Impact of Model Uncertainties on the Accuracy of Spatial Interpolation Based Coverage Estimation

Nikos Perpinias, Janne Riihijärvi, Petri Mähönen
Institute for Networked Systems, RWTH Aachen University, Aachen, Germany
email: {npe, jar, pma}@inets.rwth-aachen.de

***Abstract*—Improved propagation prediction has become important due to emerging 5G and ultra-dense wireless networks. One important aspect in this domain is to understand and model better the spatial correlations of the underlying shadowing fields. In this work, we investigate the sources of uncertainty on the predictions made by employing the optimal linear predictor, namely *kriging*. We show that kriging is robust to the estimation errors of the spatial correlation structure. Moreover, we show that the amount of the training data (known data points) has great impact on the prediction error. Our results help to quantify the trade off between number of data points collected and accuracy that can be reached by interpolation. The results especially help to design and optimize expensive measurement and test drive campaigns. Another contribution of the paper is to explicitly show the capabilities of kriging for coverage prediction.**

I. Introduction

Proliferation of wireless technologies continues with tremendous pace. The emerging 5G, ultra-dense, and dynamic spectrum access networks (DSA) will push wireless technologies to their limits. This means also that we need to continue to better understand the signal propagation environment. Due to the increased interference levels, DSA, and new MIMO approaches, we need to particularly increase our understanding and exploitation capability in the spatial domain.

In the classical approach for the propagation analysis, the signal experiences deterministic, distance-based propagation losses. The rest of the uncertainties of the propagation are modeled partially for example as Rayleigh small scale fading, and partially as a log-normal random variable also known as shadow fading [1]. A lot of effort has been put in to develop more sophisticated models for the shadow fading. One modern and promising approach is to employ spatial interpolation schemes so as to exploit the spatial structure of the shadowing or the received power [2]–[5]. This improvement can be achieved by applying such estimation methods to data gathered from spectrum sensors, drive tests or measurement campaigns even from user devices through crowd sourcing or sampling.

In particular, the received power or the extracted shadowing field can be used with a spatial interpolation method so as to provide predictors for the received power or the shadowing field in non measured locations respectively. In this work we use the optimal linear estimator i.e. *kriging*. Kriging, as shown in [6], can be applied either on the raw received power or on the extracted shadowing field. In both of these approaches, the respective covariance structure has to be extracted and provided to the predictor. Kriging treats the covariates and the correlation model as known, but in practice they must be estimated from data and this unavoidably results in model errors. Different applications of kriging use these models in slightly different ways, thus it is not clear how their performance is influenced by these errors.

In contrast to our previous work in [6], [7], where we analyze only the shadow correlations and investigate the use of kriging given certain raw measurement-based data and/or the extracted shadow correlations, here we investigate the impact of the above mentioned sources of uncertainty on the predictions done by the kriging. More specifically, we show that the number of data points used in the extraction of the spatial correlation structure and the accuracy of its estimation has only a surprisingly small impact on the prediction error. Our results suggest that even with measurement campaigns that do not succeed to capture the raw shadow correlations explicitly, the interpolation with the optimal linear predictor provides results with low estimation errors. Moreover, we verify that the most influential parameter for the estimation of the kriging is the amount of data used as training points. These results are important on quantifying the trade off between quality, in terms of accuracy of spatial interpolation, and quantity of data needed. Since the coverage estimates may depend on the model errors in very complex manner, and whether these relationships are adequately captured by simulation models is not certain, using real measurement data seems to be the only way for making such accuracy analysis.

We employed measurement data collected during two high density measurement campaigns that were conducted in urban and suburban environments at two different frequencies, namely at 485 MHz and 2600 MHz. This complements our work by providing a more extensive insight for different environments and frequencies and a broader applicability of our results.

The rest of the paper is organized as follows. In Section II we give a short description of our setup and the measurement procedure followed in data acquisition. In Section III, we explain in detail the analysis procedure adopted for this work. Finally, we provide the results in Section IV and we conclude in Section V.

