Board of the Foundation of the Scandinavian Journal of Statistics

Estimating the Prediction Mean Squared Error in Gaussian Stochastic Processes with Exponential Correlation Structure

Author(s): Markus Abt
Source: *Scandinavian Journal of Statistics*, Vol. 26, No. 4 (Dec., 1999), pp. 563-578
Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics
Stable URL: http://www.jstor.org/stable/4616579
Accessed: 10-10-2016 20:53 UTC

© Board of the Foundation of the Scandinavian Journal of Statistics 1999. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA Vol 26: 563–578, 1999

Estimating the Prediction Mean Squared Error in Gaussian Stochastic Processes with Exponential Correlation Structure

MARKUS ABT

Universität Augsburg

ABSTRACT. Given one or more realizations from the finite dimensional marginal distribution of a stochastic process, we consider the problem of estimating the squared prediction error when predicting the process at unobserved locations. An approximation taking into account the additional variability due to estimating parameters involved in the correlation structure was developed by Kackar & Harville (1984) and was revisited by Harville & Jeske (1992) as well as Zimmerman & Cressie (1992). The present paper discusses an extension of these methods. The approaches will be compared via an extensive simulation study for models with and without random error term. Effects due to the designs used for prediction and for model fitting as well as due to the strength of the correlation between neighbouring observations of the stochastic process are investigated. The results show that considering the additional variability in the predictor due to estimating the covariance structure is of great importance and should not be neglected in practical applications.

Key words: computer experiments, exponential covariance, Gaussian stochastic process, information matrix, kriging, mean squared error, spatial linear model, spatial statistics.

1. Introduction

We consider a real-valued Gaussian stochastic process $\{W(t)\colon t \in [0, 1]^d\}$ satisfying the model

$$W(t) = f(t)^{\mathrm{T}}\beta + Z(t) + \epsilon(t), \tag{1}$$

wherein $f(t) = (f_1(t), \ldots, f_k(t))^{\mathrm{T}}$ are k known regression functions and $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a vector of unknown parameters. The stochastic processes $Z(\cdot)$ and $\epsilon(\cdot)$ are assumed to be independent with zero mean and variances $\sigma^2 > 0$ and $\tau^2 > 0$, respectively. The process $\epsilon(\cdot)$ reflects uncorrelated measurement error, whereas $Z(\cdot)$ denotes the systematic deviation of $W(\cdot)$ from the mean $f(\cdot)^{\mathrm{T}}\beta$. The covariance structure of $Z(\cdot)$ is assumed to be of known form which will be specified later in section 3, but depends on a vector δ of unknown parameters describing the strength of the correlation between neighbouring observations.

Given observations $W_{\mathrm{p}} = (W(t^{(1)}), \ldots, W(t^{(n_p)}))^{\mathrm{T}}$ taken at the n_p locations in $D_p = \{t^{(1)}, \ldots, t^{(n_p)}\} \subset [0, 1]^d$, a frequently arising problem is that of predicting $W_T(t) = f(t)^{\mathrm{T}}\beta + Z(t)$. As an example, in environmental applications as in Abt *et al.* (1998), model (1) can be used to describe spatial contaminations. Based on observations taken at certain locations, the interest is typically in predicting the "true" contamination $W_T(t)$ at a location t rather than a future observation $W(t)$.

For predicting $W_T(t)$, we consider linear predictors of the form $\hat{W}(t) = c_p(t)^{\mathrm{T}} W_p$ for some vector $c_p(t) \in \mathbb{R}^{n_p}$. Following Welch *et al.* (1992), the vector $c_p(t)$ can be found by minimization of $\mathrm{MSE}[\hat{W}(t)] = \mathrm{E}[(\hat{W}(t) - W_T(t))^2]$ under the unbiasedness constraint $F_p^{\mathrm{T}} c_p(t) = f(t)$. Therein, the matrix $F_p \in \mathbb{R}^{(n_p, k)}$ is given by elements $(F_p)_{i,j} = f_j(t^{(i)})$ for $i = 1, \ldots, n_p$ and $j = 1, \ldots, k$. Furthermore, let a positive semidefinite matrix $R_p \in \mathbb{R}^{(n_p, n_p)}$ and a vector $r_p(t) \in \mathbb{R}^{n_p}$ be given through $(R_p)_{i,j} = \mathrm{cor}(Z(t^{(i)}), Z(t^{(j)}))$ and $(r_p(t))_i = \mathrm{cor}(Z(t), Z(t^{(i)}))$ for

$i, j = 1, \ldots, n_p$. Define $\Sigma_p = \sigma^2 R_p + \tau^2 I$ and $v_p(t) = \sigma^2 r_p(t)$. Then, using $z_p(t) = f(t) - F_p^{\mathrm{T}} \Sigma_p^{-1} v_p(t)$, the mean squared error of $\hat{W}(t)$ is minimized for $c_p(t) = B_p f(t) + A_p v_p(t)$, wherein $B_p = \Sigma_p^{-1} F_p (F_p^{\mathrm{T}} \Sigma_p^{-1} F_p)^{-1}$ and $A_p = \Sigma_p^{-1} - B_p F_p^{\mathrm{T}} \Sigma_p^{-1}$. This leads to

$$\mathrm{MSE}[\hat{W}(t)] = \sigma^2 - v_p(t)^{\mathrm{T}} \Sigma_p^{-1} v_p(t) + z_p(t)^{\mathrm{T}} (F_p^{\mathrm{T}} \Sigma_p^{-1} F_p)^{-1} z_p(t). \tag{2}$$

Note that the predictor $\hat{W}(t)$ as well as its mean squared error in (2) depend on parameters involved in the covariance structure of $W(\cdot)$. In practice these are unknown and need to be estimated from the data by maximum likelihood, for example. Replacing the unknown parameters by their estimates leads to additional variability in the predictor. This extra variability is not accounted for in (2). Modifications were therefore suggested by Kackar & Harville (1984) and are discussed in more detail in Harville & Jeske (1992) and Zimmerman & Cressie (1992). We briefly summarize their results in the next paragraph and refer to the original articles for more details and to Abt (1998) for their adaption to model (1) without random error term. We will denote by $|\delta|$ the number of components of the vector δ and by a superscript (k) attached to a matrix or a vector we mean the elementwise partial derivative with respect to δ_k.

Let $\hat{\delta}$ be the vector of maximum likelihood estimates for the unknown covariance parameters δ. To make the dependence of $\hat{W}(t)$ on these estimates more clear, we write it as $\hat{W}_{\hat{\delta}}(t) = c_{p,\hat{\delta}}(t)^{\mathrm{T}} W_{\mathrm{p}}$. With $h(\tau;\, t) := (c_{p,\tau}(t)^{\mathrm{T}} W_p - c_{p,\delta}(t)^{\mathrm{T}} W_p)^2$, we then have

$$\mathrm{MSE}[\hat{W}_{\hat{\delta}}(t)] = \mathrm{MSE}[\hat{W}_{\delta}(t)] + \mathrm{E}[h(\hat{\delta};\, t)].$$

A Taylor series expansion of $h(\hat{\delta};\, t)$ up to second order leads to the approximation

$$\mathrm{E}[h(\hat{\delta};\, t)] \doteq \mathrm{tr}\,(V_\delta G_\delta(t)) + \frac{1}{2} \sum_{k,l=1}^{|\delta|} \mathrm{cov}((\hat{\delta}_k - \delta_k)(\hat{\delta}_l - \delta_l),\, H_{k,l}(t)), \tag{3}$$

wherein $V_\delta \in \mathbb{R}^{(|\delta|,|\delta|)}$ denotes the mean squared error matrix $\mathrm{E}[(\hat{\delta} - \delta)(\hat{\delta} - \delta)^{\mathrm{T}}]$ or an approximation thereof, $G_\delta(t) \in \mathbb{R}^{(|\delta|,|\delta|)}$ is given by $(G_\delta(t))_{k,l} = c_{p,\delta}^{(l)}(t)^{\mathrm{T}} \Sigma_p c_{p,\delta}^{(k)}(t)$ for $k,\, l = 1, \ldots, |\delta|$, and $H_{k,l}(t) = 2 c_{p,\delta}^{(l)}(t)^{\mathrm{T}} W_p W_p^{\mathrm{T}} c_{p,\delta}^{(k)}(t)$. Note that both matrices, V_δ and $G_\delta(t)$, are symmetric.

