Assessment of uncertainty in computer experiments, from Universal Kriging to Bayesian Kriging.

C. Helbert[1], D. Dupuy[1], L. Carraro[1]

[1] ENS Mines de Saint-Etienne, Dep. of Mathematics, Cours Fauriel, 158 – 42100 St-Etienne, France

Keywords: Universal Kriging, Bayesian Kriging, Computer experiments, Monte Carlo Markov Chain.

Kriging was first introduced in the field of geostatistics. Nowadays, it is widely used to model computer experiments (Sacks, 1989). As an interpolation method, it translates the fact that the results of deterministic computer experiments have no experimental variability. Kriging is also used because it allows quantifying the prediction uncertainty which plays a major role in many applications (Jones, 1998, Oakley, 2004). Among practitioners we can distinguish those who use Universal Kriging where the parameters of the model are estimated (by maximum likelihood for example) and those who use Bayesian Kriging (Goria, 2004, Martin, 2004, Oakley, 2004) where model parameters are random variables.
The aim of this paper is to show that the prediction uncertainty has a correct interpretation only in the case of Bayesian Kriging. Different cases of prior distribution have been studied. Especially, in a specific case of prior distribution, Bayesian Kriging supplies an interpretation as a conditional variance for the prediction variance provided by Universal Kriging. Finally, we show on a simple petroleum engineering example the importance of prior information in the Bayesian approach.

1 Prediction uncertainty in Universal Kriging

Let D be included in $R^k, k \geq 1$. We suppose that the output *y* is a function of $x \in D$. We assume that *y* is the realization of a Gaussian random field $\left(Y(x)\right)_{x \in D}$ such that:

$$E\left(Y(x)\right) = f(x)\beta \tag{1}$$

$$\text{and } Cov\left(Y(x), Y(x+h)\right) = \sigma^2 R(h \mid \theta) \tag{2}$$

where $f(x) = \left(f_0(x) \ \ldots \ f_p(x)\right)^T$ is a known trend vector, $\beta = \left(\beta_0 \ \ldots \ \beta_p\right)^T$ is the vector of trend coefficients, and $\theta = \left(\theta_0 \ \ldots \ \theta_p\right)^T$ is the vector of correlation coefficients.
Note that we have used the Gaussian spatial correlation function for the examples of sections 2 and 3. This choice supposes a very smooth and infinitely differentiable surface. Depending on the characteristics of the studied response, other correlation functions such as spherical or exponential ones could be used.
Furthermore, let $Y = \left(y_1 \ \ldots \ y_n\right)^T$ be the output observed at locations $X = \left(x_1 \ \ldots \ x_n\right)^T$.

In the case where all the parameters of the model are known (trend, range and variance), the kriging predictor, also called simple Kriging $Y_{SK}(x_0)$, and the prediction variance $\sigma^2_{SK}(x_0)$ at a new location x_0 are given by (Santner, 2003):

$$Y_{SK}(x_0) = f(x_0)\beta + r_\theta^T R_\theta^{-1}\left(Y - F\beta\right) \tag{3}$$

$$\text{and } \sigma^2_{SK}(x_0) = \sigma^2\left(1 - r_\theta^T R_\theta^{-1} r_\theta\right) \tag{4}$$

where

$$R_\theta = \left(R\left(x_i - x_j \mid \theta\right)\right)_{i,j} \tag{5}$$

$$r_{\theta}^{T} = \left(R_{\theta}(x_0, x_1) \quad \dots \quad R_{\theta}(x_0, x_n) \right) \tag{6}$$

$$F = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \tag{7}$$

Here, from the theoretical point of view, the predictor and its variance can be interpreted as the expectation and the variance of $Y(x_0)$ conditional to the observations.

However in practice the parameters of the external trend and/or those of the covariance function are not known. They are usually estimated through the optimization of a criterion, like maximum likelihood or cross validation. The kriging predictor $Y_{UK}(x_0)$, called Universal Kriging, and its prediction variance $\sigma_{UK}^2(x_0)$ are then modified to take into account parameter estimation. Their expressions are the following:

$$Y_{UK}(x_0) = f(x_0)\hat{\beta} + r_{\hat{\theta}}^{T} R_{\hat{\theta}}^{-1}\left(Y - F\hat{\beta}\right) \tag{8}$$

and $$\sigma_{UK}^2(x_0) = \hat{\sigma}^2\left(1 - r_{\hat{\theta}}^{T} R_{\hat{\theta}}^{-1} r_{\hat{\theta}} + \left(f(x_0) - r_{\hat{\theta}}^{T} R_{\hat{\theta}}^{-1} F\right)\left(F^{T} R_{\hat{\theta}}^{-1} F\right)^{-1}\left(f(x_0) - r_{\hat{\theta}}^{T} R_{\hat{\theta}}^{-1} F\right)^{T}\right) \tag{9}$$

Using maximum likelihood estimation, one obtains:

$$\hat{\beta}_{ML} = \left(F^{T} R_{\hat{\theta}}^{-1} F\right)^{-1} F^{T} R_{\hat{\theta}}^{-1} Y \tag{10}$$

$$\hat{\sigma}^2{}_{ML} = \frac{\left(Y - F\hat{\beta}\right)^{T} R_{\hat{\theta}}^{-1}\left(Y - F\hat{\beta}\right)}{n - (p+1)} \tag{11}$$

$$\hat{\theta}_{ML} = \arg\min\left(\frac{n}{2} + \frac{n}{2}\log(2\pi\hat{\sigma}^2) + \frac{1}{2}\log\left(\det\left(R_{\theta}\right)\right)\right) \tag{12}$$

It can be noted that $Y_{UK}(x_0)$ of expression (8) is obtained by substituting β by its estimation in (3). Besides, variance of Universal Kriging (9) is larger than variance of Simple Kriging (4) since it includes uncertainty on β. Unfortunately expressions (8) and (9) cannot be interpreted as conditional expectation and variance. Indeed, the probability law of $\hat{\theta}$ is not known, consequently that is the same for $\hat{\beta}$ and $\hat{\sigma}^2$ which expressions depend on $\hat{\theta}$. Besides, expression (9) does neither consider uncertainty due to covariance parameters estimation nor to variance estimation.

