ENVIRONMETRICS
Environmetrics 2005; **16**: 191–203
Published online 13 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.700

Robustness in spatial studies I: minimax prediction

Douglas P. Wiens*,†

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

SUMMARY

We develop and test robust methods for estimation and for prediction in spatial studies. We assume that a stochastic process is measured, with error, at various locations. The variance/covariance structures of this process and of the measurement errors are only approximately known; in the face of these uncertainties one is to do robust estimation and prediction. We obtain a minimax linear predictor, in which mean squared error loss is first maximized over neighbourhoods quantifying the various sources of model uncertainty, and then minimized over the coefficients of the predictor subject to a constraint of unbiasedness. Robustifications of these methods are then introduced. These are based on generalized M-estimators, and are robust against contaminated error distributions. In a simulation study the procedures afford a substantial level of robustness when the model inadequacies are present, while being almost as efficient as more classical methods otherwise. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: computer experiments; environmental monitoring; generalized M-estimation; isotropic; kriging; minimax; M-estimate

1. INTRODUCTION

Consider the following problem, of interest in environmetrics. There is a set $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$ of locations at which environmental monitoring stations may be placed. The agency responsible chooses, from these, locations $\mathcal{S} = \{\mathbf{t}_{i_1}, \ldots, \mathbf{t}_{i_n}\}$. At these locations they observe, with measurement error, a stochastic process $X(\mathbf{t})$ (air quality index, pollution level, etc.): $Y(\mathbf{t}) = X(\mathbf{t}) + \epsilon(\mathbf{t})$. The purpose is to predict $X(\mathbf{t})$, possibly for $\mathbf{t} \in \mathcal{T} \setminus \mathcal{S}$. Suppose, however, that—as is typically the case in spatial studies—the spatial correlation structure of $\{X(\mathbf{t}) | \mathbf{t} \in \mathcal{T}\}$ is possibly misspecified and that $\epsilon(\mathbf{t})$, while uncorrelated with $X(\mathbf{t})$ and with $\epsilon(\mathbf{t}')$ for $\mathbf{t} \neq \mathbf{t}'$, has a possibly misspecified variance structure. In the face of these model uncertainties, one is to do (robust) estimation and prediction.

The above is an outline of the problem addressed in this article, the sole generalization being that no restriction is made on the dimension of $\mathbf{t} \in \mathbb{R}^d$. Although the need for robustness is less clear, similar scenarios are employed for the analysis of data from computer experiments—see Santner *et al.* (2003) and Sacks *et al.* (1989).

*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

We assume that:

- VAR $[\epsilon(\mathbf{t})] = f(\mathbf{t})$ for a variance function $f^2(\cdot)$. Define $\mathbf{F}_{N\times N} = \mathrm{diag}(f(\mathbf{t}_1), \ldots, f(\mathbf{t}_N))$.
- $X(\mathbf{t}) = \mathrm{E}[X(\mathbf{t})] + \delta(\mathbf{t})$, where $\mathrm{cov}[\delta(\mathbf{t}), \delta(\mathbf{t}')] = g(\mathbf{t}, \mathbf{t}')$ for a covariance function g. Define $\mathbf{G}_{N\times N}$ by $g_{ij} = g(\mathbf{t}_i, \mathbf{t}_j)$.

We aim to predict a set $\mathbf{Cx}$ of M linear functions of $\mathbf{x} = (X(\mathbf{t}_1), \ldots, X(\mathbf{t}_N))^{\mathrm{T}}$. It is our intention to obtain predictors which are robust against misspecified functions f and g, possibly misspecified as $f = f_0$ and $g = g_0$. For example, f could be erroneously specified as constant and g erroneously specified as isotropic.

We will proceed as follows. Let $\mathbf{y} = (Y(\mathbf{t}_{i_1}), \ldots, Y(\mathbf{t}_{i_n}))^{\mathrm{T}}$. For a matrix $\mathbf{A}_{M\times n}$ defining a set $\mathbf{Ay}$ of linear predictors, the mean squared prediction error matrix is

$$\textsc{mspe}\,(\mathbf{A}; f, g) = \mathrm{E}\Big[(\mathbf{Ay} - \mathbf{Cx})(\mathbf{Ay} - \mathbf{Cx})^{\mathrm{T}}\Big]$$

We shall restrict to the class $\mathcal{A}$ of $M \times n$ matrices satisfying the unbiasedness constraint

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

Our loss function is the trace of the MSPE matrix, maximized over neighbourhoods of f_0 and g_0 defined by

$$\mathcal{F}_\alpha = \left\{ f(\cdot) \big| \max_{\mathbf{t}\in\mathcal{T}} \mid f(\mathbf{t}) - f_0(\mathbf{t}) \mid \le \alpha, f(\mathbf{t}) \ge 0 \right\}$$
$$\mathcal{G}_\beta = \left\{ g(\cdot,\cdot) \mid \mathbf{0} \le \mathbf{G} \le \mathbf{G}^{(0)} + \beta\mathbf{K} \right\}$$

In these definitions, $\mathbf{F}^{(0)}$ and $\mathbf{G}^{(0)}$ are the covariance matrices under f_0 and g_0 and $\mathbf{G} \ge \mathbf{0}$ refers to the ordering by non-negative definiteness. The matrix $\mathbf{K}$ is a fixed n.n.d. matrix; natural choices to be investigated here are $\mathbf{K} = \mathbf{I}_N$ and $\mathbf{K} = \mathbf{G}^{(0)}$.

Thus the loss is

$$\mathcal{L}(\mathbf{A}) = \max_{\mathcal{F}_\alpha, \mathcal{G}_\beta}\ \mathrm{tr}\ (\textsc{mspe}(\mathbf{A}; f, g))$$

and we seek

$$\mathbf{A}_{\alpha,\beta} = \arg \min_{\mathbf{A}\in\mathcal{A}} \mathcal{L}(\mathbf{A})$$

The problem of minimizing $\mathcal{L}(\mathbf{A})$ leads to an easily interpreted and intuitively pleasing result: the minimax linear predictor turns out to be the usual ‘universal kriging’ predictor, upon adding α to each diagonal element of the variance matrix $\mathbf{F}$ of the measurement errors, and $\beta\mathbf{K}$ to the covariance matrix $\mathbf{G}$ of $\{X(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$. There is some precedent for this approach of seeking robust linear estimates after maximizing over departures from the nominal model—Marcus and Sacks (1976), and Heckman (1987) are examples. In these cited works, however, the optimal linear estimator was not a least squares estimate.