II. Experimental Spatial Data

The data used in this paper is based on raw data from a large, high spatial resolution campaign conducted by Perpinias et. al [7]. The details on the measurement equipment, location, and methodology are given in [7] and not repeated here due to the space limitations.

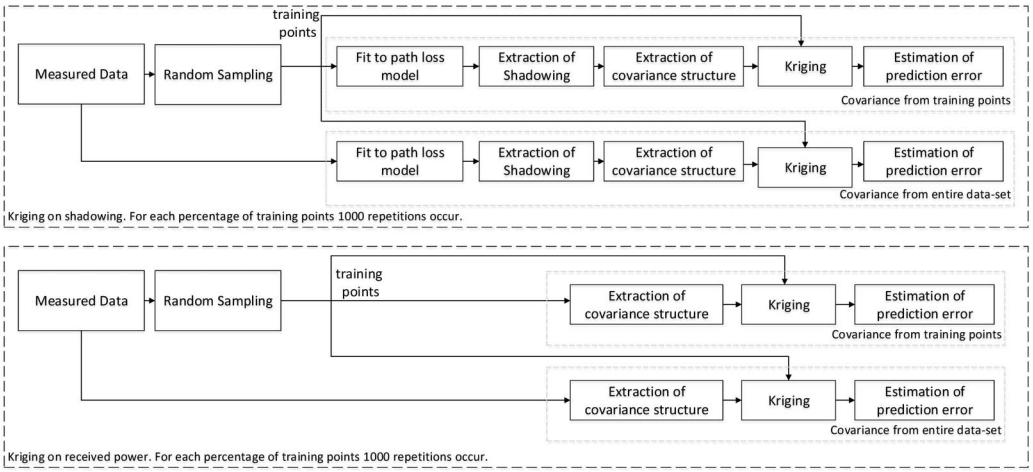


Fig. 1. Block diagram describing our analysis.

The data essentially provides high accuracy signal strength and path loss data in urban and suburban locations with 20 cm spatial accuracy. Two different frequencies, namely 485 MHz and 2600 MHz, were covered and we used the post-processed, calibrated data from several thousands of points for our uncertainty analysis.

III. Analysis

First step in the analysis is the fitting of the measured received power to the log-distance path loss model given by

$$\widehat{P}_r = c + 10\alpha \log_{10}(d), \tag{1}$$

with $\widehat{P}_r$ corresponding to the estimated received power, c incorporating the constants, such as the transmitted power, gains and losses of the setup. The estimated path loss exponent is parameter α and finally d is the distance between transmitter and receiver. The residuals of this fitting compose the shadowing field.

Various spatial statistical methods can be applied on the extracted shadowing so as to estimate the spatial correlations of the field. In literature, the empirical correlation between locations with a given separation distance is often calculated and presented in a graph (correlogram) as a measure of the spatial dependency. In this work, we calculate a function that describes the spatial dependencies of a spatial random field as:

$$\gamma(h) = \frac{1}{2}\,\mathrm{Var}(S(x_i) - S(x_i + h)), \tag{2}$$

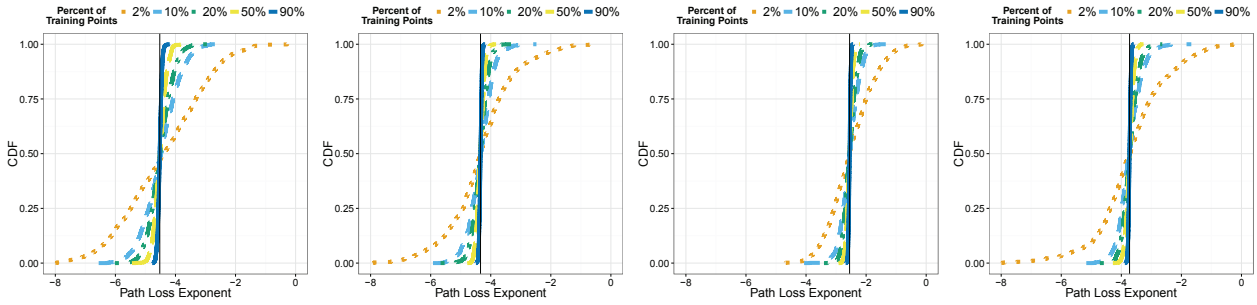
where Var is the variance, $S(*)$ is the random field, x_i is a location and h is a certain displacement from location x_i. Due to second order stationarity, the variance is only a function of the distance h. In spatial statistics this function is known as experimental semivariogram and has better statistical properties than the correlogram as shown in [8]. The experimental semivariogram is fitted to a theoretical semivariogram with a covariance structure that follows the exponential model so as to extract the optimal parameters for the given field. Let us introduce the main parameters. The value of the semivariogram in the same location is by default 0 as $\gamma(x_i - x_i) = 0$. The *nugget* parameter is the discontinuity at the origin of the semivariogram. The existence of nugget in the extracted semivariogram would in our case mean that the fast fading was not successfully averaged out. The *sill* parameter describes the maximum value of the semivariogram and is equal to the marginal variance of the random process [9]. In the case that there is a confined sill for the semivariogram the *range* parameter, ϕ, describes the distance in which two samples are not correlated any more. However, for semivariograms with asymptotic sill, the *practical range* is used to define the distance between the points that have a correlation less or equal to 0.05. For some theoretical semivariograms, like the ones with Cauchy correlation function, there is also the κ parameter that has an effect in the smoothness of the random field. From the extracted correlation parameters (specifically the range and the sill), the covariance structure of the stochastic process can be constructed and used along with the data in kriging.

Kriging is a spatial interpolation method used to predict the value of a stochastic process $S(\cdot)$ in a certain location x_0. This is done by computing a weighted average of the measured (known) values in neighbor locations $x_i,\ i \neq 0$, as well as by taking into account the covariance structure of the stochastic process. A mathematical formulation of the process is:

$$\widehat{S}(x_0) = \sum_i w_i S(x_i), \tag{3}$$

where $\widehat{S}(x_0)$ denotes the estimated value and w_i the weights. In our analysis we employ the *ordinary kriging* that tries to produce estimates for which the error variance is minimum and that is uniformly unbiased due to the fact that the sum of the weights must be equal to 1 ($\sum_i w_i = 1$). More details for kriging can be found in [8] and [10].

For both the received power and the shadowing, we can use the above mentioned analysis due to the fact that both can be modeled as stochastic processes. The shadowing is a zero mean random process. The received power is a stochastic process with mean value the large scale path loss component that is constant in a certain location/region. In more details the statistical methodology used, can be found in [8] and [11].



(a) 485 MHz, Suburban (b) 485 MHz, Urban (c) 2600 MHz, Suburban (d) 2600 MHz, Urban

Fig. 2. Path loss exponent of the log-distance model used to extract the shadowing field.

Moreover, we used the software [12] for the data analysis. In this work, two approaches, depending on the field that kriging is applied on, namely *Kriging on shadowing field* and *Kriging on received power*, are considered. In order to understand the impact of the shadowing extraction and the correlation uncertainty on the prediction accuracy of the kriging, we create two different analysis chains for each of the two approaches. That provides four different analysis procedures shown in Figure 1 and described in detail here.

A. Kriging on shadowing field with covariance structure extracted from the training points

In this process, we randomly sample our data and classify them into *training points* and *verification points*. The training points are fitted on the large scale path loss model and the limited shadowing field is derived. Then the shadow correlations are extracted from the shadowing field. Finally, the values of the shadowing field on the locations of the verification points are predicted from the kriging. The input data used in kriging are the field values measured at the training points (along with their locations) and the extracted covariance structure. The extraction of the shadowing field and the calculated parameters of the variogram from the training points are introducing uncertainties in this analysis.

