2}\right), \tag{4}$$

$$v_i \sim \mathcal{N}\left(v_i; 0, \sigma_v^2\right), \tag{5}$$

$$w_i \sim \mathcal{N}\left(w_i; 0, \sigma_w^2\right). \tag{6}$$

The attenuation variables are correlated, and for them we assume

$$\mathrm{Cov}(v_i v_j) = \sigma_v^2 \exp(-\widehat{d}_{ij}/D_{\mathrm{corr}}), \tag{7}$$

where $\widehat{d}_{ij}$ is the estimated distance between users i and j and D_{corr} is a parameter that models the correlation in the measurements.

We are given a vector of measurements (taken at the same time) obtained by N users, $z = [z_1\, z_2 \cdots z_N]^\top$, expressed by

$$z = 1P - q\alpha + u + v + w, \tag{8}$$

where 1 is an $N \times 1$ vector of 1s,

$$q = [10\log_{10}(\widehat{d}_1)\; 10\log_{10}(\widehat{d}_2) \cdots 10\log_{10}(\widehat{d}_N)]^\top, \tag{9}$$

and u, v and w are all in $\mathbb{R}^N$, and

$$u \sim \mathcal{N}\left(u; 0, \sigma_u^2 D\right), \tag{10}$$

where $D = \mathrm{diag}\left\{1/\widehat{d}_1^2, 1/\widehat{d}_2^2, \cdots, 1/\widehat{d}_N^2\right\}$, and

$$v \sim \mathcal{N}\left(v; 0, \Sigma_v\right), \tag{11}$$

$$w \sim \mathcal{N}\left(w; 0, \sigma_w^2 I_N\right), \tag{12}$$

where Σ_v is comprised of elements given by (7), and I_N is the $N \times N$ identity matrix.

In the following, we assume that P, σ_u^2, σ_v^2, D_{corr} and σ_w^2 are known. Given z, the noisy location of the users, $\widehat{x}= [\widehat{x}_1 \cdots \widehat{x}_N]$, and the model in (8) and all the assumptions, we want to estimate the values of RSS at M grid locations, $x_d = [x_{d_1} \cdots x_{d_M}] \in \mathbb{R}^{M\times 2}$.

3. ESTIMATION OF PATH LOSS EXPONENT

Given a set of measurements modeled by (8), we want to estimate α. We assume that the prior of α is a Gaussian distribution $p(\alpha|\theta) \sim \mathcal{N}\left(\alpha; \mu_\alpha, \sigma_\alpha^2\right)$, with unknown hyper-parameters $\theta = (\mu_\alpha, \sigma_\alpha^2)$. The posterior of α is

$$p(\alpha|z, \widehat{x}, \theta) = \frac{p(z|\alpha, \widehat{x})p(\alpha|\theta)}{p(z|\widehat{x}, \theta)}, \tag{13}$$

where $p(z|\alpha, \widehat{x}) \sim \mathcal{N}\left(z; \mu_{z|\alpha}, \Sigma_{z|\widehat{x},\alpha}\right)$, which is also a Gaussian whose moments are given by

$$\mu_{z|\alpha} = 1P - q\alpha, \tag{14}$$

$$\Sigma_{z|\alpha} = \sigma_u^2 D + \Sigma_v + \sigma_w^2 I_N, \tag{15}$$

and $p(z|\widehat{x}, \theta)$ is the marginalized distribution of the measurements, which can be computed by

$$p(z|\widehat{x}, \theta) = \int p(z|\alpha, \widehat{x})p(\alpha|\theta)\mathrm{d}\alpha. \tag{16}$$

ISBN 978-0-9928626-7-1

We estimate θ using the empirical Bayes method [2]. We do it by solving (16), where the result of the integral is also a Gaussian [1],

$$p(z|\widehat{x},\theta) \quad \sim \quad \mathcal{N}\left(z;\mu_z,\Sigma_z\right) \tag{17}$$

with

$$\begin{aligned} \mu_z &= 1P - q\mu_\alpha, & (18)\\ \Sigma_z &= \Sigma_{z|\alpha} + \sigma_\alpha^2 q q^\top. & (19) \end{aligned}$$