At this point, Harville & Jeske (1992) as well as Zimmerman & Cressie (1992) assume independence of the maximum likelihood estimates $\hat{\delta}$ and the observations W_{p}. As a consequence, the covariances on the right hand side of (3) vanish, which leads to $\mathrm{MSE}[\hat{W}_{\hat{\delta}}(t)] \doteq \mathrm{MSE}[\hat{W}_\delta(t)] + \mathrm{tr}\,(V_\delta G_\delta(t))$. For model (1) without random error term, this approximation was investigated in Abt (1998). Therein, for different choices of δ, the "true" prediction error $\mathrm{MSE}[\hat{W}_{\hat{\delta}}(t)]$ was obtained by simulation and then compared to $\mathrm{MSE}[\hat{W}_\delta(t)] + \mathrm{tr}\,(V_\delta G_\delta(t))$. The latter was evaluated using the values of δ chosen for simulation rather than the maximum likelihood estimates $\hat{\delta}$. The approximation turned out to be quite good, provided that the correlations between neighbouring observations are not too weak.

Rather than approximating the mean squared prediction error, its estimation is of greater practical importance when, for example, prediction intervals are to be associated with predicted values. The estimate to be discussed here is based on evaluating the above approximation at the maximum likelihood estimates $\hat{\delta}$. Furthermore, as it was not clear *a priori* whether similarly good results as in Abt (1998) could be obtained, the approximation of $\mathrm{MSE}[\hat{W}_{\hat{\delta}}(t)]$ will be extended by relaxing the assumption of independence between $\hat{\delta}$ and W_p. This is the topic of section 2. Section 3 will then discuss an exponential correlation structure we will assume for the stochastic process $Z(\cdot)$ and suggest the use of the inverse Fisher information matrix as an approximation for the mean squared error matrix of the maximum likelihood estimates $\hat{\delta}$. This has been found appropriate earlier in Abt & Welch (1998). Section 4 describes the general layout of our simulation study and the evaluation criteria. Model (1) is then investigated in sections 5 and 6. Models used for computer experimentation are the topic of section 7. Concluding remarks follow in section 8.

2. Extending the prediction error approximation

In order to extend the above approximation of the mean squared prediction error, we look at the covariances on the right hand side of (3) more closely. Our idea to evaluate this term is based on a first order Taylor series expansion of $\hat{\delta}$. To do so, we assume that a second set of data $W_f = (W(s^{(1)}), \ldots, \mathrm{W}(s^{(n_f)}))^{\mathrm{T}}$ taken at the n_f sites in $D_f = \{s^{(1)}, \ldots, s^{(n_f)}\} \subset [0, 1]^d$ is available and used for model fitting, i.e. maximum likelihood estimation of δ, only. Due to the correlation structure of the process $W(\cdot)$, the data W_f and W_p and thus $\hat{\delta}$ and W_p will not be independent. Similarly to the definitions before, quantities related to the design D_f and the data W_f will be denoted by a subscript f. Before we continue, we need some preliminary lemmas.

Lemma 1
Let $B_p = \Sigma_p^{-1} F_p (F_p^{\mathrm{T}} \Sigma_p^{-1} F_p)^{-1}$ *and* $A_p = \Sigma_p^{-1} - B_p F_p^{\mathrm{T}} \Sigma_p^{-1}$. *Then* $B_p^{\mathrm{T}} \Sigma_p A_p = 0$, $A_p^{\mathrm{T}} \Sigma_p A_p = A_p$, *and* $B_p^{\mathrm{T}} \Sigma_p B_p = (F_p^{\mathrm{T}} \Sigma_p^{-1} F_p)^{-1}$. *Further,* $A_p^{(k)} = -A_p \Sigma_p^{(k)} A_p$ *and* $B_p^{(k)} = -A_p \Sigma_p^{(k)} B_p$ *leads to* $c_{p,\delta}^{(k)}(t) = A_p(v_p^{(k)}(t) - \Sigma_p^{(k)}(B_p f(t) + A_p v_p(t)))$.

Proof. The proof can be carried out using straightforward matrix algebra.

Lemma 2
Suppose $X \sim N_n(\mu, \Sigma)$ *and let* $A, B \in \mathbb{R}^{(n,n)}$ *be two not necessarily symmetric matrices. Then* $\mathrm{E}[X^{\mathrm{T}} A X] = \mathrm{tr}\,(A\Sigma) + \mu^{\mathrm{T}} A \mu$ *and*

$$\mathrm{cov}(X^{\mathrm{T}} A X, X^{\mathrm{T}} B X) = \mathrm{tr}\,(A \Sigma B \Sigma) + \mathrm{tr}\,(A \Sigma B^{\mathrm{T}} \Sigma) + \mu^{\mathrm{T}} (A + A^{\mathrm{T}}) \Sigma (B + B^{\mathrm{T}}) \mu.$$

Furthermore, for a vector $a \in \mathbb{R}^n$, *we have* $\mathrm{cov}(a^{\mathrm{T}} X, X^{\mathrm{T}} B X) = a^{\mathrm{T}} \Sigma (B + B^{\mathrm{T}}) \mu$

Proof. For symmetric matrices A and B, the first two results can be found in Schott (1997, sect. 9.6). The last result can be derived in a similar way. For the general case we symmetrize the problem by considering $\tilde{A} = (A + A^{\mathrm{T}})/2$ and $\tilde{B} = (B + B^{\mathrm{T}})/2$.

Corollary
Let $A \in \mathbb{R}^{(n,n)}$ *and* $B \in \mathbb{R}^{(m,m)}$ *be two matrices. Suppose that*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_{n+m}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^{\mathrm{T}} & \Sigma_{YY} \end{pmatrix} \right).$$

Then

$$\mathrm{cov}(X^{\mathrm{T}} A X, Y^{\mathrm{T}} B Y) = \mathrm{tr}\,(A \Sigma_{XY} B \Sigma_{XY}^{\mathrm{T}}) + \mathrm{tr}\,(A \Sigma_{XY} B^{\mathrm{T}} \Sigma_{XY}^{\mathrm{T}}) + \mu_X^{\mathrm{T}} (A + A^{\mathrm{T}}) \Sigma_{XY} (B + B^{\mathrm{T}}) \mu_Y$$

and for a vector for $a \in \mathbb{R}^n$ *we have* $\mathrm{cov}(a^{\mathrm{T}} X, Y^{\mathrm{T}} B Y) = a^{\mathrm{T}} \Sigma_{XY} (B + B^{\mathrm{T}}) \mu_Y$.

Proof. The proof is a direct consequence of Lemma 2 applied to $(X, Y)^{\mathrm{T}}$.

We can now proceed to evaluate the covariances on the right hand side of (3). Let $\phi_{n_f}(w_f; \beta, \delta)$ be the multivariate normal density with mean $F_f \beta$ and covariance matrix Σ_f that corresponds to the observations W_f used for model fitting. Define $L(\beta, \delta; w_f) =$

$-2\ln\phi_{n_f}(w_f;\, \beta,\, \delta)$. Minimization of $L(\beta,\, \delta;\, w_f)$ with respect to β leads to the maximum likelihood estimate $\hat{\beta}_f(\delta) = (F_f^{\mathrm{T}}\Sigma_f^{-1}\mathrm{F}_f)^{-1}\mathrm{F}_f^{\mathrm{T}}\Sigma_f^{-1}W_f$. Inserting $\hat{\beta}_f(\delta)$ into $L(\beta,\, \delta;\, w_f)$ leaves dependence on δ only and gives

$$L(\delta;\, w_f) = L(\hat{\beta}_f(\delta),\, \delta;\, w_f) = n_f \ln(2\pi) + \ln\det\Sigma_f + w_f^{\mathrm{T}}A_f w_f,$$

wherein $A_f = \Sigma_f^{-1} - \Sigma_f^{-1}F_f(F_f^{\mathrm{T}}\Sigma_f^{-1}F_f)^{-1}F_f^{\mathrm{T}}\Sigma_f^{-1}$. Next let S: $\mathbb{R}^{n_f} \times \mathbb{R}^{|\delta|} \rightarrow \mathbb{R}^{|\delta|}$ be defined as

$$S(w_f,\, \delta) = \nabla_\delta L(\delta;\, w_f) = \left(\frac{\partial}{\partial\delta_1}L(\delta;\, w_f),\, \ldots,\, \frac{\partial}{\partial\delta_{|\delta|}}L(\delta;\, w_f)\right)^{\mathrm{T}} \in \mathbb{R}^{|\delta|},$$

i.e. the gradient of $L(\delta;\, w_f)$ with respect to δ. We assume $S(w_f,\, \hat{\delta}(w_f)) = 0$ for all $w_f \in \mathbb{R}^{n_f}$. Let

$$\frac{\partial S}{\partial\delta}(w_f,\, \delta) = \begin{pmatrix} \nabla_\delta S_1(w_f,\, \delta) \\ \vdots \\ \nabla_\delta S_{|\delta|}(w_f,\, \delta) \end{pmatrix} \in \mathbb{R}^{(|\delta|,|\delta|)},$$

wherein $\nabla_\delta S_k(w_f,\, \delta)$ denotes, as a row vector, the gradient of $S_k(w_f,\, \delta)$ with respect to δ. We can rewrite this as

$$\left(\frac{\partial S}{\partial\delta}(w_f,\, \delta)\right)_{k,l} = \frac{\partial^2}{\partial\delta_k\partial\delta_l}L(\delta;\, w_f), \quad k,\, l = 1,\, \ldots,\, |\delta|.$$