B. Kriging on shadowing field with covariance structure extracted from the entire data

In the second analysis method, we divide data to training points and verification points like in the first method, but we use both classes of data for the fitting of the path loss model. Thus the extraction of the shadowing field and the estimation of the semivariogram parameters are done from the entire data-set providing accurate information to the kriging process. The rest of the process is as in III-A.

C. Kriging on received power with covariance structure extracted from the training points

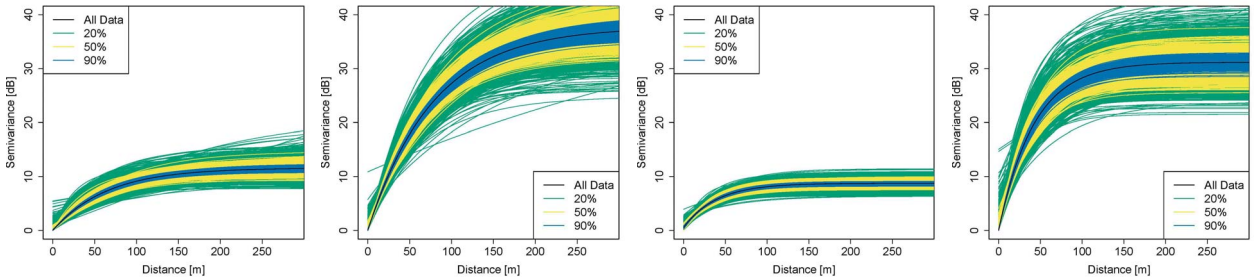
In this third analysis method, we use kriging to predict directly the received power. We utilize the randomly categorized data from III-A. Since the analysis is done on the received power, we do not fit the training data to a large scale path loss model. The covariance structure is extracted *only* from the training points introducing inaccuracies on the estimated variogram parameters used in the kriging.

D. Kriging on received power with covariance structure extracted from the entire data

Finally in this last approach, we estimate the received power from the kriging but as in III-B, the covariance structure provided to the predictor is estimated from the whole data-set, so the prediction error is only dependent on the training points. For all the above mentioned methods, the difference between the measured data of the verification points and the estimation on the same locations provided from the kriging is used as a metric for the prediction, *prediction error*. Moreover, it has to be noted that we consider the calculations and estimations for all the four methods as one iteration of our analysis implementation. We repeat our analysis for 1000 times and for different percentages of training points ($2\%, 10\%, 20\%, 50\%$ and 90%). The whole data-set is randomly sampled once for all the four methods in each iteration, so as to keep the results of each procedure relevant (in term of used locations and comparisons) to the results of the others.

