Optimal Predictive Design Augmentation for Spatial Generalised Linear Mixed Models

Evangelos Evangelou[1,*], Zhengyuan Zhu[2,†]

1. Department of Mathematical Sciences – University of Bath, Bath, UK
2. Department of Statistics – Iowa State University, Ames, Iowa, USA

Abstract: A typical model for geostatistical data when the observations are counts is the spatial generalised linear mixed model. We present a criterion for optimal sampling design under this framework which aims to minimise the error in the prediction of the underlying spatial random effects. The proposed criterion is derived by performing an asymptotic expansion to the conditional prediction variance. We argue that the mean of the spatial process needs to be taken into account in the construction of the predictive design, which we demonstrate through a simulation study where we compare the proposed criterion against the widely-used space-filling design. Furthermore, our results are applied to the Norway precipitation data and the rhizoctonia disease data.

Keywords: Generalised linear mixed models; Geostatistics; Predictive inference; Sampling design.

1 Introduction

One of the most frequently used models for the analysis of geostatistical count data is the spatial generalised linear mixed model (SGLMM) (Diggle et al., 1998). Applications of SGLMM include Diggle et al. (1998) who looked into residual contamination from nuclear weapons testing and campylobacter infections in UK, Diggle et al. (2002) who studied the risk of malaria in Gambia, Zhang (2002) who analysed a root disease in an agricultural study, and Eidsvik et al. (2009) who examined precipitation data for the purpose of weather forecasting and for operating hydropower plants. This class of models assumes the existence of an unobserved Gaussian random field over the region of interest and that the observations, drawn at fixed locations, are conditionally independent given the value of the random field. The distribution of the random field may depend on unknown parameters and among the objectives is to use the sample to predict the value of the random field at every location in the region. The "plug-in" approach is a common method for prediction in these models from a frequentist point of view (Christensen, 2004; Evangelou et al., 2011) where in the first stage an estimate of the model parameters is obtained and in the second stage the predictive distribution of the random field is constructed conditional on the data and the parameter estimates. (Alternatively see Diggle et al. (1998), Christensen and Waagepetersen (2002) and Eidsvik et al. (2009) for a Bayesian solution.)

The objective of *spatial predictive design* (Zimmerman, 2006; Zhu and Stein, 2006) is to select the sampled locations within the region of interest in order to optimise, in some sense, the predictive capability of the sample. In summary, the strategy of optimal design comes down to developing some optimality criterion, such as the average prediction variance, and then searching over all possible sampling configurations for the optimal value of the criterion. We focus on the case where the sampling has already taken place at some locations and the data are available but either because there are more resources available or because the prediction error is too high more data are to be sampled. That is, we seek to augment the current sampling scheme with new locations after using the available data to infer about the parameters of the random field.

Date: June 5, 2012

*Corresponding author: ee224@bath.ac.uk

†Partially supported by NSF DMS 0605434 grant.

This scenario was also studied by Diggle and Lophaven (2006) in which they referred to it as *retrospective design*.

There is a large volume of the literature which is concerned with optimal predictive designs for Gaussian models (e.g. Cressie et al., 1990; Martin, 2001; Müller, 2007; Heuvelink et al., 2010) which tend to be uniformly spaced. Some development towards optimal SGLMM designs has been made in Diggle and Lophaven (2006) where they discuss the minimisation of the Bayesian prediction variance for a design criterion as a natural method for incorporating parameter estimation and prediction in the design, and in Zhang and Wang (2002) where they consider minimisation of the mean square prediction error. A drawback in applying the ideas of optimal experimental design in the context of SGLMM is that in this case the optimality criteria do not exist in closed-form. One solution is to evaluate them using Monte-Carlo approximations, but this approach is too computationally intensive due to the fact that in practice the design criterion needs to be evaluated for a potentially large number of candidate designs in search for the optimal one. Alternatively, Eidsvik et al. (2009) approximated the prediction variance by combining ideas from Rue et al. (2009) and direct Monte-Carlo integration.

Motivated by the fact that optimal predictive designs for Gaussian models are uniformly spaced, Nychka and Saltzman (1998) and Royle and Nychka (1998) suggested the so-called *space-filling design*. The idea is to propose a model-free approach from a wholly geometric point of view where the sampling sites are chosen in a way that the region of interest is covered uniformly and therefore there is no need of numerically evaluating the estimation and prediction error. Moreover, space-filling designs tend to be very similar to Gaussian optimal designs (Royle and Nychka, 1998); however, they do not necessarily perform well in the context of SGLMM. The reason is that, contrary to the Gaussian model, the prediction error not only depends on the variance of the underlying process but also on its value. Hence if the random field varies highly within the region, a space-filling design would not be appropriate. To support this hypothesis, let us consider the following example. Suppose that the observations are binary with probability of "success" varying from 0 to 1 across the region of interest. If a space-filling design is implemented part of the sample will be associated with areas with very high or very low success probabilities; however, the data from these areas exhibit almost no variability and, in essence, are not very useful. We believe that, in choosing a good design, more weight should be given in areas where the probability of success is closer to half, i.e. where the variability of the data is higher.

In this article, we propose a model-based criterion for constructing optimal predictive designs based on an approximation to the conditional predictive variance. We assume that prior data are available and the objective is to augment the given sampling network. Our strategy proceeds as follows. By fitting an SGLMM to the data we construct a prediction map for the conditional mean and variance of the spatial random field using an asymptotic approximation to the conditional predictive distribution (Evangelou et al., 2011). These estimates are used in the evaluation of the proposed design criterion, and an exchange algorithm is implemented to search for the best design. We point out that although the ideal would be to minimise the conditional prediction variance, the proposed approximate criterion is able to capture the overall pattern of the optimal design.

In the next section we describe the SGLMM and derive an approximation to the conditional predictive variance which we use as our design criterion. In Section 3 we present our implementation of the exchange algorithm used for searching for the optimal design and in Section 4 we illustrate our method through a computational study and two examples. Finally, Section 5 presents a summary of the conclusions of this article.

2 An approximate predictive design criterion

2.1 Model

We assume that observations $\boldsymbol{y} = \{y_1, \ldots, y_k\}$ are taken from a spatial domain $\mathbb{S}$ which can be modelled by a SGLMM, that is, there exists a Gaussian random field $\mathcal{Z}$ defined over $\mathbb{S}$ and the observations are conditionally independent given the value of the random field with distribution from the exponential family. Our objective is to predict $\mathcal{Z}$ for given observations $\boldsymbol{y}$, drawn at k fixed, distinct locations $S = \{s_1, \ldots, s_k\} \in \mathbb{S}$ called the *sampling design*. Furthermore, each y_i corresponds to repeated sampling of size n_i from location s_i. In the binomial case n_i may be interpreted as the number of trials in a Bernoulli experiment and in the Poisson case as the length of time that the sampling is taking place.

Furthermore, it is common to express the mean of the random field at location s as a linear combination of p explanatory variables $\mathbf{f}_s$, i.e.

$$\lambda_s \coloneqq \mathsf{E}\, \mathcal{Z}(s) = \mathbf{f}_s^{\mathsf{T}} \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of regressor coefficients.