Example 1: If $p = 1$ and $z(\mathbf{t}) = 1$, then $X(\mathbf{t})$ has constant mean θ. Then if $M = 1$ and $\mathbf{C} = (1, \ldots, 1)$ we are predicting $X_{\text{Total}} = \sum_{\mathbf{t}\in\mathcal{T}} X(\mathbf{t})$ by a linear function $\hat{X}_{\text{Total}} = \mathbf{a}^{\mathrm{T}}\mathbf{y}$. The loss in this case is $\mathcal{L}(\mathbf{a}) = \max_{\mathcal{F}_\alpha,\mathcal{G}_\beta}\mathrm{E}[(\hat{X}_{\text{Total}} - X_{\text{Total}})^2]$.

Example 2: 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

$$\max_{\mathcal{F}_\alpha,\mathcal{G}_\beta} \sum_{\mathbf{t}\notin\mathcal{S}} \mathrm{E}\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right]$$

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

$$\max_{\mathcal{F}_\alpha,\mathcal{G}_\beta} \sum_{\mathbf{t}\in\mathcal{T}} \mathrm{E}\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right]$$

the total mean squared prediction error.

The minimax linear predictor is obtained in Section 2. Of course there are strict limits to the amount of robustness which can be expected of any linear procedure, and so we discuss, in Section 3, a ‘robustified’ predictor, optimized for use with a generalized M-estimate (GM-estimate) of $\boldsymbol{\theta}$. Robust estimation of the nuisance parameters is considered as well.

We also give the results of a simulation study in Section 3. A summary of our findings is that considerable benefits are to be gained by robust estimation and prediction procedures, when the error distribution is contaminated, for only a small premium in efficiency at the assumed model. In Section 4 we revisit a data set from the literature, in order to illustrate an application of our methods.

2. MINIMAX LINEAR PREDICTION

We first determine the loss $\mathcal{L}(\mathbf{A})$ for a fixed $M \times n$ matrix $\mathbf{A}$. The observed data vector $\mathbf{y}$ may be decomposed as $\mathbf{y} = \mathbf{Q}(\mathbf{x} + \boldsymbol{\varepsilon})$, where $\mathbf{Q} : n \times N$ is the incidence matrix for $\mathcal{S}$, $\mathbf{x} = (X(\mathbf{t}_1), \ldots, X(\mathbf{t}_N))^{\mathrm{T}}$ and $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{t}_1), \ldots, \varepsilon(\mathbf{t}_N))^{\mathrm{T}}$. For a generic $N \times N$ matrix $\mathbf{M}$ we write $\mathbf{M}_1$ for $\mathbf{QM}$ and $\mathbf{M}_{11}$ for $\mathbf{QMQ}^{\mathrm{T}}$. Thus $\mathbf{M}_{11}$ refers only to the elements of $\mathbf{M}$ corresponding to sample values. Define:

$$\begin{aligned} \mathbf{Z} &= (\mathbf{z}(\mathbf{t}_1), \ldots, \mathbf{z}(\mathbf{t}_N))^{\mathrm{T}} \\ \mathbf{Z}_1 &= \mathbf{QZ} \, : \, n \times p \\ \boldsymbol{\Sigma}_{11} &= \mathbf{F}_{11} + \mathbf{G}_{11} \, : \, n \times n \end{aligned}$$

In this notation the covariance matrix of $\mathbf{y}$ is $\boldsymbol{\Sigma}_{11}$, assumed positive definite. With

$$\mathbf{B}_{M\times N} \overset{\triangle}{=} \mathbf{AQ} - \mathbf{C}$$

the unbiasedness condition (1) is equivalent to

$$\mathbf{BZ} = \mathbf{0}_{M\times p}$$

The form of the maximum loss $\mathcal{L}(\mathbf{A})$ follows immediately from the definitions of $\mathcal{F}_\alpha$ and $\mathcal{G}_\beta$.

Theorem 1. *Let* $\mathbf{A}$ *be a fixed* $M \times n$ *matrix defining a linear unbiased predictor* $\mathbf{Ay}$ *of* $\mathbf{Cx}$. *For fixed functions f and g, we have*

$$\mathrm{tr}(\textsc{mspe}(\mathbf{A};f,g)) = \mathrm{tr}\big(\mathbf{A}\mathbf{F}_{11}\mathbf{A}^{\mathrm{T}}\big) + \mathrm{tr}(\mathbf{BGB}^{\mathrm{T}})$$

The maximum over $f \in \mathcal{F}_\alpha,\ g \in \mathcal{G}_\beta$ *of* $\mathrm{tr}(\textsc{mspe}(\mathbf{A};f,g))$ *is*

$$\mathcal{L}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}\big(\mathbf{F}_{11}^{(0)} + \alpha\mathbf{I}_n)\mathbf{A}^{\mathrm{T}}) + \mathrm{tr}(\mathbf{B}(\mathbf{G}^{(0)} + \beta\mathbf{K})\mathbf{B}^{\mathrm{T}})$$

where the superscript '0' denotes evaluation at f_0 *and* g_0.

