Estimating Wireless Network Properties with Spatial Statistics and Models

Janne Riihijärvi, Petri Mähönen

Estimating Wireless Network Properties with Spatial Statistics and Models

Janne Riihijärvi and Petri Mähönen
Institute for Networked Systems, RWTH Aachen University
Kackertstrasse 9, D-52072 Aachen, Germany
email: {jar, pma}@inets.rwth-aachen.de

***Abstract*—In this paper we discuss the use of spatial statistics and models for different estimation problems related to wireless networks. We focus specifically on problems of state estimation, in which measurement data on only part of the system is available, and should be used to infer the state of the rest of the system. Such problems are becoming increasingly important for network operations and diagnostics, and are also emerging in the context of resource management of future wireless networks. We provide a concise survey of existing techniques from the spatial statistics literature, show how they can be applied in the context of wireless networks, and outline key research challenges for future work.**

I. Introduction

Recently there has been significant interest in the study of spatial structure of wireless networks. However, most of the work in the literature has been specifically focusing on understanding the properties a wireless network would have, if the locations of network nodes were arising from a Poisson point process. Few authors have also studied the properties of closely related wireless network models, such as various hard core and cluster processes. For an overview of these lines of work, we refer the reader to the recent survey of Haenggi et al. [1]. On the other hand, there has been relatively little work on studying and modeling of actual network deployments, not to mention applying spatial statistics and models in the context of operational networks.

In this paper we outline some of the theoretical and practical challenges in applying spatial modeling tools into empirical network data sets, in part based on our earlier work in this domain [2]–[5]. We give an overview with new case studies of the key techniques from the spatial statistics literature that can be used for this purpose. Our focus is specifically on identifying opportunities for reasoning about the state of a wireless network at runtime. While existing spatial statistics tools provide viable approaches for this, we also identify several inference problems that have not been discussed in the statistical or engineering literature until now. We believe tackling these challenges forms a highly interesting problem for the research community. This is both due to the complex but well-defined nature of the problems involved, as well as the potentially high impact these techniques could have for operations and management of future wireless networks.

The rest of this paper is structured as follows. In Section II we briefly discuss the mathematical background needed to formulate the problems discussed in the rest of the text. We then present in Sections III and IV small case studies on applications of spatial statistics and models on reasoning problems related to location data as well as continuous phenomena related to interference and coverage. Finally, we draw conclusions and outline challenges for future work in Section V.

II. Spatial Statistics of Radio Environment

For completeness and establishing notation we briefly introduce in this section point processes and random fields.

A. Statistics and Models for Point Processes

A point process is typically defined as a random counting measure, that is, as a map

$$N : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{M}), \tag{1}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and M is the space of counting measures defined on some metric space or region of interest W [6], [7]. The sigma-algebra $\mathcal{M}$ is chosen so that for each measurable $A \subset W$, $N(A)$ is an integer-valued random variable counting the number of points or node locations within A. A point process N_λ is said to be a *homogeneous Poisson point process of intensity* λ if $N_\lambda(A)$ is Poisson with parameter $\lambda|A|$, where $|A|$ is the area of A, and $N_\lambda(A)$ and $N_\lambda(B)$ are independent for any disjoint A and B. As discussed in the introduction, the homogeneous Poisson point process is widely used to model different wireless network types. However, for our purposes here more complex models are needed for two reasons. First, as shown in [2], [3] and in Section III below, actual network deployments tend not to be well modeled by homogeneous Poisson point processes. Second, even if we were to approximate a network structure with a Poisson model, such a model would be almost entirely uninformative regarding the structure of the network in regions in which no observations are available as $N_\lambda(B)$ would be by definition independent of $N_\lambda(W \setminus B)$.

