A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments

D. Ginsbourger [1], D. Dupuy [1], A. Badea[1], L. Carraro[1], O. Roustant[1]

[1] Ecole des Mines, Departement 3MI, 158 Cours Fauriel, 42023 Saint-Etienne, France

Keywords: Kriging, External Drift, Likelihood Methods, D.A.C.E.

Our goal in the present work is to give an insight on some important questions to be asked when choosing a kriging model for the analysis of numerical experiments. We are especially concerned about the cases where the size of the design of experiments is small relatively to the algebraic dimension of the inputs. We first fix the notations and recall some basic properties of kriging. Then we expose two experimental studies on subjects that are often skipped in the field of computer simulation analysis: the lack of reliability of likelihood maximization with few data, and the consequences of a trend misspecification. We finally propose a case study, with the application of an original kriging method where an additive model is used as external trend.

1 Linear predictors for spatial interpolation of numerical simulators

Here we study a deterministic numerical simulator as a function $Z : D \subset \mathbb{R}^d \to \mathbb{R}$, where $x \in D$ is the vector of inputs variables.We denote by $\mathbf{X} = \{x_1, ..., x_n\}$ the design of experiments and by $\mathbf{Z} = \{Z(x_1), ..., Z(x_n)\}$ the set of simulator responses associated with the design of experiments. Kriging is a class of methods which comes from the field of geostatistics (Matheron (1963), Cressie (1993)), known as *linear optimal prediction* in classical statistics. It provides at each point $x \in D$ a prediction $\hat{Z}(x)$ linearly depending on $\mathbf{Z}$, where the weights depend on the design of experiments and the kriging model but not on the observations. The way the weights are defined varies as a function of the type of kriging (Simple : SK; Ordinary : OK; Universal : UK, etc...) and many parameters such as the trend functions, the covariance kernels and their own parameters (sill, scales, nugget, etc...) denoted by ψ. In the following, we will concentrate on the parameters of sill and scales, denoted respectively either by $\psi_1,\ \psi_2$ or by $\sigma^2,\ p$. Nevertheless, kriging can always be seen as an interpolation by a random process, relying on the assumption that:

$$\forall x \in D,\ Z(x) = t(x) + \varepsilon(x) \tag{1}$$

where t is a numerical deterministic function and $\varepsilon(x)$ is a path of a centered stationary gaussian process with known covariance kernel $k : h \in \mathbb{R}^d \longrightarrow k(h) \in \mathbb{R}$. t is generally known up to a set of parameters or a semi-parametric structure to be estimated within kriging.
Several founder works (such as Sacks et al. (1989), Jones et al. (1998)) on the application of kriging to computer simulations start off with an extremely simplified version of (1). They assume that the trend is constant (Ordinary Kriging, i.e. $t(x) = \mu$) and that k is a generalized exponential kernel (see Santner et al. (2003) for instance), letting the stochastic part of (1) account for the variability of Z. Then the covariance parameters ψ are estimated by maximizing the gaussian likelihood conditionally to $(\mathbf{X}, \mathbf{Z})$. On the other hand, recent approaches (such as Jourdan (2002) or Martin and Simpson (2005)) try to take advantage of a more complex trend, from linear and polynomial functions to Fourier series. In other respects, Martin and Simpson (2004) presents an application of bayesian methods to kriging interpolation (see also O'Hagan (2006) for the bayesian analysis of computer codes).

2 Fitting covariance parameters by MLE with a small sample

The Maximum Likelihood Estimation method is widely used in kriging to choose covariance parameters on the basis of the observed data. Following the assumption (1), MLE relies on the maximization of the density of $\mathbf{Z}$, seen as a function of ψ:

$$L(\psi; \mathbf{Z}) = f(\mathbf{Z}/\psi) = (2\pi)^{-\frac{n}{2}} det(K_\psi)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{Z}-\mathbf{t})^T K_\psi^{-1}(\mathbf{Z}-\mathbf{t})} \tag{2}$$

where K_ψ is the covariance matrix of $\mathbf{Z}$ provided that ψ is the true vector of covariance parameters, and $\mathbf{t}$ is the vector of values of t at $\mathbf{X}$. Obviously, the obtained result $\hat{\psi} = argmax_\psi\{L(\psi;\mathbf{Z})\}$ is closely depending on the observed sample $\mathbf{Z}$. The behaviour of $\hat{\psi}$ relatively to ψ when the sample varies is a matter of importance. We recall that we are in the case where $\mathbf{Z}$ is drawn at random following a multigaussian distribution with covariance parameters ψ. Then $L = L(\psi;\mathbf{Z})$ becomes a random function (see Figure 1), and $\hat{\psi} = argmax\{L(\psi;\mathbf{Z})\}$ becomes a random vector as well.

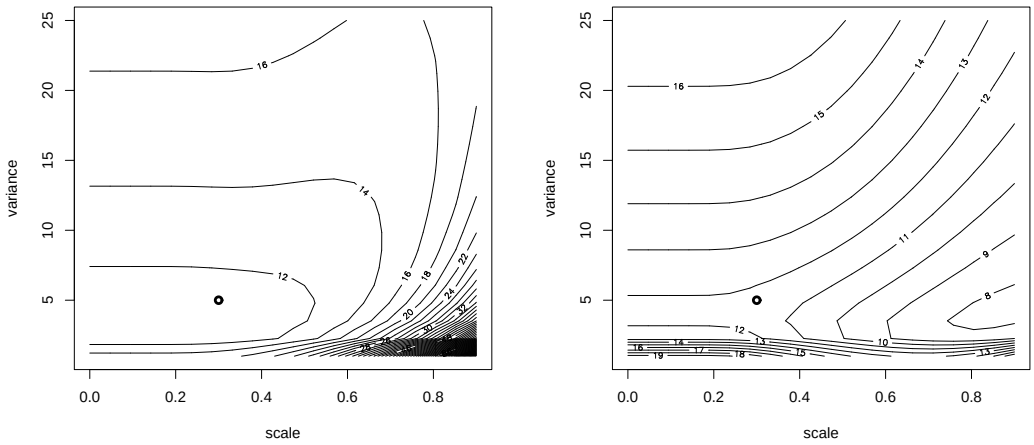

FIGURE 1. Realizations of $-2\ln L$ corresponding to two simulated Z, both with $\psi = (5,\ 0.3)$ and with gaussian covariance function; $\hat{\psi} \approx \psi$ (left) and $\hat{\psi} \neq \psi$ (right)