The geostatistical approach to SGLMM assumes further that the covariance between two components of the random field, $\mathcal{Z}(s)$ and $\mathcal{Z}(r)$, at locations $s, r \in \mathbb{S}$ is a function of the distance $\|s - r\|$. In particular we denote

$$\begin{aligned} \sigma^2 &\coloneqq \mathsf{Var}(\mathcal{Z}(s)) = \nu^2 + \tau^2, \\ \sigma_{s,r} &\coloneqq \mathsf{Cov}(\mathcal{Z}(s), \mathcal{Z}(r)) = \tau^2 c(\|s - r\|; \phi), \end{aligned}$$

where (ν^2, τ^2, ϕ) are covariance parameters. ν^2 and τ^2 are termed *nugget* and *partial sill* respectively, ϕ is the correlation parameter and $c(\cdot)$ is a positive definite function defined on $(0, \infty)$ called the *correlogram*. Two particular versions of the correlogram that we make use in the examples of Section 4 are the following.

- *Exponential*: $c(h; \phi) = \exp(-h/\phi)$,
- *Spherical*: $c(h; \phi) = 1 - 1.5\, \frac{h}{\phi} + 0.5 \left(\frac{h}{\phi}\right)^3$ if $0 < h < \phi$, and $c(h; \phi) = 0$ if $h \geq \phi$.

In these cases ϕ is also referred to as the *range* parameter. Other correlograms such as Matérn and power exponential are also popular choices for spatial modelling. Although we do not use these in our paper, the methodology presented covers these cases as well.

Let $\boldsymbol{z} = \mathcal{Z}(S) = (z_1, \ldots, z_k)$ denote the components of the random field associated with the locations in the sampling design S and let $\boldsymbol{\lambda}_S$ and Σ_S denote its mean vector and the variance-covariance matrix respectively. Also let $f(\cdot)$ denote the probability density/mass function of its arguments so that

$$f(\boldsymbol{z}; \boldsymbol{\lambda}_S, \Sigma_S) = (2\pi)^{-\frac{k}{2}} |\Sigma_S|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{z} - \boldsymbol{\lambda}_S)^{\mathsf{T}} \Sigma_S^{-1} (\boldsymbol{z} - \boldsymbol{\lambda}_S) \right\},$$

and

$$f(y_i | z_i) \propto \exp\{y_i z_i - n_i \psi(z_i)\},\ i = 1, \ldots, k,$$

where $\psi(\cdot)$ is a known function; for example under the binomial model with canonical link $\psi(z) = \log(1 + e^z)$ and for the Poisson model with canonical link $\psi(z) = e^z$ (McCullagh and Nelder, 1999). The mean, $\boldsymbol{\lambda}_S$, is expressed as a linear combination of p regressors F_S, i.e. $\boldsymbol{\lambda}_S = F_S \boldsymbol{\beta}$ where F_S is the $k \times p$ design matrix.

2.2 The conditional distribution of the random field

We assume increasing-domain asymptotics in the spirit of Mardia and Marshall (1984), i.e. that $k \to \infty$ and that the rows of the variance-covariance matrix Σ_S are absolutely summable as $k \to \infty$. Furthermore we assume that $k/n_i \to 0$ as $k \to \infty$. The latter assumption is made in order to facilitate the application of Laplace approximation to the conditional distribution of the random field given the observations (see Shun and McCullagh, 1995; Evangelou et al., 2011).

In the following, a collection of sampling locations will be denoted by the capital letters Q, R, S, while lower case q, r, s will denote a single location. Data $\boldsymbol{y}$ are sampled at locations $S = \{s_1, \ldots, s_k\}$. A symbol with a subscript Q, R, S, q, r, s will be used to refer to the elements of the symbol associated with these locations.

Consider the question of predicting the value of the random field, $\mathcal{Z}$, from $\boldsymbol{y}$ at a finite set of locations $R \subset \mathbb{S}$. Let

$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{z}}(\boldsymbol{y}) \coloneqq \operatorname*{argmax}_{\boldsymbol{z}} f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\lambda}_S, \Sigma_S), \tag{1}$$

and

$$\begin{aligned} \hat{H} &= -\frac{\partial^2}{\partial \boldsymbol{z} \partial \boldsymbol{z}^\mathsf{T}} \log f(\boldsymbol{y}, \hat{\boldsymbol{z}}; \boldsymbol{\lambda}_S, \Sigma_S) \\ &= \hat{D}_S + \Sigma_S^{-1}, \end{aligned}$$

where $\hat{D}_S = \operatorname{Diag}\{n_i \psi''(\hat{z}_i)\}$. Here and subsequently a hat over a symbol will denote that the corresponding symbol is a function of the data $\boldsymbol{y}$ through $\hat{\boldsymbol{z}}$.

Then by an application of Laplace approximation in the spirit of Shun and McCullagh (1995), the conditional distribution of $\boldsymbol{z}|\boldsymbol{y}$ is approximately normal with mean $\hat{\boldsymbol{\mu}}_S \coloneqq \hat{\boldsymbol{z}}$ and variance-covariance matrix $\hat{\Upsilon}_S \coloneqq \hat{H}^{-1}$, written

$$\boldsymbol{z}|\boldsymbol{y} \overset{\text{approx}}{\sim} N_k(\hat{\boldsymbol{z}}, \hat{H}^{-1}). \tag{2}$$

For finite $R \subset \mathbb{S}$, let $\boldsymbol{\lambda}_R = \mathsf{E}\, \mathcal{Z}(R)$, $\Sigma_R = \mathsf{Var}\, \mathcal{Z}(R)$, $\Sigma_{S,R} = \mathsf{Cov}(\mathcal{Z}(S), \mathcal{Z}(R))$, $\Sigma_{R,S} = \Sigma_{S,R}^\mathsf{T}$. Then by the joint normality of $(\mathcal{Z}(S), \mathcal{Z}(R))$, the conditional distribution of $\mathcal{Z}(R)|\boldsymbol{y}$ is approximately normal with mean

$$\hat{\boldsymbol{\mu}}_{R|S} \coloneqq \boldsymbol{\lambda}_R + \Sigma_{R,S}\Sigma_S^{-1}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\lambda}_S), \tag{3}$$

and variance (see Appendix A)

$$\hat{\Upsilon}_{R|S} \coloneqq \Sigma_R - \Sigma_{R,S}(\hat{D}_S^{-1} + \Sigma_S)^{-1}\Sigma_{S,R}. \tag{4}$$

The order of the approximations in (3) and (4) is $O(kn^{-2})$ where $n = \min\{n_1, \ldots, n_k\}$. (For details about the order of the approximation see Evangelou et al. (2011).) Also note that (4) depends on the mean of the random field through the matrix $\hat{D}_S^{-1}$ but as $n \to \infty$ the elements of this matrix tend to 0 and (4) equals the prediction variance under the Gaussian model. Thus, if we accept that a space-filling design is very similar to the optimal design under the Gaussian model, then for large n, the optimal design for SGLMM prediction is a space-filling design.