We can specify more complex point process models, called *Gibbs point processes*, by defining a density function with respect to the *distribution* (pushforward measure) $\mathbb{P}N_\lambda^{-1}$ of a Poisson point process, usually further chosen to have $\lambda = 1$. Such a density is defined on the space of counting measures on W, and in essence quantifies how much more or less likely a given pattern of node locations is to occur in the model compared to the Poisson case. We have shown in [5] that the *Geyer saturation process* [8], a generalization of the Strauss

978-1-61284-824-2/2012 - Copyright is with IFIP

process [7], yields very good fits for a wide variety of data sets on transmitter locations. For the Strauss process we have the density

$$f(N) = \alpha\beta^{\#(N)}\gamma^{s_r(N)}, \tag{2}$$

where $\#(N)$ is the number of points in N and $s_r(N)$ is the number of point pairs of N that are closer than distance r apart. For the Geyer process an additional *saturation threshold* ζ is added, bounding the contribution of the exponent of γ. The case $\zeta \to \infty$ yields the Strauss process as a limit. Especially the parameter γ has an intuitive interpretation. If $\gamma = 0$, the likelihood of a point pattern arising with any pair of points closer than distance r apart becomes zero. Therefore, the process becomes an example of a pure *hard-core process*. As γ is increased, close by point pairs become possible, but are still less likely to occur than for the Poisson case. Finally, $\gamma = 1$ yields the Poisson point process as a limit. Increasing γ further will then yield clustered distributions.

B. Characterizing Interference and Spectrum Use

Received signal strength, interference power, and shadow fading are all examples of continuous spatial phenomena. All of these phenomena can be modeled by *random fields*, which are simply stochastic processes defined over the whole of W. First two are often studied using a shot noise representation, where an underlying point process model is combined with a propagation model to yield a random field. However, we shall argue below that it is often advantageous to adopt a more direct approach. A random field Z is typically characterized by its mean

$$\mu(\boldsymbol{s}) \equiv \mathbb{E}\{Z(\boldsymbol{s})\} \tag{3}$$

and the covariance function

$$C(\boldsymbol{s},\boldsymbol{t}) \equiv \mathbb{E}\{Z(\boldsymbol{s})Z(\boldsymbol{t})\} - \mu(\boldsymbol{s})\mu(\boldsymbol{t}). \tag{4}$$

Assuming second-order stationarity the mean becomes constant and $C(\boldsymbol{s},\boldsymbol{t})$ will depend only on the *lag* $\boldsymbol{\tau} = \boldsymbol{t} - \boldsymbol{s}$. For second-order stationary random fields we write the covariance function as $C(\boldsymbol{\tau})$. A weaker stationarity condition still is *intrinsic stationarity*, defined by $\mathrm{Var}\{Z(\boldsymbol{s}) - Z(\boldsymbol{t})\}$ being dependent on the lag alone. For intrinsically stationary systems one conventionally defines the *semivariogram*

$$\gamma(\boldsymbol{s}-\boldsymbol{t}) = \frac{1}{2}\mathrm{Var}\{Z(\boldsymbol{s}) - Z(\boldsymbol{t})\} \tag{5}$$

which is related for second-order stationary processes to the covariance function by

$$\gamma(\boldsymbol{\tau}) = C(\boldsymbol{0}) - C(\boldsymbol{\tau}). \tag{6}$$

If the semivariogram further depends on $\boldsymbol{\tau}$ by its norm alone, the random field is called *isotropic*.

We shall see below how these quantities can be estimated from empirical data sets, and how the resulting estimates can be used for state estimation problems in wireless networks.

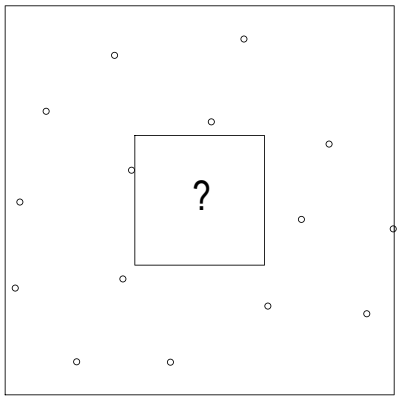

Fig. 1. Example configuration of node locations, with a region for which no observations are available located in the middle.