The asymptotic distribution of $\hat{\psi}$ has been studied in detail within the theory of likelihood, and leads to the Fisher Information Matrix (FIM) $\mathcal{I}(\psi)$:

$$\hat{\psi} \xrightarrow{\mathcal{L}} \mathcal{N}(\psi, \mathcal{I}(\psi)^{-1}) \text{ where } \mathcal{I}(\psi) \text{ is defined by } (\mathcal{I}(\psi))_{ij} = E\left[\frac{\partial \ln(L(\psi;\mathbf{Z}))}{\partial\psi_i}\frac{\partial \ln(L(\psi;\mathbf{Z}))}{\partial\psi_j}\right] \quad (3)$$

The gaussian covariance function is chosen in many computer experiments for its regularity properties. Covariance parameters are then fitted by MLE; the efficiency and robustness of this method are rarely discussed. Our concern is to check in what measure the asymptotic results hold with small samples. To do so, we computed the theoretical FIM of $\mathbf{Z}$:

$$\forall i,j \in [1,n]\ (\mathcal{I}(\psi))_{ij} = \frac{1}{2}tr\left(K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_i}K_\psi^{-1}\frac{\partial K_\psi}{\partial\psi_j}\right). \quad (4)$$

To obtain comparable results for different values of ψ, we introduced a relative inverse FIM: $(\mathcal{J}(\psi))_{ij} = (\mathcal{I}^{-1}(\psi))_{ij}/(\psi_i\psi_j)$. J is in fact the asymptotical covariance matrix of $\frac{\hat{\psi}}{\psi}$, where the division is made component by component. We conducted experiments with vectors taken from simulated monodimensional gaussian processes to compute empirical means and variances of the ML estimators. For each simulation, we computed covariance parameters estimated by MLE and the Mean Squared Error (e) between simulated ($Z(x)$) and interpolated ($\widehat{Z}(x)$) data:

$$e = \frac{1}{|D|}\int_D |Z(x) - \widehat{Z}(x)|^2 dx.$$

The latter was approximated by taking the average sum of squared errors on a fine grid (i.e. 200 points). We finally collected the averages and variance matrices of the relative values of the estimated covariance parameters, the averages and variances of e, and the covariances between both of them (the last two indicators are not presented in the following tables) . We focused on gaussian processes with covariance functions $c_g(h) = \sigma^2 e^{-\frac{h^2}{p^2}}$ $(gaussian)$ and $c_e(h) = \sigma^2 e^{-\frac{|h|}{p}}$ $(exponential)$. The covariance parameters then were reduced to $\psi = (\sigma^2, p) \in \mathbb{R}_+^\star \times \mathbb{R}_+^\star$ and the design of experiments $\mathbf{X}$ is taken as a regular subdivision of $[-1,1]$: $\mathbf{X_{n+1}} = \{-1, -1+\frac{2}{n}, \ldots, -1+\frac{2(n-1)}{n}, 1\}(n \in \mathbb{N}^\star)$. We restricted our experiments to the following designs: $\mathbf{X_5}$ and $\mathbf{X_{10}}$ with both covariance functions, $\psi_1 = \sigma^2 \in \{5, 10\}$, and $\psi_2 = p \in \{0.3, 0.4, 0.5, 0.6\}$.

TABLE 1. MLE and MSE measures on simulated realizations of gaussian processes with gaussian covariance function, for relative parameters $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, $i = 1, 2$ and for $\mathbf{X} = \mathbf{X_5}$.

ψ	$E\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotic $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$E[e]$
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.034 \\ 0.018 \end{pmatrix}$	$\begin{pmatrix} 1.105 & 0.277 \\ 0.277 & 1.270 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	4.976
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.147 \\ 0.042 \end{pmatrix}$	$\begin{pmatrix} 1.329 & 0.501 \\ 0.501 & 0.976 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	3.287
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.222 \\ 0.033 \end{pmatrix}$	$\begin{pmatrix} 4.037 & 0.757 \\ 0.757 & 0.679 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	1.947
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.187 \\ 0.027 \end{pmatrix}$	$\begin{pmatrix} 2.058 & 0.504 \\ 0.504 & 0.421 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	0.706
$\begin{pmatrix} 10 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.131 \\ 0.006 \end{pmatrix}$	$\begin{pmatrix} 3.334 & 0.867 \\ 0.867 & 1.564 \end{pmatrix}$	$\begin{pmatrix} 0.402 & 0.071 \\ 0.071 & 2.048 \end{pmatrix}$	10.138
$\begin{pmatrix} 10 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.083 \\ 0.106 \end{pmatrix}$	$\begin{pmatrix} 1.645 & 0.484 \\ 0.484 & 0.862 \end{pmatrix}$	$\begin{pmatrix} 0.427 & 0.111 \\ 0.111 & 0.452 \end{pmatrix}$	6.398
$\begin{pmatrix} 10 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.166 \\ 0.024 \end{pmatrix}$	$\begin{pmatrix} 1.343 & 0.440 \\ 0.440 & 0.629 \end{pmatrix}$	$\begin{pmatrix} 0.479 & 0.131 \\ 0.131 & 0.217 \end{pmatrix}$	3.678
$\begin{pmatrix} 10 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.256 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} 14.960 & 0.963 \\ 0.963 & 0.392 \end{pmatrix}$	$\begin{pmatrix} 0.538 & 0.135 \\ 0.135 & 0.133 \end{pmatrix}$	1.459

In the case of a gaussian covariance, we observe a negative relative bias [1] (-3.4% to -25.6% of σ^2) in the estimation of $\psi_1 = \sigma^2$. This bias is decreasing with $|\mathbf{X}|$ (see table 2 where the negative relative bias varies between -4.2% and -11.9%) , which seems in accordance with the asymptotic unbiasedness of MLE. On the other hand, the relative bias of $\hat{\psi}_2$ has a small order of magnitude when $|\mathbf{X}| = 5$ ($+0.6\%$ to 10.6%) and slightly oscillates around 0 when $|\mathbf{X}| = 10$.

