ENVIRONMETRICS
Environmetrics 2005; **16**: 205–217
Published online 13 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.701

Robustness in spatial studies II: minimax design

Douglas P. Wiens*,†

Department of Mathematical and Statistical Sciences, Statistics Centre, University of Alberta, Edmonton, Alberta T6G 2G1, Canada

SUMMARY

We consider robust methods for the construction of sampling designs in spatial studies. The designs are robust against misspecified regression responses, and are tailored for possible use with predictors which are minimax robust against misspecified variance/covariance structures. The loss function is based on the mean squared error of the predicted values. This is maximized, analytically, over a neighbourhood quantifying the departures from the fitted linear regression response. This maximum is then minimized numerically—by simulated annealing, or sequentially—in order to obtain the optimal designs. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: anisotropic; computer experiments; generalized M-estimation; isotropic; kriging; minimax; M-estimate; sequential; simulated annealing

1. INTRODUCTION

In Wiens (2004) minimax linear predictors for spatial processes were obtained. The scenario entertained there was that one would sample locations $\mathcal{S} = \{\mathbf{t}_{i_1}, \ldots, \mathbf{t}_{i_n}\}$ from a set $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\} \subset \mathbb{R}^d$. At these locations one would observe $Y(\mathbf{t})$, consisting of a stochastic process $X(\mathbf{t})$ along with uncorrelated, additive measurement error $\epsilon(\mathbf{t})$. The random variable $X(\mathbf{t})$ was assumed to consist of a deterministic mean $\mathbf{z}^{\mathrm{T}}(\mathbf{t})\boldsymbol{\theta}$ perturbed by stochastic errors $\delta(\mathbf{t})$. Thus the model for the data was

$$Y(\mathbf{t}_{i_j}) = \mathbf{z}^{\mathrm{T}}(\mathbf{t}_{i_j})\boldsymbol{\theta} + \delta(\mathbf{t}_{i_j}) + \epsilon(\mathbf{t}_{i_j}), \quad j = 1, \ldots, n$$

The purpose was to predict a set $\mathbf{Cx}$ of M linear functions of $\mathbf{x} = (X(\mathbf{t}_1), \ldots, X(\mathbf{t}_N))^{\mathrm{T}}$, in the face of uncertainly about the variance structure of the measurement errors, and of the covariance structure of $\{\delta(\mathbf{t}) \mid \mathbf{t} \in \mathcal{T}\}$.

Specifically, it was assumed that the measurement errors had a covariance matrix $\mathbf{F} = \mathrm{diag}(f(\mathbf{t}_1), \ldots, f(\mathbf{t}_N))$, with $f(\mathbf{t})$ lying within a constant α of the nominal variances $f_0(\mathbf{t})$, and that the covariance matrix $\mathbf{G}_{N\times N} = \left(g(\mathbf{t}_i, \mathbf{t}_j)\right)_{i,j=1}^{N}$ of $\{\delta(\mathbf{t}) \mid \mathbf{t} \in \mathcal{T}\}$ would be bounded above, in the sense of positive definiteness, by a matrix of the form $\mathbf{G}^{(0)} + \beta\mathbf{K}$, where $\mathbf{G}_0$ was the nominal

*Correspondence to: D. P. Wiens, Department of Mathematical and Statistical Sciences, Statistics Centre, University of Alberta, Edmonton, Alberta T6G 2G1, Canada.
†E-mail: doug.wiens@ualberta.ca

Received 24 March 2004
Accepted 10 June 2004

covariance matrix and $\mathbf{K}$ was a fixed positive definite matrix, typically $= \mathbf{I}_N$. For instance one might specify homoscedastic errors, $f_0(\mathbf{t}) \equiv \sigma_1^2$, and isotropic covariances, $g(\mathbf{t_i}, \mathbf{t_j}) = \sigma_2^2 \rho(\|\mathbf{t}_i - \mathbf{t}_j\|)$, for some correlation function $\rho(\cdot)$. With $\mathbf{y} \triangleq (Y(\mathbf{t}_{i_1}), \ldots, Y(\mathbf{t}_{i_n}))^{\mathrm{T}}$, loss was taken to be the trace of the mean squared prediction error matrix

$$\mathrm{MSPE}(\mathbf{A}) = \mathrm{E}\left[(\mathbf{Ay} - \mathbf{Cx})\ (\mathbf{Ay} - \mathbf{Cx})^{\mathrm{T}}\right]$$

maximized over f and g. The maximum loss was minimized subject to an unbiasedness constraint

$$\mathrm{E}[\mathbf{Ay}] = \mathrm{E}[\mathbf{Cx}] \quad \text{for all } \boldsymbol{\theta} \tag{1}$$

Two notable special cases are as follows:

1. Suppose that $M = N - n$ and $\mathbf{C}$ is the incidence matrix for $\mathcal{T}\backslash\mathcal{S}$, i.e. $\mathbf{C}$ is the result of omitting, from $\mathbf{I}_N$, rows $i_1, \ldots, i_n$. Then we are predicting $X(\mathbf{t})$ by a linear function $\hat{X}(\mathbf{t}) = \mathbf{a}_{\mathbf{t}}^{\mathrm{T}}\mathbf{y}$ for each $\mathbf{t} \notin \mathcal{S}$. The matrix $\mathbf{A}$ has rows $\mathbf{a}_{\mathbf{t}}^{\mathrm{T}}$, and the loss is

$$\max_{f,g} \sum_{\mathbf{t} \notin \mathcal{S}} \mathrm{E}\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right] \tag{2}$$

2. If, instead, $M = N$ and $\mathbf{C} = \mathbf{I}_N$, then the loss is

$$\max_{f,g} \sum_{\mathbf{t} \in \mathcal{T}} \mathrm{E}\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right] \tag{3}$$

the total mean squared prediction error (TMSPE).