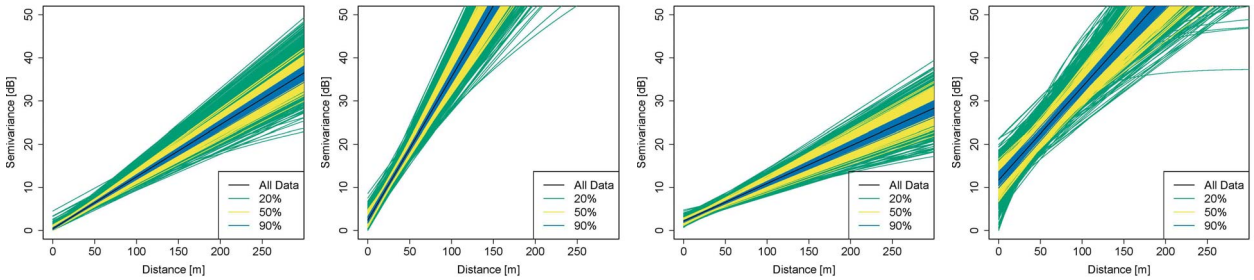
IV. Results

We start with the path loss exponent as it is estimated from the fit of the measured received power to the log-distance path loss model in analysis approaches III-A and III-B. Figure 2 depicts the empirical cumulative distribution function (ECDF) of the path loss exponent for all the frequencies and environments. The black lines in each of the plots correspond on the path loss exponent provided from III-B, where the entire data set is used in the fitting. It can be seen that the distribution is altered by the percentage of training points, i.e. from the number of data used in the fit. In particular the fit starts to provide more concise path loss exponent estimations when the percentage of the training points are over 20%. This percent corresponds to as low as 94 measurement points for the suburban environment at 485 MHz, and up to 190 measurement points for the suburban environment at 2600 MHz and for the urban environment at both frequencies. Moreover, for the urban case we can see that there is a small difference between the estimated path loss exponents for the two frequencies. On the other hand, the



(a) 485 MHz, Suburban (b) 485 MHz, Urban (c) 2600 MHz, Suburban (d) 2600 MHz, Urban

Fig. 3. Exponential fits of the shadowing field semivariograms for various percent of training points.



(a) 485 MHz, Suburban (b) 485 MHz, Urban (c) 2600 MHz, Suburban (d) 2600 MHz, Urban

Fig. 4. Exponential fits of the received power semivariograms for various percent of training points.

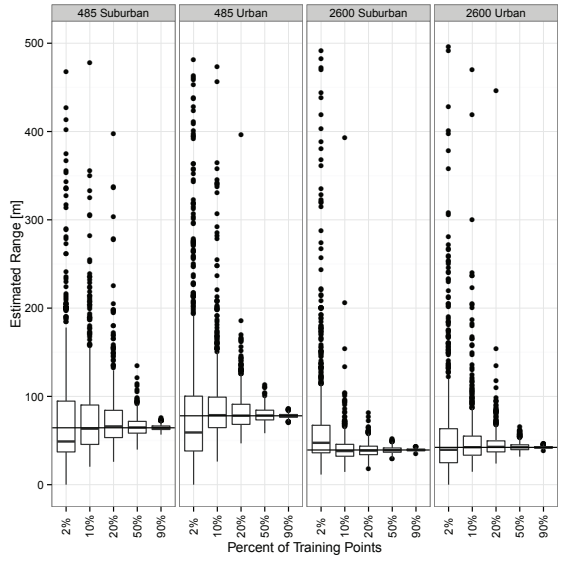


Fig. 5. Boxplots of the estimated range for the shadowing covariance structure. The black line depicts the range estimation from the entire data-set.