Consider the expectation of $\hat{\boldsymbol{\mu}}_{R|S}$ and $\hat{\Upsilon}_{R|S}$ with respect to the distribution of $\boldsymbol{y}$. Since both quantities depend on $\boldsymbol{y}$ only through $\hat{\boldsymbol{z}}$, the expectation may be taken with respect to the distribution of $\hat{\boldsymbol{z}}$. This distribution is approximately equal, up to first order, to the unconditional distribution of $\boldsymbol{z}$, namely $N_k(\boldsymbol{\lambda}_S, \Sigma_S)$. Therefore, $\mathsf{E}\, \hat{\boldsymbol{\mu}}_{R|Q} = \boldsymbol{\lambda}_R$, as one would expect. For the conditional variance, let $K_S = \mathsf{E}\, \hat{D}_S^{-1}$, and note that by an application for the formula for matrix inversion,

$$(I - A)^{-1} \approx I + A + A^2 + \ldots,$$

where I is the identity matrix, we have

$$\begin{aligned}\mathsf{E}\{(\hat{D}_S^{-1}+\Sigma_S)^{-1}\} &= \Sigma_S^{-1}\,\mathsf{E}\{(I+\hat{D}_S^{-1}\Sigma_S^{-1})^{-1}\}\\ &\approx \Sigma_S^{-1}\,\mathsf{E}\{(I-\hat{D}_S^{-1}\Sigma_S^{-1})\}\\ &= \Sigma_S^{-1}(I-K_S\Sigma_S^{-1})\\ &\approx (K_S+\Sigma_S)^{-1}.\end{aligned}$$

The remainder term in the second equality above is $A^2+A^3+\ldots$ with $A=-\hat{D}_S^{-1}\Sigma_S^{-1}$ has elements of order $O(n^{-1})$. Then A^m has elements of order $O(k^{m-1}n^{-m})$, and similar arguments hold for the last equality, so the remainder term above has order $O(kn^{-2})$. Therefore,

$$\mathsf{E}\,\hat{\Upsilon}_{R|S} \approx \Sigma_R - \Sigma_{R,S}(K_S+\Sigma_S)^{-1}\Sigma_{S,R}. \tag{5}$$

2.3 Prediction variance

A measure of the uncertainty in prediction of the random field at location s is the *conditional prediction variance*, $\mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y};S)$ where in the notation we make explicit the dependence of the variance on the sampling design S. The *integrated conditional prediction variance* over $\mathbb{S}$ for given data $\boldsymbol{y}$ is defined as

$$\frac{1}{|\mathbb{S}|}\int_{\mathbb{S}}\mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y};S)\,\mathrm{d}s, \tag{6}$$

where $|\mathbb{S}|$ denotes the volume of $\mathbb{S}$. Harville and Jeske (1992), Zimmerman and Cressie (1992), and Booth and Hobert (1998) are among those who advocate using (6) as a measure of the prediction accuracy. To that end, a suitable criterion for choosing a good sampling design is to make (6) small; however, in theory, $\mathbb{S}$ is a continuous region so prediction at all locations in it is infeasible. In practice prediction is performed over a finite, fine grid, $\bar{\mathbb{S}}$, covering $\mathbb{S}$, and the optimality criterion reduces to

$$\frac{1}{|\bar{\mathbb{S}}|}\sum_{s\in\bar{\mathbb{S}}}\mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y};S). \tag{7}$$

The criterion in (7) has been extensively used for the derivation of sampling designs for Gaussian models (e.g. Cressie et al., 1990). However, for SGLMM the prediction variance is not known in closed-form so exact calculation of (7) is impossible (Booth and Hobert, 1998). Below we derive an approximation in closed-form to the prediction variance which will be used for defining our design criterion.

For a single location $s\in S$, we know from (4) that the conditional prediction variance is approximately equal to

$$\mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y};S) \approx \sigma^2 - \boldsymbol{c}_S^{\mathsf{T}}(\hat{D}_S^{-1}+\Sigma_S)^{-1}\boldsymbol{c}_S, \tag{8}$$

where $\sigma^2=\mathsf{Var}(\mathcal{Z}(s))=\Sigma_{s,s}$ is the unconditional variance of the random field at location s, and $\boldsymbol{c}_S=\mathsf{Cov}(\boldsymbol{z},\mathcal{Z}(s))=\Sigma_{S,s}$.

Now suppose that we are looking to augment our sampling scheme with an additional set of l new locations $Q=\{q_1,\ldots,q_l\}\subset\bar{\mathbb{S}}\setminus S$ from where we will collect new data $\boldsymbol{x}$. Consider the expected prediction variance

$$\mathsf{E}_{\boldsymbol{x}}\,\mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y},\boldsymbol{x};S,Q) \approx \sigma^2 - \boldsymbol{c}_{SQ}^{\mathsf{T}}\,\mathsf{E}_{\boldsymbol{x}}(\tilde{D}_{SQ}^{-1}+\Sigma_{SQ})^{-1}\boldsymbol{c}_{SQ}, \tag{9}$$

from (8), where $\tilde{D}_{SQ}=\mathrm{BlockDiagonal}\{\tilde{D}_S,\tilde{D}_Q\}$, $\boldsymbol{c}_{SQ}$ and Σ_{SQ} denote the augmented covariance vector and variance-covariance matrix respectively, and the expectation is with respect to the

conditional distribution of $\boldsymbol{x}|\boldsymbol{y}$. Here and below a tilde over a symbol indicates that it is a function of

$$\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{z}}(\boldsymbol{y}, \boldsymbol{x}) \coloneqq \underset{\boldsymbol{z}}{\operatorname{argmax}}\, f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\lambda}_{SQ}, \Sigma_{SQ}).$$

Since $\tilde{D}_{SQ}$ depends on $\boldsymbol{x}$ only through $\tilde{\boldsymbol{z}}$, the expectation in (9) may be replaced by the expectation with respect to the distribution of $\tilde{\boldsymbol{z}}$, i.e.

$$\mathsf{E}_{\boldsymbol{x}}\, \mathsf{Var}(\mathcal{Z}(s)|\boldsymbol{y}, \boldsymbol{x}; S, Q) \approx \sigma^2 - \boldsymbol{c}_{SQ}^{\mathsf{T}}\, \mathsf{E}_{\tilde{\boldsymbol{z}}}(\tilde{D}_{SQ}^{-1} + \Sigma_{SQ})^{-1} \boldsymbol{c}_{SQ}. \tag{10}$$

By (2) the conditional distribution of $\tilde{\boldsymbol{z}}$ for given data $\boldsymbol{y}$ is approximately normal with mean and variance $\tilde{\mu}_{SQ|S}$ and $\tilde{\Upsilon}_{SQ|S}$ respectively. On the other hand, exact evaluation of the expectation in (10) is still cumbersome unless some numerical method is used, such as Monte-Carlo (MC) integration. Apart from the fact that this method is time consuming, there is the question of how large should the size of the MC sample be in order to make the error in the MC integration sufficiently small. For example, under the setting of the example in Section 4.1, an MC sample size of 1000 was highly inefficient — the correct answer was obtained only 51% of the time — while an MC sample size of 2500 always gives the correct answer. Therefore, we are not suggesting evaluating (10) using MC integration as the design criterion, and only use it as a benchmark in our simulation study.

Alternatively, an approximation may be used by the same arguments that lead to (5). We therefore propose as a measure of prediction uncertainty the quantity

$$\widehat{\mathsf{Var}}(\mathcal{Z}(s)|\boldsymbol{y}; S, Q) = \sigma^2 - \boldsymbol{c}_{SQ}^{\mathsf{T}}(\hat{K}_{SQ|S} + \Sigma_{SQ})^{-1} \boldsymbol{c}_{SQ}, \tag{11}$$

where $\hat{K}_{SQ|S} = \text{BlockDiagonal}\{\hat{D}_S^{-1}, \hat{K}_Q\}$, $\hat{K}_Q = \text{Diag}\{n_i^{-1}\, \mathsf{E}(1/\psi''(W_i)),\ i = 1, \ldots, l\}$ and $W_i \sim N(\hat{\mu}_{q_i|S}, \hat{v}^2_{q_i|S})$ with $\hat{\mu}_{q_i|S}$ being the ith element, and $\hat{v}^2_{q_i|S}$ being the ith diagonal element of $\hat{\boldsymbol{\mu}}_{Q|S}$ and $\hat{\Upsilon}_{Q|S}$ respectively, i.e. the prediction and its variance at q_i given the data $\boldsymbol{y}$ as defined in (3) and (4).