The resulting minimax linear predictor $\mathbf{A}_{\alpha,\beta}\mathbf{y}$ turned out to be a 'universal kriging' predictor computed from modified inputs. The following result summarizes the salient facts of Wiens (2004). Before stating it we require some notation. Let $\mathbf{Q}_1$ and $\mathbf{Q}_2$ be the incidence matrices for $\mathcal{S}$ and $\mathcal{T}\backslash\mathcal{S}$, respectively, so that

$$\mathbf{Q}_{N\times N} = \begin{pmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{pmatrix} \begin{matrix} n \\ n-N \end{matrix}$$

is a permutation of the rows of $\mathbf{I}_N$. For a generic $N \times N$ matrix $\mathbf{M}$ we write $\mathbf{M}_{ij}$ for $\mathbf{Q}_i\mathbf{M}\mathbf{Q}_j^{\mathrm{T}}$. Thus, for instance, $\mathbf{M}_{11}$ refers only to the elements of $\mathbf{M}$ corresponding to sampled locations. Similarly, we write $\mathbf{M}_i$ for $\mathbf{Q}_i\mathbf{M}$. A superscript '$^{(0)}$' denotes evaluation at the nominal variance/covariance functions f_0 and g_0.

Theorem 1 (Wiens, 2004)**.** *Let* $\mathbf{Z} = (\mathbf{z}(\mathbf{t}_1), \ldots, \mathbf{z}(\mathbf{t}_N))^{\mathrm{T}}$ *be the* $N \times p$ *matrix of regressors for* $\mathcal{T}$*, so that* $\mathbf{Z}_1 = \mathbf{Q}_1\mathbf{Z} : n \times p$ *is that for* $\mathcal{S}$*. Define*:

$$\boldsymbol{\Sigma}_{11} = \mathbf{F}_{11}^{(0)} + \mathbf{G}_{11}^{(0)} : n \times n, \quad \boldsymbol{\Lambda}_{\alpha,\beta} = \boldsymbol{\Sigma}_{11} + \alpha\mathbf{I}_n + \beta\mathbf{K}_{11} : n \times n$$
$$\mathbf{H}_\beta = \mathbf{G}^{(0)} + \beta\mathbf{K} : N \times N, \quad \mathbf{R}_{\alpha,\beta} = \left(\mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\Lambda}_{\alpha,\beta}^{-1}\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\Lambda}_{\alpha,\beta}^{-1} : p \times n$$

The minimax unbiased linear predictor of $\mathbf{Cx}$ *is* $(\widehat{\mathbf{Cx}})_{\mathrm{LIN}} = \mathbf{A}_{\alpha,\beta}\mathbf{y}$, *where* $\mathbf{A}_{\alpha,\beta} = \mathbf{C}\mathbf{P}_{\alpha,\beta} : M \times n$ *for*

$$\mathbf{P}_{\alpha,\beta} = \mathbf{Z}\mathbf{R}_{\alpha,\beta} + \mathbf{H}_{\beta,1}^{\mathrm{T}}\boldsymbol{\Lambda}_{\alpha,\beta}^{-1}\left(\mathbf{I}_n - \mathbf{Z}_1\mathbf{R}_{\alpha,\beta}\right) : N \times n$$

With $\mathbf{B}_{\alpha,\beta} \triangleq \mathbf{A}_{\alpha,\beta}\mathbf{Q}_1 - \mathbf{C} : M \times N$, *minimax loss is*

$$\mathcal{L}_0\left(\mathbf{A}_{\alpha,\beta}\right) = \mathrm{tr}\left[\mathbf{B}_{\alpha,\beta}\mathbf{H}_\beta\mathbf{B}_{\alpha,\beta}^{\mathrm{T}} + \mathbf{A}_{\alpha,\beta}\left(\mathbf{F}_{11}^{(0)} + \alpha\mathbf{I}_n\right)\mathbf{A}_{\alpha,\beta}^{\mathrm{T}}\right]$$

Note that $\boldsymbol{\Lambda}_{0,0} = \mathbf{F}_{11}^{(0)} + \mathbf{G}_{11}^{(0)}$ and $\mathbf{H}_0 = \mathbf{G}^{(0)}$. If $\alpha = \beta = 0$, then the loss is the trace of the MSPE matrix of $\mathbf{Ay}$, in predicting $\mathbf{Cx}$, at the nominal model. This is well known to be minimized subject to (1) by the universal kriging predictor $\mathbf{A}_{0,0}\mathbf{y}$. The minimax linear predictor is instead the universal kriging predictor computed from the modified variance/covariance matrices $\boldsymbol{\Lambda}_{\alpha,\beta}$ and $\mathbf{H}_\beta$, which result from adding $\alpha\mathbf{I}_n$ to $\mathbf{F}_{11}^{(0)}$ and $\beta\mathbf{K}$ to $\mathbf{G}^{(0)}$.

To this point only two possible sources of non-robustness—those arising from variance/covariance misspecifications—have been considered. A third source—the effect of outlying observations—was addressed in Wiens (2004) by replacing the terms in $\mathbf{A}_{\alpha,\beta}\mathbf{y}$ by robust estimates, including generalized M-estimates of $\boldsymbol{\theta}$. This resulted in a robustified predictor $(\widehat{\mathbf{Cx}})_{\mathrm{GM}}$.

In this article we continue this program by considering robustness of design, against misspecified regression response functions. We suppose that the predictor $\mathbf{A}_{\alpha,\beta}\mathbf{y}$, perhaps with $\alpha = \beta = 0$, is to be computed, but that the mean response is only approximately linear in the regressors $z_j(\mathbf{t})$. Define a parameter vector $\boldsymbol{\theta}_{p\times 1}$ to be that which makes the approximation $\mathrm{E}[X(\mathbf{t})] \approx \mathbf{z}^{\mathrm{T}}(\mathbf{t})\boldsymbol{\theta}$ most accurate, viz.

$$\boldsymbol{\theta} = \arg\min_{\mathbf{v}} \sum_{\mathbf{t}\in\mathcal{T}} \left(\mathrm{E}[X(\mathbf{t})] - \mathbf{z}^{\mathrm{T}}(\mathbf{t})\mathbf{v}\right)^2$$

Now define $h(\mathbf{t})$ by

$$\mathrm{E}[X(\mathbf{t})] = \mathbf{z}^{\mathrm{T}}(\mathbf{t})\boldsymbol{\theta} + h(\mathbf{t})$$

These definitions imply the orthogonality condition

$$\sum_{\mathbf{t}\in\mathcal{T}} \mathbf{z}(\mathbf{t})h(\mathbf{t}) = \mathbf{0}_{p\times 1} \tag{4}$$

We propose to construct minimax designs, by first maximizing $\mathrm{MSPE}\left(\mathbf{A}_{\alpha,\beta}\right)$ over

$$\mathcal{H}_\gamma = \left\{ h(\cdot) \,|\, N^{-1}\sum_{\mathbf{t}\in\mathcal{T}} h^2(\mathbf{t}) \le \gamma,\ \sum_{\mathbf{t}\in\mathcal{T}} \mathbf{z}(\mathbf{t})h(\mathbf{t}) = \mathbf{0} \right\}$$

and then minimizing the result over the design, i.e. over $\mathbf{Q}_1$. This minimization is a discrete optimization problem which we shall address in one of two ways—by simulated annealing, or sequentially.