The empirical covariance matrices of the ML estimates offer some surprising results. In particular, the relative variances of $\hat{\psi}_1$ present huge fluctuations: they vary sometimes of an order of more than 10 between two samples of 1000 realizations issued from the same process (for instance by resimulating a GP with $\psi = (10, 0.4)$ and $|\mathbf{X}| = 5$ we obtained $Var\left[\left(\psi_i^{rel}\right)_i\right] = \begin{pmatrix} 43.555 & 3.242 \\ 3.242 & 0.971 \end{pmatrix}$) .

Since it is obviously in contradiction with normality (expected in asymptotical conditions) and the order of magnitude given by (4) (and reported in the 4th column of tables 1 and 2), we analyzed this phenomenon in detail. First, we observed that the extreme values of $Var[\hat{\psi}_1]$ were caused by some outliers, highly perturbating the non-robust estimate of variance. Second, a graphical study of the cloud of $\hat{\psi}_1$s suggested that the distribution is rather log-normal than normal in small-sample regime. Finally, the comparison with the relative FIM shows that empirical variance of $\hat{\psi}_1$ is clearly bigger than predicted by the second order Fisher approximation, in particular with the smallest designs (see Figure 2 for a graphical comparison between theoretical and empirical results as a function of $|\mathbf{X}|$).

Concerning the relative variances of $\hat{\psi}_2$, the results are much more regular: they decrease monotonically with ψ_2 and with $|\mathbf{X}|$, both for the empirical and theoretical quantities. Once again, the empirical variances tend to match the theoretical variances as $|\mathbf{X}|$ grows, even if the first ones are still typically two times larger than the second ones for a sample of size 10. Nevertheless, the theoretical Fisher variance has a diverging behaviour as the range becomes small.

In other respects, both tables illustrate some fundamental properties of the mean squared error. Obviously decreasing with $|\mathbf{X}|$, the MSE is also decreasing with the range ψ_2 and linearly increasing with the variance ψ_1. Finally, we quantified the linear dependence between the underestimation of both covariance parameters by MLE and the MSE (not in the tables). It is worth noticing that ψ_1 and ψ_2 play drastically different roles here: it seems that a bad estimation of ψ_1 is weakly correlated with the MSE. This result seems natural when considering that the OK predictor is not depending on the process variance (see Cressie (1993)). Conversely, the correlation between the MSE and the relative MLE error on ψ_2 is significantly positive: it varies between 40.1% and 55.7% when $|\mathbf{X}| = 5$ and between 15% and 62.5% when $|\mathbf{X}| = 10$. This coincides with our previous qualitatives observations of bigger MSE when the range is much underestimated.

[1] Mind the fact that by negative relative bias we understood an overestimation of ψ.

TABLE 2. MLE and MSE measures on simulated realizations of gaussian processes with gaussian covariance function, for relative parameters $\psi_i^{rel} = \frac{\psi_i - \hat{\psi}_i}{\psi_i}$, $i = 1, 2$ and for $\mathbf{X} = \mathbf{X_{10}}$.

ψ	$E\left[\left(\psi_i^{rel}\right)_i\right]$	$Var\left[\left(\psi_i^{rel}\right)_i\right]$	asymptotic $Var\left[\left(\psi_i^{rel}\right)_i\right]$	$E[e]$
$\begin{pmatrix} 5 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.054 \\ 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.432 & 0.105 \\ 0.105 & 0.085 \end{pmatrix}$	$\begin{pmatrix} 0.297 & 0.057 \\ 0.057 & 0.033 \end{pmatrix}$	0.177
$\begin{pmatrix} 5 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.042 \\ -0.019 \end{pmatrix}$	$\begin{pmatrix} 0.424 & 0.058 \\ 0.058 & 0.024 \end{pmatrix}$	$\begin{pmatrix} 0.340 & 0.044 \\ 0.044 & 0.014 \end{pmatrix}$	0.009
$\begin{pmatrix} 5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.067 \\ -0.013 \end{pmatrix}$	$\begin{pmatrix} 0.46 & 0.051 \\ 0.051 & 0.013 \end{pmatrix}$	$\begin{pmatrix} 0.362 & 0.036 \\ 0.036 & 0.008 \end{pmatrix}$	0.0004
$\begin{pmatrix} 5 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.075 \\ -0.007 \end{pmatrix}$	$\begin{pmatrix} 0.728 & 0.059 \\ 0.059 & 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.375 & 0.032 \\ 0.032 & 0.005 \end{pmatrix}$	4.e-05
$\begin{pmatrix} 10 \\ 0.3 \end{pmatrix}$	$\begin{pmatrix} -0.067 \\ 0.003 \end{pmatrix}$	$\begin{pmatrix} 0.432 & 0.089 \\ 0.089 & 0.079 \end{pmatrix}$	$\begin{pmatrix} 0.297 & 0.057 \\ 0.057 & 0.033 \end{pmatrix}$	0.345
$\begin{pmatrix} 10 \\ 0.4 \end{pmatrix}$	$\begin{pmatrix} -0.097 \\ 0.015 \end{pmatrix}$	$\begin{pmatrix} 0.495 & 0.071 \\ 0.071 & 0.028 \end{pmatrix}$	$\begin{pmatrix} 0.340 & 0.044 \\ 0.044 & 0.014 \end{pmatrix}$	0.03
$\begin{pmatrix} 10 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} -0.06 \\ -0.009 \end{pmatrix}$	$\begin{pmatrix} 0.491 & 0.046 \\ 0.046 & 0.011 \end{pmatrix}$	$\begin{pmatrix} 0.362 & 0.036 \\ 0.036 & 0.008 \end{pmatrix}$	0.0008
$\begin{pmatrix} 10 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} -0.119 \\ -0.009 \end{pmatrix}$	$\begin{pmatrix} 0.582 & 0.05 \\ 0.05 & 0.011 \end{pmatrix}$	$\begin{pmatrix} 0.375 & 0.032 \\ 0.032 & 0.005 \end{pmatrix}$	0.0001

A similar study with exponential covariance function gave very different results both for the bias and the variances of ML estimates (the corresponding tables are not presented here). Indeed, we observed very regular variances of ML estimates while the bias reached impressive orders of magnitude. However, the behaviour of the MSE and the correlations between MSE and relative MLE errors followed the same sketch as in the gaussian case.
To sum up this part about ML estimation:

- Fisher's asymptotical results must be applied with much care in non-asymptotical conditions.
- More precisely, the distribution of the estimated range parameter is asymmetrical when n is very small and quickly stabilize to a gaussian when n increases (from 5 to 13).
- On the other hand, the distribution of the estimated variance parameter has a very large right tail but its shape is far from being gaussian when n is very small. Furthermore, this results still holds when n increases (from 5 to 13) and we guess that the asymptotical regime only starts for larger values of n.