The expression in (11) equals the prediction variance (9) to order $O(kn^{-2})$ and the fact that it is positive is a nice property. A computational advantage of (11) in comparison to (10) is that it maintains a block structure in components associated with S and Q which makes the inversion of the matrix in the parentheses less cumbersome during the search for the optimal design. Moreover, in certain cases the expectation in $\hat{K}_Q$ may be computed in closed-form. For example for the binomial model with canonical link

$$\mathsf{E}(1/\psi''(W_i)) = 2 + \exp\left(-\hat{\mu}_{q_i|S} + \frac{1}{2}\hat{v}^2_{q_i|S}\right) + \exp\left(\hat{\mu}_{q_i|S} + \frac{1}{2}\hat{v}^2_{q_i|S}\right),$$

and for the Poisson model with canonical link

$$\mathsf{E}(1/\psi''(W)) = \exp\left(-\hat{\mu}_{q_i|S} + \frac{1}{2}\hat{v}^2_{q_i|S}\right).$$

For models for which $\hat{K}_Q$ cannot be computed in closed-form a one-dimensional Gaussian quadrature method may be used, which is still better than the multidimensional integration required in (9) or (10).

From (11) the optimal SGLMM design criterion is: Choose $Q \subset \mathbb{S} \setminus S$ to minimise

$$\sum_{s \in \bar{\mathbb{S}}} \widehat{\mathsf{Var}}(\mathcal{Z}(s)|\boldsymbol{y}; S, Q). \tag{12}$$

We note that the criterion (12) is not exact for the expected predictive variance and the approximation is valid when n is large. However the case where n is large is the one where a space-filling design is more appropriate while the small-n case is the one where an alternative

criterion is mostly needed. We expect that the proposed criterion (12) will maintain the general structure of the optimal predictive design for SGLMM as it accounts for the effect of the mean in a similar way as in (8) and proceed by using it as the optimal design criterion even when n is small.

A criticism against the proposed criterion (12), as well as the one derived from (10), is that it does not account for parameter uncertainty in the prediction. In the typical Gaussian kriging model, the kriging prediction formula and its variance adjust for the uncertainty in the estimation of the regression coefficients. It is relatively straightforward to do the same for SGLMM if we use for the unconditional density of the random field the density

$$f(\boldsymbol{z};\Sigma_S) = (2\pi)^{\frac{k-p}{2}}|\Sigma_S|^{-\frac{1}{2}}|F_S^\intercal\Sigma_S^{-1}F_S|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{z}^\intercal V_S\boldsymbol{z}\right\},$$

with

$$V_S = (I - F_S G_S)^\intercal\Sigma_S^{-1},\ G_S = (F_S^\intercal\Sigma_S^{-1}F_S)^{-1}F_S^\intercal\Sigma_S^{-1},$$

and follow the procedure described in Sections 2.2 and 2.3. However we choose not to elaborate further on this aspect as this falls outside the scope of the present paper, which is to demonstrate that optimal designs for SGLMM are not space-filling designs and present the computational issues arising from implementing such design. On the other hand, the inclusion of uncertainty about the variance parameters is less straightforward and we will not address it in this paper.

3 The exchange algorithm

Searching for the augmented set Q that minimises (12) can be computationally challenging. If l sites are to be selected out of N possible candidates then in theory the design criterion has to be computed N-choose-l times. Even for moderate sizes this number is too large so an alternative exchange algorithm is proposed. Royle (2002) reviews the basic exchange algorithm and some of its extensions.

The basic exchange algorithm is described as follows. For a given configuration $S \cup Q \subset \bar{\mathbb{S}}$, where $\bar{\mathbb{S}}$ is a finite set denoting all feasible sampling locations in $\mathbb{S}$, and a set of candidate points $(S \cup Q)^{\mathsf{c}} := \bar{\mathbb{S}} \setminus (S \cup Q)$, i.e. the elements of $\bar{\mathbb{S}}$ which are not included in the given design, fix $q \in Q$ and compute the design criterion by exchanging q with each of the elements of $(S \cup Q)^{\mathsf{c}}$. If no better design is found then q remains in the design otherwise q is replaced by the element of $(S \cup Q)^{\mathsf{c}}$ with the best value of the design criterion. This process is repeated for every $q \in Q$ and iterated until no better design is found. The exchange algorithm is guaranteed to converge but it is worth pointing out that its limit is not necessarily the optimal design. Nevertheless we find that it is a good compromise between computational speed and quality of the result. Below we describe how an update of the design criterion is computed in our implementation of the exchange algorithm.

Let $A_{SQ} = K_{SQ|S} + \Sigma_{SQ}$ and suppose without loss of generality that we are looking to update the element q of Q associated with the last row and column of A_{SQ}.

Note first that, since σ^2 is a constant, we only need to consider $\boldsymbol{c}_{SQ}^\intercal A_{SQ}^{-1}\boldsymbol{c}_{SQ}$ in (11).

Write $\boldsymbol{c}_{SQ}$ and A as a partitioned vector and a partitioned matrix respectively in the form

$$\boldsymbol{c}_{SQ} = \begin{pmatrix}\boldsymbol{u}\\ v\end{pmatrix}, \qquad A_{SQ} = \begin{pmatrix}B & \boldsymbol{u}\\ \boldsymbol{u}^\intercal & b\end{pmatrix},$$

with $\boldsymbol{u}$ and B being $(k+l-1)$ and $(k+l-1)\times(k+l-1)$-dimensional respectively and v, b being scalars. Also let $a = b - \boldsymbol{u}^\intercal B^{-1}\boldsymbol{u}$. Then

$$A_{SQ}^{-1} = \begin{pmatrix}B & \boldsymbol{u}\\ \boldsymbol{u}^\intercal & b\end{pmatrix}^{-1} = \begin{pmatrix}B^{-1} & \boldsymbol{0}\\ \boldsymbol{0}^\intercal & 0\end{pmatrix} + \frac{1}{a}\begin{pmatrix}B^{-1}\boldsymbol{u}\\ -1\end{pmatrix}\left(\boldsymbol{u}^\intercal B^{-1}, \quad -1\right),$$

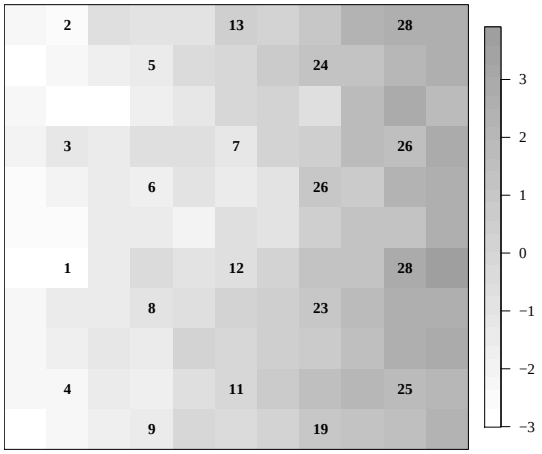

Figure 1: Simulation from the binomial SGLMM. The random field is shown by a greyscale and the binomial observations are indicated at the respective locations.