III. Reasoning about Node Locations with Point Process Models

Figure 1 illustrates a simple example of a reasoning problem we shall discuss in this section. We are given a collection $\{X_i\}$ of known locations in $W \setminus B$ for some compact region B, and we seek to estimate whether further locations are to be found in B, and how they would be distributed there. If $\{X_i\}$ are well modeled by a homogeneous Poisson process on $W \setminus B$ we simply need to estimate λ from the available observations. Usually the structure of $\{X_i\}$ is, however, more complex. For example, in the locations depicted in Figure 1, a significant degree of regularity can be observed.

For studying whether the homogeneous Poisson point process model is a feasible one, we can employ several statistical techniques [7], [9], [10]. In classical spatial statistics literature *nearest neighbor distances* were widely used. More detailed characterization of location data can be done by estimating various correlation functions between locations. These can be formally defined as follows. Any point process can be written as $N = \sum_{i=1}^{n} \varepsilon_{X_i}$, where X_i are W-valued random variables and ε_x denotes a point mass at x. Then, denoting

$$N^{(k)} \equiv \sum_{(i_1,\ldots,i_k)}^{\neq} \varepsilon_{(X_{i_1},\ldots,X_{i_k})}, \tag{7}$$

we can define the *factorial moment measures* of N by

$$\mu_N^{(k)} \equiv \mathbb{E}\{N^{(k)}\}. \tag{8}$$

The measures $\mu_N^{(k)}$ often admit density functions on W^n, called the *product density functions* $\lambda^{(k)}$. Finally, we can use product densities in normalized form to define various correlation functions. For example, the *pair correlation function* ξ is defined by

$$1 + \xi(x,y) \equiv \frac{\lambda^{(2)}(x,y)}{\lambda(x)\lambda(y)}. \tag{9}$$

If N is both stationary and isotropic, the pair correlation function will depend only on the distance r between x and y. For the homogeneous Poisson point process $\xi(r) \equiv 0$.

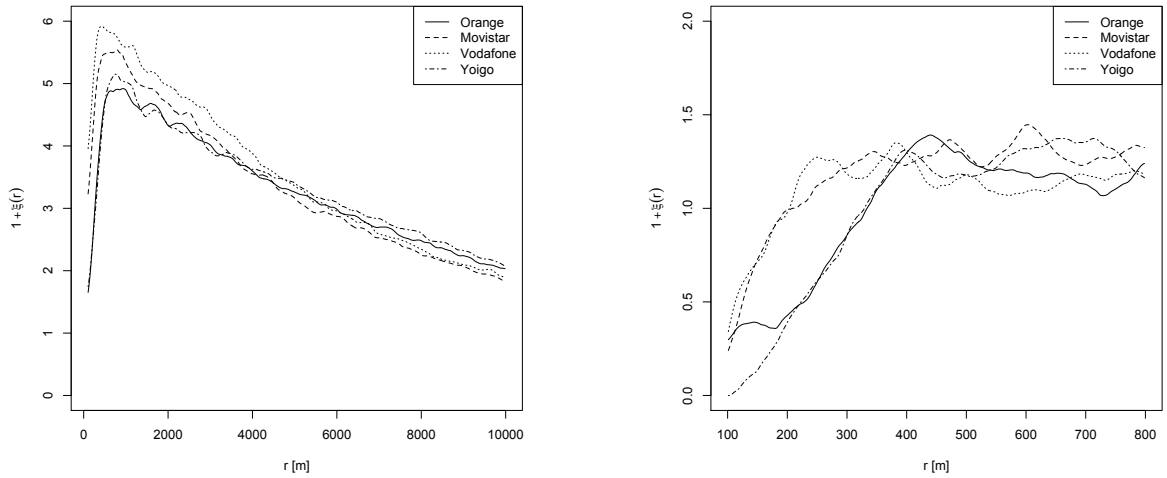

Fig. 2. The pair correlation function estimates for the UMTS cellular network sites in suburbs (left) and downtown Madrid (right) for different operators.