Now we wish to examine another difficulty encountered when kriging with few data: the selection and the estimation of deterministic trends.

3 Kriging with trends: a bless or a curse?

The most commonly used kriging methods are simple and ordinary kriging. However, they reach their limits when the stationarity assumption does not hold any longer, i.e. when non constant trends $t(x)$ are impossible to ignore.
In this case, we are back to the general decomposition of (1), where Z is assumed to be the sum of a deterministic trend t and a centered gaussian process ε. At this stage, we may consider several subcases. If t is known and ε is to be estimated, a straightforward solution is to perform simple kriging of the residuals $\{Z(x) - t(x)\}$. If t is unknown, we distinguish between linear and most general non-linear frames. The case in which t depends linearly on its parameters and ε has a kwown structure has been intensively studied: it is well known as *universal* kriging (UK, Martin and Simpson (2005)). Indeed, when the covariance parameters ψ are known and the trend is a linear combination of some known functions f_j of the input variables $t(x) = \sum_{j=1}^{p} \beta_j f_j(x)$, then the only unknowns are the parameters of the trend β_j, and they are estimated by generalized least squares (GLS):

$$\hat{\beta}(\psi_2) = (\mathbf{F}^T R_{\psi_2}^{-1} \mathbf{F})^{-1} \mathbf{F}^T R_{\psi_2}^{-1} \mathbf{Z} \tag{5}$$

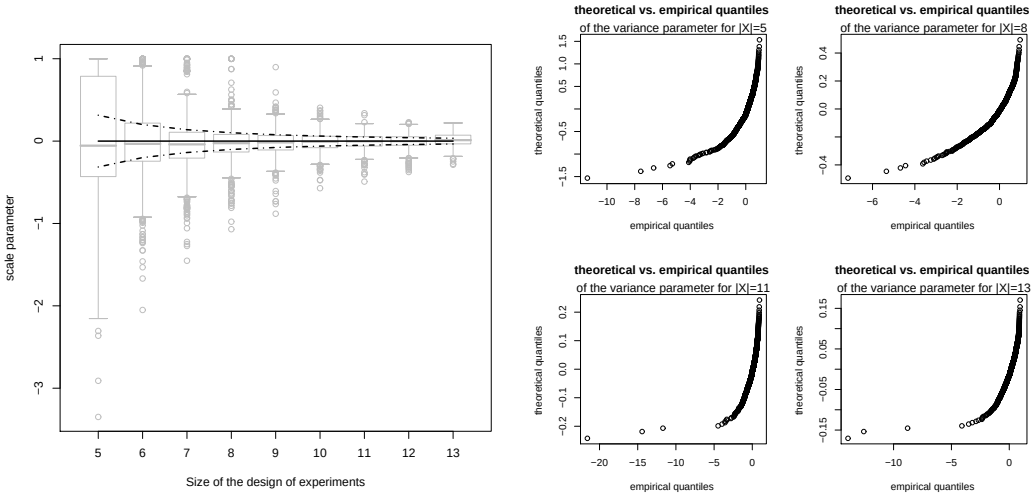

FIGURE 2. **Left:** Comparison between the experimental law (gray boxplots) and the asymptotic law (black lines) for the scale parameter for increasing size of the experimental design. The boxplots for the experimental laws have been done using 1000 simulations, with gaussian covariance function of parameters $\psi = (5, 0.5)$. For the asymptotic law the median is represented in continuous line and the first and third quartiles in dashed lines. **Right:** Comparison between the theoretical and experimental quantiles of the variance parameter.

where $\mathbf{F}$ denotes the evaluation of $\mathbf{f}(x) = [f_1(x), \ldots, f_p(x)]$ at the n points of the design $\mathbf{X}$ and $R_{\psi_2} = (1/\psi_1)K_\psi$ is the correlation matrix of $\mathbf{Z}$.
In practice, however, one has seldom in hand the value of the covariance parameters previous to performing UK. So one has to estimate a model with linear trend and unknown covariance parameters ψ (in the following we will also refer to this case as "UK", like many practitioners do). Hence ψ and β have to be estimated within kriging. At a first sight, this is likely to create a circularity problem: on the one hand, one needs a known trend to work on the residuals and thus estimate ψ. On the other hand, estimating t without taking the residual into account may lead to unadapted trends (the estimation of the parameters of the trend would be done by Ordinary Least Squares instead of GLS). Fortunately, the ML estimation gives a way to escape this vicious circle in the linear case. Assuming, like in section 2 that the covariance parameters to be estimated are $\psi = (\sigma^2, p)$, and using the MLE method (and the same formula (5) for $\hat{\beta}$), we obtain a straightforward formula for $\widehat{\sigma^2}$:

$$\widehat{\sigma^2}(\psi_2) = (1/n)(\mathbf{Z} - \mathbf{F}\hat{\beta})^T R_{\psi_2}^{-1}(\mathbf{Z} - \mathbf{F}\hat{\beta}). \tag{6}$$