Furthermore, let

$$\frac{\partial S}{\partial w_f}(w_f,\, \delta) = \begin{pmatrix} \nabla_{w_f} S_1(w_f,\, \delta) \\ \vdots \\ \nabla_{w_f} S_{|\delta|}(w_f,\, \delta) \end{pmatrix} \in \mathbb{R}^{(|\delta|,n_f)},$$

wherein $\nabla_{w_f}\mathrm{S}_k(w_f,\, \delta)$, $k = 1,\, \ldots,\, |\delta|$, denotes the gradient (again as a row vector) of $S_k(w_f,\, \delta)$ with respect to w_f. Suppose now that $w_f^{(0)}$ are the current data we used for model fitting. According to the earlier assumption, we have $S(w_f^{(0)},\, \hat{\delta}(w_f^{(0)})) = 0$ and thus

$$M := \frac{\partial\hat{\delta}}{\partial w_f}(w_f^{(0)}) = (-1)\left(\frac{\partial S}{\partial\delta}(w_f^{(0)},\, \hat{\delta}(w_f^{(0)}))\right)^{-1}\left(\frac{\partial S}{\partial w_f}(w_f^{(0)},\, \hat{\delta}(w_f^{(0)}))\right) \in \mathbb{R}^{(|\delta|,n_f)}.$$

Therein

$$\frac{\partial\hat{\delta}}{\partial w_f}(w_f) = \begin{pmatrix} \nabla_{w_f}\hat{\delta}_1(w_f) \\ \vdots \\ \nabla_{w_f}\hat{\delta}_{|\delta|}(w_f) \end{pmatrix} \in \mathbb{R}^{(|\delta|,n_f)},$$

and gradients again appear as row vectors. Using a Taylor series expansion of $\hat{\delta}$, we find $\hat{\delta}(w_f) \doteq \hat{\delta}(w_f^{(0)}) + M(w_f - w_f^{(0)}) +$ higher order terms $= \omega + Mw_f +$ higher order terms, wherein $\omega = \hat{\delta}(w_f^{(0)}) - Mw_f^{(0)}$ does not depend on w_f. For simplicity, we will later drop the argument w_f and write $\hat{\delta}$ instead of $\hat{\delta}(w_f)$ and $\hat{\delta}_k$ for the kth component. In the above, terms of order higher than 1 will be neglected. Under these assumptions, lemma 3 now evaluates the covariance terms in (3).

Lemma 3

Let $\Sigma_{f,p} = \mathrm{cov}(W_f,\, W_p^{\mathrm{T}}) \in \mathbb{R}^{(n_f,n_p)}$. *With* $Q_\delta(t) = M\Sigma_{f,p}(c_{p,\delta}^{(1)}(t),\, \ldots,\, c_{p,\delta}^{(|\delta|)}(t))$, *we have*

$$\frac{1}{2}\sum_{k,l=1}^{|\delta|}\mathrm{cov}((\hat{\delta}_k - \delta_k)(\hat{\delta}_l - \delta_l),\, H_{k,l}(t)) \doteq (\mathrm{tr}\,(Q_\delta(t)))^2 + \mathrm{tr}\,(Q_\delta(t)^2).$$

Proof. With $C_{kl}=c^{(k)}_{p,\delta}(t)c^{(l)}_{p,\delta}(t)^{\mathrm{T}}$, we can write $H_{k,l}(t)=2W^{\mathrm{T}}_{p}C_{kl}W_{\mathrm{p}}$. Also, $\hat{\delta}_k=\omega_k+M_{k\cdot}w_f$, wherein $M_{k\cdot}$ denotes the kth row of the matrix M. From this and the corollary to lemma 2,

$$\operatorname{cov}(\hat{\delta}_k,\,H_{k,l}(t))=\operatorname{cov}(M_{k\cdot}W_f,\,2W^{\mathrm{T}}_{p}C_{kl}W_{\mathrm{p}})=2M_{k\cdot}\Sigma_{f,p}(C_{kl}+C^{\mathrm{T}}_{kl})F_p\beta=0,$$

because $C^{\mathrm{T}}_{kl}F_p=c^{(l)}_{p,\delta}(t)c^{(k)}_{p,\delta}(t)^{\mathrm{T}}F_p=c^{(l)}_{p,\delta}(t)(c_{p,\delta}(t)^{\mathrm{T}}F_p)^{(k)}=0$, and, by a similar argument, $C_{kl}F_p=0$ as well. Similarly, $\operatorname{cov}(\hat{\delta}_l,\,H_{k,l}(t))=0$. Using these results, the second order parts involving $\hat{\delta}_k\hat{\delta}_l$ can be written as

$$\begin{aligned}\operatorname{cov}(\hat{\delta}_k\hat{\delta}_l,\,H_{k,l}(t))&=2\operatorname{cov}(W^{\mathrm{T}}_{f}M^{\mathrm{T}}_{l\cdot}M_{k\cdot}W_f,\,W^{\mathrm{T}}_{p}C_{kl}W_{\mathrm{p}})\\&=2[\operatorname{tr}(M^{\mathrm{T}}_{l\cdot}M_{k\cdot}\Sigma_{f,p}C_{kl}\Sigma^{\mathrm{T}}_{f,p})+\operatorname{tr}(M^{\mathrm{T}}_{l\cdot}M_{k\cdot}\Sigma_{f,p}C^{\mathrm{T}}_{kl}\Sigma^{\mathrm{T}}_{f,p})]\\&=2M_{k\cdot}\Sigma_{f,p}c^{(k)}_{p,\delta}(t)M_{l\cdot}\Sigma_{f,p}c^{(l)}_{p,\delta}(t)+2M_{k\cdot}\Sigma_{f,p}c^{(l)}_{p,\delta}(t)M_{l\cdot}\Sigma_{f,p}c^{(k)}_{p,\delta}(t).\end{aligned}$$

With $q_{kl}=M_{k\cdot}\Sigma_{f,p}c^{(l)}_{p,\delta}(t)$ being the $(k,\,l)$th element of $Q_\delta(t)$, $k,\,l=1,\,\ldots,\,|\delta|$, we have

$$\frac{1}{2}\sum_{k,l=1}^{|\delta|}\operatorname{cov}((\hat{\delta}_k-\delta_k)(\hat{\delta}_l-\delta_l),\,H_{k,l}(t))=\sum_{k,l=1}^{|\delta|}(q_{kk}q_{ll}+q_{kl}q_{lk})=(\operatorname{tr}(Q_\delta(t)))^2+\operatorname{tr}(Q_\delta(t)^2).$$

Altogether, we obtain the following approximation

$$\operatorname{MSE}[\hat{W}_{\hat{\delta}}(t)]\doteq\underbrace{\operatorname{MSE}[\hat{W}_\delta(t)]}_{T_1(t)}+\underbrace{\operatorname{tr}(V_\delta G_\delta(t))}_{T_2(t)}+\underbrace{(\operatorname{tr}(Q_\delta(t)))^2+\operatorname{tr}(Q_\delta(t)^2)}_{T_3(t)}.\tag{4}$$

The mean squared error can thus be approximated by a sum of three terms. The first, $T_1(t)$, is the mean squared error when assuming the covariance parameters to be known. The additional uncertainty due to them being unknown is taken into account by $T_2(t)$. The term $T_3(t)$ considers the correlation between W_{p} and $\hat{\delta}$. We define $T_{12}(t)=T_1(t)+T_2(t)$ and $T_{123}(t)=T_{12}(t)+T_3(t)$.