so

$$\boldsymbol{c}_{SQ}^{\mathsf{T}} A_{SQ}^{-1} \boldsymbol{c}_{SQ} = \boldsymbol{u}^{\mathsf{T}} B^{-1} \boldsymbol{u} + \frac{1}{a}(v - \boldsymbol{u}^{\mathsf{T}} B^{-1} \boldsymbol{u})^2,$$

therefore, in the exchange of a single element we may only consider maximising the quantity

$$\sum_{s \in \bar{\mathbb{S}}} \frac{(v - \boldsymbol{u}^{\mathsf{T}} B^{-1} \boldsymbol{u})^2}{b - \boldsymbol{u}^{\mathsf{T}} B^{-1} \boldsymbol{u}},$$

where the component $\boldsymbol{u}^{\mathsf{T}} B^{-1} \boldsymbol{u}$ is computed only once for each element of Q. In fact, by considering the block structure of B in terms of components associated with S or Q, its inverse can be computed by employing the well-known formula for the inverse of a block matrix (see Seber, 2008, section 14.2). Therefore the inverse of the block associated with S is computed once for the whole execution of the algorithm and for each update the inverse of a matrix of dimension $(l-1) \times (l-1)$ is needed each time.

4 Examples

4.1 A simulated example

We consider a regularly-spaced, 11×11 square grid within $[0,1] \times [0,1]$ and a Gaussian random field having exponential covariance function with parameters $\nu^2 = 0.1$, $\tau^2 = 0.3$, and $\phi = 0.2$ corresponding to nugget, partial sill and range. We also set the mean of the random field at coordinate (x, y) to $-2.5 + 5.0 \times x$, i.e. the mean increases linearly as we move from left to right from -2.5 to 2.5. The current sampling design consists of $k = 20$ locations, indicated by a "$\circ$" in Figure 2, and the observations follow the binomial SGLMM with canonical link and the number of trials is $n = 30$ at each location.

We simulate once from the Gaussian random field and, conditionally, from the binomial model at the sampled locations. The sample is shown in Figure 1. Note that the conditional variance of the observations given the value of the random field is $n\, e^z (1 + e^z)^{-2}$ and, due to the non-constant mean, the observations which are closer to the left or the right sides of the grid have very low variability while the observations that are equally far from the left and the right sides have the highest variability.

We are interested in augmenting the current sampling scheme by sampling at $l = 6$ more locations where the sample size at each location is $n = 30$. We implement an approximate

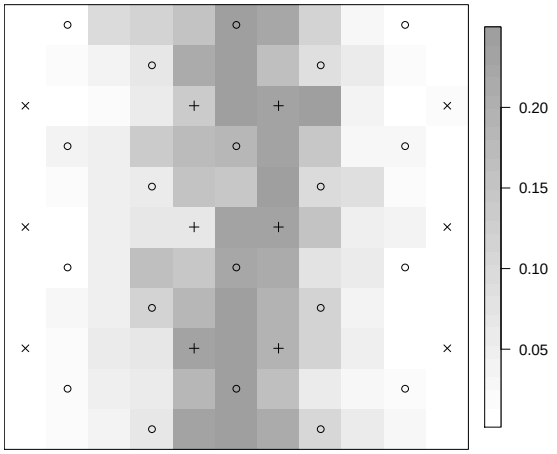

Figure 2: Predictive design for the simulated example. Showing current design (∘), space-filling augmented locations (×), and approximate SGLMM augmented locations (+). The greyscale shows the Bernoulli variance.

SGLMM design by minimising (12) and compare it with the space-filling design. The space-filling design suggests sampling from locations at the far left and the far right of the grid, the ones indicated by a "×" in Figure 2 while the SGLMM design suggests locations from the centre of the grid, indicated by a "+". As we would expect, the locations suggested by the approximate SGLMM design correspond to areas with higher variability while in this case the space-filling design recommends sampling from locations with very little variability. In fact a random sample from the locations suggested by the space-filling method were $1, 1, 5, 29, 29, 27$ (ordered from bottom to top and then left to right) while a random sample from the locations suggested by the approximate SGLMM design gave respectively $16, 1, 10, 19, 6, 17$. Evidently, the second set of observations is far more valuable in predicting the random field than the first. In particular, the average prediction variance with the original data is 0.2986. If we augment the data with the observations from the space-filling design the average prediction variance reduces to 0.2872, a 3.8% improvement over the original design. If, instead, we use the data from the approximate SGLMM design the average prediction variance becomes 0.2826, corresponding to a 5.3% improvement which is substantially more efficient than the space-filling design. Our approximate design criterion (12) for the two designs are 0.2876 and 0.2811 respectively which is close enough to the actual prediction variance and is good enough for the purpose of the design.

4.2 Assessing the approximation error

The design criterion (12) is derived from two asymptotic approximations. The first approximation is the use of the conditional distribution of the random field, namely equations (2), (3), and (4). These approximations have been used elsewhere in the literature, e.g. Eidsvik et al. (2009) and Evangelou et al. (2011), and found to be very accurate. The second approximation is to the expectation in (10). An interesting question is to assess how different the design would have been had we used that equation instead as our design criterion.

Criterion (10) requires the numerical evaluation of the expectation with respect to an l-dimensional normally distributed random variable with mean $\hat{\mu}_{Q|S}$ and variance $\hat{\Upsilon}_{Q|S}$. To that end, we followed a Monte-Carlo integration approach with simulation size of 5000. The large simulation size was to ensure consistency among different runs.

In the first scenario we assume that the same data displayed in Figure 1 are observed but the augmented data will be binomial counts with number of trials n. The approximate and simulated designs are computed for $n = 1$, 2, 5, 10, 30, and 50 and are shown in Figure 3. In