By injecting (5) and (6) in the expression of the likelihood, we obtain a function $L(\psi_2, \widehat{\sigma^2}(\psi_2), \hat{\beta}(\psi_2))$ which clearly depends on one single parameter ψ_2 and which has to be maximized to get $\hat{\psi}_2$.
In both those cases we obtain the same form of the kriging predictor

$$\hat{Z}(x) = \mathbf{f}^T(x)\hat{\beta} + r^T(x)R_{\hat{\psi}_2}^{-1}(\mathbf{Z} - \mathbf{F}\hat{\beta}). \tag{7}$$

where $r(x)$ is a vector representing the correlation between an unknown point x and the design $\mathbf{X}$.
We will see in the next part that no such simple solving can be done in most general non-linear cases. Now we would like to go one step deeper in the application and ask a naive (but complex) question which has to be handled in practice: how can one come back to a trend from raw data $(\mathbf{X}, \mathbf{Z})$? As soon as the modelizer finds himself in a situation where neither prior information nor obvious clue is available, he has indeed to select a trend on the basis of $(\mathbf{X}, \mathbf{Z})$. What means does he have to do so, and what risk does he run in case of a bad choice?
In order to show that the answer to these questions is important let us first perform some experiments. The set-up is the following. A realization of a one dimensional Gaussian process with known covariance function and parameters is simulated on a regular grid (401 points) on $[-1, 1]$ and a affine trend (of the form $a + bx$) is added; those are the data ($\mathbf{X}$, $\mathbf{Z}$). From this set we chose a subset of 5 regularly distributed points and we perform three types of kriging : OK, UK with linear trend and UK with quadratic trend (of the form $a + bx + cx^2$). Due to the fact that the

points were taken regularly on the grid all the three kriging give similar good results, even if in two (OK and UK with quadratic trend) of the three cases the trend was misspecified. This may lead to the conclusion that specifying the trend is not very important and we could obtain good results using OK. But if we perform the same krigings on different designs, where there are few points concentrated either on the boundaries or in the center of the domain, then the results are very bad (due to the ratio between the parameter ψ_2 and the subdivision length) when the trend is misspecified, see Figure 3. They are even worse if we use the kriging predictor given by OK or by UK with quadratic trend in extrapolation, see Figure 4 here after. (The covariance parameters used for the simulated process in Figures 3 and 4 are $\psi = (5, 0.2)$.)

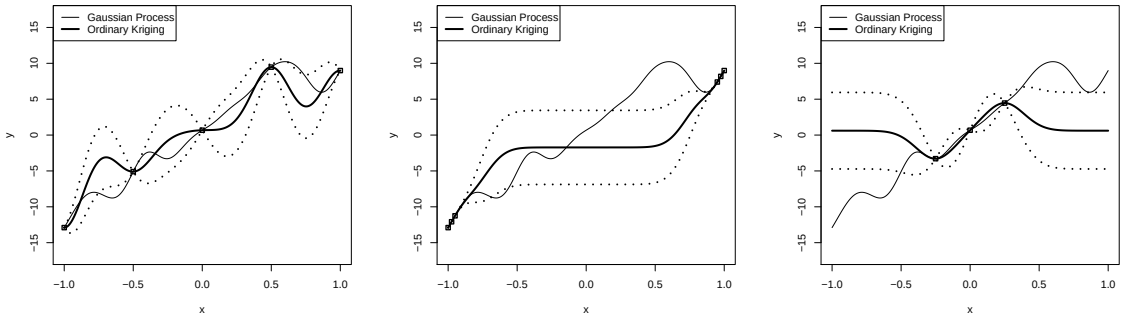

FIGURE 3. OK on a regular grid (left), on a grid concentrated on the boundaries (middle); on a grid concentrated in the center of the domain (right). The different lines represent the real process, the kriging predictor and the prediction kriging interval at 95%.

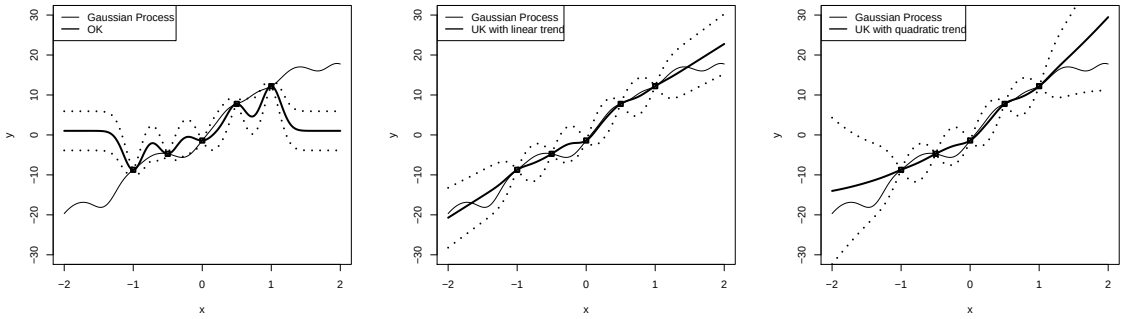

FIGURE 4. Extrapolation to a $[-2, 2]$ domain by three different kriging methods. The different lines represent the real process, the kriging predictor and the prediction kriging interval at 95%.

In the one dimensional case the choice of the trend doesn't seem to be essential while interpolating data which are not very distant one from another with respect to the frequency of the variation of the process. On the contrary, when the design is not regular and we are in extrapolation, the kriging performances are very sensitive to the adequacy between the real trend of the process and the kriging trend. Hence it is enough to fill the space in order to avoid the risk caused by the choice of the trend. But what is possible in one or two dimensions becomes unrealistic with increasing dimension: a design with only one point at each vertex of a cubic domain $[0, 1]^d$ has 2^d points, i.e. 1024 points in 10 dimensions and more than a billion points in 30 dimensions (30-d is frequently encountered in the applications). As we usually dispose of 10 data per dimension (which is already an optimistic case) the data based choice of the trend is a very difficult task.
Let us see nevertheless what it would be possible in order to choose the trend starting from a data set $(\mathbf{X},\ \mathbf{Z})$:
The classical frame of linear regression offers a panel of diagnostic tools dedicated to validating both assumptions on the trends and on the model of residuals. For instance, commonly used indicators include R^2 (and R^2 adjusted), the F-ratio and the p-values for each estimated regression coefficients, and numerous criteria to check the adequacy of the residuals to the underlying model.

In most cases the gaussian likelihood of the residuals is considered among the relevant criteria of model selection (some model testing techniques or even based upon it).

Now it seems necessary to recall that the latter measures are exclusively done at the design of experiments, also called "training sample" or "learning sets" in the literature of statistical learning (see Hastie et al. (2001)). Selecting only on the basis of a R^2 fit would lead for instance to the systematic choice of models interpolating ($\mathbf{X}$, $\mathbf{Z}$). However such models are not meant to be good in prediction outside the design of experiments. This warning leads to the double message:

- model complexity must be taken into account in selection procedures
- testing the model at some points not used in the model fitting could be worth: this is for instance what cross-validation does.