The pair-correlation does not, of course, determine the distribution of a point process uniquely. An illustrative example is the Baddeley and Silverman process [11] having the same second-order structure as the Poisson point processes while being structurally substantially different. Nevertheless, it has been found extremely powerful tool in a number of applications (especially so in astrophysics [12]) and by the virtue of being straightforward to estimate numerically warrants close attention. Figure 2 illustrates by way of example the estimates of the pair correlation function for various operator networks around Madrid, Spain at two different length scales. We see that the large-scale structure of the node locations is clearly clustered, due to inhomogeneities in the underlying population distribution, whereas on shorter length scales the locations are clearly highly regular, due to the careful planning and deployment optimization of the underlying network.

Given that we have ruled out the homogeneous Poisson model for our data set, a more refined model is needed. The local regularity in node locations can be easily captured by the Geyer and Strauss models introduced in the previous section. The model parameters can be estimated and the resulting fits validated using the techniques developed by Baddeley et al. [13]–[15]. Applications of these techniques for modeling wireless network node locations are described in [3], [5]. Once a satisfactory fit is obtained, the resulting model can be used to compute various estimates of the conditional distributions of locations in the middle region in Figure 1. Unfortunately many of these estimates are not tractable analytically, and numerical techniques are needed. One of the key advantages of using Gibbs models as our reasoning framework is precisely that such numerical techniques become rather straightforward to implement [16].

We shall now return to our original example given in Figure 1. Given the regular structure of the known node locations, a Strauss model was fitted to the location data, with parameter r estimated from the pair correlation function. Figure 3 shows the estimate of the conditional structure of locations within the central square, obtained using conditional simulation with the fitted model. The model-based approach clearly yields more insight to the possible location structure than what could be obtained by simple Poisson assumption.

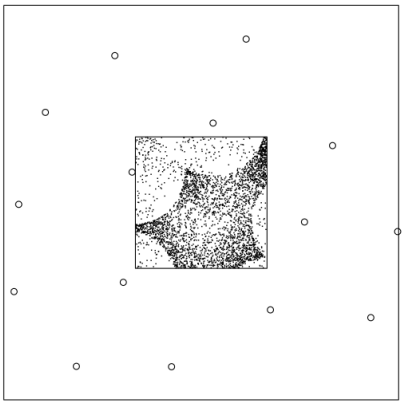

Fig. 3. Outcomes of 10^3 conditional simulation runs for the interior region of the data set of Figure 1 using the Strauss model for the node locations. The complex spatial structure of the conditional distribution is clearly visible.

IV. Coverage and Interference Estimation with Random Fields

We shall now move on to the case of reasoning about continuous phenomena over W, using interference power or received signal strength as an example. We shall treat the received power as a random field $Z(\boldsymbol{s})$, and assume that from measurements we are given a collection of samples $\{Z(\boldsymbol{s}_k)\}$. Our objective is to estimate based on these samples

the value of Z at some location $\boldsymbol{s}_0$ for which a measurement is not directly available. Such estimation problems are highly relevant for a variety of network management, diagnostics and planning problems.

Denoting our estimate by $\widehat{Z}(\boldsymbol{s}_0)$, we focus here on *linear* estimators due to their simplicity. Each such estimator can be written in the form

$$\widehat{Z}(\boldsymbol{s}_0) = \sum_{k=1}^{n} \lambda_k(\boldsymbol{s}_0) Z(\boldsymbol{s}_k) \tag{10}$$

with differences between the various estimators arising from the way the *weights* λ_k are determined. Natural conditions to impose on $\widehat{Z}(\boldsymbol{s}_0)$ are that the estimate is unbiased, that is, we have $\mathbb{E}\{\widehat{Z}(\boldsymbol{s})\} = \mathbb{E}\{Z(\boldsymbol{s})\}$ for all $\boldsymbol{s}$, and that amongst all the unbiased estimators it should have the minimum prediction error in terms of the variance of the estimator. In order to achieve these properties, we need to estimate the second order structure of Z from the available samples. We shall focus here on the estimation and use of the semivariogram instead of the covariance function due to the better robustness properties of the former [9].