Define:

$$\begin{aligned}\mathbf{\Lambda}_{\alpha,\beta} &= \mathbf{\Sigma}_{11}^{(0)} + \alpha\mathbf{I}_n + \beta\mathbf{K}_{11}\ :\ n \times n\\ \mathbf{H}_\beta &= \mathbf{G}^{(0)} + \beta\mathbf{K}\ :\ N \times N\end{aligned}$$

Note that $\mathcal{L}(\mathbf{A})$ is the trace of the MSPE matrix of $\mathbf{Ay}$, in predicting $\mathbf{Cx}$, if $\mathbf{F}_{11}^{(0)}$ is replaced by $\mathbf{F}_{11}^{(0)} + \alpha\mathbf{I}_n$, $\mathbf{G}^{(0)}$ by $\mathbf{H}_\beta$ and hence $\mathbf{\Sigma}_{11}^{(0)}$ by $\mathbf{\Lambda}_{\alpha,\beta}$. The minimax linear predictor, subject to (1), is then the universal kriging predictor computed from $\mathbf{\Lambda}_{\alpha,\beta}$ and $\mathbf{H}_\beta$. It can be described in terms of

$$\mathbf{R}_{\alpha,\beta} = \left(\mathbf{Z}_1^{\mathrm{T}}\mathbf{\Lambda}_{\alpha,\beta}^{-1}\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^{\mathrm{T}}\mathbf{\Lambda}_{\alpha,\beta}^{-1}\ :\ p \times n$$

Theorem 2. *The minimax unbiased linear predictor of* $\mathbf{Cx}$ *is* $(\widehat{\mathbf{Cx}})_{\mathrm{LIN}} = \mathbf{A}_{\alpha,\beta}\mathbf{y}$, *where* $\mathbf{A}_{\alpha,\beta} = \mathbf{CP}_{\alpha,\beta}\ :\ M \times n$ for

$$\mathbf{P}_{\alpha,\beta} = \mathbf{ZR}_{\alpha,\beta} + \mathbf{H}_{\beta,1}^{\mathrm{T}}\mathbf{\Lambda}_{\alpha,\beta}^{-1}\big(\mathbf{I}_n - \mathbf{Z}_1\mathbf{R}_{\alpha,\beta}\big)\ :\ N \times n \tag{2}$$

Minimax loss is

$$\mathcal{L}\big(\mathbf{A}_{\alpha,\beta}\big) = \mathrm{tr}\Big\{\mathbf{C}\Big[\mathbf{P}_{\alpha,\beta}\mathbf{\Lambda}\mathbf{P}_{\alpha,\beta}^{\mathrm{T}} - \mathbf{P}_{\alpha,\beta}\mathbf{H}_{\beta,1} - \mathbf{H}_{\beta,1}^{\mathrm{T}}\mathbf{P}_{\alpha,\beta}^{\mathrm{T}} + \mathbf{H}_\beta\Big]\mathbf{C}^{\mathrm{T}}\Big\}$$

Remarks:

1. It can be shown that Theorems 1 and 2 continue to hold, with $\mathbf{K} = \mathbf{I}$, if $\mathcal{G}_\beta$ is changed to

$$\mathcal{G}'_\beta = \left\{g(\cdot,\cdot)\middle|\left(\mathrm{tr}(\mathbf{G} - \mathbf{G}^{(0)})^2\right)^{1/2} \le \beta, \mathbf{G} \ge \mathbf{0}\right\}$$

and $\mathcal{L}(\mathbf{A})$ to

$$\mathcal{L}'(\mathbf{A}) = \sum_{i=1}^{M} \max_{\mathcal{F}_\alpha,\mathcal{G}_\beta} \textsc{mspe}(\mathbf{A};f,g)_{ii}$$

the trace of the MSPE matrix, with each diagonal element maximized independently.

2. In terms of $\mathbf{R}_{0,0} = \big(\mathbf{Z}_1^{\mathrm{T}}[\boldsymbol{\Sigma}_{11}^{(0)}]^{-1}\mathbf{Z}_1\big)^{-1}\mathbf{Z}_1^{\mathrm{T}}[\boldsymbol{\Sigma}_{11}^{(0)}]^{-1}$, the generalized least squares (GLS) estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{\mathrm{GLS}} = \mathbf{R}_{0,0}\mathbf{y}$. Then, with fitted values $\hat{\mathbf{x}} = \mathbf{Z}\hat{\boldsymbol{\theta}}_{\mathrm{GLS}}$, $\hat{\mathbf{y}} = \mathbf{Z}_1\hat{\boldsymbol{\theta}}_{\mathrm{GLS}}$ and residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, an alternate expression for the predictor is

$$(\widehat{\mathbf{Cx}})_{\mathrm{LIN}} = \mathbf{C}\big(\mathbf{P}_{\alpha,\beta}\mathbf{e} + \hat{\mathbf{x}}\big)$$

(Here we use the fact that $\mathbf{R}_{\alpha,\beta}\mathbf{Z}_1 = \mathbf{I}_p$, so that $\mathbf{I}_n - \mathbf{Z}_1\mathbf{R}_{\alpha,\beta}$ is orthogonal to $\hat{\mathbf{y}}$ and $\mathbf{P}_{\alpha,\beta}\hat{\mathbf{y}} = \hat{\mathbf{x}}$.) This highlights further possible sources of non-robustness of the predictor—outlying observations may have an undue influence both through their effect on $\hat{\mathbf{x}}$ (through $\hat{\boldsymbol{\theta}}$) and through their effect on the residual vector. In the next section we propose to address these points by replacing $\hat{\boldsymbol{\theta}}_{\mathrm{GLS}}$ by a generalized M-estimate $\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$, and $\mathbf{e}$ by a corresponding vector of robust residuals. We do this first for known values of $\mathbf{F}^{(0)}$ and $\mathbf{G}^{(0)}$, and then propose a robust method of estimation for these quantities as well.

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

3. GENERALIZED M-ESTIMATION AND PREDICTION

3.1. $\mathbf{F}^{(0)}$ *and* $\mathbf{G}^{(0)}$ *completely specified*

Let σ_n be a scale functional for $\boldsymbol{\varepsilon}$ such as $(n^{-1}\mathrm{tr}\mathbf{F}_{11}^{(0)})^{1/2}$, and put $\mathbf{S} = \boldsymbol{\Sigma}_{11}^{(0)}/\sigma_n^2$. Then we have

$$\mathbf{v} = \mathbf{U}\boldsymbol{\theta} + \boldsymbol{\eta}$$

where $\mathbf{v} = \mathbf{S}^{-1/2}\mathbf{y}$, $\mathbf{U} = \mathbf{S}^{-1/2}\mathbf{Z}_1$ and where $\boldsymbol{\eta} = \mathbf{S}^{-1/2}\mathbf{Q}(\boldsymbol{\varepsilon} + \boldsymbol{\delta})$ has covariance matrix $\sigma_n^2\mathbf{I}_n$. Here $\mathbf{S}^{1/2}$ is any matrix satisfying $\mathbf{S}^{1/2}\mathbf{S}^{1/2^{\mathrm{T}}} = \mathbf{S}$. For the numerical work we use lower triangular square roots obtained from the Choleski decomposition. Denote by $\{\mathbf{u}_i^{\mathrm{T}}\}_{i=1}^n$ the rows of $\mathbf{U}$.