The following experiment was performed in an intent to illustrate the first point. The second point will be illustrated in the next section.
Here we investigate on a simple case how trend selection may be misleading when likelihood is the only criterion (without any consideration of model complexity). To do so, we compute, for each trend form of the kriging model, the optimal parameter $\hat{p}$ by ML, we compare the corresponding values of the likelihoods and we select the kriging model having the highest value of likelihood. We compare in the next table three kriging models (OK, UK with linear trend, UK with quadratic trend) for three different processes: a realization of a one dimensional Gaussian process with 11 points and with Gaussian covariance function (of parameters $\psi = (5, 0.4)$), the same realization plus a linear trend (equal to $0.5 + 5x$) and the same realization plus a quadratic trend (equal to $0.5 + 5x + 5x^2$).

TABLE 3. Comparison of minimum values of $-2\log(L)$ for different processes:

	GP		GP +linear $\mathbf{t}$		GP+quadratic $\mathbf{t}$	
kriging type	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$	$\hat{p}$	$-2\ln(L(\hat{p}))$
OK	0.4082	32.07	0.4445	36.90	0.4595	38.80
UK, linear $\mathbf{t}$	0.4085	31.89	0.4085	31.89	0.4387	35.80
UK, quadratic $\mathbf{t}$	0.4084	31.89	0.4084	31.89	0.4084	31.89

Here it is essential to point out that the likelihood values are necessarily larger when a model is applied with more degrees of freedom (which happens for instance between a first order and a second order polynomial trend), and hence L will always increase (decreasing values of $-2\ln(L(\hat{p}))$) with the complexity of the model. (What we could really compare are maximum values of the likelihood with the same number of degrees of freedom.) On the last line of Table 3, in the cases of the process without trend (GP) and of the process with linear trend (GP +linear $\mathbf{t}$) for which we compute the optimal likelihood, the estimated values $\hat{\beta}$ are very close to zero, hence one of the best kriging predictors in the neighbourhood the 11 points, but which will perform badly in extrapolation.

Kriging with external trend (see Cressie (1993)) seems to constitute a good alternative for solving both the problem of the "general" form of trend and the one of the circularity. However it raises other problems such as providing no uncertainty in estimation, and hence preventing from having a global view of model uncertainty.

4 On the use of additive models as external drift

Linear models are often used by practitioners of quantitative disciplines since they are simple to interprete and to assess. Additive models (AM) are an extension of linear models. A description of this method can be found in Hastie and Tibshirani (1991). The advantage of AM is to conserve the feature of non-interacting predictors, but they allow much more flexible inference for each univariate problem, using splines for instance. The generic additive decomposition of Z(x) can be written in the following way:

$$Z(x) = Z_{am}(x) + \varepsilon \text{ where } Z_{am}(x) = \alpha + \sum_{j=1}^{d} f_j(x_j) \text{ and } \epsilon \text{ is } n.i.i.d. \tag{8}$$

and the f_js are arbitrary univariate functions, one for each predictor. Once the nature of the f_js is chosen (possibly not the same for every dimension), there exist a unique solution, which can be estimated using a powerful iterative procedure called *backfitting algorithm* (see Hastie et al. (2001)).

In this section, we propose a combination of additive model and kriging that would offer the great flexibility of *AMs* and yet interpolate the data. The most obvious decomposition to achieve such a model is:

$$Z(x) = Z_{am}(x) + \varepsilon_{SK}(x), \text{ where } \varepsilon_{SK} \text{ is a process like in (1)} \tag{9}$$

This identity seems similar to the equation of Universal Kriging. However in this case, the non-linear (and even possibly non-parametric) nature of the trend prevents from solving the estimation globally. Indeed, a likelihood maximization would lead to an optimization problem in infinite dimension. On the other hand, the backfitting algorithm is not suited anymore if we take the kriging part into account. Consequently, we develop here a two-step approach: first, the additive trend $Z_{am}(x)$ is estimated using backfitting algorithm, and then Simple Kriging is applied to the residuals $Z(x) - Z_{am}(x)$ (and the correlation parameters are given by likelihood maximization). Unfortunately, there are significant drawbacks in the latter procedure, mainly related to the uncontrolled trade-off between determinist and stochastic parts. Hence, the whole uncertainty reduces here to the kriging variance estimated on the residuals (there is indeed no global uncertainty on the trend unless we use only splines in the AM). This is likely to cause an underestimation of the prediction variance associated to the model. Furthermore, these residuals may be not very well suited to estimate the gaussian process part: since the additive model is constructed to fit Z accurately on the design of experiments -possibly leading to *overfitting*-, the residuals may vary with a small magnitude that would prevent a reasonable generalization outside the design of experiments.

The previous approach is applied to a 3-dimensional industrial case. The data are obtained by simulation and the numerical response Z is studied as a function of three physical parameters characterizing the porous media and denoted by X_1, X_2 and X_3. The surface is simulated at the 1331 locations corresponding to a 11-level complete factorial design (denoted by "F" in the sequel). Our goal is to provide a metamodel of the simulator on the basis of a poor design of experiments. The metamodel should interpolate the data (to respect the determinism of the underlying simulation) and provide a prediction uncertainty that allows statistical-based exploration (for instance to solve optimization problems). Furthermore, it should take into account a prior knowledge inherited from a previous study: the phenomenon is almost additive in its parameters.