The simplest estimator for $\gamma(\boldsymbol{\tau})$ is obtained by estimating the variance of $Z(\boldsymbol{s}) - Z(\boldsymbol{t})$ by the method of moments. While simple to implement the method of moments estimator is not robust against the presence of outliers. The most important of the alternatives is the *Cressie-Hawkins robust estimator* [17]

$$\hat{\gamma}(\boldsymbol{\tau}) \equiv \frac{1}{2}\left(\frac{1}{|N(\boldsymbol{\tau})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in N(\boldsymbol{\tau})} \{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\}^{1/2}\right)^4 \Bigg/ \left(0.457 + \frac{0.494}{|N(\boldsymbol{\tau})|}\right), \tag{11}$$

which can be further modified by replacing the averaging with the taking of the median, or applying robust estimators of scale [18]. We shall use this estimator in our case study below.

The estimated semivariogram is not yet sufficient for our estimation problem since it might fail to be negative-semidefinite. Therefore a parametric *semivariogram model* is usually required. The classical example is the *exponential semivariogram model* given by

$$\gamma(\tau; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1\left(1 - \exp\left(\frac{-\tau}{\alpha_2}\right)\right). \tag{12}$$

The parameter α_0 is used to account for the *nugget effect* $\lim_{\tau\to 0}\gamma(\tau) \neq 0$ possibly arising from measurement inaccuracies, short-range variations, or superimposed noise. The parameter α_1 controls the *sill*, that is, the asymptotic value of $\gamma(\tau;\boldsymbol{\alpha})$ when $\tau \to \infty$. Finally, α_2 is the *range parameter*, scaling the correlation distance of the model. For fitting the model parameters to the semivariogram estimate the method of choice is the *weighted least squares* scheme, for which the fitting criterion is given by

$$L_{\mathrm{WLS}}(\boldsymbol{\alpha}) = \big(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\alpha})\big)^T \boldsymbol{W}^{-1} \big(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\alpha})\big), \tag{13}$$

where $\boldsymbol{W}$ is now a diagonal matrix consisting of the variances of the entries of $\hat{\boldsymbol{\gamma}}$ [19].

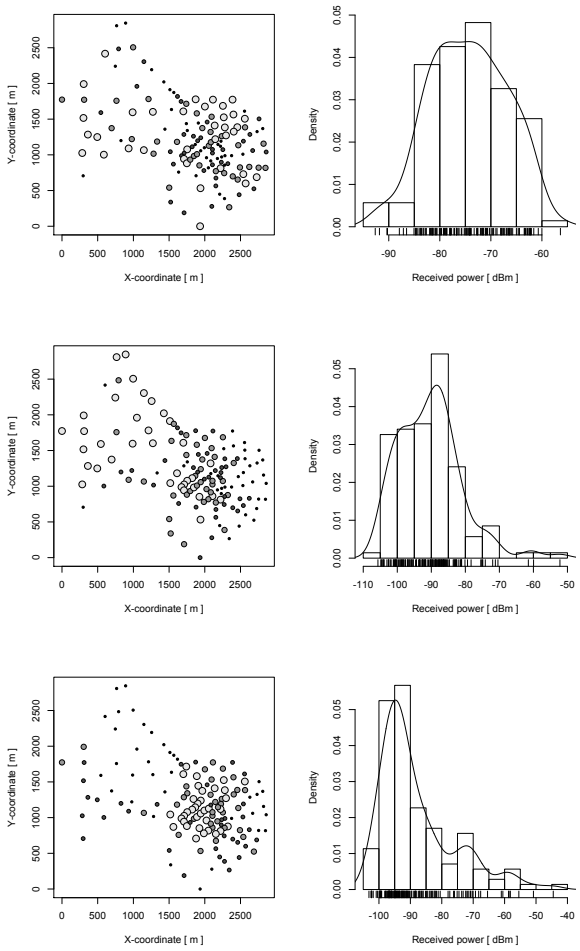

Fig. 4. Results from spatial spectrum use measurements for frequency bands of three cellular operators, denoted as operators A (top), B (middle) and C (bottom) in the following. Leftmost panels illustrate the spatial structure of the data sets while the rightmost panels show the marginal distributions of the data.