In lemma 1 we provided a formula for $c^{(k)}_{p,\delta}(t)$ needed to evaluate $G_\delta(t)$ and $Q_\delta(t)$. For the evaluation of M, we need to find formulas for

$$\frac{\partial^2}{\partial\delta_k\partial\delta_l}L(\delta;\,w_f)\quad\text{and}\quad\nabla_{w_f}S_k(w_f,\,\delta).$$

Using the result

$$\frac{\partial}{\partial\delta_k}\ln\det\Sigma_f=\operatorname{tr}(\Sigma_f^{-1}\Sigma_f^{(k)}),$$

and lemma 1, we obtain

$$\begin{aligned}\frac{\partial^2}{\partial\delta_k\partial\delta_l}L(\delta;\,w_f)&=\frac{\partial}{\partial\delta_k}[\operatorname{tr}(\Sigma_f^{-1}\Sigma_f^{(l)})-w^{\mathrm{T}}_fA_f\Sigma_f^{(l)}\mathrm{A}_fw_f]\\&=\operatorname{tr}(\Sigma_f^{-1}\Sigma_f^{(kl)})-\operatorname{tr}(\Sigma_f^{-1}\Sigma_f^{(k)}\Sigma_f^{-1}\Sigma_f^{(l)})\\&\quad+2w^{\mathrm{T}}_fA_f\Sigma_f^{(k)}A_f\Sigma_f^{(l)}A_fw_f-w^{\mathrm{T}}_fA_f\Sigma_f^{(kl)}A_fw_f.\end{aligned}$$

Second,

$$S_k(w_f,\,\delta)=\frac{\partial}{\partial\delta_k}L(\delta;\,w_f)=\operatorname{tr}(\Sigma_f^{-1}\Sigma_f^{(k)})-w^{\mathrm{T}}_fA_f\Sigma_f^{(k)}A_fw_f.$$

Taking the gradient (as a row vector) of $S_k(w_f,\,\delta)$ gives $\nabla_{w_f}\mathrm{S}_k(w_f,\,\delta)=-2w^{\mathrm{T}}_fA_f\Sigma_f^{(k)}A_f$.

This provides all formulas necessary to evaluate $Q_\delta(t)$ and thus the three parts in approximation (4). To estimate the mean squared error, we will later replace the unknown parameters δ in (4) by the maximum likelihood estimates $\hat{\delta}$. The resulting quantities will then be denoted by $\hat{T}_1(t)$, $\hat{T}_{12}(t)$, and $\hat{T}_{123}(t)$, respectively.

3. The exponential correlation structure

Our discussion so far applies to any parametric form for the covariance structure of $Z(\cdot)$. In order to evaluate the approximation (4), we will from now on assume the correlation structure of $Z(\cdot)$ to be given by

$$\operatorname{cor}(Z(s),\, Z(t)) = \prod_{k=1}^{d} \exp(-\theta_k |s_k - t_k|^{2-\alpha_k}), \qquad s,\, t \in [0,\, 1]^d. \tag{5}$$

The vector δ of covariance parameters will thus be of the form

$$\delta = (\theta_1, \ldots, \theta_d, \alpha_1, \ldots, \alpha_d, \sigma^2, \tau^2) \in (0, \infty)^d \times [0, 2)^d \times (0, \infty)^2, \quad |\delta| = 2d + 2,$$

if we assume all parameters $\alpha_1, \ldots, \alpha_d$ to be unknown. If these parameters are assumed to be known, we have $\delta = (\theta_1, \ldots, \theta_d, \sigma^2, \tau^2) \in (0, \infty)^{d+2}$, $|\delta| = d + 2$. The family (5) allows substantial flexibility in modelling correlations. Small values of θ_k generally indicate strong dependencies of the observations along the kth direction. Weaker correlations are obtained for larger values of these parameters. The exponents α_k control the smoothness of the process. If $\alpha_k = 0$, the paths of the stochastic process are infinitely differentiable along the kth coordinate axis. Differentiability is lost along directions with α_k greater than zero. Based on (5), formulas for $\Sigma_p^{(k)}$, $\Sigma_f^{(k)}$, $\Sigma_f^{(kl)}$, and $v_p^{(k)}(t)$ needed in the evaluation of (4) can be readily derived by differentiation of the elements in the corresponding matrices.

The term $T_2(t)$ on the right hand side of (4) depends on V_δ, the mean squared error matrix of $\hat{\delta}$. Other than in rather simple situations, its exact computation will be impossible. We therefore resort to approximation. The inverse of the Fisher information matrix has been found helpful in this context in Abt & Welch (1998) and its derivation is thus briefly outlined in the following. Let $\tilde{L}(\beta, \delta; w_f) = \ln \phi_{n_f}(w_f; \beta, \delta)$ be the log-likelihood function. In order to derive the $(k + 2d + 2) \times (k + 2d + 2)$ joint information matrix $\mathscr{I}_{\beta,\delta}$ of (β, δ), denote by $\nabla_\beta \tilde{L}$ and $\nabla_\delta \tilde{L}$ the gradients of $\tilde{L}(\beta, \delta; w_f)$ with respect to β and δ, respectively. Then it can be seen that $\mathrm{E}[\nabla_\beta \tilde{L}\ \nabla_\beta^{\mathrm{T}} \tilde{L}] = F_f^{\mathrm{T}} \Sigma_f^{-1} F_f$, $\mathrm{E}[\nabla_\beta \tilde{L}\ \nabla_\delta^{\mathrm{T}} \tilde{L}] = 0$, and the $(k,\, l)$th element of the matrix $\mathrm{E}[\nabla_\delta \tilde{L} \nabla_\delta^{\mathrm{T}} \tilde{L}]$ is given by $\operatorname{tr}(\Sigma_f^{-1} \Sigma_f^{(k)}\ \Sigma_f^{-1} \Sigma_f^{(l)})/2$. Let $\mathscr{I}_\delta$ be the submatrix of $\mathscr{I}_{\beta,\delta}$ that corresponds to δ. Then, due to the fact that $\mathscr{I}_{\beta,\delta}$ is block-diagonal, we suggest using $\mathscr{I}_\delta^{-1}$ as an approximation of V_δ. When $\alpha_1, \ldots, \alpha_d$ are known, the information matrix corresponding to the remaining covariance parameters $\theta_1, \ldots, \theta_d, \sigma^2, \tau^2$ is obtained by omitting rows and columns $d + 1$ to $2d$ from $\mathscr{I}_\delta$.

Using the inverse of the information matrix as an approximation for the mean squared error matrix of $\hat{\delta}$ requires that the true parameter δ is in the interior of the parameter space. If any of the parameters $\alpha_1, \ldots, \alpha_d$ are on the boundary zero, the variability in the maximum likelihood estimates is expected to be smaller than the approximation provided by the inverse information. Based on work by Moran (1971), modifications of $\mathscr{I}_\delta$ that account for these boundary effects were adapted to the exponential correlation (5) in Abt (1998). We will not repeat the details here. In all simulations carried out below under the assumption that some of the α_ks are equal to zero, the inverse of the accordingly modified information matrix is used to approximate V_δ.

4. Evaluation criteria

In the following we describe the layout of a simulation study that has been used to investigate how good each of $\hat{T}_1(t)$, $\hat{T}_{12}(t)$, and $\hat{T}_{123}(t)$ performs as an estimate of MSE$[\hat{W}_{\hat{\delta}}(t)]$. The goal is to identify the most important factors on the estimation accuracy. Among those are possible effects due to different designs D_f and D_p, but also the effect of different underlying covariance parameters. For a first step, simulations will be obtained under model (1), as this can serve as a benchmark on what quality we can expect the estimates to be. We will later consider a situation where model (1) is fitted but the data were generated by a different mechanism.

In all that follows, we assume $f(t) \equiv 1$ and the unknown mean β to be equal to zero. This assumption is somewhat extreme but has been successfully used in a variety of practical applications (Gao *et al.*, 1996; McMillan *et al.*, 1999; Jones *et al.*, 1998), where it is often difficult to assume a more detailed linear model for the mean. At the current moment we do not have any experience whether the results reported in the following will carry over to non-stationary means as well. Also the sensitivity due to misspecification of the mean has not yet been investigated.