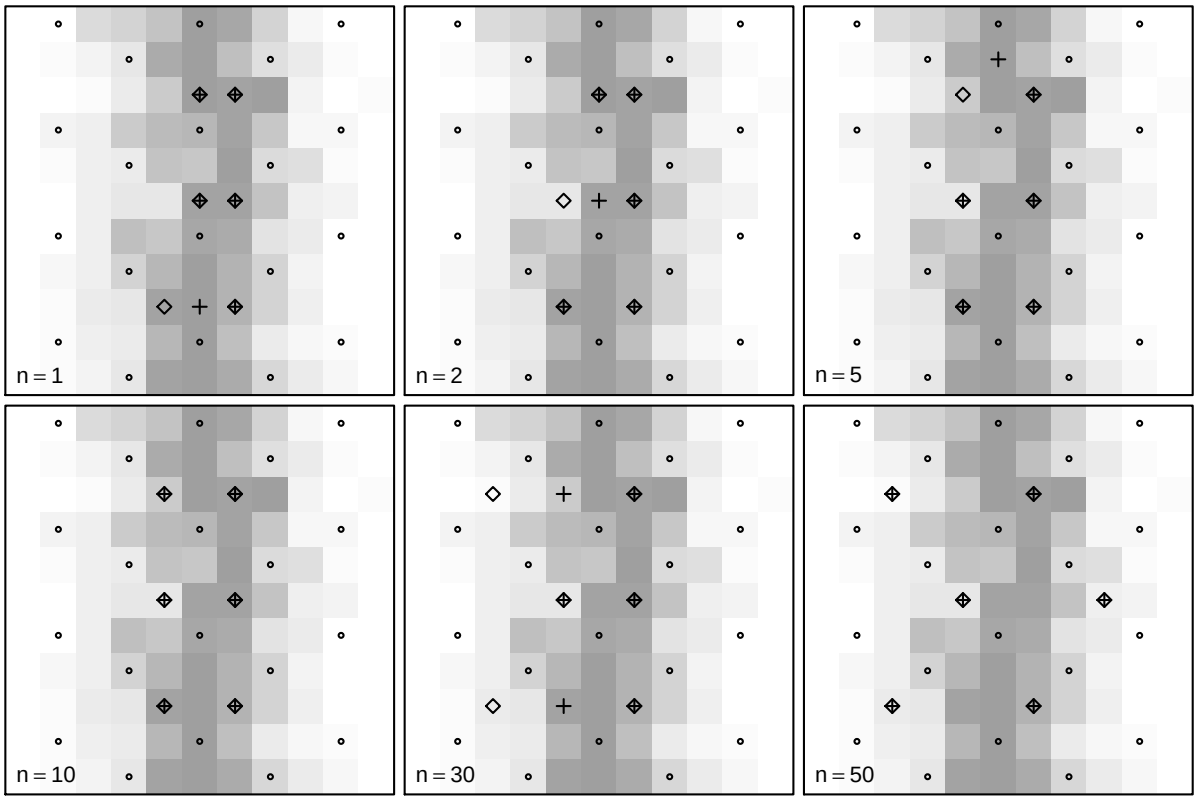

Figure 3: Approximate (+) and simulated ($\diamond$) designs for the binomial model as the sample size of the augmented data (n) varies.

n	1	2	5	10	30	50
Space-filling	0.00327	0.00589	0.01134	0.01585	0.01705	0.01517
Approximate	0.00001	0.00001	0.00009	0.00000	0.00028	0.00000

Table 1: Relative increase in prediction variance for space-filling and approximate augmented designs over simulated design as the sample size of the augmented data (n) varies. The entries are computed by dividing the increase in the average prediction variance from the simulated design by the average prediction variance of the simulated design.

two cases ($n = 10$ and $n = 50$) the two designs are identical and in the other cases there are only few differences among them. The relative increase in prediction variance for space-filling and approximate augmented designs over the simulated design is shown in Table 1. In the worst case, when $n = 40$, the increase in the average prediction variance for the approximate design, compared to the simulated design, is in the magnitude of 10^{-4}. In comparison, the relative increase for the space-filling design is in the magnitude of 10^{-2}.

The effect of the sample size at the augmented locations on the sampling design is also apparent in Figure 3. The design tends to depart from the region with the highest variability as n increases. For small n, both, the approximated and simulated designs, are highly concentrated in the middle, where the variability is higher, but gradually spread out as n increases, although the approximate design is slower to respond to this increase. This pattern is not surprising since, as we discussed in Section 2.2, the optimal SGLMM design comes closer to the space-filling design when n is large.

In the second scenario we assume the same model as before but with the difference that the observations at the initial sampled locations consist of binomial counts with number of trials $n = 1$, 2, 5, 10, 30, and 50, while the augmented data are binomial counts with number of

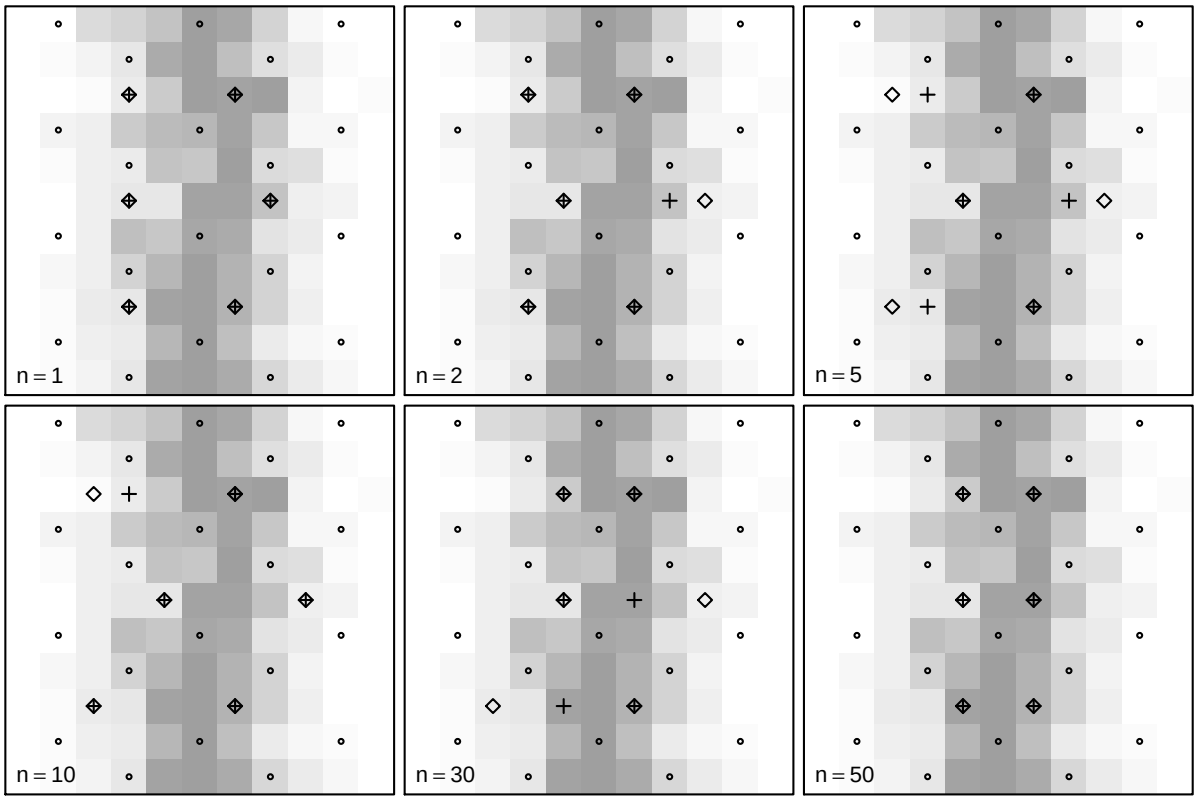

Figure 4: Approximate (+) and simulated ($\diamond$) designs as the sample size of the observed data (n) varies.

n	1	2	5	10	30	50
Space-filling	0.05457	0.04777	0.03499	0.02578	0.01684	0.01575
Approximate	0.00000	0.00012	0.00034	0.00013	0.00048	0.00000

Table 2: Relative increase in prediction variance for space-filling and approximate augmented designs over simulated design as the sample size of the observed data (n) varies. The entries are computed by dividing the increase in the average prediction variance from the simulated design by the average prediction variance of the simulated design.

trials $n = 30$. For each value of n, we simulate the vector of observations $\boldsymbol{y}$ conditioned on the observed random field, and compute the approximate and simulated design. The two designs are shown in Figure 4. As before, we observe only a few discrepancies between the two designs. The relative increase in the prediction variance using the approximate design is, by Table 2, in the magnitude of 10^{-4}, while for the space-filling design, it is about 10^{-2}.

The sample size n has the opposite effect on the sampling designs compared to our earlier setting. Here, both designs start further from the centre, where the variability is higher, but converge to the same limit for large n, with the approximate design being faster to respond to the increase in n.

The two simulation studies described in this section show that, even when n is very small, the approximate design is very close to the simulated design with only a fraction of the computation time.