Here we present an experimental work we conducted on the industrial case in an intent to identify and estimate an accurate interpolation metamodel addressing the needs previously exposed, and yet constructed it with only 11 data per dimension. We took at first a 20-elements Hammersley sequence ("H") and then completed it for intermediate validation and re-estimation with 14 additional points ("A") taken from a 40-elements D-optimal design. We finally proceeded to a posterior validation on a 11-level complete factorial design ("F", with $11^3 = 1331$ elements.). Concerning the models, we considered isotropic kriging with gaussian covariance and several external drifts: linear models (first and second order polynomials), and additive models. At first we performed the 2-step procedure explained hereabove: straight linear or additive regression followed by a simple kriging of the residuals with MLE for the covariance parameters. Some quantitative results are presented in table 4. Then we proceeded to an intermediate validation on the additional design "A" and proposed an alternative method for kriging parameters fitting. We finally compared the different approaches on the biggest design "F".
A graphical analysis of the coplots on the design "H" (see Figure 5) confirmed the prior belief of additivity. A first additive decomposition was then estimated using splines. We observed that we could take a linear trend in the inputs X_1, X_3 and a non-linear trend in X_2 without loosing much accuracy. Hence, we could try a simplified additive model and considered both sets of additive components described below (see Figure 5 – right).
Different krigings with external trend were fitted to the observed data of the design "H". We focused on four trends : a first and second order regression model and two additive models. The model "GAM splines" was constructed with splines in all directions, and the model "GAM mixed" is the one we customized with a spline only in the direction X_2 and first order linear trends in the others directions. For each model, we fitted trend models (respectively by OLS and backfitting) and measured their relevance using indicators (residuals deviance and p-values when available)

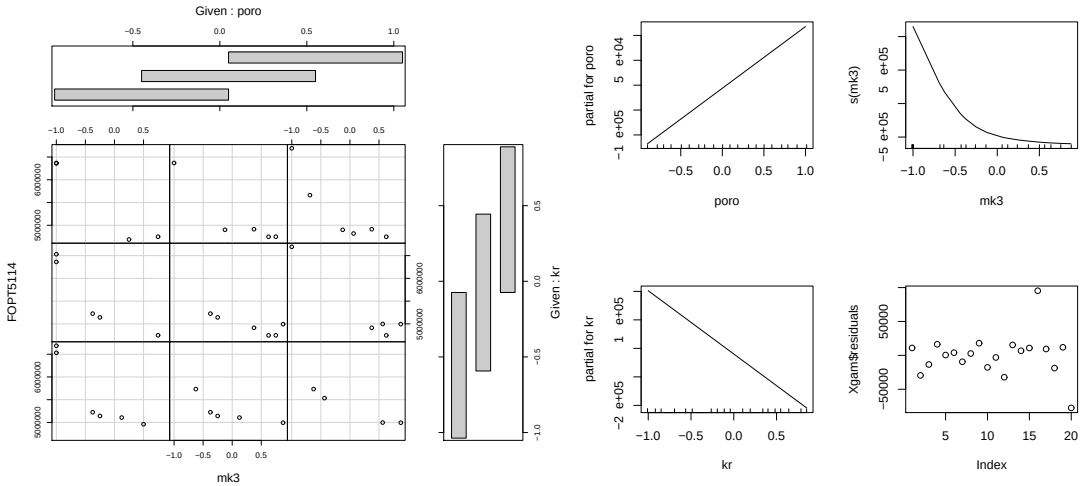

FIGURE 5. Coplots of Z on the Hammersley design (left) and summary of the additive components and the residuals given by the backfitting algorithm within GAM estimation (right).

computed with the residuals at the design "H". Then we fitted a kriging model to the residuals, as explained earlier. For each kriging model, we stored the maximum reached value of the likelihood and the corresponding range and variance values. The results are listed in Table 4.

TABLE 4. Estimated model parameter and loglikelihood obtained by detrended kriging

Model	Loglikelihood	Range	σ^2	R^2_{adj}	R^2	p-value
1st order Linear + SK	-276.46	1.04	1.36 e+11	0.78	0.82	4.03e-06
2st order Linear + SK	-255.34	0.048	7.19e+09	0.97	0.98	6.44e-08
GAM splines + SK	-230.67	0.048	6.1e+08	-	0.99	-
GAM mixed +SK	-235.28	0.16	9.67e+08	-	0.99	-

These results support the belief that a general additive trend is adapted for these data: both the variance of residuals and the values of their likelihood (compared to the 2nd order linear model, which uses more degrees of freedom) indicate their good fit to the data.

In practice, however, we care more about the model's abilities to make correct predictions outside the design of experiments than about its mean squared error at $\mathbf{X}$. Hence, model validation should not be blindly supported by the indicator R^2 or the likelihood of the residuals at $\mathbf{X}$. First, we should consider the number of degrees of freedom of the model. Second, it may be worth validating the model outside the design of experiments. Indeed, the residuals drawn from Figure 6 were computed in the same locations as those used to fit the model.

Concerning the first point, we compared the degrees of freedom of both "2nd order linear" and "GAM mixed": respectively 7 and 6. Concerning the second point, we conducted a validation test on some additional data, inspired by the cross-validation procedure. In order to valid the adjusted trend it seemed to us meaningful to make the comparison of the trend and the real response outside of the design of experiments. Hence, the design "A" was used to valid (and then update) the parameters associated to the model fitted at $\mathbf{X}$. The 14 locations of this design were used to test the validity of the covariance parameters previously found by MLE. Figure 6 shows the residuals $\varepsilon = Z - Z_{am}$ standardized by the MLE variance and the variation of the MSE as a function of the scale parameter p. We recall that the residuals must satisfy the assumption of normality in order to get relevant kriging variances and hence predictions. Figure 6 (left) shows that the MLE variance hasn't the right value for the residuals at the design "A" to be compatible with the model assumptions: $[-1.96, 1.96]$ should be a 95% confidence interval for the standardized residuals. In other respects, Figure 6 (right) shows that the mean square error at the test design could be significantly reduced by increasing the range.
After both those observations, we decided to re-fit the parameters on the basis of those new

residuals. Instead of a MLE , we chose at first to exploit the work done to compute the MSE as a function of the range. We already knew indeed the range that gives the best fit on the test data: $p_A = 1.5$. Concerning the variance, we obtained satisfying standardized residuals with $\sigma_A^2 = 4 \times \sigma_{MLE}^2$. So we took σ_A^2 as kriging variance.
Remark: A MLE with these residuals gave $\sigma^2 = 5.9 \times 10^9$ and $p = 0.97$.