Once the semivariogram model has been obtained, we can write down our estimator (which corresponds to the *ordinary kriging* used extensively in geostatistics, see [9] for discussion and the derivation). Denoting $\tau_{ij} \equiv |\boldsymbol{s}_i - \boldsymbol{s}_j|$ the weights λ_k can be obtained from the matrix equation

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(\tau_{11}) & \dots & \gamma(\tau_{n1}) & 1 \\ \gamma(\tau_{12}) & \dots & \gamma(\tau_{n2}) & 1 \\ \vdots & \ddots & \vdots & \\ \gamma(\tau_{1n}) & \dots & \gamma(\tau_{nn}) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma(\tau_{01}) \\ \gamma(\tau_{02}) \\ \vdots \\ \gamma(\tau_{0n}) \\ 1 \end{pmatrix}$$

where the parameter μ is a Lagrange multiplier used in the minimization of the estimator variance.

We shall now illustrate how these methods work in practice, using as an example the data sets shown in Figure 4. These data sets were gathered during a measurement campaign using several spectrum analyzers, carried out in downtown Aachen,

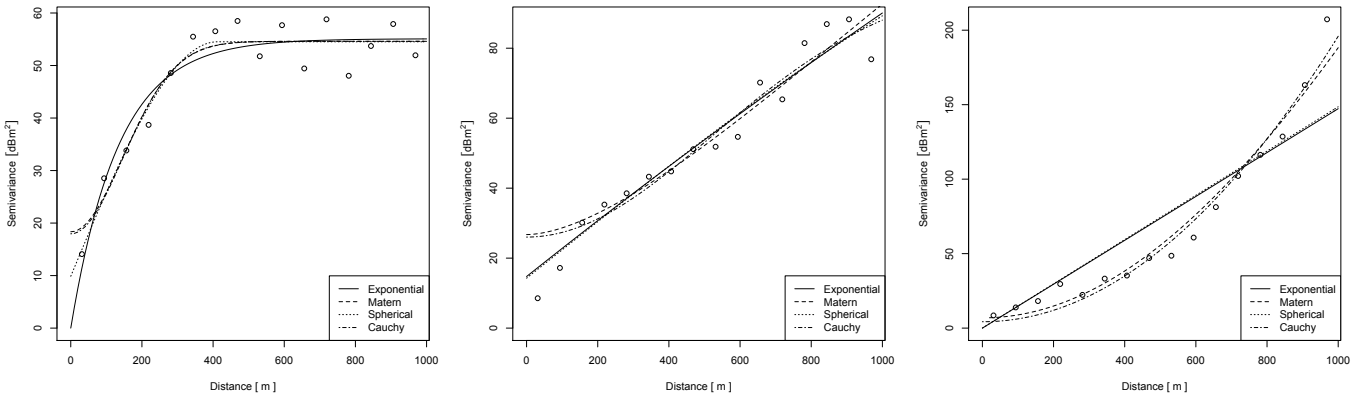

Fig. 5. Semivariogram fits for the operator A (left), B (middle) and C (right) data sets.