4.3 Rhizoctonia disease

The rhizoctonia root rot is a disease affecting the roots of plants and hinders the absorption of water and nutrients by them. In this study 15 plants were pulled out of each of 100 randomly

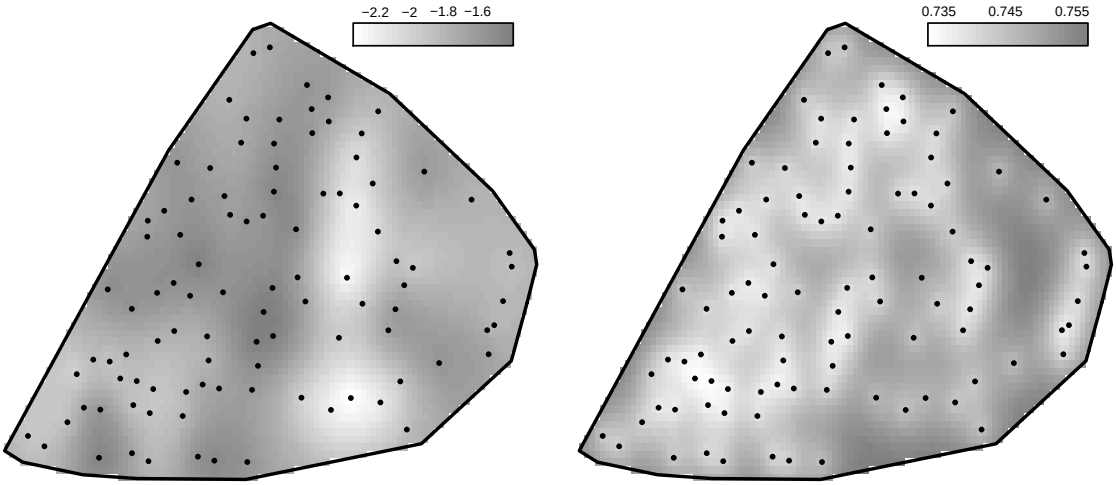

Figure 5: Prediction of the random field in the rhizoctonia example. Showing the prediction (left) and the predictive standard deviation (right).

chosen locations in a farm and the number of crown roots and infected crown roots were counted for each plant. The number of total roots at each location varies from 80 to 197 with an average of 138 suggesting that each plant has on average nine crown roots.

Zhang (2002) treated the data as binomial with the random field having constant mean and spherical covariance structure. This example was also analysed by Evangelou et al. (2011) using approximate Laplace approximation. Figure 5 shows the prediction and prediction standard deviation using their method with the sampled locations indicated by the points. Note that, due to the assumption that the mean being constant and the fact that the variance of the spatial random effects is not very high, the variation in the predicted random field is relatively small. As a result, the regions with the highest variability tend to be the ones which are the least represented in the sampling design.

We consider the question of augmenting the current network by sampling 15 plants from each of $l = 8$ new locations ($n = 135$ at each location) chosen from a regularly-spaced, square grid of 242 points. The locations suggested by the space-filling design and the approximate SGLMM design using the criterion in (12) are shown by a "$\times$" and a "+" respectively in Figure 6. The two designs appear very close except for two points where they are very different. This is not surprising in this case since, as we explain below, due to the large sample size at each location, the variation due to uncertainty at each sampled location ($K_{SQ|S}$) has very little effect in the design criterion. Therefore the variability is higher at the most isolated locations which is also what the space-filling design tends to select. However consider two alternative scenarios where the available resources allow us to sample only 5 plants ($n = 45$) from each new location and in the other case only one plant ($n = 9$). The two designs are shown in Figure 6 with a "V" and a "I" respectively. In the case of 5 plants there is more departure from the space-filling design and in the case of 1 plant there is a big difference. The explanation for this is because the sample size at each location enters into our design criterion as $n^{-1}e^{-\hat{z}}(1 + e^{\hat{z}})^2$; therefore, when n is large the effect of the mean of the random field in the design criterion is very small so the approximate design tends to be close to the space-filling design. When n is small, the mean of the random field has a more significant effect. This is also verified by comparing the total prediction variance under the space-filling design with the approximate SGLMM design for the different sample sizes as shown in Table 3. We observe that as the sample size increases the difference in the design criterion is reduced.

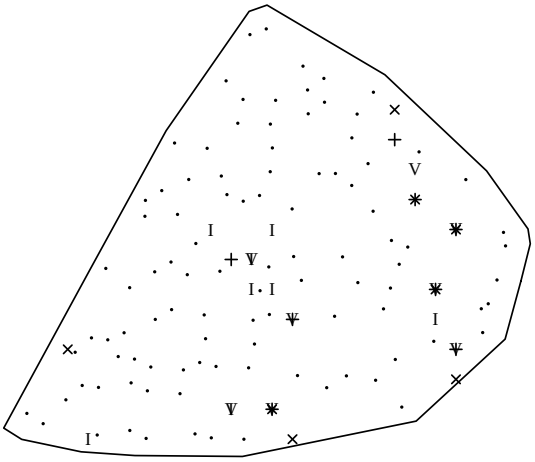

Figure 6: Predictive design for the rhizoctonia example. Showing current design ($\cdot$), space-filling augmented locations ($\times$), and approximate SGLMM augmented locations when sampling 15 plants (+), 5 plants (V), and 1 plant (I).

Number of plants	15	5	1
Space-filling	136.95	137.75	139.57
Approximate	136.92	137.68	139.45

Table 3: Total prediction variance under the approximate SGLMM and space-filling designs for different samples for the rhizoctonia example. The total prediction variance under the original design is 141.20.

4.4 Precipitation in central Norway

The number of rainy days for a region in central Norway were recorded for a period of $n = 61$ days at 92 monitoring stations. The current monitoring network is shown in Figure 7. The data were analysed by Eidsvik et al. (2009) who used a binomial SGLMM with the random effect assumed to be a Gaussian random field having constant mean and exponential covariance function with no nugget effect. In their analysis the authors derived that the observations from 4 stations were outliers (not shown in Figure 7) and only used data from $k = 88$ stations. Within the network there were also 6 stations not in operation (marked by a "$\diamond$") and the authors considered the improvement in prediction had data being obtained from these stations as well.

Using the available data we estimate, using the approximate likelihood method, the constant mean of the random field to be $\hat{\lambda} = 0.75$, and $\sigma^2 = 0.09$ and $\phi = 0.75$ for the partial sill and range respectively. We then consider prediction at a fine grid of 587 points. Using the data from the 88 locations the total prediction variance is 27.33 while assuming that data from the additional 6 stations were provided the total prediction variance from (12) reduces to 26.73.

Assuming that we have the power to place the 6 monitoring stations not in operation at different sites, the space-filling and the approximate SGLMM augmented designs were derived. These are marked by a "$\times$" and a "+" respectively in Figure 7. All three designs agree at one location at coordinate $(8.0, 63.5)$ but apart from that there are significant differences. The total prediction variance under the space-filling design is 26.68 and for the approximate SGLMM it is 26.66. Both suggestions are an improvement compared to the current network.