For absolutely continuous, bounded, odd, weakly increasing functions ψ_i and a continuous, bounded, even function χ, the GM-estimate $\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$ and corresponding scale estimate $\hat{\sigma}_n$ are defined as members of any sequence satisfying:

$$\sum_{i=1}^n \psi_i\left(\frac{\hat{\eta}_i}{\hat{\sigma}_n}\right)\mathbf{u}_i = o_P(n^{1/2}) \tag{3}$$

$$\sum_{i=1}^n \left[\chi\left(\frac{\hat{\eta}_i}{\hat{\sigma}_n}\right) - \tau_n\right] = o_P(n^{1/2}) \tag{4}$$

for $\hat{\eta}_i = v_i - \mathbf{u}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$ and a bounded sequence of constants $\{\tau_n\}$. For consistency at the normal distribution we take $\tau_n = \mathrm{E}_\Phi[\chi(\eta/\sigma_n)]$. Common choices of ψ_i are $\psi_i(r) = w(\mathbf{u}_i)\psi(r/s(\mathbf{u}_i))$ for positive function $w(\cdot), s(\cdot)$. With $s(\mathbf{u}) \equiv 1$ this describes a Mallows-type GM estimate (Hill, 1977). Schweppe (Merrill and Schweppe, 1971) proposed $s(\mathbf{u}) \equiv w(\mathbf{u})$. Corresponding to these proposals, one-step estimates for linear models with independent errors were investigated by Simpson *et al.* (1992), and Coakley and Hettmansperger (1993). For ordinary M-estimators ($s(\mathbf{u}) \equiv w(\mathbf{u}) \equiv 1$, $\psi_i \equiv \psi$), Silvapullé (1992) proved asymptotic normality for fixed (by design) carriers in such models; Wiens (1996) extended these result to GM-estimators, and to approximately linear models.

See Field and Wiens (1994) for a study of one-step ordinary M-estimators of regression, under dependence.

To robustify the predictor, we replace the GLSE by $\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$, and the residuals $\mathbf{e} = \mathbf{S}^{1/2}(\mathbf{v} - \mathbf{U}\hat{\boldsymbol{\theta}}_{\mathrm{GM}})$ by robustified residuals $\mathbf{S}^{1/2}\hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ has elements $\hat{\sigma}_n\psi_i(\hat{\eta}_i/\hat{\sigma}_n)$. This results in the predictor

$$(\widehat{\mathbf{Cx}})_{\mathrm{GM}} = \mathbf{C}(\mathbf{PS}^{1/2}\hat{\mathbf{p}} + \hat{\mathbf{x}}_{\mathrm{GM}}) \tag{5}$$

where $\hat{\mathbf{x}}_{\mathrm{GM}} = \mathbf{Z}\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$.

3.2. $\mathbf{F}_0$ *and* $\mathbf{G}_0$ *unknown*

The preceding development has assumed a known covariance structure. In practice, one would typically posit a parametric form for this structure, and substitute parameter estimates. Robust methods for variogram estimation were studied by Cressie and Hawkins (1980) and by Genton (2001; see also references to earlier work therein). Both employ judiciously chosen order statistics of the differences $|X(\mathbf{t}) - X(\mathbf{t}')|$, or the second differences. (In particular, it is assumed in the references of this paragraph that $X(\mathbf{t})$ is observable, i.e. that $\sigma_1^2 = 0$.) Non-parametric covariogram estimation is studied by Genton and Gorsich (2002). Militino and Ugarte (1997) propose 1-step Schweppe-type GM-estimation of regression, after applying a transformation, based on the residuals from an initial least trimmed squares fit which ignores the dependence structure, to achieve an approximately diagonal covariance matrix.

We propose here a method of GM-estimation of the regression parameters, and correspondingly a robust method of prediction, appropriate when measurement errors are present. We suppose that the experimenter assumes the measurement errors $\varepsilon(\mathbf{t})$ to have common variance σ_1^2, and the covariance function to be of the form

$$g(\mathbf{t}, \mathbf{t}') = \sigma_2^2\rho_\lambda(\|\mathbf{t} - \mathbf{t}'\|)$$

for an isotropic correlation function ρ_λ depending on a, possibly multidimensional, parameter λ. In the notation of previous sections, and with $\boldsymbol{\Phi} = (\rho_\lambda(\|\mathbf{t}_j - \mathbf{t}_k\|))_{j,k=1}^N$, we have

$$\mathbf{F}_0 = \sigma_1^2\mathbf{I}_N,\ \mathbf{G}_0 = \sigma_2^2\boldsymbol{\Phi}$$

Under this model the data vector $\mathbf{y}$ has covariance matrix $\sigma_1^2\mathbf{S}$, where $\mathbf{S} = \boldsymbol{\Sigma}_{11}^{(0)}/\sigma_1^2 = \mathbf{I}_n + \zeta\boldsymbol{\Phi}_{11}$ for $\zeta = \sigma_2^2/\sigma_1^2$. Our rationale in what follows is that if the errors were Gaussian then the elements of $\mathbf{S}^{-1/2}(\mathbf{y} - \mathbf{Z}_1\boldsymbol{\theta})$, being uncorrelated, would be independent and so standard asymptotics would apply. We propose a series of iterations where, at each step, current parameter estimates are used to transform to approximately uncorrelated data values, from which revised estimates are computed.

Regression/scale step ('R/S'): Given a trial value $\hat{\mathbf{S}} = \mathbf{I}_n + \hat{\zeta}\hat{\boldsymbol{\Phi}}_{11}$, put $\mathbf{v} = \hat{\mathbf{S}}^{-1/2}\mathbf{y}$, $\mathbf{U} = \hat{\mathbf{S}}^{-1/2}\mathbf{Z}_1$ and obtain trial estimates $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{GM}}$ and $\hat{\sigma}_1 = \hat{\sigma}_n$ as at (3) and (4), with $\psi_i(r) = w(\mathbf{u}_i)\psi_c(r)$, $\chi(r) = \psi_c^2(r)$, $\psi_c(r) = \mathrm{sign}(r)\cdot\min(c, |r|)$ and

$$\tau_n = E_\Phi\left[\psi_c^2(\cdot)\right] = 1 - 2\Phi(-c) - 2c(\phi(c) - c\Phi(-c))$$

Specifically, we update the estimate by carrying out three iterations of reweighted least squares—see Huber (1981, §7.8):

$$\boldsymbol{\theta} \leftarrow \left(\mathbf{U}^{\mathrm{T}}\mathbf{W}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{W}\mathbf{v}$$

$$\hat{\sigma}_1^2 \leftarrow \frac{1}{(n-p)\tau_n}\sum \psi_c^2\left(\frac{v_i - \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}_1}\right)\hat{\sigma}_1^2$$

where

$$\mathbf{W} = \mathrm{diag}\left(\ldots, \frac{\psi_i\left(\frac{v_i - \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}_1}\right)}{v_i - \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\theta}}, \ldots\right)$$

Covariance step (‘C’): Put $\mathbf{r} = \mathbf{y} - \mathbf{Z}_1\boldsymbol{\theta}$ and write $r_i = r(\mathbf{s}_i)$, where $\mathbf{s}_i$ is the location at which r_i is obtained. Under the assumptions of homoscedasticity and isotropy the variogram is

$$\begin{aligned} 2\gamma_r(\|\mathbf{h}\|) &= \mathrm{VAR}[r(\mathbf{s}+\mathbf{h}) - r(\mathbf{s})] \\ &= 2(C(0) - C(\|\mathbf{h}\|)) \end{aligned}$$

say. The correlogram is then

$$\rho_r(\|\mathbf{h}\|) = 1 - \frac{\gamma_r(\|\mathbf{h}\|)}{C(0)}$$

In terms of ρ_λ, we have

$$\rho_\lambda(\|\mathbf{h}\|) = \frac{1+\zeta}{\zeta}\rho_r(\|\mathbf{h}\|)$$

We replace $\mathbf{r}$ by the residual vector $\hat{\mathbf{r}} = \mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\theta}}$, use the differences $\hat{r}_i - \hat{r}_j (i \neq j)$ to compute robust estimates $\hat{\gamma}_r(d_k)$ for selected distances $d_1, \ldots, d_K$, and then estimate λ by

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{k=1}^{K} w_k[\rho_\lambda(d_k) - \hat{\rho}(d_k)]^2$$

where:

$$\begin{aligned} \hat{\rho}(d_k) &= \left[\frac{1+\hat{\zeta}}{\hat{\zeta}}\hat{\rho}_r(d_k)\right]_0^1 \\ \hat{\rho}_r(d_k) &= 1 - \frac{\hat{\gamma}_r(d_k)}{\mathrm{MAD}(\mathbf{r})^2} \\ w_k &= 0.5 + 0.5I(\hat{\rho}(d_k) > 0) \end{aligned}$$

Here $[\cdot]_0^1$ denotes truncation at 0 and 1 and $\textsc{mad}(\mathbf{r})$ is the median absolute deviation of the residuals around their median, normalized through division by $\Phi^{-1}(0.75) = 0.6745$. From $\hat{\lambda}$ we compute

$$\hat{\mathbf{\Phi}} = (\rho_{\hat{\lambda}}(\|\mathbf{t}_j - \mathbf{t}_k\|))_{j,k=1}^{N}$$

To update ζ we exploit the relationship $\mathbf{S} - \mathbf{I} = \zeta\mathbf{\Phi}_{11}$ to carry out a one parameter least squares fit of the elements of $\hat{\mathbf{S}} - \mathbf{I}$ on those of $\hat{\mathbf{\Phi}}_{11}$, under the constraint that the slope $\hat{\zeta}$ be non-negative.

Finally, $\hat{\mathbf{S}}$ is updated to $\mathbf{I}_n + \hat{\zeta}\hat{\mathbf{\Phi}}_{11}$ and $\hat{\sigma}_2^2$ to $\hat{\sigma}_1^2\hat{\zeta}$, whence we return to R/S.

To estimate the variogram, we use a modification of a proposal due to Cressie and Hawkins (1980). Write $r_i = r(\mathbf{s}_i)$, where $\mathbf{s}_i$ is the location at which r_i is obtained. Divide the range of $\{\|\mathbf{s}_i - \mathbf{s}_j\|| 1 \leq i < j \leq n\}$ into K intervals I_k, of cardinality $L = n(n-1)/2K$ each and with medians $d_1, \ldots, d_K$. Compute

$$2\hat{\gamma}(d_k) = \left[\text{median}\left\{|r(\mathbf{s}_i) - r(\mathbf{s}_j)| : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k\right\}/\Phi^{-1}(0.75)\right]^2$$

Computational details: It seems generally sufficient to iterate between R/S and C about four times. We have used $c = 1.5$. In our simulations the only non-constant regressors are the locations $\mathbf{t}$, so that influential carriers are not an issue. Thus we took constant weights $w(\mathbf{u}_i) \equiv 1$ in R/S. We chose $L = [n/2]$, so that $K = n - I$ (n is even).

Robust prediction: The robust predictor can now be computed by substituting estimates into (2) and then (5):

$$\hat{\mathbf{H}} = \begin{cases} \hat{\sigma}_1^2\hat{\zeta} + \beta\mathbf{I}_N, & \mathbf{K} = \mathbf{I} \\ (1+\beta)\hat{\sigma}_2^2\hat{\mathbf{\Phi}}, & \mathbf{K} = \mathbf{G}_0 \end{cases}$$

$$\hat{\mathbf{\Lambda}} = \begin{cases} \hat{\sigma}_1^2\mathbf{S} + (\alpha+\beta)\hat{\mathbf{I}}_n, & \mathbf{K} = \mathbf{I} \\ \hat{\sigma}_1^2\hat{\mathbf{S}} + \alpha\mathbf{I}_n + \beta\hat{\sigma}_2^2\hat{\mathbf{\Phi}}_{11}, & \mathbf{K} = \mathbf{G}_0 \end{cases}$$

$$(\widehat{\mathbf{Cx}})_{\text{GM}} = \mathbf{C}(\hat{\mathbf{P}}\hat{\mathbf{S}}^{1/2}\hat{\mathbf{p}} + \mathbf{Z}\hat{\boldsymbol{\theta}}_{\text{GM}})$$

3.3. *Simulation study*

A simulation study was carried out to test these methods. The main conclusions remained approximately constant across a range of inputs, and are reported in detail here for two situations. In each, we considered an $N = 10 \times 20$ grid of equally spaced locations. We simulated 100 populations of size N, and from each randomly chose a sample of size $n = 20$. Various combinations of estimation and prediction methods were applied to each sample.