In the following part, we show that the Bayesian context allows interpreting $\sigma_{UK}^2(x_0)$ as a conditional variance.

2 The Bayesian approach to interpret Universal Kriging's prediction variance

This section will be illustrated with the set of data of Martin and Simpson, 2004, in which the output is the temperature of a chemical reaction (Figure 1). The mass ratio of oxidant to fuel being burned (the input) is increased from no oxidant to an excess of oxidant. In this process, the reaction increases in temperature to a maximum and then decreases as excess of oxidant is added. The output is observed on 11 regularly distributed values on the interval [0,1].

From here, we will assume that $\left(Y(x) \mid \beta, \sigma^2, \theta\right)_{x \in D}$ is a Gaussian random field such that expectation and spatial covariance function are equal to $f(x)\beta$ and $\sigma^2 R(. \mid \theta)$ respectively. Moreover, model parameters are considered random which prior joint density denoted by π.

In this Bayesian context, the predicted value at any point x of domain D and the prediction variance are simply given by the expectation and the variance of the output conditionally to the observations, *i.e.* $E\left(Y\left(x\right)|Y\right)$ and $\mathrm{var}\left(Y\left(x\right)|Y\right)$.

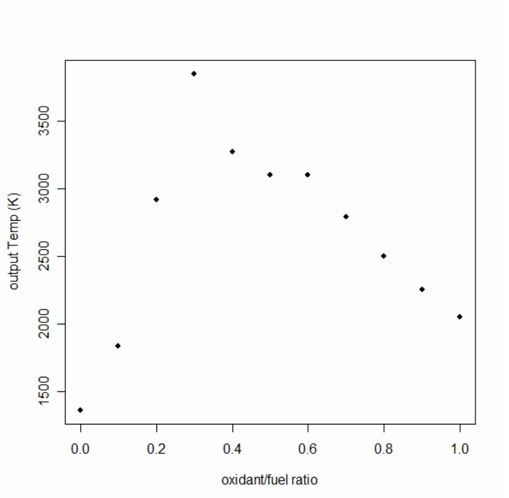

FIGURE 1 data set.

The Bayesian rules give the following general expression for any function *g*:

$$E\left(g\left(Y\left(x\right)\right)|Y\right)=\iiint_{\beta\sigma^2\theta} E\left(g\left(Y\left(x\right)|Y,\beta,\sigma^2,\theta\right)\right)\pi\left(\beta,\sigma^2,\theta|Y\right)d\beta d\sigma^2 d\theta \qquad (13)$$

The conditional variance, the conditional density etc. derive from this formula.
In the right term of expression (13) one can recognize the Simple Kriging: $Y\left(x\right)|Y,\beta,\sigma^2,\theta$ is indeed a Gaussian random variable with mean $Y_{SK}\left(x\right)$ and variance $\sigma^2_{SK}\left(x\right)$.

a. *Known variance, known correlation parameters and conjugate prior for trend parameters.*

This case is interesting because analytical calculations can be conducted when the prior law of trend parameters is assumed to be Gaussian.
Let β be a Gaussian random vector with mean μ and variance $\lambda\Sigma$, where λ is a positive scalar and Σ is a symmetric definite positive matrix. The posterior distribution of β is also Gaussian with the following parameters:

$$E\left(\beta|Y\right)=\mu+\lambda\Sigma F^T\left(\lambda F\Sigma F^T+\sigma^2 R_\theta\right)^{-1}\left(Y-F\mu\right) \qquad (14)$$

$$Var\left(\beta|Y\right)=\lambda\Sigma-\lambda^2\Sigma F^T\left(\lambda F\Sigma F^T+\sigma^2 R_\theta\right)^{-1}F\Sigma \qquad (15)$$

As mentioned before, $Y\left(x_0\right)_{|Y,\beta}$ is Gaussian with the same parameters as the Simple Kriging:

$$E\left(Y\left(x_0\right)|Y,\beta\right)=\left(f\left(x_0\right)-r_\theta^T R_\theta^{-1}F\right)\beta+r_\theta^T R_\theta^{-1}Y \qquad (16)$$

$$Var\left(Y\left(x_0\right)|Y,\beta\right)=\sigma^2\left(1-r_\theta^T R_\theta^{-1}r_\theta\right) \qquad (17)$$

The posterior distribution for the output is also Gaussian:

$$E\left(Y\left(x_0\right)|Y\right)=\left(f\left(x_0\right)-r_\theta^T R_\theta^{-1}F\right)\left(\mu+\Sigma F^T\left(\lambda F\Sigma F^T+\sigma^2 R_\theta\right)^{-1}\left(Y-F\mu\right)\right)+r_\theta^T R_\theta^{-1}Y \qquad (18)$$

$$Var\left(Y\left(x_0\right)|Y\right)=\left(f\left(x_0\right)-r_\theta^T R_\theta^{-1}F\right)\left[\lambda\Sigma-\lambda^2\Sigma F^T\left(\lambda F\Sigma F^T+\sigma^2 R_\theta\right)^{-1}F\Gamma\right]\left(f\left(x_0\right)-r_\theta^T R_\theta^{-1}F\right)^T$$
$$+\sigma^2\left(1-r_\theta^T R_\theta^{-1}r_\theta\right) \qquad (19)$$

Two particular cases can be noticed. Firstly, when λ tends to zero, we obtain equations of Simple Kriging, (3) and (4).