Our framework is sufficiently broad as to encompass several scenarios:

1. No locations have yet been chosen, and we are free to choose any n sites. In the case of (3) our robust optimality criterion is then analogous to the classical notion of I-optimality. In a similar vein, Sacks and Schiller (1988) considered the construction of designs, assuming that $f_0(\cdot)$ and $g_0(\cdot,\cdot)$ were correctly specified and that $\mathrm{E}[X(\mathbf{t})]$ was known and $\equiv 0$. They used the loss function $\max_{\mathbf{t}} \mathrm{E}[(\mathbf{a}_{\mathbf{t}}^{\mathrm{T}}\mathbf{y} - X(\mathbf{t}))^2]$, and remarked upon the lack of robustness, to changes in g_0, of their procedures.
2. There is an existing network of n_0 sites at locations $\mathcal{S}_0$ and we are to choose $n - n_0$ further sites. Scenario 1 is this case with $n_0 = 0$. See Thompson (1997, p. 18) for a discussion.
3. There is an existing network of n_1 sites at locations $\mathcal{S}_1$ and we must *eliminate* $n_1 - n$ of them. This is equivalent to setting $\mathcal{T} = \mathcal{S}_1$ and then finding the n best sites which are to remain.

Spatial design problems for correctly specified models have been studied by Martin (1986), Fedorov and Hackl (1994), Stein (1995), and Thompson (1997), among others. Schilling (1992) and McArthur (1987) assess some particular sampling designs. Santner *et al.* (2003) and Sacks *et al.* (1989) utilize ideas of spatial prediction and design for computer based experimentation. For a discussion see Martin (2001), who also contrasts spatial designs with those for unreplicated field trials. To our knowledge this is the first work to explicitly seek robustness of spatial design against model uncertainties.

The maximum loss is exhibited in Theorem 2 of Section 2. In Section 3 we obtain optimal robust designs by minimizing this maximum loss. For small values of N and n this can be done exactly, by performing an exhaustive search of all $\binom{N}{n}$ designs. For more realistic values we investigate two algorithms, both of which seem to give at least nearly optimal solutions in reasonable amounts of time. The first is a simulated annealing algorithm, whereas the second employs a sequential search technique.

All derivations are in the Appendix.

2. ROBUST DESIGN

In this section we obtain the maximum loss of $(\widehat{\mathbf{Cx}})_{\mathrm{LIN}} = \mathbf{A}_{\alpha,\beta}\mathbf{y}$, for $h \in \mathcal{H}_\gamma$, and discuss numerical approximations to be used in the repeated evaluation of this maximum, leading to minimax designs.

We propose to use those designs, optimized for use with the GLS estimate, even when $\hat{\boldsymbol{\theta}}$ is a GM-estimate. One reason for this is, of course, the relative intractability of the MSE of the GM-estimate, making it very difficult to maximize. A more compelling reason is that our experience has been that the robust designs seem to depend very little on the choice of the estimate—see, for example, Sinha and Wiens (2002).

Theorem 2. *For the minimax linear predictor of Theorem 1 the maximum value of* $\mathcal{L}(\mathbf{A}_{\alpha,\beta})$, *for* $h \in \mathcal{H}_\gamma$, *is*

$$\max_{\mathcal{H}_\gamma} \mathcal{L}(\mathbf{A}_{\alpha,\beta}) = \mathcal{L}_0(\mathbf{A}_{\alpha,\beta}) + N\gamma\lambda_{\max}\left(\mathbf{B}_{\alpha,\beta}\mathbf{B}_{\alpha,\beta}^{\mathrm{T}}\right) \tag{5}$$

where $\lambda_{\max}$ *denotes the largest eigenvalue. The maximum is attained if* $\mathbf{h}$ *is an eigenvector of* $\mathbf{B}_{\alpha,\beta}^{\mathrm{T}}\mathbf{B}_{\alpha,\beta}$, *with* $N^{-1}\|\mathbf{h}\|^2 = \gamma$, *corresponding to* $\lambda_{\max}$.

Note that, if $M = 1$, as when $X_{\mathrm{Total}} = \sum_{\mathbf{t}\in\mathcal{T}} X(\mathbf{t})$ is being predicted, then $\mathbf{B}_{\alpha,\beta}$ is a row vector and in (5), $\lambda_{\max}(\mathbf{B}_{\alpha,\beta}\mathbf{B}_{\alpha,\beta}^{\mathrm{T}}) = \mathbf{B}_{\alpha,\beta}\mathbf{B}_{\alpha,\beta}^{\mathrm{T}}$.

Where possible we will now drop the subscripts β and α, β.

2.1. Modifications for large M

The algorithms described in the next section call for the repeated calculation of the loss (5), and hence of the eigenvalues of the $M \times M$ matrix $\mathbf{B}\mathbf{B}^{\mathrm{T}}$. For large values of M this is not feasible in any reasonable amount of time. This in particular is a problem when the loss is (2) or (3), so that M is $N-n$ or N, respectively, if N is realistically large. We have noticed, however, that in these cases $\lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)$ is typically very close to the largest eigenvalue of the $n \times n$ matrix $\mathbf{P}^{\mathrm{T}}\mathbf{P}$.

To explain this closeness, we first note that for loss (2), $\mathbf{C} = \mathbf{Q}_2$ and so

$$\mathbf{Q}\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathbf{Q}^{\mathrm{T}} = \mathbf{0}_{n\times n} \oplus \mathbf{I}_{N-n} \tag{6}$$

For loss (3), $\mathbf{C} = \mathbf{I}_N$ and

$$\mathbf{Q}\mathbf{C}^{\mathrm{T}}\mathbf{C}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}_N \tag{7}$$

Lemma 3 shows that our approximation of the largest eigenvalue becomes exact as $\|\mathbf{F}_{11}^{(0)} + \alpha\mathbf{I}_n\| \to \infty, 0$.

Lemma 3. *Assume that one of (6), (7) holds. Then*

$$\lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right) = \lambda_{\max}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right) + \mathrm{o}(1)$$

as either:
A1. $\lambda_{\min}(\mathbf{\Lambda}) \to \infty$ *in such a way that* $\mathbf{R} \to \left(\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^{\mathrm{T}}$, *or*
A2. $\lambda_{\max}(\mathbf{F}_{11}^{(0)} + \alpha\mathbf{I}_n) \to 0$.

If $\mathbf{F}^{(0)} = \sigma_1^2\mathbf{I}_N$, then conditions A1 and A2 hold if $\sigma_1^2 \to \infty$, or $\sigma_1^2, \alpha \to 0$, respectively.

When (6) or (7) hold and $M > 25$, we replace $\lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)$ by the much more easily computed $\lambda_{\max}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)$ in (5). This approximation is surprisingly accurate. As examples, we computed the relative error

$$re = \left|1 - \frac{\lambda_{\max}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)}{\lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)}\right|$$

for various choices of N, and for a number of randomly chosen n-element subsets of $\mathcal{T}$. For $\mathbf{C} = \mathbf{I}_N$ and $(N, n) = (49, 7)$, (100,10) and (400,20), the average values of *re* over 100 trials were 1.6, 1.1 and 0.34 per cent, respectively, when the regressors were $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^{\mathrm{T}}$ and $\mathbf{F}^{(0)}$, $\mathbf{G}^{(0)}$ were as described in Section 3.1 below, with isotropic correlations and $\alpha = \beta = 0$. Varying the regression function or the choices of $\mathbf{F}^{(0)}$ and $\mathbf{G}^{(0)}$ typically resulted in even smaller relative errors.

A further simplification in evaluating $\mathcal{L}_0\left(\mathbf{A}_{\alpha,\beta}\right)$, if the loss is (2) or (3), is to write

$$\mathcal{L}_0\left(\mathbf{A}_{\alpha,\beta}\right) = \mathrm{tr}\left(\mathbf{B}\mathbf{H}\mathbf{B}^{\mathrm{T}}\right) + \mathrm{tr}\left(\mathbf{F}_{11}^{(0)}\left[\mathbf{A}^{\mathrm{T}}\mathbf{A}\right]\right) + \alpha\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)$$

where

$$\mathrm{tr}\left(\mathbf{B}\mathbf{H}\mathbf{B}^{\mathrm{T}}\right) = \mathrm{tr}(\mathbf{H}) + \begin{cases} \mathrm{tr}\left(\mathbf{H}_{11}\left[\mathbf{A}^{\mathrm{T}}\mathbf{A}\right] - 2\mathbf{H}_{12}\mathbf{A} - \mathbf{H}_{11}\right), & \text{if } \mathbf{C} = \mathbf{Q}_2 \\ \mathrm{tr}\left(\mathbf{H}_{11}\left[\mathbf{A}^{\mathrm{T}}\mathbf{A}\right] - 2\mathbf{H}_1\mathbf{A}\right) & \text{if } \mathbf{C} = \mathbf{I}_N \end{cases}$$

With the exception of $\mathrm{tr}(\mathbf{H})$, which must only be calculated once, all traces are now of $n \times n$ matrices.

3. DESIGNS: ALGORITHMS AND EXAMPLES

3.1. Test cases

We begin by exhibiting some optimal designs in situations in which N and n are small enough that the optimization can be carried out by an exhaustive search of all $\binom{N}{n}$ possible designs. We consider two types of correlation structures. The first—isotropic Gaussian correlations—employs nominal covariances $g_0(\mathbf{t},\mathbf{t}') = \sigma_2^2 \exp\{-\lambda d^2\}$ for $d = \|\mathbf{t}-\mathbf{t}'\|$. The second—anisotropic Gaussian correlations—uses instead $d = ((\mathbf{t}-\mathbf{t}')^{\mathrm{T}}\mathrm{diag}(1,5)(\mathbf{t}-\mathbf{t}'))^{1/2}$. In our examples we set $\sigma_2^2 = 2$, and choose λ so that the nearest neighbour correlation is 0.9. The ideal error variances are taken to be $\sigma_1^2 \equiv 1$.

Figure 1 exhibits optimal designs with $N = 25$ and $n = 7$. In all cases we take $\mathbf{C} = \mathbf{I}_N$, so that we are predicting $X(\mathbf{t})$ for all $\mathbf{t} \in \mathcal{T}$ and aim for minimization of TMSPE as at (3). Each of the designs took about 460 s to compute (using MATLAB, on a 2200 MHz PC with 1 Gbyte of RAM). The modifications of Section 2.1 led to the same designs, in about 275 s. In all four cases, the same designs were obtained when the loss was given by (2).

Figure 2 gives designs in the same situations as those in Figure 1, except that in these cases $\mathbf{C} = \mathbf{1}_N^{\mathrm{T}}$, so that we are predicting X_{Total} with minimum MSPE.

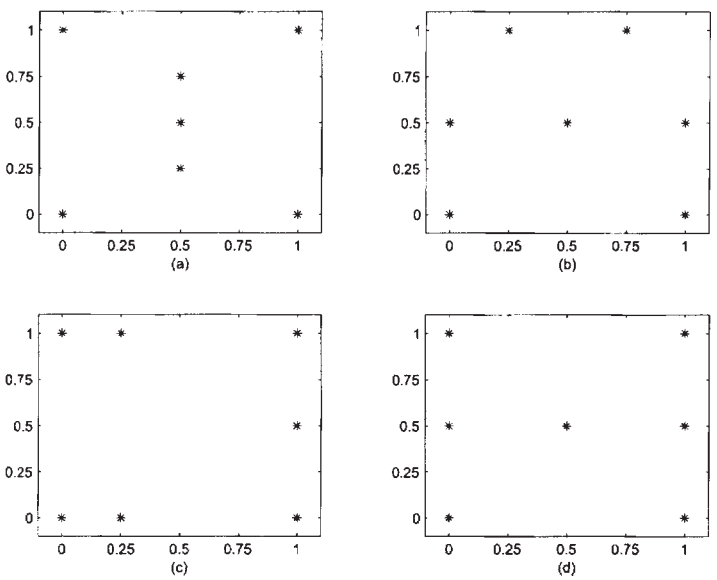

Figure 1. Optimal designs on an $N = 25$ point square grid in the unit square with $n = 7$ sites chosen. Regressors are $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^{\mathrm{T}}$ and $\mathbf{C} = \mathbf{I}_N$. (a) Isotropic correlations; $(\alpha,\beta,\gamma) = (0,0,3)$. The optimal sites are $\{1,5,8,13,18,21,25\}$, respectively—counting left to right across row 1, then row 2, etc. An equivalent design with the same minimum loss is obtained by rotating the design shown through 270°, obtaining sites $\{1,5,12,13,14,21,25\}$. (b) As in (a), but with anisotropic correlations. The optimal design shown has sites $\{1,5,11,13,15,22,24\}$; a rotation through 180° gives another. (c) As in (a), but with $(\alpha,\beta,\gamma) = (\sigma_1^2,\sigma_2^2,3)$. The optimal design shown has sites $\{1,2,5,15,21,22,25\}$; equivalent designs are obtained by rotation through 90°, 180°, 270°. (d) As in (b), but with $(\alpha,\beta,\gamma) = (\sigma_1^2,\sigma_2^2,3)$

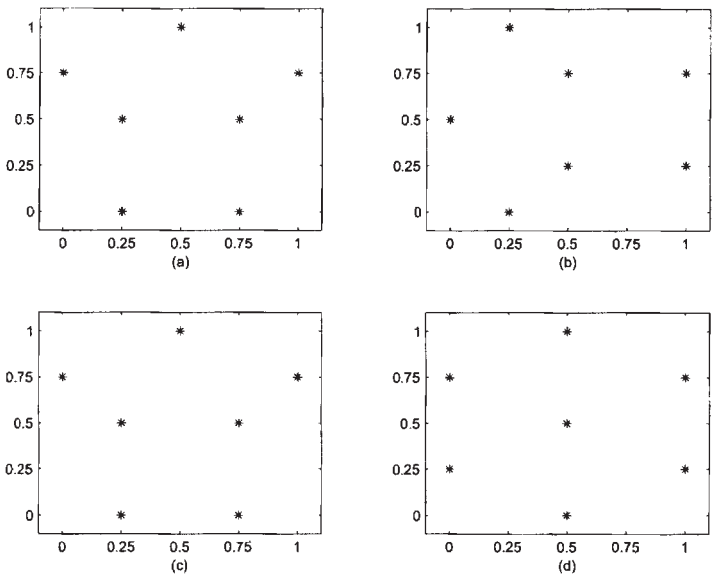

Figure 2. Optimal designs for the same inputs as in Figure 1. In this case the aim is to predict X_{Total}. The designs can be rotated in the same manner as those in Figure 1, to obtain equivalent optimal designs

3.2. Simulated annealing

We have found that simulated annealing can be quite successful in determining the optimal robust designs. Van Groenigen and Stein (1998) used simulated annealing to construct designs optimal for variogram estimation. Benedetti and Palma (1995) construct designs for the estimation of a mean, over lattice sites, in this manner.

Our algorithm is a modification of that of Sacks and Schiller (1988). The Sacks and Schiller algorithm was also used, again with minor modifications, by Chao and Thompson (2001). As we apply it the algorithm depends on a sequence $\{\pi_j\}$ of acceptance probabilities and parameters $\{n_0, \delta_0, \delta_1, \nu, m\}$, and is described as follows. Suppose that at the jth stage ($j = 0, 1, 2, \ldots$) of the process we are considering a configuration $S^{(j)} = \{\mathbf{t}_{i_1}^{(j)}, \ldots, \mathbf{t}_{i_n}^{(j)}\}$, with loss $L^{(j)}$ as at (5). Pick, at random, a location $\mathbf{t} \in \mathcal{T} \backslash S^{(j)}$ and determine in sequence the loss that arises if one of the $\mathbf{t}_{i_k}^{(j)}$ in $S^{(j)}$ is replaced by $\mathbf{t}$. If the least of these is less than $L^{(j)}$, then the corresponding configuration is accepted. Otherwise, this configuration is accepted with probability π_j.

Note that we are specifying that the same location must not appear more than once in the design. Many practical situations, e.g. geological sampling, require this. It is not a crucial point, however, and the method works equally well when replicates are allowed.

For large values of N we sometimes modify this step by employing a suggestion of Royle (2002). The suggestion was made in connection with exchange algorithms but is also applicable to the current situation. In this '100 per cent nearest neighbour' modification we test only those $\mathbf{t}_{i_k}^{(j)}$ for which $\|\mathbf{t}_{i_k}^{(j)} - \mathbf{t}\| < p \max_{\mathbf{t}' \in \mathcal{T}} \|\mathbf{t}' - \mathbf{t}\|$.

The preceding step is repeated n_0 times, or until a new configuration is accepted, whichever comes first. If a new configuration is accepted it is labelled $S^{(j+1)}$, and its loss is labelled $L^{(j+1)}$. If no new configuration is accepted, then $S^{(j)}$ is relabelled as $S^{(j+1)}$, $L^{(j)}$ as $L^{(j+1)}$. One then moves on to the next stage.

The sequence of acceptance probabilities is defined by $\pi_0 = 0.7$ and

$$\pi_{j+1} = \begin{cases} \min\left(1, \frac{\pi_j}{1-\delta_0}\right), & \text{if no new configuration was accepted at the } j\text{th stage} \\ (1-\delta_1)\pi_j, & \text{if a new configuration was accepted and } L^{(j+1)} < (1-\nu)\min_{i\le j} L^{(i)} \\ \pi_j, & \text{otherwise} \end{cases}$$

Iterations cease when there have been m evaluations of the loss since the last change in the value of the acceptance probability. The initial state $S^{(0)}$ is the best of m randomly chosen configurations.

This describes one 'run' of the annealing algorithm. We have found it best to carry out a number of runs, with different parameter values. Typically, in a run we will randomly choose $n_0 \in [0.1(N-n), 0.5(N-n)]$, $\delta_0 \in [0.1, 0.5]$, $\delta_1 \in [0.3, 0.5]$, $\nu \in [0.01, 0.05]$ and $m \in [50, 200]$. For the situations illustrated in Figure 1, the optimal configuration was generally found in no more than 10 runs. To carry out 10 runs requires about 30 s of computing time.