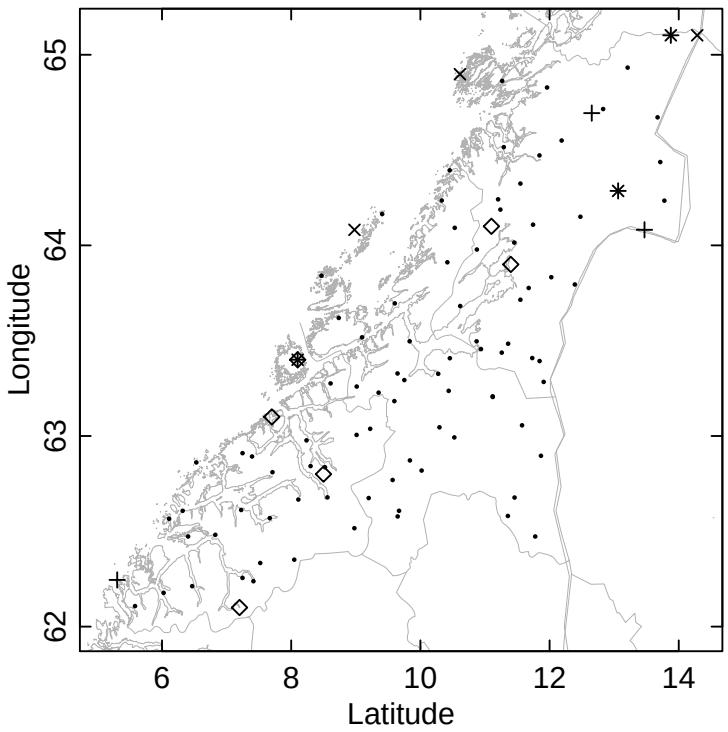

Figure 7: Norway precipitation network. Samples are obtained from locations marked by a "·" and the ones in circle are considered outliers. All locations that did not register data are marked by a "⋄". Also showing the space-filling augmented locations (×) and the approximate SGLMM augmented locations (+).

5 Discussion

In this article we propose a criterion for augmenting a spatial design under the SGLMM framework when data is available. Our design criterion is derived as an approximation to the expected predictive variance of the random field and it can be evaluated with little computational cost as it avoids making use of the typical MC methods used for these models. Our approximation is based on Laplace approximation and is valid under the increasing-domain asymptotic framework.

We demonstrate theoretically and through our examples that, contrary to the more popular Gaussian model, optimal designs for SGLMM may not be regularly spaced. The level of non-uniformity depends on the structure of the conditional mean of the random field and on the sample size.

A number of issues still remain. From a computational point of view, searching for the optimal design can be very hard if the number of candidate locations is large. The exchange algorithm we presented works well for moderate sample sizes but more sophisticated techniques are needed for large data.

Moreover, the ideas presented here can be used to derive alternative criteria, such as maximising the gain in information about the random field, as in Caselton and Zidek (1984) and Shewry and Wynn (1987) (see Evangelou and Zhu (2012) for a more elaborate discussion on this area).

We discussed how uncertainty in the regressor coefficients can be incorporated in the design criterion. Questions related to the construction of predictive designs by accounting for uncertainty in the covariance parameters in the spirit of Smith and Zhu (2004); Zhu and Stein (2006); Zimmerman (2006) as well as viewing the problem from a Bayesian context as in Diggle and Lophaven (2006) still remain.

Acknowledgements

The Norway precipitation data were kindly provided by Jo Eidsvik (NTNU).

A Derivation of the approximate predictive variance

Here we derive the approximation to the predictive variance given in (8).

First note that by an application of the formula for matrix inversion

$$(A+B)^{-1} = A^{-1} - A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1},$$

we have

$$\hat{\Upsilon}_S = \hat{H}^{-1} = \Sigma_S - \Sigma_S(\hat{D}_S^{-1} + \Sigma_S)^{-1}\Sigma_S. \tag{13}$$

Using the rule of iterated expectations and the fact that $\mathcal{Z}(R)|\mathcal{Z}(S), \boldsymbol{y} = \mathcal{Z}(R)|\mathcal{Z}(S)$ in distribution,

$$\begin{aligned}
\mathsf{Var}(\mathcal{Z}(R)|\boldsymbol{y}; S) &= \mathsf{E}\,\mathsf{Var}(\mathcal{Z}(R)|\mathcal{Z}(S), \boldsymbol{y}) + \mathsf{Var}\,\mathsf{E}(\mathcal{Z}(R)|\mathcal{Z}(S), \boldsymbol{y}) \\
&= \Sigma_R - \Sigma_{R,S}\Sigma_S^{-1}\Sigma_{S,R} + \Sigma_{R,S}\Sigma_S^{-1}\hat{\Upsilon}_S\Sigma_S^{-1}\Sigma_{S,R} \\
&= \Sigma_R - \Sigma_{R,S}(\hat{D}_S^{-1} + \Sigma_S)^{-1}\Sigma_{S,R}
\end{aligned}$$

by (13), as given in (4).

References

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93(441):262–272.

Caselton, W. F. and Zidek, J. V. (1984). Optimal monitoring network designs. *Statistics & Probability Letters*, 2(4):223–227.

Christensen, O. F. (2004). Monte carlo maximum likelihood in model-based geostatistics. *Journal of computational and graphical statistics*, 13(3):702–718.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58(2):280–286.

Cressie, N., Gotway, C. A., and Grondona, M. O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems*, 7(3):251–271.

Diggle, P. J. and Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1):53–64.

Diggle, P. J., Moyeed, R. A., Rowlingson, B., and Thomson, M. (2002). Childhood malaria in the gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506.

Diggle, P. J., Moyeed, R. A., and Tawn, J. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Eidsvik, J., Martino, S., and Rue, H. (2009). Approximate bayesian inference in spatial generalized linear mixed models. *Scandinavian journal of statistics*, 36(1):1–22.

Evangelou, E. and Zhu, Z. (2012). Maximum-information predictive designs for model-based geostatistics. *Preprint*.

Evangelou, E., Zhu, Z., and Smith, R. L. (2011). Estimation and prediction for spatial generalized linear mixed models using high order laplace approximation. *Journal of Statistical Planning and Inference*, 141(11):3564–3577.

Harville, D. A. and Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, pages 724–731.

Heuvelink, G. B. M., Jiang, Z., De Bruin, S., and Twenhöfel, C. J. W. (2010). Optimization of mobile radioactivity monitoring networks. *International Journal of Geographical Information Science*, 24(3):365–382.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.

Martin, R. J. (2001). Comparing and contrasting some environmental and experimental design problems. *Environmetrics*, 12(3):273–287.

McCullagh, P. and Nelder, J. A. (1999). *Generalized Linear Models*. Chapman & Hall Ltd.

Müller, W. G. (2007). *Collecting spatial data: optimum design of experiments for random fields*. Springer Verlag.

Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks. In Nychka, D., Piegorsch, W., and Cox, L., editors, *Case Studies in Environmental Statistics*, number 132 in Lecture Notes in Statistics, pages 51–76. Springer Verlag, New York.

Royle, J. A. (2002). Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, 100(2):121–134.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, 24(5):479–488.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 319–392.

Seber, G. A. F. (2008). *A matrix handbook for statisticians*. Wiley series in probability and statistics. Wiley-Interscience.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):749–760.

Smith, R. L. and Zhu, Z. (2004). Asymptotic theory for kriging with estimated parameters and its application to network design. Preliminary version, available from http://www.stat.unc.edu/postscript/rs/supp5.pdf.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.

Zhang, H. and Wang, H. H. (2002). A study on prediction of spatial binomial probabilities with an application to spatial design. In Wegman, E. J. and Braverman, A., editors, *Computing Science and Statistics*, volume 34, pages 263–276. Interface Foundation of North America.

Zhu, Z. and Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of agricultural, biological, and environmental statistics*, 11(1):24–44.

Zimmerman, D. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6):635–652.

Zimmerman, D. L. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the institute of statistical mathematics*, 44(1):27–43.