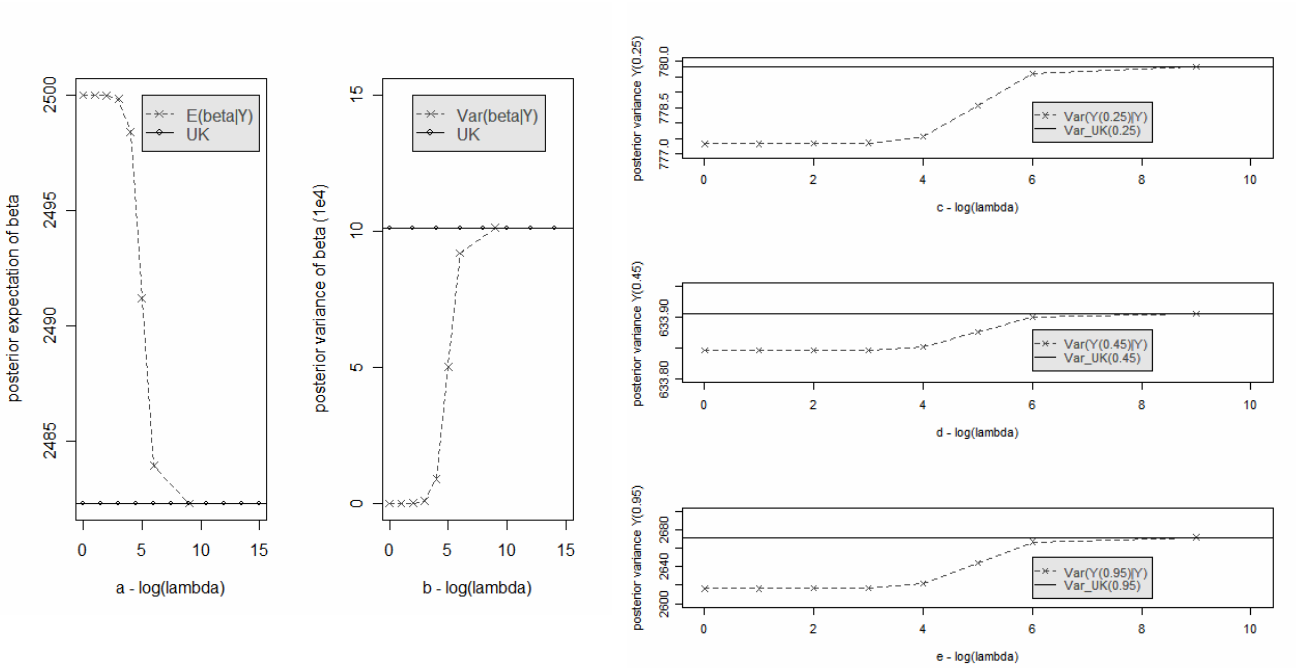

FIGURE 2 Evolution of the expectation (a) and the variance (b) of the posterior distribution of β and of the variance of the posterior distribution of $Y(x)$, $x \in \{0.25(c), 0.45(d), 0.95(e)\}$ (c,d,e), when the prior distribution is Gaussian with mean m = 2500 and variance λ, $\lambda \in \{10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^9\}$

Secondly, when λ tends to infinity, the first moments of the posterior distribution of β tend to the expectation and the variance of the maximum likelihood estimator (Figure 2):

$$E(\beta|Y)_{\lambda=+\infty} = \left(F^T R_\theta^{-1} F\right)^{-1} F^T R_\theta^{-1} Y = \hat{\beta}_{ML} \quad (20)$$

$$Var(\beta|Y)_{\lambda=+\infty} = \sigma^2 \left(F^T R_\theta^{-1} F\right)^{-1} = Var\left(\hat{\beta}_{ML}\right) \quad (21)$$

Figure 2 (a) and (b) presents the evolution of $E(\beta|Y)$ and $Var(\beta|Y)$ with λ varying from 1 to 10^9. It can be observed that when the variance is high (non informative prior), the posterior distribution is the ML's one. In such a case, the moments of the posterior distribution of $Y(x_0)$ are:

$$E(Y(x_0)|Y) = \left(f(x_0) - r_\theta^T R_\theta^{-1} F\right)\hat{\beta}_{ML} + r_\theta^T R_\theta^{-1} Y \quad (22)$$

$$Var(Y(x_0)|Y) = \sigma^2 \left[\left(f(x_0) - r_\theta^T R_\theta^{-1} F\right)\left(F^T R_\theta^{-1} F\right)^{-1}\left(f(x_0) - r_\theta^T R_\theta^{-1} F\right)^T + \left(1 - r_\theta^T R_\theta^{-1} r_\theta\right)\right] \quad (23)$$

This result can be observed on Figure 2 c (resp. d and e) which presents the evolution of $Var(Y(x_0)|Y)$ with λ for $x_0 = 0.25$ (resp. $x_0 = 0.45$ and $x_0 = 0.95$). In expression (22) and (23) one recognizes the predicted value and the prediction variance supplied by Universal Kriging. Hence, Universal Kriging is confounded with Bayesian Kriging in the particular case of an uniform prior distribution for β, and when σ^2 and θ are constants. This is not the case for other prior distributions anymore, as shown in next part.

b. *Known correlation parameters and non informative priors for trend and variance parameters.*

This case is interesting because the posterior distributions are centred on the maximum likelihood estimators: nevertheless, the prediction variance of Universal Kriging does not correspond to a conditional variance.

Let us define the joint prior density as $\pi(\beta,\alpha)=\frac{1}{\alpha}$, where $\alpha=\frac{1}{\sigma^2}$.

Thus, theorical results (Goria,2004) give

$$\pi(\beta,\alpha\,|\,Y)=\phi\left(\beta\,|\,\sigma^2\right)\gamma(\alpha) \qquad (24)$$

In expression (24), ϕ is the density of the normal distribution centred on $\hat{\beta}_{ML}$ and with variance $\sigma^2\left(F^T R^{-1} F\right)^{-1}$, and γ is the density of the gamma distribution with a shape of $\frac{n-(p+1)}{2}$ and a scale of $\frac{2}{\left(n-(p+1)\right)\hat{\sigma}^2_{ML}}$. Hence, the mean of this distribution is exactly $\frac{1}{\hat{\sigma}^2_{ML}}$. Thus, in Bayesian context with non informative prior (defined as above), the joint posterior distribution of the trend parameters and the variance is centred on the ML estimators, that is to say on the same parameters as those used in Universal Kriging. Therefore, it is interesting to compare the two approaches into details (Figure 3).

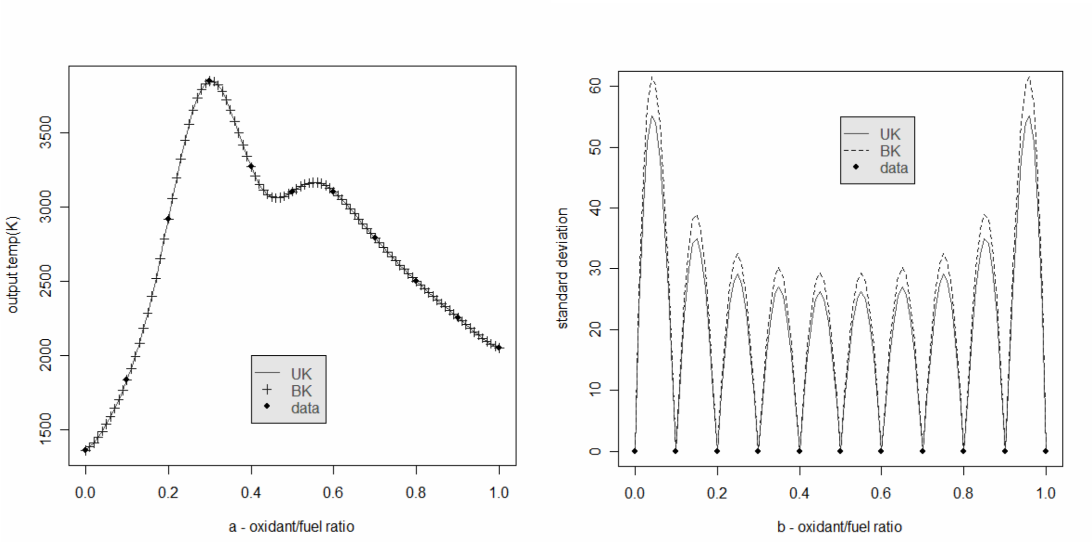

FIGURE 3 Comparison between $Y_{UK}(x)$ and $Y_{BK}(x)=E\left(Y(x)\,|\,Y\right)$ (plot (a)) and between $\sigma_{UK}(x)$ and $\sigma_{BK}(x)=\sqrt{Var\left(Y(x)\,|\,Y\right)}$ when the prior is non informative, i.e. $\pi(\beta,\alpha)=\frac{1}{\alpha},\left(\alpha=1/\sigma^2\right)$.

One can observe on Figure 3 (a) that $Y_{UK}(x)$ and $Y_{BK}(x)=E\left(Y(x)\,|\,Y\right)$ give the same results. In this particular case, the Universal Kriging estimator can be interpreted as a conditional expectation. Nevertheless, it is not the same for the prediction variance of Universal Kriging $\sigma^2_{UK}(x)$ which is inferior to Bayesian variance $\sigma^2_{BK}(x)=Var\left(Y(x)\,|\,Y\right)$. This difference is mainly explained by the fact that $\sigma^2_{UK}(x)$ takes only into account the uncertainty due to the estimation of β and not the uncertainty due to the one of σ^2 .

Thus, this short example shows that the $\sigma^2_{UK}(x)$ can not be interpreted as a conditional variance and that Universal Kriging underestimates uncertainty when it is compared to uncertainty of non informative Bayesian Kriging.

c. Known correlation parameters and conjugate priors for trend and variance parameters.

The aim of this case is to validate the Monte Carlo Markov chain simulations (MCMC), useful in practice to get samples from posterior distributions which are not explicitly known (for example, when there is a prior distribution on θ).

Let the prior distribution be the conjugate prior (Gaussian for β and Gamma for $1/\sigma^2$):

$$\pi\left(\beta,\frac{1}{\sigma^2}\right)=\mathrm{N}\left(\mu,\sigma^2\Sigma\right)\Gamma\left(a_1,\,1/a_2\right) \tag{27}$$

The posterior distribution is then well known (Gaussian for β and Gamma for $1/\sigma^2$):

$$\pi\left(\beta,\frac{1}{\sigma^2}|Y\right)=\mathrm{N}\left(\mu_Y,\sigma^2\Sigma_Y\right)\Gamma\left(a_{11},\,1/a_{22}\right) \tag{28}$$

where

$$\mu_Y=\Sigma_Y.\left(F^TR^{-1}Y+\Sigma^{-1}\mu\right)\text{ and }\Sigma_Y=\left(\Sigma^{-1}+F^TR^{-1}F\right)^{-1} \tag{29}$$

$$a_{11}=a_1+\frac{n}{2}\text{ and }a_{22}=a_2+\frac{\mu^T\Sigma^{-1}\mu+Y^TR^{-1}Y-\mu_Y{}^T\Sigma_Y{}^{-1}\mu_Y}{2} \tag{30}$$

The Metropolis Hastings algorithm (Robert, 1996) is used to compute MCMC simulations with a Gaussian random walk.

Table 1 and Figure 4 show that the resulting posterior distributions are closed to the theoretical ones. The short difference comes from imprecision of the sampling method.

Thus, simulations will be used to compute the distribution of the output at any point of the domain conditionally to the observations and for every kind of prior distributions, proper or improper. Note that the surfaces are generally compared through the first two moments of the distribution: posterior expectation and posterior variance. In this Bayesian context, the posterior variance includes all sources of uncertainty that comes from the trend, the variance and the correlation function.

TABLE 1 Comparison of the parameters of the sample simulated by MCMC and theoretical parameters in the case of the conjugate prior.

	beta		**1/sigma2**	
prior distribution	μ	Σ	a_1/a_2	a_1/a_2^2
	2500	1000^2	10^{-5}	10^{-10}
posterior distribution	μ_Y	Σ_Y	a_{11}/a_{22}	a_{11}/a_{22}^2
Theoretical results	2482	0.22	$2.50\ 10^{-6}$	$9.61\ 10^{-13}$
Simulation results (MCMC)	2477	0.23	$2.50\ 10^{-6}$	$9.17\ 10^{-13}$

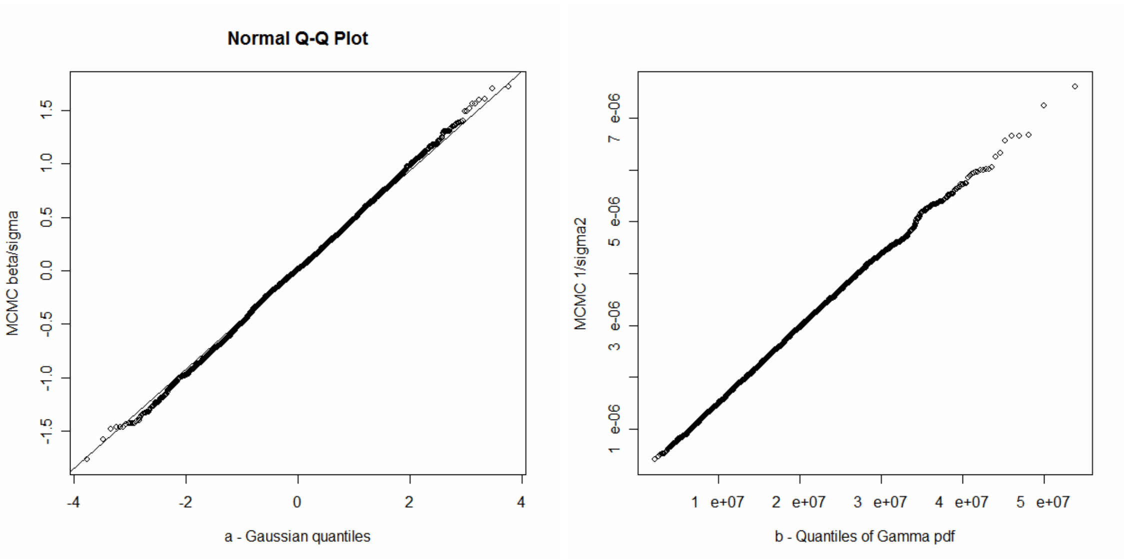

FIGURE 4 Checking of law adequacy between theoretical posterior distributions and data simulated using a MCMC method in the case of the conjugate prior.
(a) Posterior distribution of $\beta | \sigma^2$ (b) Posterior distribution of σ^{-2}

d. *Prior distribution on β, σ^2 and θ*

Let us consider the same experimental case as Martin and Simpson, 2004, where the prior is:

$$\pi\left(\beta, \frac{1}{\sigma^2}, \theta\right) = \frac{1}{\left(\sigma^2\right)^{3/2}} \tag{31}$$

The posterior distributions are here sampled for all model parameters (trend, variance and also correlation) using MCMC techniques. The results are validated by a comparison to the posterior distributions of Martin and Simpson's paper.

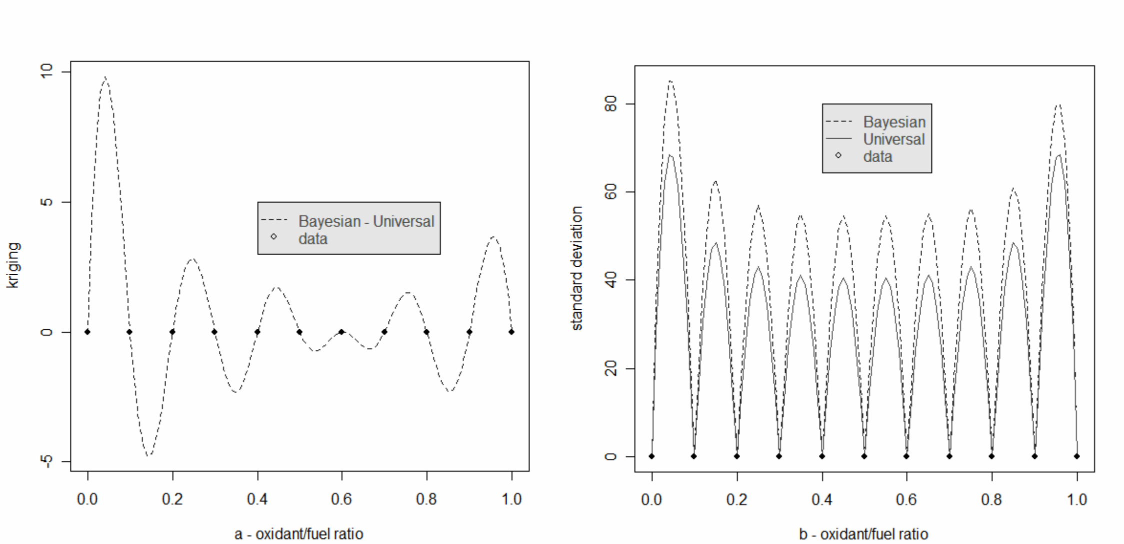

FIGURE 7 Comparison between Universal Kriging and Bayesian Kriging in the non informative case where the prior is $\pi\left(\beta, 1/\sigma^2, \theta\right) = 1/\sigma^3$

On this example, one can compare Universal Kriging where the model parameters are estimated by maximum likelihood and Bayesian Kriging which is a mixture of Kriging models where parameters follow the posterior distribution. The left part of the Figure 7, which presents

$Y_{BK}(x) - Y_{UK}(x)$, shows that estimators are different, especially near the origin of the domain. Concerning uncertainty (Figure 7 (b)) standard deviation of Universal Kriging is always smaller than the one of Bayesian Kriging. Thus, expectation and variance provided by Universal Kriging and Bayesian Kriging are different. In particular, Universal Kriging underestimates uncertainty, a result already observed previously.

Another advantage of the Bayesian approach is the assessment of the whole distribution of the predicted values. For example, a very asymmetric posterior distribution will not be detected by universal approach.
At the same time, Bayesian Kriging avoids the optimization of the likelihood function which is often badly conditioned, especially in high dimension, when little information is available. Besides, the difference between approaches increases with dimension. Several works (Oakley and O'Hagan, 2004, Goria, 2004) are carried out in this context. However, these authors only use "classical" prior distributions: conjugate or non informative. The next part addresses a novel way to get prior information.

3 Case study: impact of the choice of the prior

A streamlines 3D oil production simulator (namely3DSL®) was used to illustrate the impact of the choice of the prior. The inputs we deal with are: two permeability factors (LMULTKZ and KRWMAX) and the well's bottom hole pressure (LBHP). They were transformed to belong to the range [-1, 1]. The simulator output is the field oil production (FOPT) after 7000 days. In oil industry, managers need to quantify precisely uncertainty which lays on the expected oil reserves, before taking any decision on the field's exploitation. The simulator is then used to propagate uncertainty from geophysics parameters (inputs) to the oil production (output). Nevertheless, a single run of 3DSL® takes a long time (2.5 CPU time). Consequently, very few simulations are available. A surrogate, obtained by Kriging for example, is built and uncertainty is propagating through the surrogate, instead of through the simulator itself.

The idea presented here consists in using simplified simulations as they require much less time consuming, to derive prior information for Bayesian Kriging. We study the impact of two different ways used to degrade simulations. These degraded simulators (Degraded 1) and (Degraded 2) use a number of nodes on each stream line ten times smaller than 3DSL®. Moreover, some constraints like the actualization of the field pressure are relaxed. The main difference between Degraded 1 and Degraded 2 comes from the time step used to do the calculus which is more accurate with Degraded 1 than with Degraded 2.

TABLE 3 Correlation between the 3 sets of simulations (3DSL®, Degraded1 and Degraded 2)

Correlation Coefficient	ALL	LMULTKZ	KRWMAX	LBHP
3DSL®/Degraded 1	0.95	0.61	0.97	0.98
3DSL®/Degraded 2	0.80	-0.19	0.89	0.91
Degraded 1/Degraded 2	0.89	0.56	0.96	0.94

In order to compare the simulators a full factorial design at 11 levels in each direction has been done. The whole surface contains 1331 points. Table 3 shows that 3DSL® and Degraded 1 are highly correlated with a correlation coefficient of 0.95. As expected, Degraded 1 is closer to 3DSL® than Degraded 2. It can be noticed that the correlation is very good in the directions of KRWMAX and LBHP (coefficient higher than 0.97). Figure 8 also shows this accuracy. It is not the case in the direction of LMULTKZ with a correlation coefficient of 0.61. Degraded 2 is less correlated with 3DSL® than Degraded 1. This can be observed by lower figures in Table 3 and also on Figure 8, which shows the mean of the response in each direction. Let us focus on two points of Figure 8:

- curve's level: Degraded 2 seems to be closer to 3DSL® than to Degraded 1,

- curve's variations: Degraded 1 seems to be closer to 3DSL® than to Degraded 2, a result which corresponds to the correlation coefficients of Table 3.

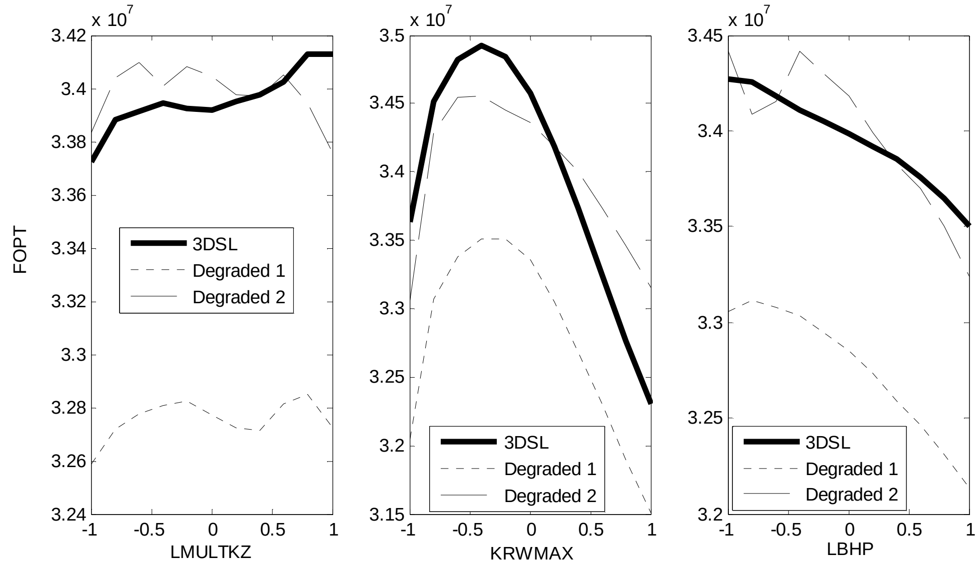

FIGURE 8 Comparison between 3DSL®, Degraded 1 and Degraded 2 in directions LMULTKZ, KRWMAX and LBHP.

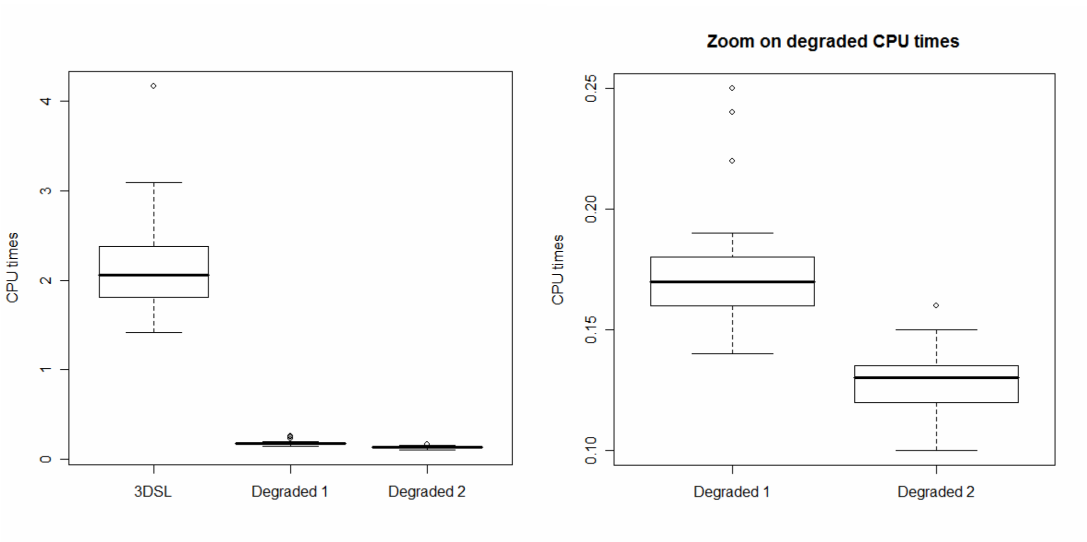

Figure 9 (left) Comparison of CPU times (3DSL®, Degraded 1, Degraded 2) for the same set of simulations (full factorial 3^3). (right) Zoom in on Degraded 1 and Degraded 2 CPU times.

Regarding CPU time[1], Figure 9 exhibits that one run takes around 2.5 CPU time with 3DSL® against 0.25 with Degraded 1 and only 0.18 with Degraded 2. Thus, Degraded 1 and Degraded 2 are quite close to 3DSL® but much less time consuming. These simulators will quickly deliver useful information to compute Bayesian Kriging.
As already mentioned, we want to build a predictive metamodel for 3DSL® in order to propagate uncertainty. We limit our study to surrogate obtained by Kriging (Universal or Bayesian) estimated using very few runs. The trend of the kriging model will supposed to be linear with respect to the factors.[2].

[1] The CPU time been higly dependant to the hardware, we provide here relative values of the time answering of the simulators.

[2] $Y\left(x \mid \beta, \sigma^2, \theta\right) = \beta_0 + \beta_1 Multkz + \beta_2 Krwmax + \beta_3 Lbhp + Z(x)$ where $Cov\left(Z(x), Z(x+h)\right) = \sigma^2 R\left(h \mid \theta\right)$

We will compare four different strategies equivalent with respect to CPU time consuming:

"**UK**"	the classical one using Universal Kriging on 20 simulations with 3DSL®,
"**no info BK**"	the Bayesian approach using a non informative Kriging on the same 20 runs,
"**info 1 BK**"	the Bayesian approach on 17 simulations with 3DSL® using a prior distribution obtained from 24 simulations of "Degraded 1",
"**info 2 BK**"	the Bayesian approach on 18 simulations with 3DSL® using a prior distribution obtained from 24 simulations of "Degraded 2".

Indeed, the four strategies almost consume the same amount of CPU time: 41.55 units of time for the first two strategies, 39.20 for the third one and 40.36 for the last one.

Note that the non informative law (see § 2b) used to compute strategy "**no info BK**" is the following:

$$\pi(\beta,\sigma,\theta)=\frac{\pi(\theta)}{\sigma} \quad (32)$$

To the best of our knowledge, little information is known about θ. We will use a uniform distribution on [0,10]. Longer range would not be coherent with the size of the domain, where the maximal estrangement between two points is equal to 2.
Note also that Bayesian Kriging with the same non informative prior has been computed on the 24 degraded simulations in order to extract the prior distribution needed by "**info 1 BK**" and "**info 2 BK**".

TABLE 4 Parameters posterior distributions on the 24 runs of Degraded 1 and 2

		β_0	β_1	β_2	β_4	σ^2	θ_1	θ_2	θ_3
24 runs of degraded 1	**E(.|Y)**	-0.40	-0.23	-0.52	-0.64	2.12	3.57	0.76	2.71
	Std(.|Y)	0.78	0.57	0.88	0.53	0.88	1.93	0.31	1.30
24 runs of degraded 2	**E(.|Y)**	-0.56	0.00	-0.63	-0.58	1.69	2.25	0.48	0.97
	Std(.|Y)	0.61	0.42	0.68	0.70	0.68	0.78	0.19	0.35

Expectations and variances of posterior distributions obtained on both simulators - Degraded 1 and Degraded 2 - are summed up in Table 4[3]. There are several differences between these two sets of data, especially on β_1, θ and σ^2: β_1 is equal to zero with Degraded 2 (LMULKZ has no influence on FOPT), the range parameters are smaller with Degraded2 and the total variance appears also smaller with Degraded 2.

The prior used for **"info 1 BK"** (resp. **"info 2 BK"**) is centred on parameters shown in the first two lines (resp. the last two lines) of Table 4. For example, mean and variance of σ^2 are *a priori* equal to 2.12 and 0.88 in strategy **"info 1 BK"**, whereas they are equal to 1.69 and 0.68 in **"info 2 BK"**.
Concerning distributions, the Normal law is chosen for β and θ and Lognormal for σ^2.

Table 5 presents the comparison of the four strategies which are evaluated through 4 indicators computed on the whole surface of N=1331 points :

- the **R**oot **M**ean **S**quare **E**rror: $RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Fopt(x_i)-\hat{Y}(x_i)\right)^2}$
- the **M**ean **A**bsolute **E**rror: $MAE=\frac{1}{N}\sum_{i=1}^{N}\left|Fopt(x_i)-\hat{Y}(x_i)\right|$

[3] Note that the values of parameters are not expressed in the same scale as the output's one (see Figure 8). Actually, data have been centred and reduced before modelling.

- the **A**verage **S**tandard **D**eviation: $ASD = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sigma_{\hat{Y}}^{2}(x_i)}$

- and the **Pr**obability of exceeding: $PR = \frac{1}{N}\sum_{i=1}^{N} 1_{\left|Fopt(x_i)-\hat{Y}(x_i)\right|>2\sigma_{\hat{Y}}(x_i)}$.

The RMSE and the MAE are average distances between the real surface composed of 1331 runs of 3DSL® and the surface which has been estimated by one of the four strategies.
The ASD represents the average uncertainty provided by each strategy.
The Pr is the proportion of the points which are outside the interval: $\left[\hat{Y}(x)-2\sigma_{\hat{Y}}(x);\hat{Y}(x)+2\sigma_{\hat{Y}}(x)\right]$.

TABLE 5 Comparison of the four strategies

	"UK"	"no info BK"	"info 1 BK"	"info 2 BK"
RMSE	205 042	204042	166770	223845
MAE	159 402	143190	125748	171786
ASD	99457	141573	188411	258394
Pr	0.40	0.19	0.08	0.04

"info 1 BK" appears the best strategy. Indeed, its RMSE and its MAE are smaller than those of other strategies'. The average uncertainty provided by this strategy is well estimated: the interval $\left[\hat{Y}(x)-2\sigma_{\hat{Y}}(x);\hat{Y}(x)+2\sigma_{\hat{Y}}(x)\right]$ contains 92% of the output of 3DSL®. Thus, the information that is obtained from Degraded 1 and introduced through Bayesian Kriging is useful: the prediction surface and its uncertainty are accurate.

We can observe that **"info 2 BK"** gives a RMSE and a MAE higher than other strategies. "Degraded 2" is quite far from 3DSL®, therefore the introduced information degrades the estimation of the surface. The difference between the surfaces obtained by the three Bayesian strategies can be observed through the difference of posterior distributions (see Table 6). The posterior distribution obtained after using the most degraded simulator ("Degraded 2") appears remarkable: mean of σ^2, teta2 and teta3 are smaller in this distribution than in the others. Note also that this distribution is less dispersed.

TABLE 6 : Parameters posterior distributions for the three Bayesian strategies

		beta0	beta1	beta2	beta3	sigma2	teta1	teta2	teta3
"info 1 BK"	**E(.|Y)**	-0.41	0.19	-0.41	-0.47	1.88	3.74	0.87	3.32
	Std(.|Y)	0.55	0.30	0.64	0.32	0.40	1.16	0.19	0.88
"info 2 BK"	**E(.|Y)**	-0.41	0.17	-0.60	-0.21	1.00	2.98	0.70	1.48
	Std(.|Y)	0.42	0.22	0.51	0.37	0.14	1.21	0.18	0.96
"no info BK"	**E(.|Y)**	-0.50	0.54	0.32	-0.18	2.64	3.54	1.05	4.75
	Std(.|Y)	0.99	0.46	1.07	0.41	0.75	1.37	0.25	2.00

What can also be noticed in Table 5 is that **"no info BK"** and **"UK"** give similar results according to RMSE and MAE. Indeed, when the prior is non informative, information used with Bayesian Kriging is only given by data. Thus, this case is close to Universal Kriging. However the average standard deviation AStD appears very smaller with UK than with BK. Thus, the interval $\left[\hat{Y}(x)-2\sigma_{\hat{Y}}(x);\hat{Y}(x)+2\sigma_{\hat{Y}}(x)\right]$ contains only 60% of the output with UK against 81% with BK. The uncertainty announced by UK is widely underestimated, especially because it does not take into account the uncertainty on correlation parameters.
One last remark must be added relating to the impact of the design. The good results obtained with strategy **"info 1 BK"** are not only due to good prior information but also to the impact of the design. For example (not presented here in details) it is very strange that the 20 runs design

gives worse results than the 17 runs (see Table 7) design with all kind of Kriging methods, whereas the former includes the latter. A deeper study is needed here in order to provide a better understanding of the phenomenon.
However, the results mentioned above are still the same, even when the source of variability coming from the design is removed.

TABLE 7 Comparison between Universal Kriging and Bayesian Kriging on the 17 runs design

	UK	no info	info 1	info 2
RMSE	175 595	186976	166770	200200
MAE	132 935	135422	125748	151905
AStD	136832	199505	188411	281317
Pr	0.22	0.07	0.08	0.02

4 Conclusion

The first aim of this paper is to show that in most cases, prediction variance of Universal Kriging can not be interpreted as the variance of the response conditionally to the observations. Indeed, it only takes into account the uncertainty induce by the estimation of trend parameters and not the one induce by the estimation of variance and correlation parameters. Therefore, it underestimates the resulting uncertainty on the response. This result has also been observed in the 3D case study.
The second aim of this paper is to propose different strategies to get informative prior information. Experimentations tend to show that Bayesian Kriging gives a good prediction of the response and its uncertainty. However, special care must be taken while using this kind of prior: as Bayesian Kriging with false prior information can give worse results than Bayesian Kriging with a non informative prior.

Bibliography

Goria, S., Armstrong, M., Galli, A. 2002. Using a Bayesian approach to incorporate new information when estimating resources, IAMG 2002, pp; 75-80.

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

Martin, J., Simpson, T. 2004. A Monte Carlo simulation of the kriging model. American Institute of Aeronautics and Astronautics (4483).

Oakley, J., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal of the Royal Statistical Society, (66), 751-769.

Robert, C. 1996. Méthodes de Monte Carlo par chaînes de Markov. Econimica.

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.

3DSL® http://www.streamsim.com