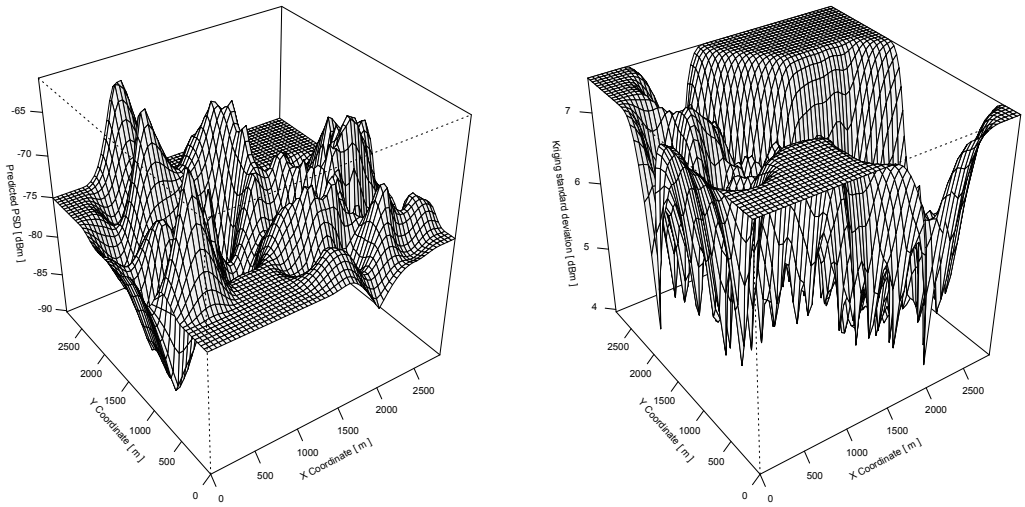

Fig. 6. Kriged estimates of the mean PSD field (left) and the standard deviations of the estimator (right) for the operator A data set.

Germany, in Summer of 2010. They consist of measured mean power levels at 134 locations for a number of cellular network frequency bands. Typical network management problem would be to estimate the coverage properties of the network in question based on such a collection of samples.

Figure 5 shows the semivariogram model fits for these three data sets, with the best fitting model used for estimation purposes in the following. For a more detailed description of the involved semivariogram models we refer the reader to [9]. Figure 6 illustrates the results of the estimation process for the operator A data set. Using the fitted semivariogram model kriging interpolation was applied, resulting in the estimate of the mean received signal strength surface shown on the left panel of the figure. As per usual properties of the kriging estimator, the interpolated surface is faithful to the data in the sense that each data point is also point of the interpolated surface, and at regions far enough of any data point the interpolated surface plateaus towards the mean of the estimates. Another useful property of the kriging estimator is that the variance or standard deviation of the estimates is available theoretically. These are shown in the right panel of the figure. Near the locations at which measurements were made the standard deviation approaches zero, whereas far away it asymptotically approaches the standard deviation of the underlying random field.

Another very interesting question is how to characterize the performance of such estimators. One approach is to use cross-validation statistics, or "leave one out" approach. This is based on estimating the value of the field at a location where a measurement is available, *without* using that particular measurement point in the estimation process (otherwise, e.g., the kriging estimator would always have zero error). The process is then repeated for all the measurement points. The results for the operator A data set are shown in Figure 7. The marginal distribution of the estimation errors follows rather well the expected normal distribution, and the accuracy

of the interpolated results is comparable to or even better than those of classical propagation models. However, the use of the marginal distribution alone might result in too pessimistic estimates, as much of the contribution at the tails is induced by points at the edge of the measurement areas, which would be expected to be difficult to estimate accurately in any case. Further insight into the error distribution and other characteristics of the estimates of the underlying random field can again be obtained by means of conditional simulation.

V. Conclusions and Research Challenges

We strongly believe that the model-based approaches discussed above form a highly promising starting point for future work on reasoning about the various properties of the radio environment. Nevertheless, there are several research challenges that should be met in order to make these methods practical for on-line applications. First, we did not consider the role of measurement errors in our location data. Especially if locations of network nodes are inferred using signal strength based localization techniques, significant residual error is unavoidable. There is a clear need to incorporate such measurement errors into the fitting process, as well as in the evaluation of the goodness of fit of the arising models. Also, we carried out all estimates in a batch manner, with complete data set at our disposal. In a number of applications these locations would become known either one at a time, or we would only be able to observe some functional of the observed part of the point process. Some theoretical foundations for the related inference problems are available [6], but no applications to problems in the wireless domain have been worked out in detail to the best of our knowledge. Especially the problem of fitting a Gibbs model using only such observations is a highly interesting problem both from the statistical viewpoint as well as regarding impact in potential applications.

Another interesting challenge is to bridge the two problem domains discussed above. Often the samples of the random fields of interest are of the form $Z(X_i)$, where the X_i are points in a realization of a point process N. While some work exists on state estimators for Z based on Poisson samples, rather little seems to be known on the properties of samples obtained using a more general point process. It is clear that by exploiting the structure of N and Z improved sampling procedures can be devised. Such problems are of significant interest especially related to diagnostics of wireless networks, such as discussed in the Minimization of Drive Tests (MDT) work item in 3GPP.

Acknowledgment

The authors would like to thank RWTH Aachen University and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for providing financial support through the UMIC research centre. We would also like to thank the European Union for providing partial funding of this work through the FARAMIR project.

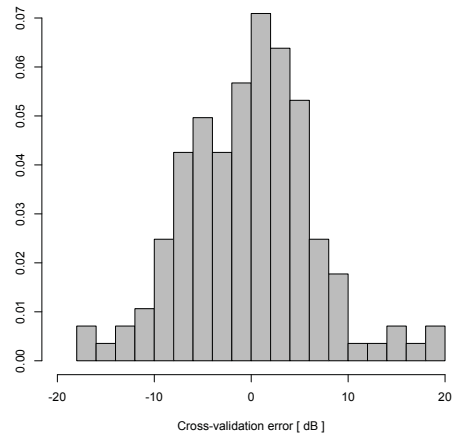

Fig. 7. The distribution of cross-validation errors for operator A data set.

References

[1] M. Haenggi et al., "Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Networks," *IEEE Journal on Selected Areas in Communications*, Sep. 2009.

[2] J. Riihijärvi, P. Mähönen, and M. Rübsamen, "Characterizing wireless networks by spatial correlations," *IEEE Communications Letters*, vol. 11, no. 1, pp. 37–39, January 2007.

[3] J. Riihijärvi and P. Mähönen, "Exploiting Spatial Statistics of Primary and Secondary Users towards Improved Cognitive Radio Networks," in *Proc. of CROWNCOM'08*, Singapore, May 2008.

[4] M. Wellens, J. Riihijärvi, and P. Mähönen, "Spatial statistics and models of spectrum use," *Elsevier Computer Communications Journal*, in press.

[5] J. Riihijärvi and P. Mähönen, "Modeling Spatial Structure of Wireless Communication Networks," in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, 2010, pp. 1–6.

[6] A. F. Karr, *Point Processes and Their Statistical Inference*, 2nd ed. Marcel Dekker, 1991.

[7] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. Wiley, 1995.

[8] C. Geyer, "Likelihood inference for spatial point processes," *Stochastic Geometry: Likelihood and Computation*, vol. 80, pp. 79–140, 1999.

[9] N. Cressie, *Statistics for spatial data*. Wiley-Interscience, 1993.

[10] B. D. Ripley, *Statistical Inference for Spatial Processes*. Cambridge University Press, 1988.

[11] A. Baddeley and B. Silverman, "A Cautionary Example on the Use of Second-Order Methods for Analyzing Point Patterns," *Biometrics*, vol. 40, no. 4, pp. 1089–1093, 1984.

[12] P. Peebles, *The large-scale structure of the universe*. Princeton Univ Pr, 1980.

[13] A. Baddeley, J. Møller, and A. Pakes, "Properties of residuals for spatial point processes," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 3, pp. 627–649, 2008.

[14] A. Baddeley, R. Turner, J. Møller, and M. Hazelton, "Residual analysis for spatial point processes," *Journal of the Royal Statistical Society (series B)*, vol. 67, pp. 1–35, 2005.

[15] A. Baddeley and R. Turner, "Practical Maximum Pseudolikelihood for Spatial Point Patterns (with Discussion)," *Australian & New Zealand Journal of Statistics*, vol. 42, no. 3, pp. 283–322, 2000.

[16] J. Møller and R. Waagepetersen, *Statistical inference and simulation for spatial point processes*. CRC Press, 2004, vol. 100.

[17] N. Cressie and D. Hawkins, "Robust estimation of the variogram: I," *Mathematical Geology*, vol. 12, no. 2, pp. 115–125, 1980.

[18] M. Genton, "Highly robust variogram estimation," *Mathematical Geology*, vol. 30, no. 2, pp. 213–221, 1998.

[19] N. Cressie, "Fitting variogram models by weighted least squares," *Mathematical Geology*, vol. 17, no. 5, pp. 563–586, 1985.