The distribution of the measurement error ε was either $\text{N}(0, \sigma_1^2 = 1)$ ('clean data') or ('contaminated data'), this distribution was mixed with 10% $\text{N}(0, \sigma_1^2 = 25)$ variables per sample. In the case of contaminated data we also introduced heteroscedasticity, by replacing each ε by $U\varepsilon$, with U being a randomly generated Uniform (1,2) variable. The marginal distribution of $\delta(\mathbf{t})$ was $\text{N}(0, \sigma_2^2 = 1)$. Thus $\zeta = 1$. The regression response was $\text{E}[X(\mathbf{t})] = \mathbf{z}^{\text{T}}(\mathbf{t})\boldsymbol{\theta}$, with $\boldsymbol{\theta} = (10, 10, 10)$.

In the first situation the true correlations were of an exponential form, with $\rho_\lambda(d) = e^{-\lambda d}$ for $d(\mathbf{t}) = \|\mathbf{t}\|$ and $\lambda = 30.58$ chosen for a nearest neighbour correlation of 0.2. In the 'clean data' simulations it was assumed that the experimenter would fit the true correlation model. In the case of contaminated data we assumed that the correlation model would be misspecified, and that the experimenter would fit Gaussian correlations $\rho_\lambda(d) = e^{-\lambda d^2}$.

In the second situation the true correlations were anisotropic, with $\rho_\lambda(d) = e^{-\lambda d}$ for $d(\mathbf{t}) = \|\mathbf{A}^{1/2}\mathbf{t}\|$ with $\mathbf{A} = \text{diag}(1, 2)$. We took $\lambda = 4.24$, for a nearest neighbour correlation of 0.8. We assumed that the experimenter would fit an (isotropic) exponential correlation model.

In both cases we took $\mathbf{K} = \mathbf{I}$, $\alpha = \hat{\sigma}_1^2$, $\beta = \hat{\sigma}_2^2$. With contaminated errors or covariances the 'best' parameters $(\sigma_1^2, \sigma_2^2, \lambda)$ are no longer those used in the simulations. Thus we base our comparisons on the accuracy in the estimation of θ and on the accuracy of the predictions of $X(\mathbf{t})$. The former is gauged by the percent relative norm of the bias of $\hat{\boldsymbol{\theta}}$:

$$\%RNB = 100 \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|}{\|\boldsymbol{\theta}\|}$$

The latter is gauged by the percent average relative prediction error:

$$\%ARPE = 100 \ \text{aver}_{\mathbf{t}\in\mathcal{T}} \left(\left| \frac{\hat{X}(\mathbf{t}) - X(\mathbf{t})}{X(\mathbf{t})} \right| \right)$$

with $\hat{X}(\mathbf{t})$ computed as at (5) with $\mathbf{C} = \mathbf{I}_N$ and by the relative total prediction error,

$$\%RTPE = 100 \frac{\sum_{\mathbf{t}\in\mathcal{T}} \|\hat{X}(\mathbf{t}) - X(\mathbf{t})\|}{\| \sum_{\mathbf{t}\in\mathcal{T}} X(\mathbf{t})\|}$$

The results from the simulations are presented graphically in Figures 1 and 2. In Figure 1 we give the $\%RNB$ for the classical and robust estimation methods. To assess the sampling variability we carried out paired sample t-tests of (equality versus) the hypothesis that the $\%RNB$ resulting from the GM estimates was significantly less than that resulting from the GLS estimates. The p-values are given at the ends of the bars. As one would hope, with respect to this measure the robust estimation method

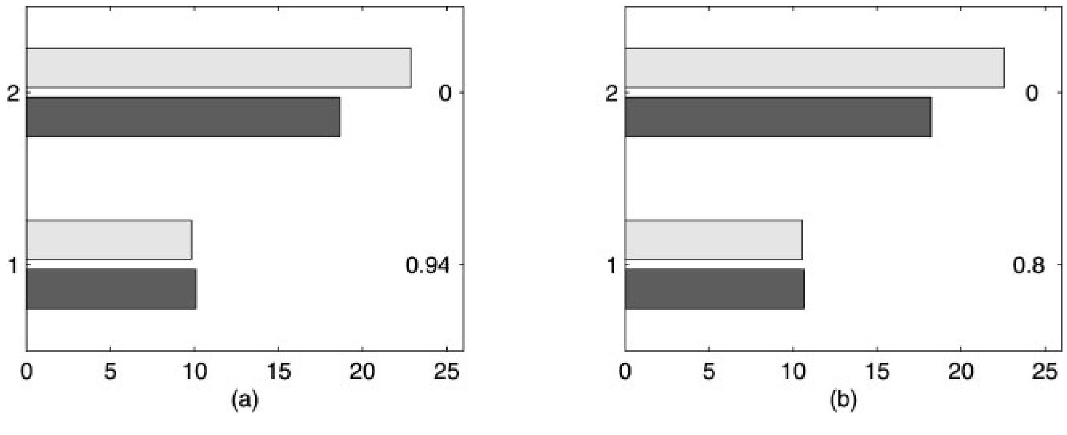

Figure 1. $\%RNB$ and accompanying p-values for the GLS estimator (light shading) and for the GM estimate (darkshading). Plot (a) 1: clean data, situation 1; 2: contaminated data, situation 1. Plot (b) 1: clean data, situation 2; 2: contaminated data, situation 2