Concerning the covariance parameters, we restrict attention to the cases $\alpha_1 = \cdots = \alpha_d$ and $\theta_1 = \cdots = \theta_d$. Reference to α_1 and θ_1 is thus sufficient and we will therefore generally drop the subscript. For $d > 1$, all designs selected will be (at least approximately) symmetric in the d dimensions so that there is no loss due to this assumption. We chose $\alpha = 0, 1$ and $\theta = 0.01, 0.1, 1, 4, 10, 15, 20, 50$. The values of α correspond to two well-known stochastic processes. For $\alpha = 0$, the sample paths of $Z(t)$ are analytic functions of t, see Belyaev (1959). The resulting correlation function is sometimes referred to as the Gaussian correlation in the literature. Selecting $\alpha = 1$ leads to the Ornstein-Uhlenbeck process. As we have no differentiability of the sample paths for any $\alpha \neq 0$, attention is given to $\alpha = 1$ only. The values of θ are chosen rather arbitrarily to cover correlations ranging from very strong for $\theta = 0.01$ to very weak for $\theta = 50$. Note that for any two corner points of $[0, 1]^d$, a given value of θ leads to equal correlations under $\alpha = 0$ and $\alpha = 1$. This does not hold for two interior points, in which case the correlation between any two observations taken at such points is stronger for $\alpha = 0$ than for $\alpha = 1$.

For the covariance parameters σ^2 and τ^2 we chose $\sigma^2 = 0.5, 1.5, 5$ and $\tau^2 = 0.05, 0.2, 2$. The first two values for each of these two parameters were taken as those obtained when fitting model (1) to dioxin contamination data in Abt *et al.* (1999). They correspond to the variabilities observed in areas of rather low and high contamination. The third values were selected as being more extreme. Considering all possibilities, we obtain $2 \times 8 \times 3 \times 3 = 144$ different parameter combinations under each of which 500 realizations according to model (1) were generated.

The d-dimensional cube $[0, 1]^d$ will be discretized in a grid T of m equally spaced points $p_1, \ldots, p_m$. The different fitting designs D_f and prediction designs D_p described later will be subsets of T. It is thus sufficient to generate the process $W(\cdot)$ over T and then select the observations corresponding to the various designs. The stochastic processes $Z(\cdot)$ and $\epsilon(\cdot)$ will be simulated over T using independent $N(0, 1)$ variates. For simulation of $Z(\cdot)$ we use the eigenvalue decomposition $R_T = U\Lambda U^{\mathrm{T}}$ for the $m \times m$ correlation matrix R_T corresponding to the grid T. Premultiplying an m-vector of $N(0, 1)$ variates by $U\Lambda^{1/2}$ then leads to the desired process $Z(\cdot)$. We used the same independent $N(0, 1)$ random samples for simulation under each of the 144 parameter combinations from above. Although this increases the susceptibility of the results to an unusual set of 500 realizations, it makes comparisons between the different levels of the factors investigated easier. Maximum likelihood estimates of the covariance parameters were computed using the software GaSP developed by Welch (1995b). Evaluation of the three terms in (4) was carried out with the GAUSS software package (Aptech, 1996).

For $j = 1, \dots, 500$, denote by $\{W^{(j)}(t)\colon t \in T\}$ the jth simulated sample from the marginal distribution of the stochastic process $W(\cdot)$. Fitting the model to the observations corresponding to D_f leads to the maximum likelihood estimate $\hat{\delta}$. Together with the part of the sample corresponding to D_p we can compute $\hat{W}^{(j)}_{\hat{\delta}}(p_i)$ as being the predicted value for $W^{(j)}_T(p_i)$. Comparing the two, we evaluate the integrated empirical mean squared error

$$\text{IEMSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{500}\sum_{j=1}^{500}(\hat{W}^{(j)}_{\hat{\delta}}(p_i) - W^{(j)}_T(p_i))^2$$

over the region of interest. This evaluates the accuracy of the predictor only. If any of $\hat{T}_a(t)$ for $a \in \{1, 12, 123\}$ accurately estimates the mean squared error of $\hat{W}_{\hat{\delta}}(t)$, the expected value of $(\hat{W}_{\hat{\delta}}(t) - W_T(t))^2/\hat{T}_a(t)$ should be close to one. We will thus look at

$$\text{IERMSE}_a = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{500}\sum_{j=1}^{500}\frac{(\hat{W}^{(j)}_{\hat{\delta}}(p_i) - W^{(j)}_T(p_i))^2}{\hat{T}^{(j)}_a(p_i)}$$

for $a \in \{1, 12, 123\}$, and call it the integrated empirical relative mean squared error. These quantities will be reported in the following sections. Values close to one are desirable and indicate, on average, good performance of the mean squared error estimate $\hat{T}_a(t)$.

Results in Abt *et al.* (1999) showed that when using a model with random error term as given in (1), inclusion of replicated observations is essential in order to ensure reasonable prediction accuracy. We did some preliminary experiments to investigate this when estimation of the prediction error is of interest. As it turns out, here replicates are also required, especially when $\hat{T}_1(t)$ is used, whereas $\hat{T}_{12}(t)$ or $\hat{T}_{123}(t)$ can compensate for missing replicates to some extent. In the present case, all designs used for model fitting below will contain replicates.

5. Simulation results

In this and the following sections we will evaluate the accuracy of the proposed estimates for the mean squared prediction error in two dimensions. We do not consider the one-dimensional situation, as it leads to theoretical problems concerning the identifiability and thus estimability of the covariance parameters that do not occur in dimensions higher than one. For example, even if an entire sample path of an Ornstein–Uhlenbeck process in one dimension were available, only the product $\theta\sigma^2$, but not the two parameters individually, could be identified, see Ibragimov & Rozanov (1978, p. 100). In case of the Gaussian correlation, as the sample paths are analytic, no identifiability problems arise (Ibragimov & Rozanov, 1978, p. 95). Due to the minor practical relevance of the one-dimensional situation, we will not present any results here.

Figure 1 shows the various designs D_p and D_f used for prediction as well as for model fitting in two dimensions. The two prediction designs are numbered P-DES1 and P-DES2, for fitting designs we will replace the leading letter P by F. Four different fitting designs are considered. The digit '2' means that a pair of replicates was obtained at the corresponding location, while '1' refers to a single observation. Together with the above 144 combinations of covariance parameters, considering four fitting designs and two prediction designs leads to $144 \times 4 \times 2 = 1152$ different combinations overall.

There are different goals underlying prediction and model fitting. For prediction, we expect that a good design should uniformly represent the entire area we are interested in. So-called space filling designs are very popular in practical applications, see Haaland *et al.* (1994). Although regular grid designs like P-DES1 and P-DES2 in Fig. 1 are not necessarily optimal in this sense, they appear plausible and are easier to deal with for our purpose. On the other hand,

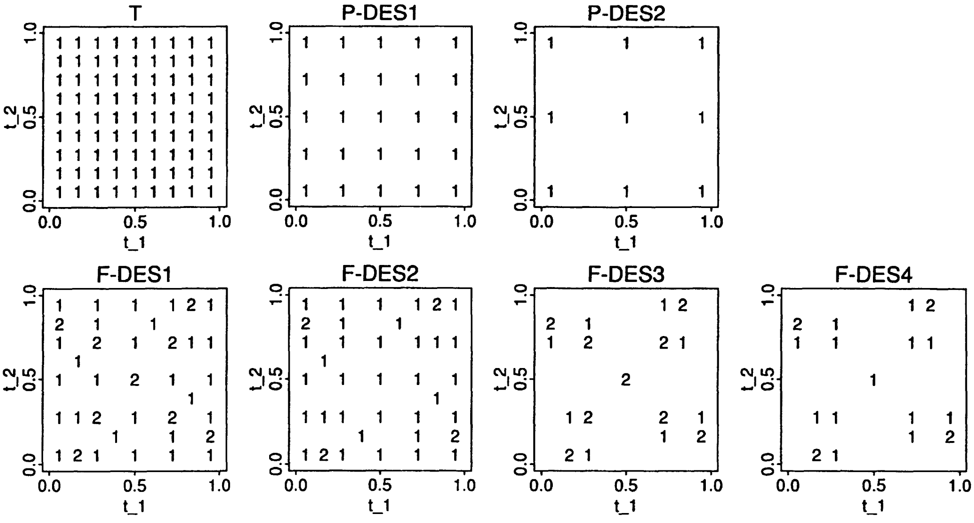

Fig. 1. Region of interest (*T*), prediction designs (P-DES), and fitting designs (F-DES) used for prediction and model fitting in two dimensions. In the fitting designs, a pair of replicated observations is taken at four or nine locations.

when fitting the model, we need to learn about the spatial correlations. To this end, the locations do not need to be equally spaced, but we aim to cover as many different distances in each direction as possible with few points only. Moreover, to estimate the measurement error variance, replicated observations are required at some locations. Fitting designs 2 and 4 have four replicated locations, whereas there are nine pairs of replicates in fitting designs 1 and 3.