Figure 3 illustrates the output from a set of 300 runs of the annealing algorithm, with $N = 441$ points arranged in a square grid, from which $n = 16$ locations are chosen. There are $\binom{441}{16} \approx 7 \times 10^{28}$ possible designs. The parameters used were $\alpha, \beta, \gamma = (0, 0, 3)$, the correlations were of the exponential type: $g_0(\mathbf{t}, \mathbf{t}') = \sigma_2^2 \exp\{-\lambda\|\mathbf{t} - \mathbf{t}'\|\}$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, loss was TMSPE and the regressors were

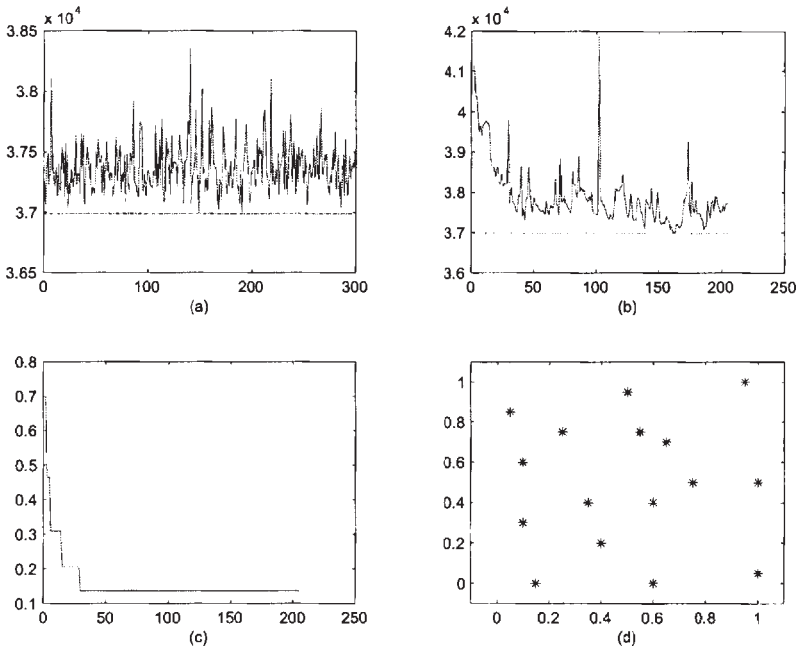

Figure 3. Output from 300 runs of the simulated annealing algorithm with the 20% nearest neighbour modification. Design space $\mathcal{T}$ is a 21×21 point grid from which $n = 16$ sites are chosen. Regressors are $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^{\mathrm{T}}$ and the parameters are $\alpha = 0, \sigma_1^2 = 1, \beta = 0, \sigma_2^2 = 2, \gamma = 3$. (a) Minimum loss (TMSPE) vs. run. (b) Accepted loss vs. stage in best run. (c) Acceptance probabilities vs. stage in best run. (d) Best design found has sites {4, 13, 42, 93, 129, 176, 181, 226, 231, 255, 308, 321, 327, 359, 410, 440}

$\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^{\mathrm{T}}$. The 20% nearest neighbour modification to the annealing algorithm was employed, as were the simplifications of Section 2.1. Each run required 14.4 s of computing time and 427 evaluations of the loss, on average.

Although it is time consuming, we can successfully run the annealing algorithm with inputs at least as large as $n = 100$, $N = 5000$. Of course, designs for predicting X_{Total} can be obtained much more quickly.

3.3. Sequential design

Choosing the design points sequentially is an obvious alternative, and one which we have also investigated. One 'run' of our algorithm consists of randomly choosing p points from $\mathcal{T}$, then finding that $(p+1)$th point which minimizes the loss when appended to the current p-point design, and repeating until n points have been determined. We run the algorithm numerous times, and choose the best of the resulting designs.

For each of the scenarios of Figure 1, we ran the sequential procedure 350 times, thus using about the same amount of time as 10 runs of the annealing algorithm. Although the optimal designs were never found in this way, the sequentially determined designs were at least close to optimal. The amounts by which their loss exceeded the minimum loss, expressed as a percentage of the minimum loss, were typically <1.5 per cent.

For large values of N it does not seem feasible to carry out many runs of the sequential approach. There is no substitute here for the nearest neighbour modification, which drastically reduces the number of evaluations of the loss which are required in the annealing algorithm. Figure 4 shows the result of 30 runs of the sequential procedure using the same inputs as in Figure 3. Each run required 5617 evaluations of the loss, and the total sequence of runs required 74 min of computing time—2 min more than 300 runs of the simulated annealing algorithm. The loss for the best design found exceeded that in Figure 3 by 2 per cent.

As we are applying it here, our algorithm is sequential but not adaptive. It could, however, be applied adaptively, with estimates updated at each stage as information accrues. See Chao and Thompson (2001) for a discussion of optimal adaptive sampling.

4. EXAMPLE

This example continues that of Wiens (2004), in which we restudy the 'coal-ash' data, given by Gomez and Hazen (1970) and described in Cressie (1993). There are 208 coal-ash core measurements

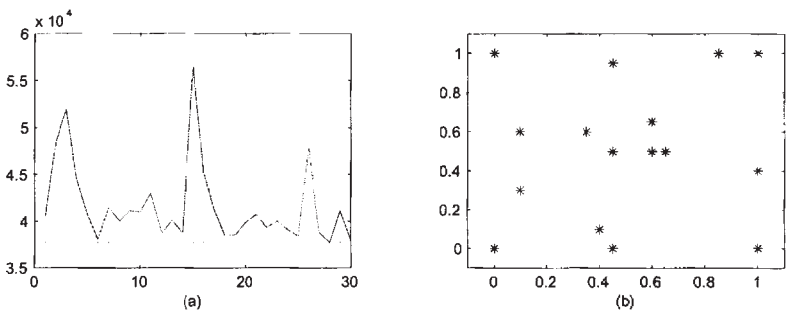

Figure 4. Output from 30 runs of the sequential procedure. Inputs are the same as for Figure 3. (a) Loss vs. run. (b) Best design found has sites {1, 10, 21, 51, 129, 189, 220, 223, 224, 255, 260, 286, 409, 421, 438,441}