We estimate μ_α and σ_α from the data and proceed as follows. We approximate μ_z with z and write

$$1P - z \quad = \quad q\mu_\alpha, \tag{20}$$

and μ_α is obtained from

$$\widehat{\mu}_\alpha \quad = \quad \frac{q^\top(1P-z)}{q^\top q}. \tag{21}$$

We estimate the variance from

$$a \quad = \quad \sigma_\alpha^2 b, \tag{22}$$

where a and b are diagonal elements of $A = (z-\widehat{\mu}_z)(z-\widehat{\mu}_z)^\top - \Sigma_{z|\alpha}$ and $B = qq^\top$, respectively, where $\widehat{\mu}_z = 1P - q\widehat{\mu}_\alpha$. If a diagonal element of A (or B) is negative, it is set to zero. Then

$$\widehat{\sigma}_\alpha^2 \quad = \quad \frac{b^\top a}{b^\top b}. \tag{23}$$

Finally, following (13), the estimated posterior distribution for α is obtained as

$$p(\alpha|z,\widehat{x},\widehat{\mu}_\alpha,\widehat{\sigma}_\alpha^2) \quad \sim \quad \mathcal{N}\left(\alpha;\mu_{\alpha|z},\sigma_{\alpha|z}^2\right), \tag{24}$$

where

$$\begin{aligned} \sigma_{\alpha|z}^2 &= (\widehat{\sigma}_\alpha^{-2} + q^\top\Sigma_{z|\alpha}^{-1}q)^{-1}, & (25)\\ \mu_{\alpha|z} &= \sigma_{\alpha|z}^2(-q^\top\Sigma_{z|\alpha}^{-1}(z-1P) + \widehat{\sigma}_\alpha^{-2}\widehat{\mu}_\alpha). & (26) \end{aligned}$$

4. RSS ESTIMATION WITH GAUSSIAN PROCESSES

The system in (8) can be seen as

$$z \quad = \quad f(\widehat{x}) + \epsilon \tag{27}$$

where $\widehat{x}$ is an $N \times 2$ matrix with the noisy locations of the users, ϵ is a noise vector formed by

$$\epsilon = u + w \quad \sim \quad \mathcal{N}\left(\epsilon; 0, \Sigma_\epsilon = \sigma_w^2 I_N + \sigma_u^2 D\right), \tag{28}$$

and the function $f(\widehat{x}) = 1P - q\alpha + v$ is modeled as a GP $f(\widehat{x}) \sim \mathcal{GP}\left(f(\widehat{x}); m_{\widehat{x}}, K_{\widehat{x}}\right)$, where $m_{\widehat{x}}$ and $K_{\widehat{x}}$ are the mean and covariance functions, respectively, given by

$$\begin{aligned} m_{\widehat{x}} &= [m(\widehat{x}_1)\cdots m(\widehat{x}_N)]^\top, & (29)\\ K_{\widehat{x}} &= K(\widehat{x},\widehat{x}) + \widehat{\sigma}_\alpha^2 qq^\top, & (30) \end{aligned}$$

where $K(\widehat{x},\widehat{x})$ is an $N \times N$ matrix comprised of elements $k(\widehat{x}_i,\widehat{x}_j)$ given by a specific kernel function. A possible choice of this kernel function will be seen in Subsection 4.1. Usually, the elements of the mean function, $m(\widehat{x}_i)$, are assumed to be zero, which means that in the absence of training data, the model would tend to zero. In order to avoid this behavior, we set each element of the mean vector as

$$m(\widehat{x}_i) \quad = \quad P - 10\widehat{\mu}_\alpha \log_{10}(\widehat{d}_i). \tag{31}$$

The joint distribution of the observed values and the estimated RSS at the main grid locations, f_d, can be computed as

$$\begin{bmatrix} z \\ f_d \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} z \\ f_d \end{bmatrix}; \begin{bmatrix} m_{\widehat{x}} \\ m_{x_d} \end{bmatrix}, \begin{bmatrix} K_{\widehat{x}} + \Sigma_\epsilon & K_{\widehat{x},x_d} \\ K_{x_d,\widehat{x}} & K_{x_d,x_d} \end{bmatrix}\right), \tag{32}$$

where

$$\begin{aligned} m_{x_d} &= [m(x_{d_1})\cdots m(x_{d_M})]^\top, & (33)\\ K_{\widehat{x},x_d} &= K(\widehat{x},x_d) + \widehat{\sigma}_\alpha^2 q q_d^\top, & (34)\\ K_{x_d,\widehat{x}} &= K_{\widehat{x},x_d}^\top, & (35)\\ K_{x_d,x_d} &= K(x_d,x_d) + \widehat{\sigma}_\alpha^2 q_d q_d^\top, & (36)\\ q_d &= [10\log_{10}(d_{d_1})\cdots 10\log_{10}(d_{d_M})]^\top, & (37) \end{aligned}$$

and d_{d_i} is the distance between the grid node i and the transmitter.

The conditional distribution of the RSS at the grid locations given the measurements and the estimated locations of the users is

$$p(f_d|x_d,\widehat{x},z,\widehat{\mu}_\alpha,\widehat{\sigma}_\alpha) \quad \sim \quad \mathcal{N}\left(f_d;\mu_{f_d},\Sigma_{f_d}\right), \tag{38}$$

where

$$\begin{aligned} \mu_{f_d} &= m_{x_d} + K_{x_d,\widehat{x}}\left(K_{\widehat{x}} + \Sigma_\epsilon\right)^{-1}(z - m_{\widehat{x}}), & (39)\\ \Sigma_{f_d} &= K_{x_d,x_d} - K_{x_d,\widehat{x}}\left(K_{\widehat{x}} + \Sigma_\epsilon\right)^{-1}K_{\widehat{x},x_d}. & (40) \end{aligned}$$

The estimates of the RSS at locations x_d can be obtained by $\widehat{f}_d = \mu_{f_d}$. We call this approach as GP-based approach.

4.1. In the kernel selection

The choice of the kernel function is typically left to the user. One commonly used kernel is the squared exponential (SE) kernel

$$k(x_i,x_j) \quad = \quad \sigma_k^2 \exp\left(-\frac{1}{2l^2}(x_i-x_j)^\top(x_i-x_j)\right), \tag{41}$$

where the hyper-parameters, σ_k and l, are tuned according to the training data. We used this kernel to show the good performance of the algorithm even with a kernel that does not fit exactly the model in (7). In this paper, the hyper-parameters will be set to $l = D_{corr}$ and $\sigma_k^2 = \sigma_v^2$.

ISBN 978-0-9928626-7-1 © EURASIP 2017

5. EXPERIMENTAL RESULTS

We considered an area of 500 m × 500 m with a transmitter placed at its center and with 100 users at random locations within the area. The RSS was generated following (2). The density of users was four per 100 m × 100 m. The noise variance was set to $\sigma_w^2 = 0.8$ dB, $D_{corr} = 50$ m and $P = -10$ dBm. We considered 16 fixed uniformly placed nodes where we estimate the RSS and conducted 1000 different experiments. In all the simulations, we show the performance of our GP approach, the inverse distance weighting (IDW), the ordinary Kriging with detrending (OKD) technique and an the implementation of the graph signal inpainting via total variation regularization (GTVR) [4]. This last approach assumes smooth graph signals corrupted by noise and recovers the inaccessible graph signals from the accessible ones by minimizing the graph total variation based on the second norm. For this algorithm, we used an adjacency matrix defined by

$$A_{ij} = \exp\left(-\frac{N^2 d_{ij}}{\sum_i \sum_j d_{ij}}\right). \tag{42}$$

In Figure 1, we plotted the mean square error (MSE) of all the approaches, where

$$MSE = \frac{1}{M} \sum (\mu_{f_d} - f_d)^2, \tag{43}$$

where μ_{f_d} is given by (39). It is shown that our approach is quite stable even when we use the estimated locations of the users. We observed that the OKD technique is quite sensitive on the location error of the users, while IDW is more stable but less accurate than our approach.

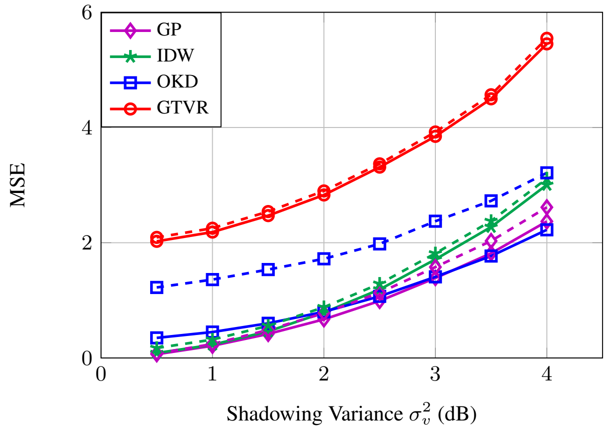

Fig. 1: Mean square error for varying shadowing variance and $\sigma_u = 0$ (solid) and $\sigma_u = 200$ (dashed).

Figure 2 shows the median estimation errors for varying σ_v^2. The solid lines correspond to the simulations with $\sigma_u = 0$, while the dashed lines to $\sigma_u = 200$ m. The estimation error of each node is given as

$$\frac{|\mu_{f_d} - f_d|}{f_d} \times 100\%, \tag{44}$$

where μ_{f_d} is given by (39). We observed that the median estimation error increased when we introduced uncertainty in the locations of the users (i.e., when σ_u^2 grew). The GTVR technique did not work correctly in this scenario – it had the highest error. IDW performs similarly to our approach but with a somewhat higher error, and OKD deteriorates considerably when errors were added to the estimated locations of the users. In this setting, the approach proposed in this paper outperformed the interpolation techniques and GTVR.

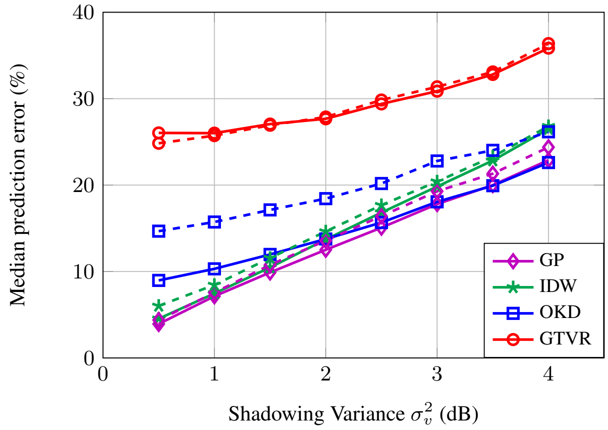

Fig. 2: Median estimation errors for varying shadowing variance and $\sigma_u = 0$ (solid) and $\sigma_u = 200$ (dashed).

Finally, in Figure 3 we show how the median estimation error varies with the density of users in the system, for $\sigma_v = \sqrt{2}$ dB. The filled bars were obtained with $\sigma_u = 0$, while the pattern bars with $\sigma_u = 200$ m. It can be observed that the median estimation error decreased remarkably when the density of the users increased. Similarly to Figure 2, GTVR had the highest error, while our approach was the best. The performance of OKD and IDW was in-between that of GTVR and of our approach.