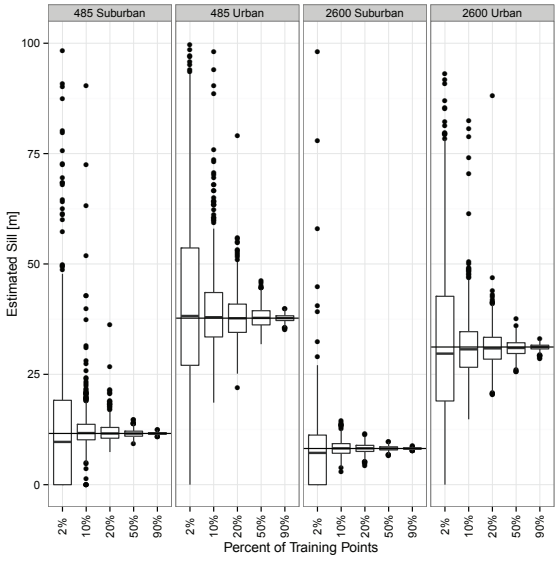
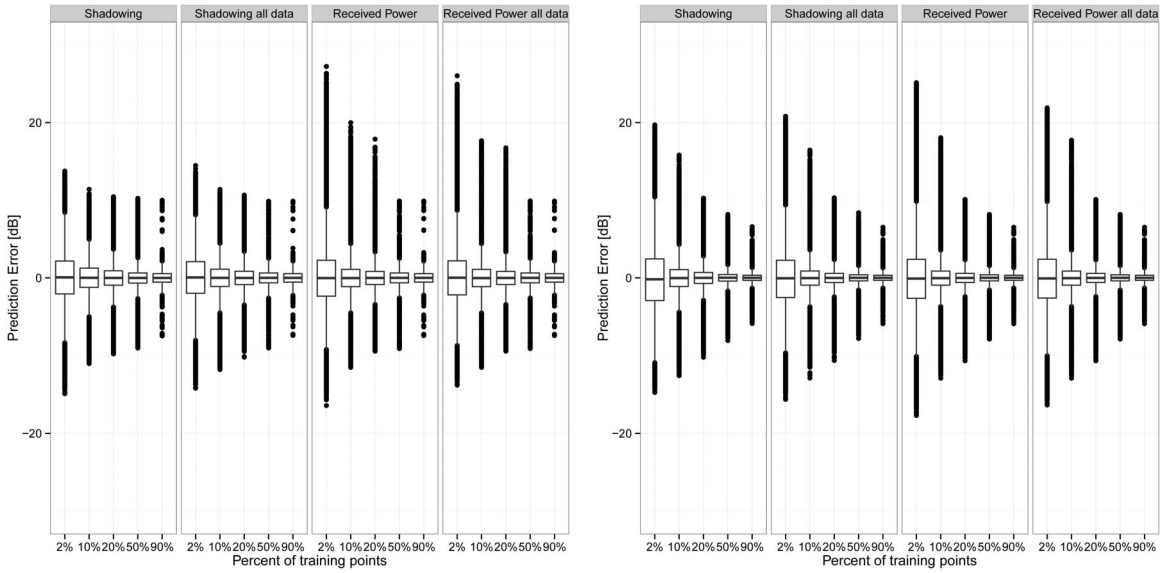


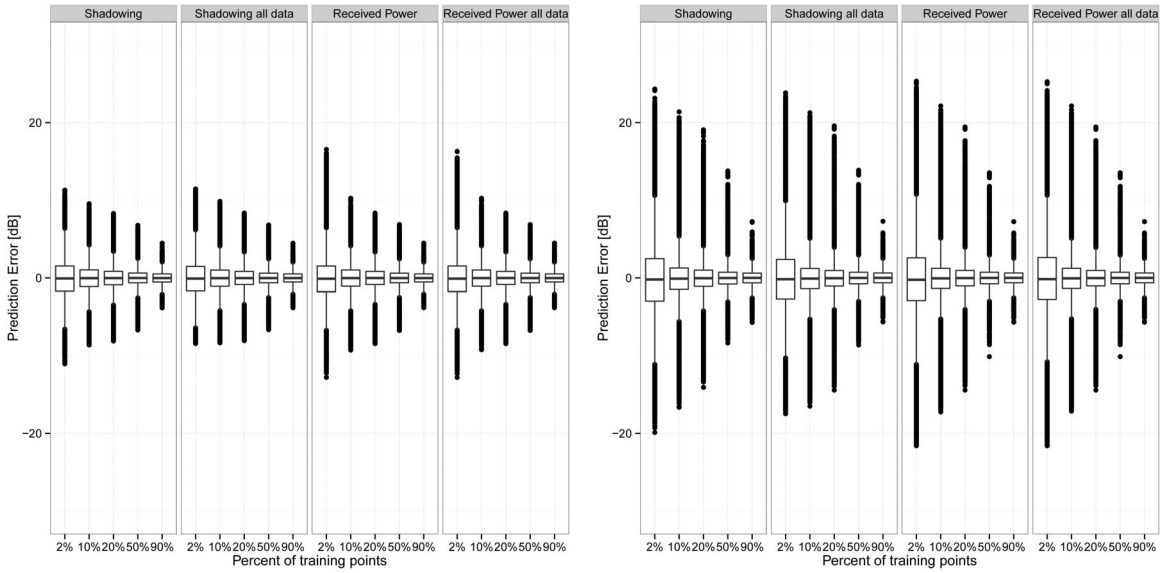
Fig. 6. Boxplots of the estimated sill for the shadowing covariance structure. The black line depicts the sill estimation from the entire data-set.