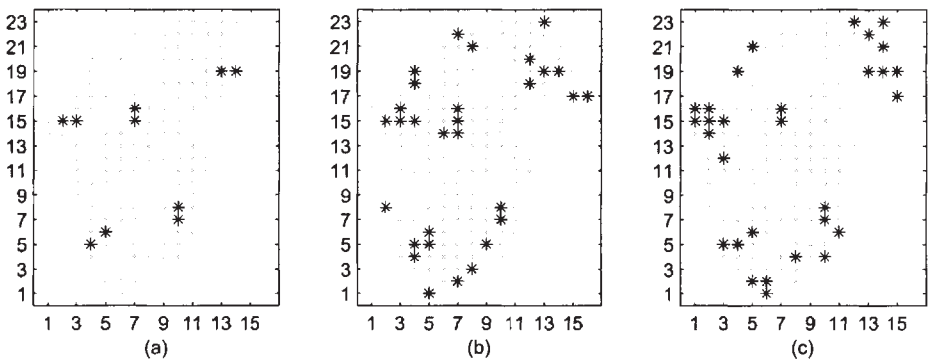

Figure 5. Designs for the example of Section 4: (a) initial design; (b) final design based on robust initial estimates and procedures; (c) final design using initial GLS estimates and $\alpha = \beta = \gamma = 0$

obtained from the Pittsburgh coal seam, at sites throughout a grid, as displayed in Figure 5. We seek an efficient and robust design upon which to base regression estimates of the effects in the east–west and north–south directions, corresponding to positive and negative values of t_1 and t_2, respectively. An initial ten-point design, given in Figure 5(a), was chosen. We intentionally included the location $(5, 6)$ of an acknowledged outlier (Cressie, 1993). An isotropic correlation model with $\rho_\lambda(d) = (1 + \lambda d)^{-2}$ was decided upon. We then fitted a regression model, with regressors $(1, t_1, t_2)$, to these data. We did this in two ways. In the first, M-estimates as described in Wiens (2004) were computed. In the second, GLS estimates were used. The resulting initial estimates are given in Table 1.

We then determined, by simulated annealing, designs consisting of a further 20 points chosen to minimize the maximum (estimated) loss (5). Corresponding to the robust initial estimates we used $\alpha = \hat{\sigma}_1^2$, $\beta = \hat{\sigma}_2^2$, $\gamma = 3\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)$. With the GLS estimates we used $\alpha = \beta = \gamma = 0$. The resulting designs—see Figures 5(b), (c)—were somewhat different. The parameters were then re-estimated and predictions made—after updating α and β in the robust case. Note the implausibly large GLS estimate $\hat{\lambda} = 31.38$, implying that even observations from nearest neighbours are essentially uncorrelated. See Figure 6, where the GLS predicted values, and those obtained from the M-estimates, are plotted against each other. The outlier at $(5, 6)$, with $Y_{5,6} = 17.61$, clearly has a much greater influence on the GLS fit ($\hat{X}_{5,6} = 14.67$) than on the robust fit ($\hat{X}_{5,6} = 10.80$).

Table 1. Parameter estimates (standard errors in parentheses) for coal-ash study

n	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
			M-estimates			
10	11.57 (0.279)	−0.20 (0.026)	0.03 (0.021)	0.20	0.76	24.20
30	10.87 (0.186)	−0.24 (0.016)	0.06 (0.011)	0.94	0.77	0.027
			GLS estimates			
10	11.78 (0.139)	−0.21 (0.015)	−0.01 (0.009)	1.11	1.46	27.07
30	10.71 (0.133)	−0.16 (0.013)	0.05 (0.009)	1.08	1.64	31.38

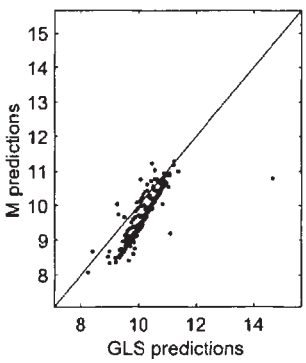

Figure 6. M-estimation based coal-ash predictions against the sorted GLS predictions

5. SUMMARY

In this article we have derived robust methods for the construction of sampling designs for spatial processes. Our framework assumes that the stochastic process of interest is itself subject to measurement error, and has a mean structure relying on a set of regressors. The measurement error variances, the correlations between observations made at differing locations, and the regression structure are all typically only partially known, and may be incorrectly specified by the experimenter. We have maximized a loss function, based on the mean squared error of this predictor, over neighbourhoods quantifying the uncertainty in the specification of the regression function. Two algorithms—one using simulated annealing and the other sequential in nature—have been introduced in order to minimize the maximum loss, leading to minimax designs.

APPENDIX: DERIVATIONS

Proof of Theorem 2: To carry out the maximization we shall first ignore the orthogonality constraint in the definition of $\mathcal{H}_\gamma$. We will then verify that the unconstrained maximizer also satisfies the constraints. Let $\mathbf{h}, \boldsymbol{\delta}, \boldsymbol{\varepsilon}$ be the $N \times 1$ vectors with elements $h(\mathbf{t}), \delta(\mathbf{t}), \varepsilon(\mathbf{t})$, $\mathbf{t} \in \mathcal{T}$. Then the data vector is

$$\mathbf{y} = \mathbf{Q}_1(\mathbf{x} + \boldsymbol{\varepsilon}) = \mathbf{Q}_1(\mathbf{Z}\boldsymbol{\theta} + \mathbf{h} + \boldsymbol{\delta} + \boldsymbol{\varepsilon})$$

and so for any linear predictor $\mathbf{Ay}$ we have that

$$\mathbf{Ay} - \mathbf{Cx} = \mathbf{BZ}\boldsymbol{\theta} + \mathbf{Bh} + \mathbf{B}\boldsymbol{\delta} + \mathbf{AQ}_1\boldsymbol{\varepsilon}$$

Note that the unbiasedness condition (1) is equivalent to the requirement $\mathbf{BZ} = \mathbf{0}_{M\times p}$, whence $\mathbf{Ay} - \mathbf{Cx} = \mathbf{Bh} + \mathbf{B}\boldsymbol{\delta} + \mathbf{AQ}_1\boldsymbol{\varepsilon}$, with

$$\mathrm{tr}\Big(\mathrm{E}\Big[(\mathbf{Ay} - \mathbf{Cx})(\mathbf{Ay} - \mathbf{Cx})^{\mathrm{T}}\Big]\Big) = \mathrm{E}\Big[\|\mathbf{B}\boldsymbol{\delta} + \mathbf{AQ}_1\boldsymbol{\varepsilon}\|^2\Big] + \|\mathbf{Bh}\|^2$$

The first summand above is $\mathcal{L}_0(\mathbf{A})$ and the second is maximized, subject to $N^{-1}\|\mathbf{h}\|^2 \leq \gamma$, iff $\mathbf{h}$ is an eigenvector of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$, with $N^{-1}\|\mathbf{h}\|^2 = \gamma$, corresponding to the largest eigenvalue $\lambda_{\max}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right) = \lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)$. But then $\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{h} = \lambda_{\max}\mathbf{h}$ and so

$$\mathbf{Z}^{\mathrm{T}}\mathbf{h} = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{h}}{\lambda_{\max}} = \mathbf{0}$$

since $\mathbf{B}\mathbf{Z} = \mathbf{0}$. Thus the orthogonality constraint (4) of $\mathcal{H}_\gamma$ is also satisfied. □

Proof of Lemma 3: (i) First assume A1. Then $\mathbf{Q}_1\mathbf{P} \to \mathbf{V} \triangleq \mathbf{Z}_1\left(\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^{\mathrm{T}}$. Disregarding terms which are $o(1)$, we have $\mathbf{Q}_1\left(\mathbf{P} - \mathbf{Q}_1^{\mathrm{T}}\mathbf{V}\right) = \mathbf{0}$. Thus the columns of $\mathbf{P} - \mathbf{Q}_1^{\mathrm{T}}\mathbf{V}$ are orthogonal to the rows of $\mathbf{Q}_1$, and hence are linear combinations of the columns of $\mathbf{Q}_2^{\mathrm{T}}$, i.e. $\mathbf{P} = \mathbf{Q}_1^{\mathrm{T}}\mathbf{V} + \mathbf{Q}_2^{\mathrm{T}}\mathbf{M}$ for some $\mathbf{M}_{(N-n)\times n}$ (necessarily $= \mathbf{Q}_2\mathbf{P}$, implying that $\mathbf{M}\mathbf{V} = \mathbf{M}$). Thus

$$\mathbf{B} = \mathbf{C}\mathbf{Q}^{\mathrm{T}}\begin{pmatrix} -(\mathbf{I}-\mathbf{V})\mathbf{Q}_1 \\ \mathbf{M}\mathbf{Q}_1 - \mathbf{Q}_2 \end{pmatrix}$$

Note that $(\mathbf{I}-\mathbf{V})\mathbf{Q}_1\left(\mathbf{Q}_1^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} - \mathbf{Q}_2^{\mathrm{T}}\right) = \mathbf{0}$. Using this, we calculate that

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{C}\mathbf{Q}^{\mathrm{T}}\left[(\mathbf{I}-\mathbf{V}) \oplus \left(\mathbf{I}_{N-n} + \mathbf{M}\mathbf{M}^{\mathrm{T}}\right)\right]\mathbf{Q}\mathbf{C}^{\mathrm{T}}$$

It now follows from either (6) or (7), and the fact that $\mathbf{I}-\mathbf{V}$ has eigenvalues 0 and 1, that $\lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right) = 1 + \lambda_{\max}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right)$.

Write $\mathbf{V} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_p$. Then,

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{V} + \mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{U}\left(\mathbf{I}_p + \mathbf{U}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{U}\right)\mathbf{U}^{\mathrm{T}}$$

has the same non-zero eigenvalues as $\mathbf{I}_p + \mathbf{U}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{U}$, so that the maximum eigenvalue is

$$\lambda_{\max}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right) = 1 + \lambda_{\max}\left(\mathbf{U}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{U}\right) = 1 + \lambda_{\max}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{V}\right) = 1 + \lambda_{\max}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right) = \lambda_{\max}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right)$$

(ii) Under A2 we have that $\mathbf{Q}_1\mathbf{P} \to \mathbf{I}_n$ and the remainder of the derivation is very similar to, but simpler than, that under A1. □

ACKNOWLEDGEMENTS

This work has benefited from discussions with Julie Zhou, University of Victoria and Subhash Lele, University of Alberta. The research is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Benedetti R, Palma D. 1995. Optimal sampling designs for dependent spatial units. *Environmetrics* **6**: 101–114.
Chao C-T, Thompson SK. 2001. Optimal adaptive selection of sampling sites. *Environmetrics* **12**: 517–538.

Cressie N. 1993. *Statistics for Spatial Data*. Wiley: New York.
Fedorov VV, Hackl P. 1994. Optimal experimental design: spatial sampling. *Calcutta Statistical Association Bulletin* **44**: 57–81.
Gomez M, Hazen K. 1970. Evaluating sulphur and ash distribution in coal seams by statistical response surface regression analysis, *U.S. Bureau of Mines Report R1 7377*.
Martin RJ. 1986. On the design of experiments under spatial correlation, (corr: **75**, p. 396). *Biometrika* **73**: 247–277.
Martin RJ. 2001. Comparing and contrasting some environmental and experimental design problems. *Environmetrics* **12**: 272–287.
McArthur RD. 1987. An evaluation of sample designs for estimating a locally concentrated pollutant. *Communications in Statistics, Part B—Simulation and Computation* **16**: 735–759.
Royle JA. 2002. Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference* **100**: 121–134.
Sacks J, Schiller S. 1988. Spatial designs. In *Statistical Decision Theory and Related Topics IV*, Vol. 2, Gupta SS, Berger JO (eds). Springer-Verlag: New York; 385–395.
Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989. Design and analysis of computer experiments. *Statistical Science* **4**: 409–423.
Santner TJ, Williams BJ, Notz WI. 2003. *The Design and Analysis of Computer Experiments*. Springer Verlag: New York.
Schilling, Mark F. 1992. Spatial designs when the observations are correlated. *Communications in Statistics, Part B—Simulation and Computation* **21**: 243–267.
Sinha S, Wiens DP. 2002. Robust sequential designs for nonlinear regression. *The Canadian Journal of Statistics* **30**: 601–618.
Stein ML. 1995. Locally lattice sampling designs for isotropic random fields. *The Annals of Statistics* **23**: 1991–2012.
Thompson SK. 1997. Effective sampling strategies for spatial studies. *Metron* **55**: 3–21.
van Groenigen J-W, Stein A. 1998. Constrained optimisation of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality* **27**: 1078–1086.
Wiens DP. 2005. Robustness in spatial studies I: minimax prediction. *Environmetrics*, accepted.