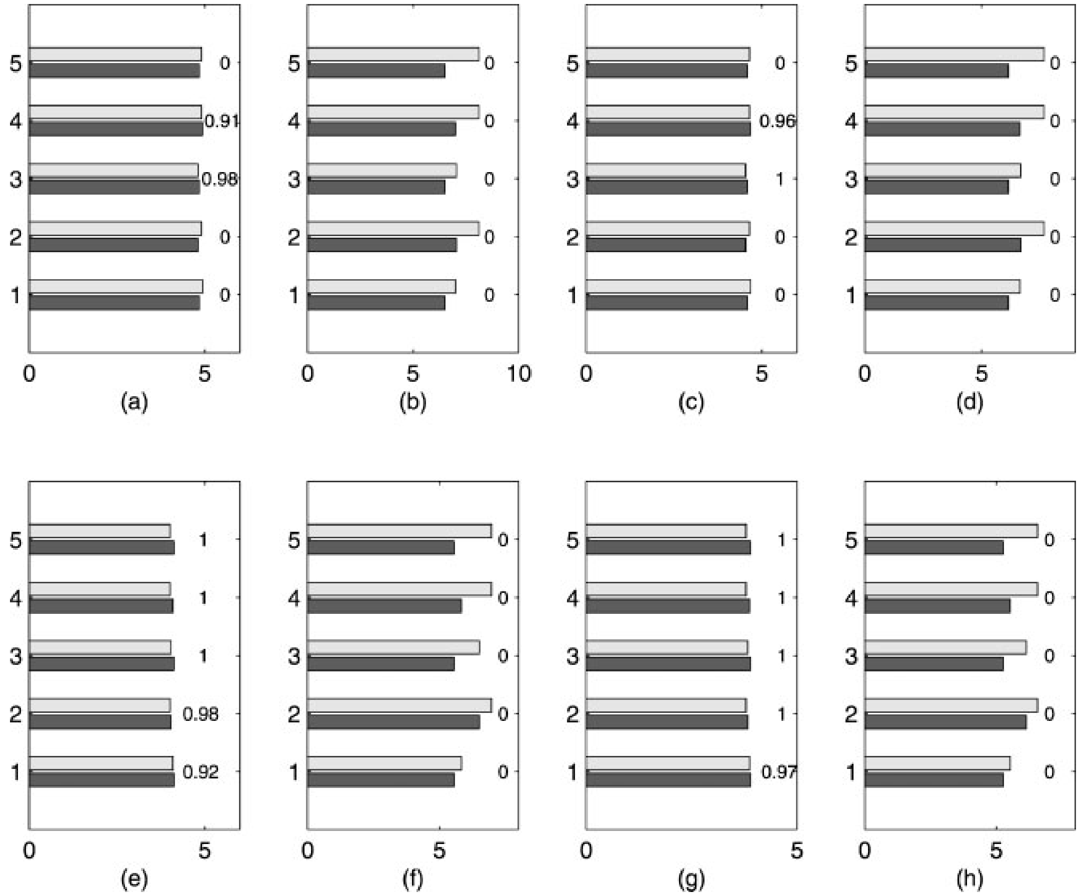

Figure 2. Measures of predictive accuracy: (a)–(d), situation 1; (e)–(h), situation 2. (a), (e): %*ARPE*, clean data; (b), (f): %*ARPE*, contaminated data; (c), (g): %*RTPE*, clean data; (d), (h): %*RTPE*, contaminated data. Contrast 1 compares the minimax predictions (darker shading) with the classical kriging predictions (lighter shading) when GM estimation was used. Contrast 2 compares instead the methods when the GLS estimate was used. Contrast 3 compares the effects of GM estimation (darker shading) and GLS estimation (lighter shading) on the minimax predictor; Contrast 4 instead assumes the use of kriging. Contrast 5 compares GM estimation + minimax prediction (darker shading) with GLS estimation + kriging

performs almost as well as least squares on clean data, and significantly better–both statistically and practically–when the data are perturbed.

In Figure 2 we give the %*ARPE* and %*RTPE* for five contrasting methods. As in Figure 1, the *p*-values measure the effectiveness of the new methods. Again the general message is clear–on clean data there is at most a small premium to be paid for the use of the robust methods, with significant protection obtained on contaminated data.

4. EXAMPLE

We have restudied the ‘coal-ash’ data given by Gomez and Hazen (1970) and described in Cressie (1993). There are 208 coal-ash core measurements obtained from the Pittsburgh coal seam, at sites

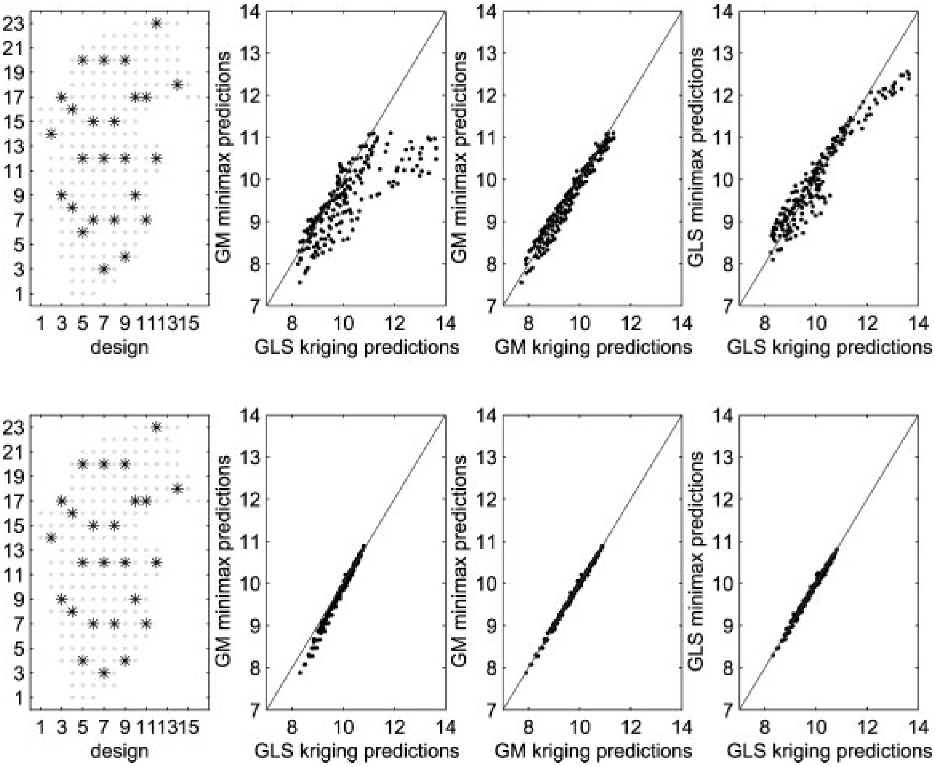

Figure 3. Designs and predictions for coal-ash study. Top row: design contains location (5, 6) of outlier. Bottom row: Location (5, 6) replaced by (3, 6). Predictions from various methods are plotted against each other