Without going into any details, we briefly consider the effect of the various factors on the integrated empirical mean squared error, IEMSE. We did not find any differences in the size of the IEMSE for known or unknown α, except that for α being unknown and θ being large fitting designs 1 and 2 lead to slightly better predictions. However, these improvements are practically negligible and thus we do not distinguish the cases of known and unknown α.

P-DES1 leads to more accurate predictions than P-DES2. Overall, except for the situation mentioned above, no effects due to the fitting design were observed. Prediction accuracy strongly depends on the spatial correlations. For $\alpha = 1$ and $\theta \geqslant 10$, the IEMSE levels off at about two; observations seem to be practically uncorrelated. For $\alpha = 0$, due to the stronger correlations, prediction errors are generally much smaller and do not level off for the range of values for θ we considered. Larger values for θ would be required to see this effect. For the Ornstein–Uhlenbeck process as well as the Gaussian correlation, the prediction errors increase as a function of σ^2, the increase being more severe as θ gets large. No interaction between θ and τ^2 was observed, as for all values of θ predictions get uniformly worse as τ^2 increases.

Figure 2 investigates the main effects of the various factors on the IERMSE. The results for known and unknown α do not appear to be different. As α is more likely to be unknown in practice, we confine ourselves to this assumption. The left column refers to simulations carried out using $\alpha = 0$, for $\alpha = 1$ the graphs appear in the right column.

It is seen that using $\hat{T}_1(t)$ to estimate the mean squared prediction error can lead to severe underestimation. On the other hand, no substantial difference between $\hat{T}_{12}(t)$ and $\hat{T}_{123}(t)$ is found. It thus appears that the extra efforts in considering the correlation between the maximum likelihood estimates and the data used for prediction does not pay off even in cases where underestimation occurs when using $\hat{T}_{12}(t)$.

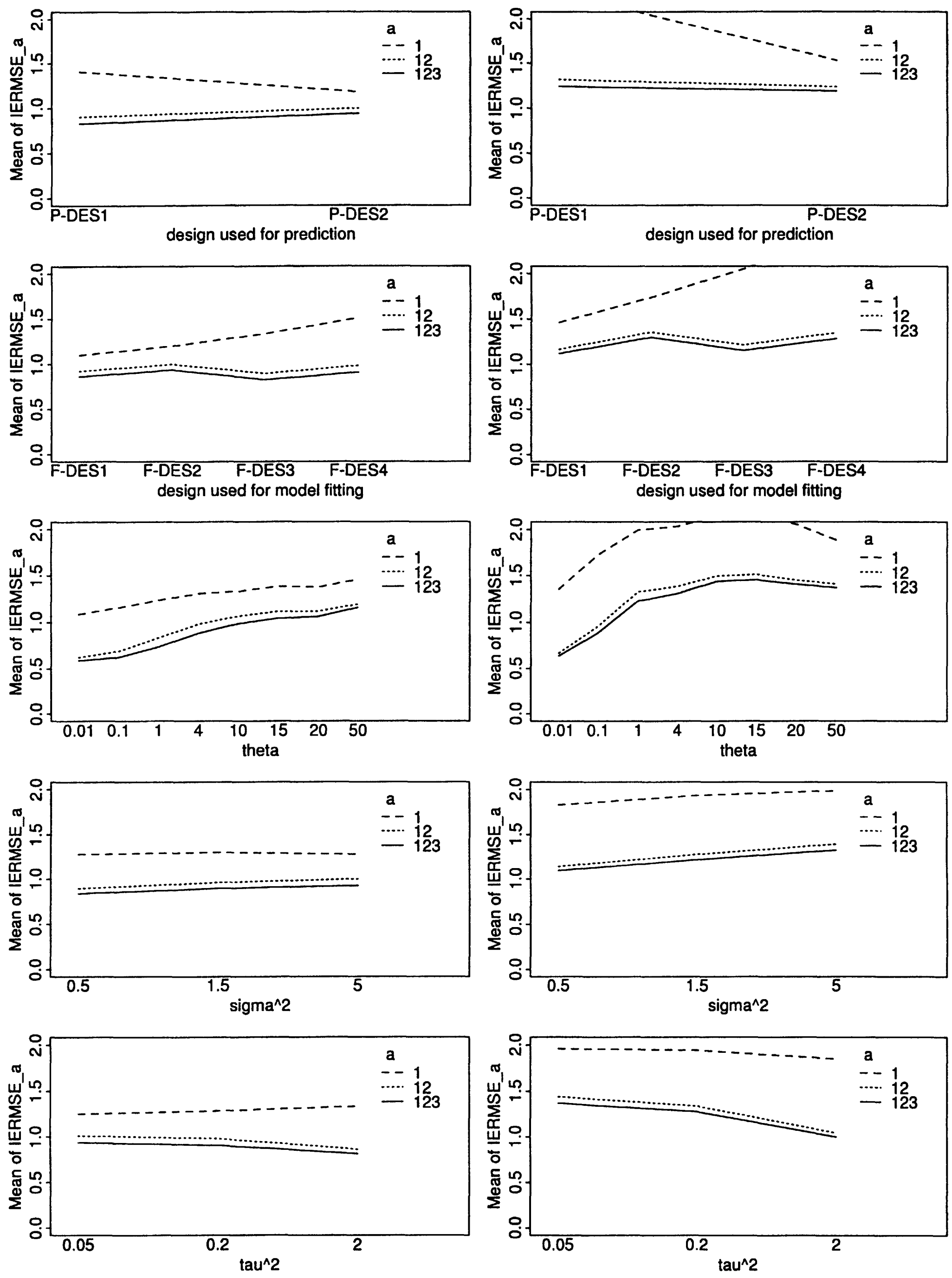

Fig. 2. Effects due to changes in the designs for predicting and model fitting as well as effects due to changes in the model parameters θ, σ^2, and τ^2 on the accuracy of the prediction error estimates in two dimensions for unknown α. Left: $\alpha = 0$, Right $\alpha = 1$.

Overall, in Fig. 2, when using $\hat{T}_{12}(t)$, minor effects due to the prediction design and the fitting design are found. However, strong effects appear to be due to the strength of the underlying correlation. If θ is small, the mean squared prediction error is generally overestimated by $\hat{T}_{12}(t)$, whereas slight underestimation occurs for $\alpha = 1$ and large values of θ. The variability as given by σ^2 and τ^2 shows minor influence only.

Considering $\hat{T}_{12}(t)$ only, we also looked at possible interaction effects. Not all graphs will be shown here. Neither for $\alpha = 0$ nor for $\alpha = 1$ were interactions discovered between θ, σ^2, and τ^2. Minor interactions were observed between the fitting design and σ^2 as well as τ^2. Fitting designs with more observations (in terms of locations as well as replicates) make the estimate $\hat{T}_{12}(t)$ more robust with respect to changes in these two parameters. Somewhat stronger interactions occur between the fitting design and θ, see Fig. 3. For small values of θ, the estimate $\hat{T}_{12}(t)$ slightly overestimates the mean squared error. Fitting designs 1 and 2 do better than fitting designs 3 and 4, i.e. for strong correlations locations are more beneficial than replicates. For large values of θ, as fitting designs 1 and 3 are doing better now, the number of replicates appears to be more important than the number of locations. As a practical guideline, along directions where strong correlations are expected, including more locations at the model fitting stage can be beneficial, whereas for weak correlations, replicates are preferable over locations.

6. Simultaneous fitting and prediction designs

Due to cost considerations it is hardly feasible in practice to have different designs for prediction and model fitting. One would rather collect one set of data and use it for both purposes. We will make this assumption in the present section. Minimax or maximin designs have been often used in applications and were discussed in Johnson *et al.* (1990). Minimax designs minimize the maximum distance of any candidate point to the design, whereas maximin designs maximize the minimum distance between design points. Both types of designs can be computed with the software ACED by Welch (1995a). As opposed to maximin designs, minimax designs are more expensive to compute but have the advantage that they avoid locations that are on the boundary of the region of interest and thus lead to a better coverage of the space. Figure 4 shows minimax designs of sizes 32, 24, 16, and 8 that were selected from the discretization T of $[0, 1]^2$ into the 81 points shown in Fig. 1. The smallest design, DES4, was generated first. Larger designs were always forced to contain the smaller designs. As model (1) contains a random error term, the four designs in Fig. 4 were augmented by allowing a pair of replicates at either four or eight locations. The eight locations to be replicated correspond to those of DES4, which are then replicated under each of the three other designs as well. The four corner points in DES4 are selected when designs with four replicates only are desired. We thus have

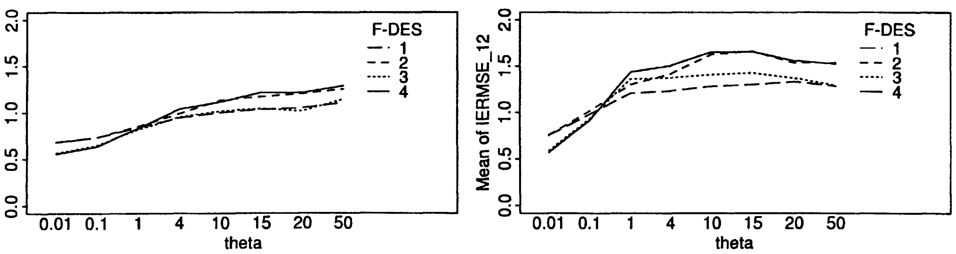

Fig. 3. Interaction effect on IERMSE$_{12}$ between θ and the fitting design in two dimensions for unknown α. Left: $\alpha = 0$, Right: $\alpha = 1$.