difference of the path loss exponent in the suburban case, is higher. The signal at 485 MHz experiences a higher path loss exponent than the signal at 2600 MHz. This is counter-intuitive as in higher frequencies we expect a more severe impact of the propagation environment. One explanation could be the unavoidable differences in the propagation environment, such as the foliage of the trees in the area, as the measurements for the two frequencies were conducted in different seasons of the year. More analysis has to be carried out though so as to come to conclusions for this matter.



(a) 485 MHz, Suburban

(b) 485 MHz, Urban



(c) 2600 MHz, Suburban

(d) 2600 MHz, Urban

Fig. 7. Prediction error for various percent of training points.

Kriging is using the estimated sill and range parameters, so as to achieve the spatial interpolation. In Figure 3 we present the semivariograms estimated *from the shadowing field* extracted from various percentage of training points ($20\%, 50\%$ and 90%), along with the semivariogram estimated from the entire data-set. First, it can be seen that even with small amount of data (20%) the semivariograms converge towards the semivariogram estimated from all the data. Moreover, the range (distance that the sill value is estimated for the first time and thus the distance after which the points are uncorrelated) is quite stable along all the cases of training points. This means that even with low number of observations the variogram provides quite good estimates of the range. On the other hand we see that the sill parameter can have a variance between 10 and 20 dB. The results show that the uncertainty on the sill does not have a high impact on the kriging process, even though there are estimations ranging more than 10 dB from the whole-data estimated value.

In Figure 4, the semivariograms estimated *from the received power* for $20\%, 50\%$ and 90% of training points are depicted as well as the semivariogram estimated from the entire data-set. As in the previous results, we have a convergence. Moreover, from these plots we can see the difference of the semivariograms extracted from the shadowing and the received power respectively. The semivariograms provided from the received power seem to follow the linear covariance model.

Nevertheless, we used the exponential model for the fitting as the linear covariance model, does not provide us with meaningful sill and range parameters.

Figures 5 and 6 depict the distribution of the range and the sill parameters respectively in the form of boxplots. The boxplot is a representation of the main characteristics of a distribution. The white box represents the first and third quartile of the distribution. The horizontal line in the square represents the median value of the data. The whiskers, i.e. the lines extended out of the box, define 1.5 times the inter-quartile distance. Last, the points outside the whiskers correspond to the outliers of the data. This is done for all the areas and frequencies and for different percentages of training points. The black lines correspond to the estimated value from the entire data-set for the shadowing cases. It can be seen that even for 10% of training points the estimation results in a distribution around the values estimated from the cases when all the data points where used. This is an indicator that the used data in most of the cases provide good enough estimates for the spatial correlation, even with less measurement points. As it is expected the number of data used for the estimation of the semivariogram has an impact on the number of outliers of the two parameters.

Finally, in Figure 7 the prediction errors for all the methods used in the analysis are depicted. By comparing the results of the kriging on shadowing (i.e. the results of the III-A under the *Shadowing* facet) with the results of the kriging on the shadowing of all the data (III-B, under the *Shadowing all data* facet) we see that more accurately estimated covariance structure does not lead to any significant improvement of the prediction error. On the contrary, even a small increase on the number of training points contributes in decreasing the first and third quartile distances as well as the predicted outliers. Similar are the results of the analysis for the received power (*Received Power* and *Received Power all data* facets). The variations of the prediction error for the kriging on the received power approaches (III-C and III-D) are marginal. Moreover, as the number of training points increase, the outliers along with the variation of the prediction error decrease. Another interesting point is that although from Figure 4 can be seen that the semivariograms follow the linear model and there is no bound on the estimated parameters, the prediction error is similar to the case of interpolating the shadowing field. This can be explained by the fact that even though the locations with higher distances have also high semivariance values for the received power case, kriging is heavily penalizing the points with higher distances from the point to be predicted and thus it is able to compensate for that and still predict correctly.

One last point to be made is that from Figure 7 we can confirm the results of [6] that the two different processes of kriging on the received power and kriging on the shadowing field, converge to the same prediction error distribution, that could be a point of interest for bounding the shadowing uncertainty. We note also that the kriging on received power does not make use of the information of the transmitter position and it has a lower overhead, since no mechanism for the extraction of the shadowing field is utilized.

V. Conclusions

In this paper, we have studied the impact of various spatial correlation and propagation parameters on the prediction error of the optimal linear predictor, that could be used as a tool for coverage planning and estimation for modern wireless systems. Since it is not clear if the simulation models can capture adequately the complex dependencies between the model errors and the coverage estimates, we used real measurement data, covering different environments (urban and suburban) and frequencies (485 MHz and 2600 MHz) for our analysis.

We showed that even with 20% of training points, the path loss exponent of the log-distance model, can be most of the times correctly estimated, for areas of $430\,\mathrm{m} \times 470\,\mathrm{m}$ and $370\,\mathrm{m} \times 410\,\mathrm{m}$. We have also demonstrated that the semivariogram is robust against the propagation uncertainties induced in the extraction of the shadowing field when low number of data points are available for the fitting of the path loss model. Finally, we illustrated that the prediction error of the kriging depend only marginally on the uncertainties of the covariance structure, but even small changes on the number of training points used, lead to high variances of the interpolation error. In other words, kriging due to the penalization of more distant points is able to compensate even for inaccurate covariance structures as the ones extracted from the received power (Figure 3) and still be able to perform adequately. In the future, we plan to analyze in detail the characteristics of the data that provide minimum prediction errors, when used as training points and investigate any trends they may have.

Acknowledgment

We would like to thank Alexandros Palaios for his valuable help on measurements and insightful discussions.

References

[1] T. S. Rappaport, *Wireless Communications: Principles and Practice 2nd Edition.* Prentice Hall, 2002.

[2] A. Alaya-Feki et al., "Interference cartography for hierarchical dynamic spectrum access," in *Proc. of IEEE DySPAN 2008*, 2008, pp. 1–5.

[3] C.Phillips et al., "Practical radio environment mapping with geostatistics," in *Proc. of IEEE DYSPAN 2012*. IEEE, 2012, pp. 422–433.

[4] J. Riihijärvi et al., "Characterization and modelling of spectrum for dynamic spectrum access with spatial statistics and random fields," in *Proc. of IEEE PIMRC 2008*. IEEE, 2008, pp. 1–6.

[5] H. Braham et al., "Coverage mapping using spatial interpolation with field measurements," in *Proc. of PIMRC 2014*, Sept 2014, pp. 1743–1747.

[6] N. Perpinias et al., "A measurement-based study on the use of spatial interpolation for propagation estimation," in *Proc. of IEEE ICC 2015*, June 2015, pp. 2715–2720.

[7] ——, "Measurements of Shadow Correlations in a Suburban Environment on the 485 MHz Band," in *Proc. of Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th.* IEEE, 2013.

[8] N. A. Cressie, *Statistics for Spatial Data.* John Wiley and Sons., 1993. Revised Edition.

[9] M. Wellens et al., "Spatial statistics and models of spectrum use," *Comput. Commun.*, vol. 32, no. 18, pp. 1998–2011, 2009.

[10] E. H. Isaaks and R. Srivastava, *Applied Geostatistics*. Oxford University Press, 1989.

[11] A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, *Handbook of Spatial Statistics.* Chapman & Hall/CRC, 2010.

[12] P. Ribeiro Jr. and P. J. Diggle, "geoR: A package for geostatistical analysis," *R-NEWS*, vol. 1, 2001.