6. CONCLUSIONS

In this paper, we proposed a novel Bayesian framework based on Gaussian Processes to obtain estimates of received signal strength values at predefined locations in an area of interest. These estimates are obtained from measurements acquired by users at random locations and equipped with inexpensive sensors. The locations of the users are estimates with errors. In our model, we account for both the substantial inaccuracy of the measurements and the errors in the locations

ISBN 978-0-9928626-7-1 © EURASIP 2017

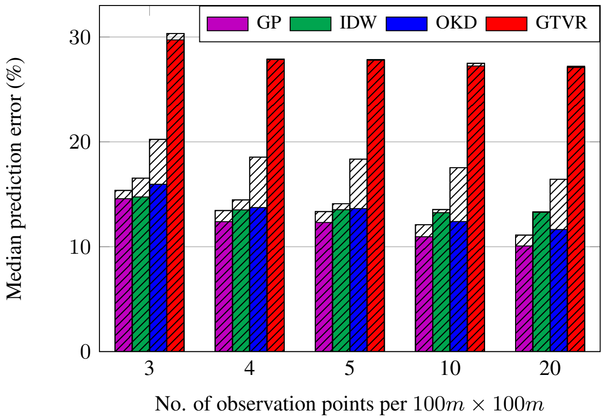

Fig. 3: Median estimation errors for varying density of observation points and $\sigma_u = 0$ (filled bars) and $\sigma_u = 200$ (pattern bars).

of the users. In addition, our model contains a path loss exponent which is assumed unknown and is modeled as a Gaussian distributed random variable. We estimated the hyper-parameters of this distribution from the available data. We showed in the simulations that our approach is quite stable even when we introduce large errors in the locations of the users. Our method outperformed the ordinary Kriging, the inverse distance weighting methods, and GTVR from [4].

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, New York, Secaucus, NJ, USA, 2006.

[2] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & HALL/CRC, 2000.

[3] F. Chan, Y. T. Chan, and R. Inkol, "Path loss exponent estimation and RSS localization using the linearizing variable constraint," in *MILCOM 2016 - 2016 IEEE Military Communications Conference*, Nov 2016, pp. 225–229.

[4] S. Chen, A. Sandryhaila, J. M. F. Moura, and J. Kovacevic, "Signal recovery on graphs: Variation minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4609–4624, 2015.

[5] M. Dashti, S. Yiu, S. Yousefi, F. Pérez-Cruz, and H. Claussen, "RSSI localization with Gaussian processes and tracking," in *2015 IEEE Globecom Workshops (GC Workshops)*, 2015, pp. 1–6.

[6] B. Ferris, C. Hähnel, and D. Fox, "Gaussian processes for signal strength-based location estimation," in *Proceeding of Robotics: Science and Systems*. Citeseer, 2006.

[7] J. A. Garrido-Castellano and J. J. Murillo-Fuentes, "On the implementation of distributed asynchronous nonlinear kernel methods over wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, pp. 171–185, 2015.

[8] X. Li and P. Mähönen, "Grid based cooperative spectrum sensing in cognitive networks under correlated shadowing," in *2012 7th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, 2012, pp. 350–355.

[9] R. K. Mahapatra and N. S. V. Shet, "Experimental analysis of RSSI-based distance estimation for wireless sensor networks," in *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, 2016, pp. 211–215.

[10] M. Molinari, M. R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," in *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*, New York, NY, USA, 2015, C2B(1)D '15, pp. 33–38, ACM.

[11] J. M. Montero, G. Fernández-Avilés, and J. Mateu, *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*, Wiley Series in Probability and Statistics. Wiley, 2015.

[12] C. Pak and A. Huang, "Increasing indoor localization accuracy using WiFi signals," *Project Report*, 2011.

[13] D. Pfammatter, D. Giustiniano, and V. Lenders, "A software-defined sensor architecture for large-scale wideband spectrum monitoring," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, New York, NY, USA, 2015, IPSN '15, pp. 71–82, ACM.

[14] J. Portelinha, F. Martins, and P. Cardieri, "Effects of correlated shadowing on cooperative spectrum sensing," in *International Workshop on Telecommunications (IWT)*, 2013.

ISBN 978-0-9928626-7-1 © EURASIP 2017