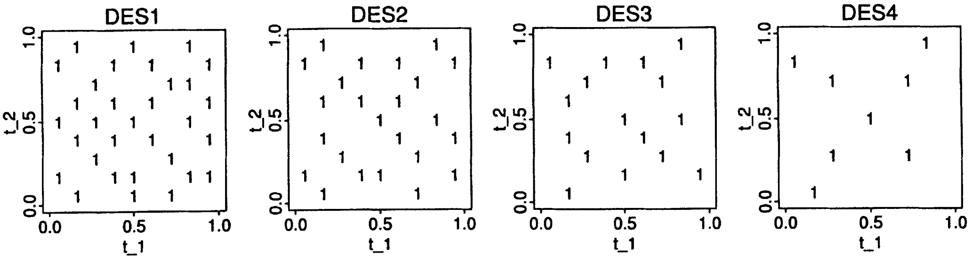

Fig. 4. Minimax designs for model fitting and prediction in two dimensions.

$4 \times 2 = 8$ different designs, each of which was used for model fitting as well as prediction at the 81 locations in T.

Real data hardly follow a simple model as in (1). We therefore attempt to consider the effect of using the estimates $\hat{T}_1(t)$, $\hat{T}_{12}(t)$, and $\hat{T}_{123}(t)$, when model (1) is fitted but the data are generated by a different mechanism. To simulate the observations, we use

$$W(t) = \beta + \sqrt{1-\gamma}Z_0(t) + \sqrt{\gamma}Z_1(t) + \epsilon(t), \tag{6}$$

where $Z_0(\cdot)$ and $Z_1(\cdot)$ are taken as independent Gaussian stochastic processes with correlation function given by (5). Therein $\alpha_1 = \alpha_2 = 0$ is selected for $Z_0(\cdot)$ and $\alpha_1 = \alpha_2 = 1$ is taken for $Z_1(\cdot)$. Both processes are assumed to have the same variance σ^2 and again the random error term $\epsilon(\cdot)$ with variance τ^2 is independent of $Z_0(\cdot)$ and $Z_1(\cdot)$. Through the parameter $\gamma \in [0, 1]$, the process $W(\cdot)$ becomes a mixture of $Z_0(\cdot)$ and $Z_1(\cdot)$. For s, $t \in T$, $s \neq t$, we have

$$\operatorname{cov}(W(s), W(t)) = (1-\gamma)\operatorname{cov}(Z_0(s), Z_0(t)) + \gamma \operatorname{cov}(Z_1(s), Z_1(t)),$$

which translates into a convex combination of the covariance functions of $Z_0(\cdot)$ and $Z_1(\cdot)$. Model (1) with α being unknown is fitted to the data generated by (6). Note that (6) still preserves the normal distribution. As this can often be achieved at least approximately by suitable transformations, we have not looked at any non-normal distributions.

We select $\gamma = 0, 0.25, 0.5, 0.75, 1$. To reduce computing time required for maximum likelihood estimation as well as for the computation of the three prediction error estimates, we restrict the number of different covariance parameter combinations and look only at $\theta = 0.01, 1, 15, 50$, $\sigma^2 = 0.5, 5$, and $\tau^2 = 0.05, 2$. This gives $8 \times 5 \times 4 \times 2 \times 2 = 640$ combinations altogether. The same values of θ were chosen for $Z_0(\cdot)$ and $Z_1(\cdot)$. In all figures in this as well as the next section, we have used the approximation of the information matrix mentioned at the end of section 3 whenever $\gamma = 0$.

As observed earlier, Fig. 5 shows again that using $\hat{T}_1(t)$ to estimate the mean squared prediction error can lead to severe underestimation, whereas $\hat{T}_{12}(t)$ and $\hat{T}_{123}(t)$ perform equally well. Strong effects are due to the number of locations in the design, with larger designs doing better overall than smaller designs. The number of replicated locations is of minor importance. Although not shown here, it is worth noting that for a design with few locations as in DES4, increasing the number of replicates does not improve the estimation accuracy.

The strength of the underlying correlation as described by θ has a huge effect on the performance of the estimated mean squared errors. Using $\hat{T}_{12}(t)$, slight overestimation may occur, but ignoring the additional variability due to estimating the covariance structure leads to underestimated mean squared errors. From a conservative point of view, the latter is more likely to be unacceptable in practical applications. The remaining factors show minor effects. As there is no effect due to variation of γ, distorting the covariance structure by the mixture (6) seems not to affect the accuracy of the estimates $\hat{T}_{12}(t)$ and $\hat{T}_{123}(t)$.

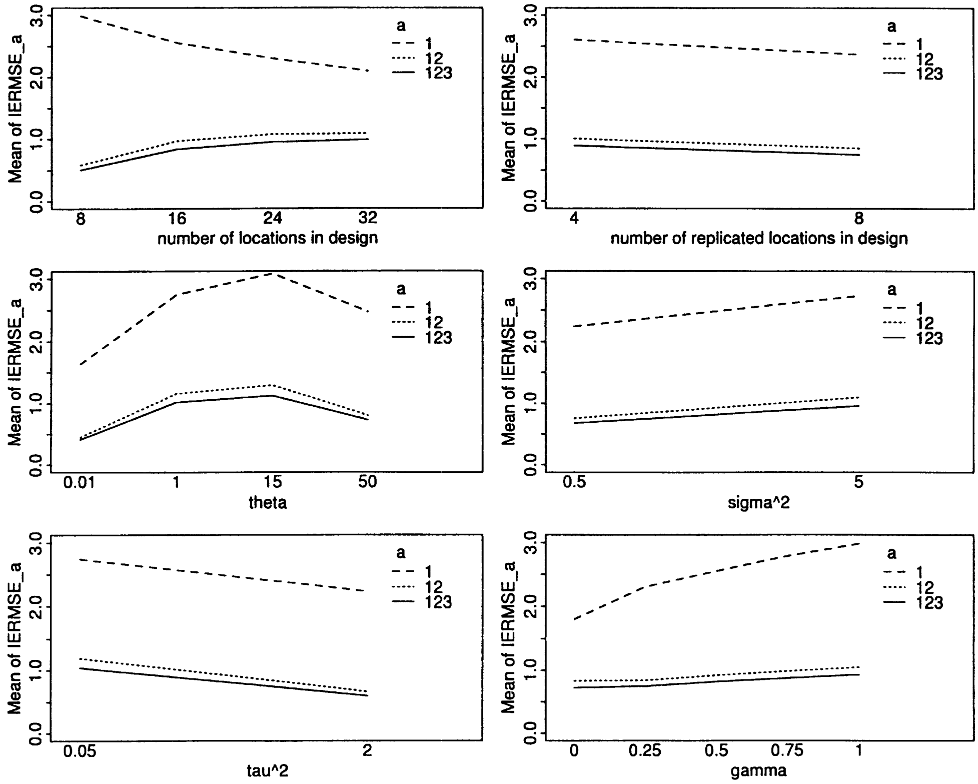

Fig. 5. Effects due to changes in the total number of locations and the number or replicated locations as well as effects due to changes in θ, σ^2, τ^2, and the mixture parameter γ on the accuracy of the prediction error estimates when using minimax designs for predicting and model fitting.

Among the interactions, only the one between the design and θ is worth noting. For small values of θ, the accuracy of $\hat{T}_{12}(t)$ as an estimate of the mean squared error seems not to depend on the size of the design used. However, as θ increases, larger designs give the better results. Again, this is consistent with earlier findings.