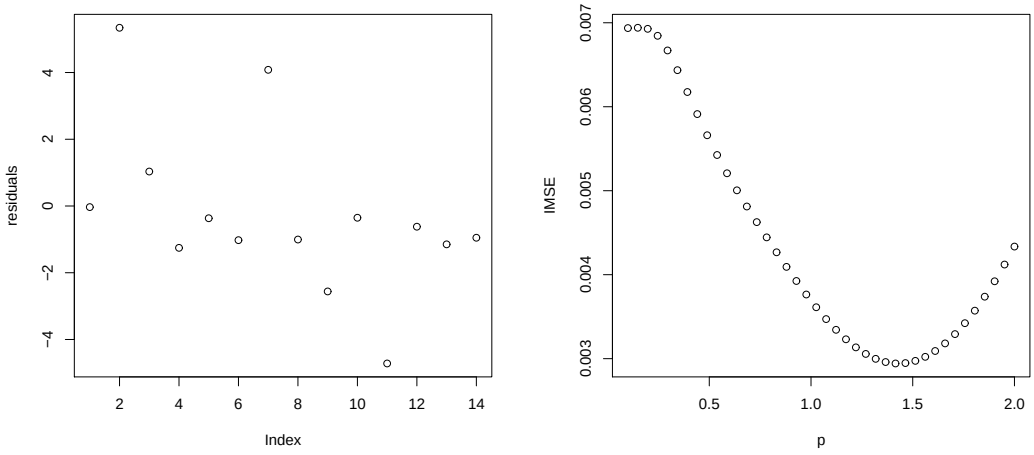

FIGURE 6. Standardized residuals plot on the validation design and evolution of the MSE with respect to the scale parameter p

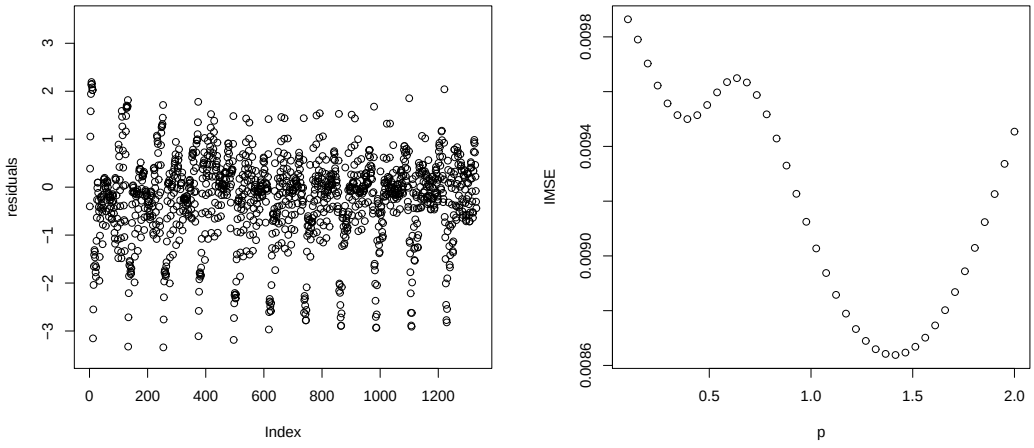

FIGURE 7. Standardized residuals plot on the full design with re-estimated residual variance Validation of the model on the factorial design 11^3

We finally tested the 2-step model on the complete design "F" (see Figure 7). The standardized residuals (with the variance σ_A^2) and the MSE as a function of p validated our empirical decisions made on the basis of the test design "A" (note that MLE on "A" gave also better results than MLE on "H" but the cross-validating strategy minimizing the MSE at "A" remained the best). To conclude with, the model investigated performed well in this case study: kriging seems to constitute a good complement to additive models, in an intent to interpolate data and also possibly explain a non-additive part. The method we used here allows inference of covariance parameters with values suited for a correct quantification of uncertainty. This seems encouraging to develop further "cross-validation-like" methods for the combination *Additive model + kriging*.

5 Conclusion

We observed in a 1-dimensional frame that MLE could behave very differently from Fisher asymptotical results when n is small. This result have to be kept in mind when dealing with higher dimensions, and further studies should be done in this latter context.
Further experiments on the topic of trend selection illustrated the fact that the likelihood cannot be considered as only criterion when comparing different functional families. This is suggesting methods penalizing complexity (like in AIC and BIC). But we mainly wish to emphazise on the risks took when predicting with trended kriging: in higher dimensions, we will allways be in an extrapolation situation. Choosing a trend with the help of a small design then seems very risky. This is an argument to consider ordinary kriging in the cases where no prior information on the trend is available.
In other respects, we proposed a model combining an additive model and simple kriging. The application to an industrial case confirmed that directly kriging the residuals by MLE gives a poor result. Our attempt to adapt a method inspired by cross-validation with a single test set gave here a kriging with different features from MLE, apparently accounting well for the non-additive part of the response. However, the question of the robustness to a change of design has not been raised yet. This is a subject to be treated in further works.

Aknowledgements

All the computations have been performed using *R* (R Development Core Team (2006)) and the packages *geoR* (Ribeiro Jr. and Diggle (2001)), *RandomFields* (Schlather (2001)), *gam* and *gstat*.

Bibliography

Cressie, N. A., 1993. Statistics for spatial data. Wiley series in probability and mathematical statistics.

Hastie, T., Tibshirani, R., 1991. Generalized Additive Models. Chapman and Hall.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

Jones, D., M., S., W.J., W., 1998. Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13, 455–492.

Jourdan, A., 2002. Approches statistiques des expériences simulées. Revue de Statistiques Appliquées 50, 49–64.

Martin, J., Simpson, T., 2005. Use of kriging models to approximate deterministic computer models. AIAA Journal 43 (4), 853–863.

Martin, J. D., Simpson, T. W., 2004. A monte carlo simulation of the kriging model. AIAA Journal.

Matheron, G., 1963. Principles of geostatistics. Economic Geology 58, 1246–1266.

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety (91), 1290–1300.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Ribeiro Jr., P., Diggle, P., 2001. geor: A package for geostatistical analysis. ISSN 1609-3631.
URL `http://CRAN.R-project.org/doc/FAQ/R-FAQ.html`

Sacks, J., Welch, W., Mitchell, T., H.P.Wynn, 1989. Design and analysis of computer experiments. Statistical Science (4), 409–435.

Santner, T., Williams, B., Notz, W., 2003. The Design and Analysis of Computer Experiments. Springer.

Schlather, M., 2001. Simulation and analysis of random fields.
URL `http://www2.hsu-hh.de/schlath/index.html`