throughout a grid, as displayed in Figure 3. The object of our study is to obtain efficient and robust 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. We considered two sampling designs, of sizes $n = 25$ each and shown in the first column of plots in Figure 3. The first contains the location (5, 6) of an apparent outlier (Cressie, 1993). The second replaces this location by (5, 4). An isotropic correlation model with $\rho_\lambda(d) = (1 + \lambda d)^{-2}$ was decided upon. Using both GM and GLS estimates we obtained the values shown in Table 1. The corresponding standard errors of $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$, determined from the approximate covariance matrix

$$\text{approx.cov}\big[\hat{\boldsymbol{\theta}}\big] = \frac{\sigma_1^2}{n}\frac{\mathrm{E}\left[\psi^2\left(\frac{\eta}{\sigma_1}\right)\right]}{\left(\mathrm{E}\left[\psi'\left(\frac{\eta}{\sigma_1}\right)\right]\right)^2}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1} \tag{6}$$

are also given in Table 1. We call this an approximate covariance matrix since it is asymptotically correct only for independent v_i–see Section 3.1. In fact the v_i are at most only approximately uncorrelated, and hence approximately independent if Gaussian. The expectations in (6) were estimated by the corresponding sample averages (using '$n - p$' as the divisor for the expectation in

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

Design	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
			GM-estimates			
1	11.58 (0.29)	−0.24 (0.02)	0.01 (0.02)	1.49	0.79	0.02
2	11.33 (0.19)	−0.20 (0.02)	−0.01 (0.01)	1.34	0.26	0.00
			GLS estimates			
1	12.66 (0.39)	−0.21 (0.03)	−0.06 (0.02)	1.64	2.76	0.15
2	11.21 (0.21)	−0.17 (0.02)	−0.00 (0.01)	1.49	0.32	0.00

Table 2. Predictive measures for coal-ash study

Design	$\hat{X}_{50}$		$\sum_{\mathbf{t}}\left(\hat{X}(\mathbf{t}) - \mathbf{Y}(\mathbf{t})\right)^2$	
	Minimax	Kriging	Minimax	Kriging
		GM-estimates		
1	10.85	10.63	275	311
2	10.22	10.25	262	266
		GLS estimates		
1	14.91	14.22	324	489
2	10.24	10.28	254	260

the numerator). Bootstrapping, as studied recently for dependent data by Lahiri (2003), is a possibility which we have not yet explored.

The effect of the outlier on the GLS estimates was considerable; that on the GM-estimates was much less so.

We then obtained predictions, using various methods. They are plotted against each other in Figure 3. Some measures of the effect of the outlier $(Y_{50} = 17.61)$ are given in Table 2; again this effect is seen to be much more severe on the least-squares-based predictions.

Robust design strategies are studied and applied to these data in Wiens (2004).

5. SUMMARY

In this article we have derived robust methods for the estimation and prediction of 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 and the correlations between observations made at differing locations are typically only partially known, and may be incorrectly specified by the experimenter. We have exhibited a minimax linear predictor, in which mean squared error loss is first maximized over neighbourhoods quantifying the various sources of model uncertainty, and then minimized over the coefficients of the predictor subject to a constraint of unbiasedness. Robustifications of these methods have also been introduced. These are based on Generalized M-estimators, and are robust against contaminated error distributions.

A simulation study has shown that the procedures perform much as hoped, affording a substantial level of robustness when these model inadequacies are present, while being almost as efficient as more classical methods otherwise.

ACKNOWLEDGEMENTS

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

REFERENCES

Coakley CW, Hettmansperger TP. 1993. A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* **88**: 872–880.

Cressie NAC, Hawkins DM. 1980. Robust estimation of the variogram I. *Mathematical Geology* **12**: 115–125.

Cressie N. 1993. *Statistics for Spatial Data*. Wiley: New York.

Field CA, Wiens DP. 1994. One-step M-estimators in the linear model, with dependent errors. *Canadian Journal of Statistics* **22**: 219–231.

Genton MG. 2001. Robustness problems in the analysis of spatial data. In *Spatial Statistics: Methodological Aspects and Applications*, Moore M (ed.). Springer: New York; 21–37.

Genton MG, Gorsich DJ. 2002. Nonparametric variogram and covariogram estimation with Fourier–Bessel matrices. *Computational Statistics and Data. Analysis* **41**: 47–57.

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*.

Heckman NE. 1987. Robust design in a two treatment comparison in the presence of a covariate. *Journal of Statistical Planning and Inference* **16**: 75–81.

Hill RW. 1977. Robust regression when there are outliers in the carriers, *Ph.D. dissertation*, Harvard University, Cambridge, Mass.

Huber PJ. 1981. *Robust Statistics*. Wiley: New York.

Lahiri SN. 2003. *Resampling Methods for Dependent Data*. New York: Springer Verlag.

Marcus MB, Sacks J. 1976. Robust designs for regression problems. In *Statistical Theory and Related Topics II*, Gupta SS, Moore DS (eds). Academic Press: New York; 245–268.

Merrill HM, Schweppe FC. 1971. Bad data suppression in power system static state estimation. *IEEE Transactions on Power Applications and Systems* **90**: 2718–2725.

Militino AF, Ugarte MD. 1997. A GM estimation of the location parameters in a spatial linear model. *Communications in Statistics A* **26**: 1701–1725.

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.

Silvapullé MJ. 1985. Asymptotic behavior of robust estimators of regression and scale parameters with fixed carriers. *Annals of Statistics* **13**: 1490–1497.

Simpson DG, Ruppert D, Carroll RJ. 1992. On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* **87**: 439–450.

Wiens DP. 1996. Asymptotics of generalized M-estimation of regression and scale with fixed carriers, in an approximately linear model. *Statistics and Probability Letters* **30**: 271–285.

Wiens DP. 2005. Robustness in spatial studies II: minimax design. *Environmetrics*, accepted.