7. Computer experiments

Models similar to (1) are also used in the design and analysis of computer experiments as in Welch *et al.* (1992). Therein, the deterministic output of a computer code is often taken as a realization of a stochastic process and assumed to follow the model

$$W(t) = f(t)^{\mathrm{T}}\beta + Z(t), \qquad t \in [0, 1]^d. \tag{7}$$

Prediction of $W_T(t)$ for unobserved locations t is of interest and estimation of the squared prediction error is important. As an example, applications can be found in the area of global optimization, see Schonlau (1997) and Jones *et al.* (1998).

We set up a simulation study similar to the one in section 6. Simulated surfaces at the 81 locations in T (see Fig. 1) were generated using (6) without random error term. The four designs in Fig. 4 were used for model fitting and prediction without including any replicated locations. As the information matrix evaluated at the maximum-likelihood estimates was frequently found to be singular for $\theta = 0.01$, we only considered $\theta = 1, 15, 50$. For γ and σ^2, the same values as in section 6 were taken. Altogether we thus have four designs, three values of θ, two values of σ^2, and five levels for γ, leading to 120 combinations overall. Note that, different from the situation in sections 5 and 6, we do not predict at all locations in T, but only at the $81 - 32 = 49$ locations in T that do not belong to DES1. This is because of the interpolating nature of $\hat{W}(t)$ under model (7) which gives $\hat{T}_a(t) = 0$ for $a \in \{1, 12, 123\}$ when evaluated at any observed location t.

Figure 6 shows main effects due to the design and θ. Different from the results in section 5, clear benefits from using $\hat{T}_{123}(t)$ over $\hat{T}_{12}(t)$ are apparent. Ignoring the effect due to estimating the covariance structure can lead to very severe underestimation of the mean squared prediction error by $\hat{T}_1(t)$, which can be up to a factor of 60 for DES3 and DES4. The effect due to variation of θ is quite opposite to what we observed in Fig. 2. Large values of θ result in better approximations. No effect due to changes in σ^2 was observed.

Interactions are depicted in Fig. 7 using $\hat{T}_{123}(t)$ to estimate the prediction error. Again, the larger the design, the more insensitive is the accuracy of the estimate with respect to variations in θ. The same holds true for γ, thus leading to the conclusion that a design large enough can compensate for misspecifications in the model. As the prediction error estimate does not appear to perform worse for $\gamma \in \{0.25, 0.5, 0.75\}$ than for $\gamma = 0$ or $\gamma = 1$, it seems to be robust against the above alteration of the covariance structure.

8. Concluding remarks

The investigations in this paper show that plugging in the covariance parameter estimates in the formula derived for the mean squared error under the assumption that these parameters are known can lead to very serious underestimation of the squared prediction

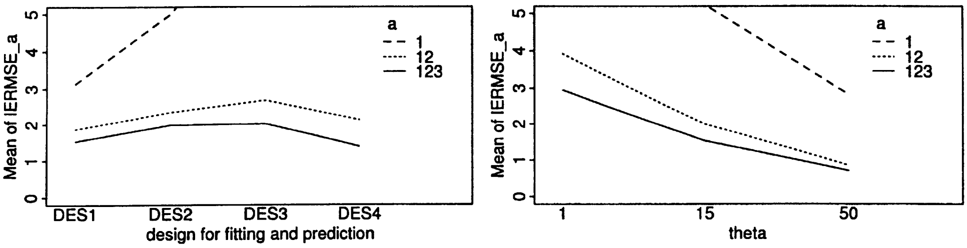

Fig. 6. Main effects of the design and θ on the accuracy of the prediction error estimates for a model without random error term.

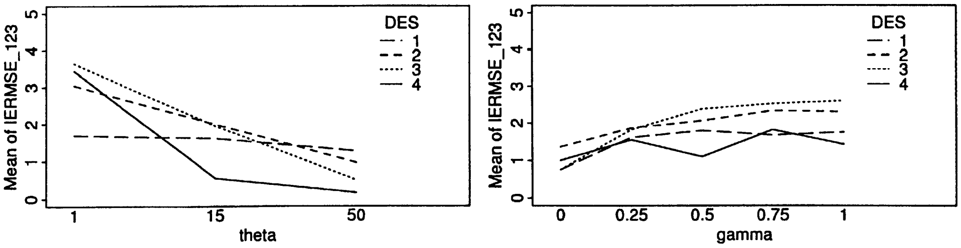

Fig. 7. Interaction effects on IERMSE_{123} between the design and θ as well as between the design and the mixture parameter γ for a model without random error term.

error. This has been demonstrated previously by Zimmerman & Cressie (1992), for example. Modifications based on work by these authors as well as by Kackar & Harville (1984) and Harville & Jeske (1992) were presented and perform rather well. As a general recommendation, we suggest to use $\hat{T}_{12}(t)$ in practice. Apart from models without random error, benefits from using $\hat{T}_{123}(t)$ seem not to justify the extra efforts required for its computation. The reason for this might be in the rather crude first order approximation of $\hat{\delta}$ as well as in the assumption that the maximum likelihood estimates are a zero of the derivative of the likelihood function. Obviously, the latter is not necessarily true.

The simulations here were carried out in two dimensions only. Increasing the dimensionality quickly increases the computing time required for maximum likelihood estimation as well as for the computation of (4).

We did assume Gaussian processes only, but also considered cases where the covariance structure is different from the one assumed in the model. The performance of the prediction error estimates seems to be unaffected by this and we believe that, under normality, the results carry over when predicting averages is of interest. However, difficulties generally arise when the data require a transformation to achieve normality in order to justify the model assumptions, but averages are to be predicted on the untransformed scale. As an example, this arises when designing schemes for contamination cleanup, see Abt *et al.* (1998). To the best of our knowledge, this problem, although of great practical relevance, has not yet been dealt with satisfactorily. Maybe some of the ideas presented here can be carried over to this situation.

Acknowledgements

I am very grateful to Professor William J. Welch for sharing his experience in this area with me and for making his software packages GaSP and ACED available to me. For financial support I like to thank the Deutsche Forschungsgemeinschaft as well as the Department of Statistics and Actuarial Science at the University of Waterloo. Furthermore, comments of two referees are gratefully acknowledged and helped to improve the presentation of the paper.

References

Abt, M. (1998). Approximating the mean squared prediction error in linear models under the family of exponential correlations. *Statist. Sinica* **8**, 511–526.

Abt, M. & Welch, W. J. (1998). Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canad. J. Statist.* **26**, 127–137.

Abt, M., Welch, W. J. & Sacks, J. (1998). Prediction of log-Gaussian processes and application to a cleanup scheme for environmental contaminations, unpublished manuscript.

Abt, M., Welch, W. J. & Sacks, J. (1999). Design and analysis for modeling and predicting spatial contamination. *Math. Geol.* **31**, 1.

Aptech (1996). *GAUSS mathematical and statistical system, system and graphics manual: version 3.2.33.* Apech Systems, Maple Valley, WA.

Belyaev, Yu. K. (1959). Analytic random processes. *Theory Probab. Appl.* **4**, 402–409.

Gao, F., Sacks, J. & Welch, W. J. (1996). Predicting urban ozone levels and trends with semiparametric modelling. *J. Agricultural Biol. Environ. Statist.* **1**, 404–425.

Haaland, P., McMillan, N. J., Nychka, D. & Welch, W. J. (1994). Analysis of space-filling designs. *Comput. Sci. Statist.* **26**, 111–120.

Harville, D. A. & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724–731.

Ibragimov, I. A. & Rozanov, Y. A. (1978). *Gaussian random processes*. Springer, New York.

Johnson, M. E., Moore, L. M. & Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26**, 131–148.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions, *J. Global Optim.* **13**, 455–492.

Kackar, R. N. & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.* **79**, 853–862.

McMillan, N. J., Sacks, J., Welch, W. J. & Gao, F. (1999). Analysis of protein activity data by Gaussian stochastic process models. *J. Biopharm. Statist.* **9**, to appear.

Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Proc. Cambridge Philos. Soc.* **70**, 441–450.

Schonlau, M. (1997). Computer experiments and global optimization. PhD dissertation, University of Waterloo, Canada.

Schott, J. R. (1997). *Matrix analysis for statistics*. Wiley, New York.

Welch, W. J. (1995a). ACED (Algorithms for the Construction of Experimental Designs), version 2.01 beta, unpublished software.

Welch, W. J. (1995b). GaSP (Gaussian Stochastic Process), version 2.02 beta, unpublished software.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J. & Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics* **34**, 15–25.

Zimmerman, D. L. & Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27–43.

Received December 1997; in final form November 1998

Markus Abt, Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany