Computers & Geosciences 54 (2013) 99–106

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Speeding up Kriging through fast estimation of the hyperparameters in the frequency-domain

J.H.S. de Baar*, R.P. Dwight, H. Bijl

TU Delft, Department of Aerodynamics, Kluyverweg 2, 2629 HT Delft, The Netherlands

ARTICLE INFO

Article history:
Received 13 November 2012
Received in revised form
25 January 2013
Accepted 28 January 2013
Available online 12 February 2013

Keywords:
Kriging
Hyperparameter
Fast
NUFFT
FMLE
FSV

ABSTRACT

Kriging is a widely applied data assimilation technique. The computational cost of a conventional Kriging analysis of N data points is dominated by the m iterations of the maximum likelihood estimate (MLE) optimization, resulting in a computational cost of $\mathcal{O}(mN^3)$. We propose two fast methods for estimating the hyperparameters in the frequency domain: frequency-domain maximum likelihood estimate (FMLE) and frequency-domain sample variogram (FSV), both of which reduce the cost of the optimization to $\mathcal{O}(N^2+mN)$ in the case of a regular Fourier transform (FT), and to $\mathcal{O}(N\ln N+mN)$ in the case of a fast Fourier transform (FFT). In addition to this speed up, problems concerning positive definiteness of the gain matrix – which limit the robustness of the conventional approach – vanish in the proposed methods.

1. Introduction

Kriging is a powerful data assimilation technique, which was developed independently in Geology by Matheron (1963) and in Meteorology by Gandin (1965). An overview of the history and application of Kriging is provided by Cressie (1990).

As a data assimilation technique, Kriging has been applied in various fields of research. One of our present interests is the application of Kriging as a surrogate model for expensive computer models (Kennedy and O'Hagan, 2000a), where it is especially useful since it enables multi-dimensional non-uniform sampling, (*i.e.* non-uniform distribution of sample locations) and can include observation errors as well as gradients (Chung and Alonso, 2002; Dwight and Han, 2009; de Baar et al., submitted for publication-a). Another of our interests is the spatial interpolation of large data sets which are sampled non-uniformly, see for example Percin et al. (2012) and de Baar et al. (submitted for publication-b). For such data sets we can still perform a Kriging analysis, using a sparse covariance matrix, however, the evaluation of the conventional maximum likelihood estimate (MLE) of the correlation ranges (with either full or sparse covariance matrices) is so costly that it is impracticable. Therefore, for such applications the development of a cheaper method of estimating the hyperparameters is required.

Several attempts have been made to speed up the Kriging analysis for large data sets. Mardia and Marshall (1984) and Mardia (1989) present a frequency-domain Torus conditional autoregression (T-CAR) model for a periodic rectangular lattice, *i.e.* for uniformly sampled data under periodic boundary conditions. This approach reduces the cost of the complete Kriging analysis, although Mardia and Marshall do not explicitly state this cost reduction as an objective. A number of other techniques aim to reduce the overall cost by restating the complete Kriging analysis: among others Chiles and Delfiner (2012) present several methods of sparse Kriging (which we use in Percin et al., 2012; de Baar et al., submitted for publication-b and Section 3.3), Cressie and Johannesson (2008) present fixed rank Kriging, and Fritz et al. (2009) present an FFT-based technique that alleviates the conditions of T-CAR. In their work Mardia and Marshall (1984) cite Whittle (1954), who already gives a frequency-domain maximum likelihood estimate (FMLE) approximation, however, with strict conditions on periodicity and uniform sampling of the data. A recent attempt to alleviate these conditions has been made in the closely related work of Fuentes (2007), who generalizes the FMLE to lattices with missing data (which involves a weight function to handle the missing data) and to non-uniformly sampled data (which involves regridding the domain in rectangular blocks). In de Baar et al. (2011) we improve FMLE by simulating the effect of aliasing and spectral leakage, while for non-uniform sampling we improve the quality of the spectrum using a low pass filter, however, we conclude that the choice of the cut-off frequency remains somewhat arbitrary.

* Corresponding author. Tel.: +31 152782596; fax: +31 152787077.
E-mail address: j.h.s.debaar@tudelft.nl (J.H.S. de Baar).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.01.016

In the present work, we aim to avoid the complications which arise with FMLE for non-uniformly sampled data by introducing the frequency-domain sample variogram (FSV). The implementation of FSV is straightforward, it relies on estimating the hyperparameters by fitting the Fourier transform of the generator function of the covariance matrix to the power spectrum of the data, where we obtain the power spectrum from a non-uniform fast Fourier transform or NUFFT (Greengard and Lee, 2004).

1.1. Bayesian derivation of Kriging

From a Bayesian perspective, Kriging predicts a set of M values $\boldsymbol{x}$, conditional on N data $\boldsymbol{y}$, where the data $\boldsymbol{y}$ are a subset of $\boldsymbol{x}$ selected by the observation matrix H. We assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalized such that $\boldsymbol{y}$ has zero mean and unit variance. The values $\boldsymbol{x}$ are located at the spatial locations $\boldsymbol{\xi}$, which are normalized such that $\boldsymbol{\xi}_{\min}=-1$ and $\boldsymbol{\xi}_{\max}=1$. We will make our derivation for one-dimensional $\boldsymbol{\xi}$, and generalize it to the multi-dimensional case in Section 2.4.

Consider the covariance matrix P_θ of the prior $p(\boldsymbol{x})$, where in the following the elements of P_θ are generated by a Gaussian function:

$$p_{ij}(\theta)=\exp\left(-\frac{h_{ij}^2}{2\theta^2}\right), \tag{1}$$

with lag $h_{ij}=\xi_j-\xi_i$ and correlation range θ (which acts as a hyperparameter). Next, consider the covariance matrix R of the likelihood $p(\boldsymbol{y})$

$$R_\epsilon=\epsilon^2 I, \tag{2}$$

for a uniform and uncorrelated observation error ϵ. From Bayes' theorem

$$p(\boldsymbol{x}|\boldsymbol{y})=\frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}, \tag{3}$$

the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is Gaussian and is defined by the Kriging predictor and variance (Wikle and Berliner, 2007)

$$\begin{aligned} E(\boldsymbol{x}|\boldsymbol{y})&=P_\theta H^T(R_\epsilon+HP_\theta H^T)^{-1}\boldsymbol{y},\\ \mathrm{var}(\boldsymbol{x}|\boldsymbol{y})&=c_{\mathrm{var}}[I-P_\theta H^T(R_\epsilon+HP_\theta H^T)^{-1}]P_\theta. \end{aligned} \tag{4}$$

The prefactor (Kitanidis and Lane, 1985; Kitanidis, 1986)

$$c_{\mathrm{var}}\coloneqq\frac{N}{N-n_{\mathrm{hyp}}-2} \tag{5}$$

of the variance shows that the posterior is broader if the prior depends on estimating n_{hyp} hyperparameters from the data. For sufficiently large N this prefactor can be neglected, therefore it is often not mentioned. Although formally the predictor and variance in (4) define a Gaussian process, we will presently focus only on the prediction of the mean, which is often used for interpolation.

1.2. Conventional estimate of the hyperparameters

The hyperparameter θ is generally unknown. However, it can be estimated from the data. There are two cardinal methods of estimating θ: the maximum likelihood estimate (MLE) and the sample variogram (SV). As the MLE is more versatile, we will consider this to be the reference method for comparison.

In the MLE, θ is defined as the maximum of the likelihood, which is equivalent to the minimum of Mardia and Marshall (1984), Kitanidis and Lane (1985), Kitanidis (1986) and Mardia (1989)

$$L_{\mathrm{c}}(\theta)\coloneqq\ln|A_{\theta\epsilon}|+\boldsymbol{y}^T A_{\theta\epsilon}^{-1}\boldsymbol{y}, \tag{6}$$

where $A_{\theta\epsilon}$ is a short hand notation for

$$A_{\theta\epsilon}\coloneqq R_\epsilon+HP_\theta H^T, \tag{7}$$

such that, using (1) and (2), the elements of $A_{\theta\epsilon}$ are given by the generator

$$\alpha(h)=\epsilon^2\delta_h+\exp\left(-\frac{h_{ij}^2}{2\theta^2}\right). \tag{8}$$

For ease of notation, we will suppress the explicit dependencies of R, P, and A on θ and ϵ in the following. Note that instead of (6) we might also use a MAP estimator (Kitanidis and Lane, 1985; Kitanidis, 1986).

The sample variogram (SV) approach can be more efficient for large data sets. Details on the SV can be found in Webster and Oliver (2007), Eqs. (4.41) and (4.43).

1.3. Computational cost of the estimate

The computational cost of a conventional Kriging prediction can be segmented into three consecutive steps. The first step is minimizing $L(\boldsymbol{\theta})$, as given in (6), using m optimization steps. The computational cost of the minimization is $\mathcal{O}(mN^3)$, since each optimization step requires solving a linear system and an eigenvalue problem. The second step is finding the vector $\boldsymbol{y}_0$ of Kriging dual weights, by solving the linear system $\boldsymbol{y}_0=A^{-1}\boldsymbol{y}$ at cost $\mathcal{O}(N^3)$. The third step is to multiply $PH^T\boldsymbol{y}_0$ in order to provide n values, at a computational cost of $\mathcal{O}(nN)$. An overview of the expected computational cost is provided in Table 1. Clearly, for large N the MLE of the hyperparameters constitutes the bottleneck of the computation. Note that the cost of an SV estimate is lower, however, the implementation of this approach is not straightforward for multi-dimensional, non-uniform sampling, and involves the introduction of new parameters.

In Section 2 we propose to target the bottleneck of expensive tuning of the hyperparameters by estimating these hyperparameters in the frequency domain, at a computational cost of $\mathcal{O}(N\ln N+mN)$. We propose two different methods, frequency-domain maximum likelihood estimate (FMLE, de Baar et al., 2011) and frequency-domain sample variogram (FSV).

1.4. Robustness of the estimate

Another issue with the MLE is that, for a combination of a small observation error ϵ and long correlation range θ, it is impossible to solve the linear system $\boldsymbol{y}_0=A^{-1}\boldsymbol{y}$ accurately, as A is not longer numerically positive definite (de Baar et al., submitted for publication-a). Another advantage of both proposed

Table 1
Computational cost of the Kriging predictor.

Subroutine		Conventional MLE	Conventional SV	Proposed FMLE / FSV
Hyperparameters	$\theta_0=\arg\min L(\theta)$	$\boldsymbol{\mathcal{O}(mN^3)}$	$\boldsymbol{\mathcal{O}(N^2+mN)}$	$\boldsymbol{\mathcal{O}(N\ln N+mN)}$
Solve system	$y_0=A^{-1}y$	$\mathcal{O}(N^3)$	$\mathcal{O}(N^3)$	$\mathcal{O}(N^3)$
Multiplication	$\overline{x}=PH^Ty_0$	$\mathcal{O}(MN)$	$\mathcal{O}(MN)$	$\mathcal{O}(MN)$

methods – FMLE and FSV – is that we do not deal with such matrix operations, and expect this problem to vanish.

2. Two fast methods for estimating the hyperparameters

Let us have a closer look at the computational procedure of estimating the hyperparameters. In the present work, we use a Nelder-Mead minimization routine, although gradients of the likelihood might readily be included (Kitanidis and Lane, 1985). In the conventional approach, the estimation procedure is illustrated by the flow chart in Fig. 1(a). The optimization loop includes the expensive evaluation of L. It seems computationally demanding to first construct matrix A from a known generating function, and then find its eigenvalues and solve a linear system without using any of our knowledge of its generator.

We propose to reduce the computational cost of the Kriging prediction by estimating the hyperparamaters θ in the frequency domain, considering the Fourier transforms of the data $\boldsymbol{y}$ and of the matrix generator a. This changes the estimation procedure to that of the existing frequency-domain maximum likelihood estimate (FMLE) for uniformly sampled data or that of the new frequency-domain sample variogram (FSV) for the more general case of non-uniformly sampled data. Both methods are illustrated in the flow charts in Fig. 1(b,c). These routines reduce the computational cost significantly, since the most expensive operation is now computing the power spectrum using non-uniform fast Fourier transform NUFFT at cost $\mathcal{O}(N \ln N)$, instead of the Cholesky decomposition at cost $\mathcal{O}(N^3)$, while at the same time it is now placed *outside* the optimization loop. For large $N > m$ the minimization is now virtually independent of m (which is typically at least $m \approx 100$). During the minimization, the problem of robustness vanishes, as numerical positive definiteness of A during matrix operations is not an issue.

In the following sections we will discuss both methods in more detail. As an illustration, we will consider the test-function

$$y(\xi) = \exp\left(-\frac{(\xi+c_0)^2}{2c_1^2}\right) - \exp\left(-\frac{(\xi-c_0)^2}{2c_1^2}\right),$$
$$c_0 = 0.5,$$
$$c_1 = 0.1, \tag{9}$$

which we sample at $N=64$ uniformly distributed locations, with an observation error $\epsilon = 0.001$. An example of the acquired data and Kriging predictor is shown in Fig. 2.

Both methods use the continuous equivalent of the matrix generator given in (8)

$$a(h) = \epsilon_c^2 \delta(h) + \exp\left(-\frac{h_{ij}^2}{2\theta^2}\right), \tag{10}$$

which has the Fourier transform

$$\hat{a}(k) = \epsilon_f^2 + \theta N \sqrt{\frac{\pi}{2}} \exp\left(-\frac{\pi^2 k^2 \theta^2}{2}\right). \tag{11}$$

The error ϵ_f represents the sum of the noise in the data and the noise that might be introduced by the (non-uniform) Fourier transform. This term ϵ_f will be found by optimization in the frequency domain, simultaneously with θ. In the present derivation we consider a single hyperparameter θ. However, both FMLE and FSV are readily generalized to multiple hyperparameters $\boldsymbol{\theta}$, as will be illustrated in Section 3. For FMLE, instead of optimizing ϵ_f, one might replace ϵ_f with an exact noise term and an aliasing term, as is treated in de Baar et al. (2011).

At this point it should be noted, that although we use a Gaussian covariance function, one can in fact use any parameterized matrix generator $a(h)$, provided that one can find a Fourier transform $\hat{a}(k)$, either analytically or numerically.

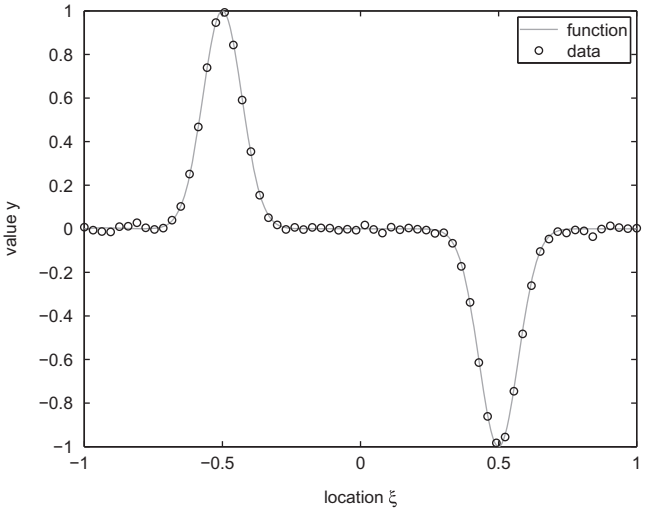

Fig. 2. Data, sampled uniformly from the test-function.

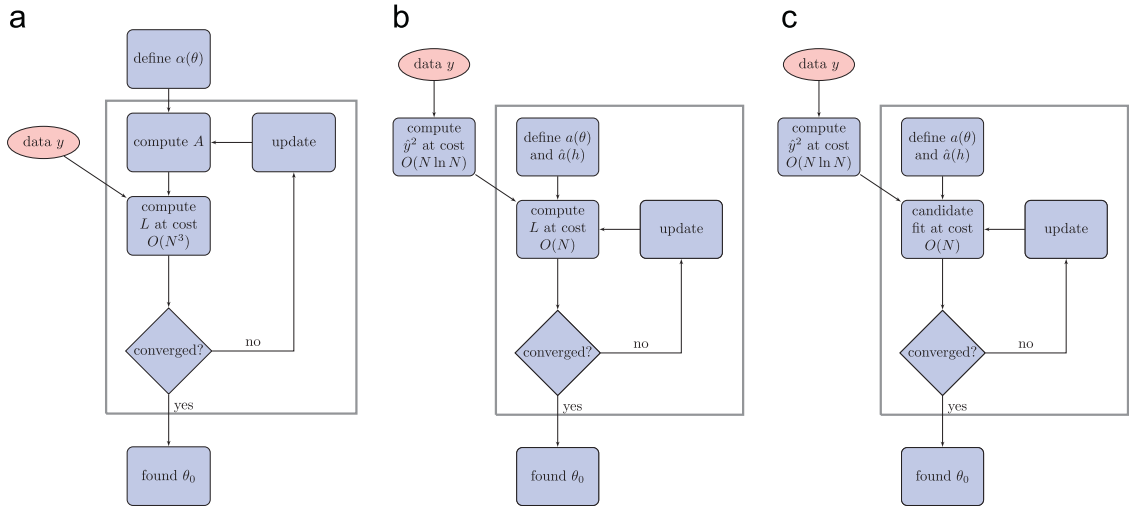

Fig. 1. The conventional MLE algorithm: (a) places an expensive evaluation of L in the optimization loop, indicated by the gray box, the proposed algorithms FMLE (b) and FSV (c) reduce the cost of the optimization loop.

We will apply the non-uniform fast Fourier transform (NUFFT, Greengard and Lee, 2004) and the following properties of the Fourier transform: Parseval's theorem, the eigenvalue theorem, the convolution theorem, and the Wiener–Khinchin–Einstein theorem. A detailed description of these properties is given in Appendix A.

2.1. Method 1: frequency-domain maximum likelihood estimate (FMLE)

It would be interesting to compare the simple Kriging weights and/or estimates obtained in the spatial domain and those obtained with the spectral version (under the periodicity assumption). The periodic assumption is neither obvious or natural for geostatisticians used to work in the spatial domain.

Recall that we would like to minimize (6)

$$L(\theta) = \ln|A| + \boldsymbol{y}^T A^{-1}\boldsymbol{y},$$

which we can restate, under the assumption of circularity, (*i.e.* the assumption of a circular or periodic domain)

$$L(\theta) = \ln|A| + \boldsymbol{y}_{\text{circ}}^T \boldsymbol{b} * \boldsymbol{y}_{\text{circ}}, \tag{12}$$

where $\boldsymbol{b}$ is the function that generates $B = A^{-1}$. Although the convolution $a(h) * b(h) \equiv \delta(h)$ does not define $\boldsymbol{b}$ properly in the spatial domain, we apply Parseval's theorem to the last term to find

$$L_f(\theta) = \ln|A| + \frac{1}{N}\hat{\boldsymbol{y}}^T \widehat{\boldsymbol{b} * \boldsymbol{y}}_{\text{circ}}, \tag{13}$$

where the hat indicates a Fourier transform. We then apply the convolution theorem to the last term to find

$$L_f(\theta,\epsilon_f) = \ln|A| + \frac{1}{N}\sum_n \hat{b}_n \hat{y}_n^2, \tag{14}$$

where the assumed periodicity of $\boldsymbol{y}$ is now implicitly contained in the spectrum $\hat{\boldsymbol{y}}$. Due to the convolution theorem we simply have $\hat{b}_n \hat{a}_n \equiv 1$, such that $\hat{\boldsymbol{b}}$ is properly defined in the frequency domain as long as $\epsilon_f^2 > 0$. Finally, since A is positive definite we can rewrite the first term

$$L_f(\theta,\epsilon_f) = \sum_n \ln \lambda_n(A) + \frac{1'}{N}\sum_n \frac{\hat{y}_n^2}{\hat{a}_n}, \tag{15}$$

where $\hat{a}_n = \hat{a}(k_n)$, with $k_n = n/L$ for a certain domain size L. Note that since L has dimension (m), the frequency k has dimension (m^{-1}), while the DFT-frequency n is dimensionless.

Although we know the spectrum of the eigenvalues $\lambda_n(A)$ from the eigenvalue theorem, it is not clear how we should sample this spectrum to find the actual eigenvalues, especially in the case of uniform sampling (Whittle, 1954). In the following section, we will use the FSV estimate of the hyperparameters, which does not suffer from this problem.

The central step in FMLE is to approximate under the assumption of periodicity – without inverting the covariance matrix A – the vector of dual weights $\boldsymbol{y}_0 \approx \boldsymbol{b} * \boldsymbol{y}_{\text{circ}}$. The latter can be evaluated in the spectral domain. Apart from using this approximation to estimate the hyperparameters, in some cases it is attractive to use the inverse Fourier transform of the spectral approximation of the dual weights in the Kriging predictor (4) which results in methods like T-CAR, although the assumptions of periodicity and regular sampling are quite strict and exclude a large number of practical applications (Mardia and Marshall, 1984; Fritz et al., 2009).

2.2. Method 2: frequency-domain sample variogram (FSV)

The FMLE is derived directly from the likelihood (6). For non-uniformly sampled data, we propose a different approach: from the Wiener–Khinchin–Einstein theorem, we fit the parameterized matrix generator to the power spectrum of the data. This is equivalent to minimizing

$$G_{\text{L2}}(\theta,\epsilon_f) = \sum_n \left|\ln(\hat{a}_n^{-1}\hat{y}_n^2)\right|^2 = \sum_n \left|\ln \hat{y}_n^2 - \ln \hat{a}_n\right|^2. \tag{16}$$

There are two alternative interpretations of this method. Firstly, the method is similar to the Fourier transform of the conventional sample variogram. Secondly, the Gaussian process $\boldsymbol{y} = A\boldsymbol{y}_0$ can be interpreted as the result of diffusion acting on the initial state $\boldsymbol{y}_0$ of the Gaussian process.

2.3. Performance

First of all, we would like the Kriging prediction to be accurate. For increasing N we would expect that the Kriging prediction converges to the true function. Therefore we define the prediction rms error e_N to track the convergence

$$e_N^2 := \frac{1}{M}\sum_{n=1}^{M} \left|E(x_n|\boldsymbol{y}) - f(\xi_n)\right|^2, \tag{17}$$

which we can estimate as (see Appendix B)

$$e_N^2 \approx 4\sigma^2\left[1 - \text{erf}\left(\frac{N\pi}{2}\frac{\theta}{\sqrt{2}}\right)\right] + \frac{N_c}{N}\epsilon^2, \tag{18}$$

where in Appendix B we define N_c such that for $e_N^2 < \epsilon^2$ the above estimate shows Monte Carlo convergence. Fig. 3 displays the prediction error of the conventional and the proposed methods applied to the test-function (9), together with the above estimate (18). Both proposed methods are as accurate as the conventional approach.

Given that these methods are accurate, our main interest is the CPU time required by the different methods for the estimation of the hyperparameters. Fig. 4 shows a significant reduction of the CPU time for both proposed methods. For $N = 300$ data, the CPU time for FMLE and FSV is roughly 30 times lower than for conventional MLE. Note that for large N, the CPU time scales with mN^3 for the conventional MLE and with $N^2 + mN$ for SV, while it only scales with $N \ln N + mN$ for the two proposed methods.

Finally, Fig. 5 illustrates how a small observation error ϵ and a large correlation range θ can influence the robustness of the minimization. The gray lines show the goal function for the conventional approach, while the black lines are for FMLE (FSV gives similar results). The dotted lines have $\epsilon = 10^{-2}$, the dashed

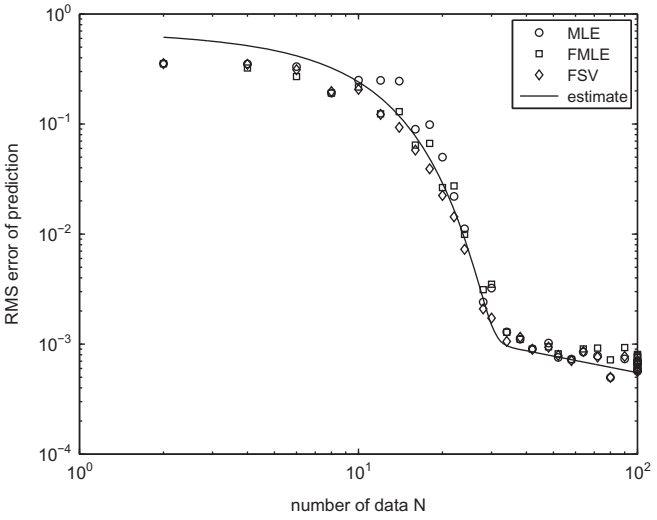

Fig. 3. Convergence of the rms error of the Kriging prediction (4) using either MLE, FMLE, or FSV. Estimate from (18).

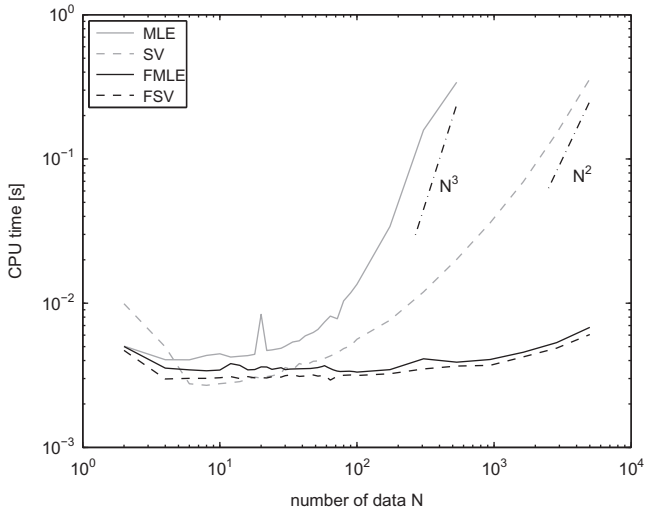

Fig. 4. Both methods reduce the CPU time of estimating the hyperparameters

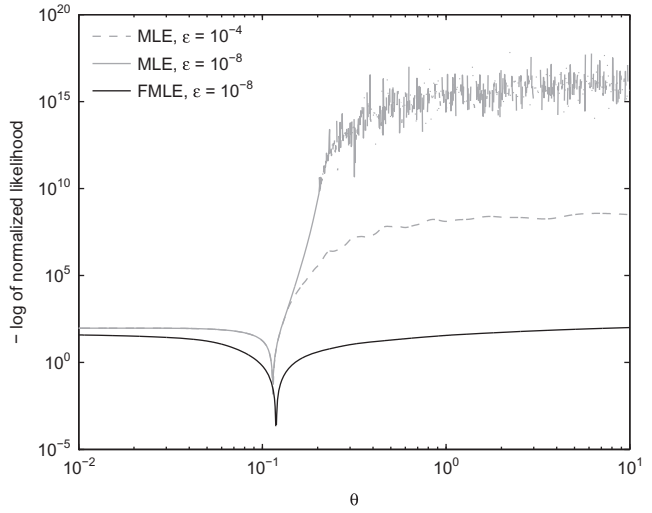

Fig. 5. Robustness of the minimization: goal functions for MLE show oscillatory behavior, depending on observation error. The goal function for FMLE is smooth, arrives at the same minimum, and coincides for different observation errors.

lines have $\epsilon = 10^{-4}$, and the continuous lines have $\epsilon = 10^{-8}$. As we see, when we decrease the observation error, evaluation of the conventional goal function is problematic for larger correlation ranges. This is due to the fact that A is not numerically positive definite, which leads to inaccurate results and a ragged goal function; this makes optimization very difficult, especially when it is gradient-based (de Baar et al., submitted for publication-a). As expected, the proposed methods FMLE and FSV do not suffer from this problem, as it does not involve the matrix operation $\boldsymbol{y}_0 = A^{-1}\boldsymbol{y}$, resulting in a smooth goal function.

2.4. Extension to multiple dimensions

The following equation extends (11) to multiple dimensions:

$$\hat{a}(\boldsymbol{k}) = \epsilon_f^2 + a_0 \exp\left(-\frac{\pi^2(\boldsymbol{k}^t\boldsymbol{\theta})^2}{2}\right). \tag{19}$$

In FSV, the pre-factor a_0 is now chosen such that the multi-dimensional power spectrum and fit (19) contain an equal amount of power, in order to satisfy Parseval's theorem (23).

3. Examples of fast hyperparameter estimation for large data sets

In the previous section we have derived FMLE and FSV and have illustrated the accuracy, efficiency, and robustness of these methods for a uniformly-sampled one-dimensional test-function. We have found that FSV is more general than FMLE, as it can be applied to data obtained by sampling at random locations. We apply FMLE or FSV in the following three examples. In Section 3.1 we illustrate the efficiency of FMLE for a vertical mixing layer with dual fidelity oxygen data, acquired from titration and sensor readings. In Section 3.2 we illustrate the flexibility of FSV for randomly sampled terrain elevation data. Finally, in Section 3.3 we illustrate the efficiency of FSV for dual fidelity terrain elevation data, acquired from satellite radar data and F-16 imagery. When given, CPU timings are from a 3.4 GHz Intel Pentium 4 processor.

3.1. Dual fidelity oxygen data from the northern Atlantic Ocean: titration data augmented with sensor readings

We consider the high fidelity bottle titration data ($N=12$) and the uncorrected low fidelity sensor data ($N=2815$) of an oxygen measurement at one of the stations of a Northern Atlantic transect, shown in Fig. 6, made by the Royal Netherlands Institute for Sea Research vessel *Pelagia*. A detailed description of the experiments is given in van Aken (2011). Normally the sensor data would be corrected for pressure, temperature, and station number by means of linear regression. Here, we augment the bottle data with the uncorrected sensor data using the autoregressive model presented by Kennedy and O'Hagan (2000), which can be considered to be a form of co-Kriging under a Markov property.

In order to use Kriging, we have to estimate the following hyperparameters: the correlation length θ_{sensor} of the sensor data, the correlation length θ_{bottle} of the bottle data, and the regression parameter ρ. Kennedy and O'Hagan (2000) show that we can estimate the hyperparameters $(\theta_{\text{bottle}},\rho)$ independently of θ_{sensor}. Since the large number of sensor data is sampled uniformly, we use FFT-based FMLE to estimate θ_{sensor}, while for the small number of bottle data we use conventional MLE to estimate θ_{bottle} and ρ.

The uncorrected data as well as the resulting augmented data are shown in Fig. 7(a), which is a good interpolation result that combines the accuracy of the bottle data with the detail of the sensor data. Fig. 7(b) illustrates the reduction of CPU time for an increasing number of sensor data points, as a result of using FMLE instead of MLE. Clearly, the cost of MLE increases with $\mathcal{O}(N^3)$, while the cost of FMLE increases much slower. Both MLE and FMLE use a Nelder-Mead minimization and for an increasing number of samples they converge to roughly the same value ($\theta = 6.7$ for MLE and $\theta = 5.9$ for FMLE), however, for the full data set the conventional MLE minimization routine does not reach a minimum due to erratic behavior of the goal function (the routine used to compute the eigenvalues of the correlation matrix – already a more robust alternative to computing the determinant – does not converge).

3.2. Terrain elevation data of a flood barrier, obtained from photogrammetric matching of UAV acquired images

In the present section we consider $M=75{,}000$ randomly sampled terrain elevation data, which have been acquired by photogrammetric matching of images from an unmanned aerial vehicle (UAV).

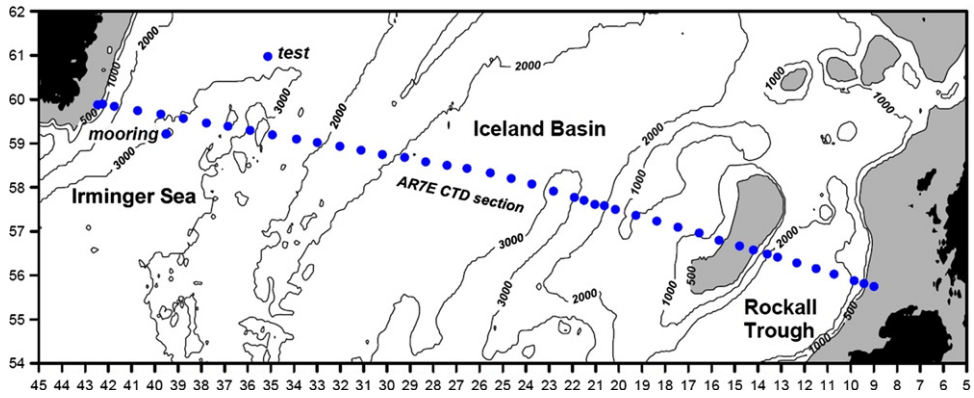

Fig. 6. Map of the Northern Atlantic, displaying the stations where oxygen measurements were made. Figure from van Aken (2011).

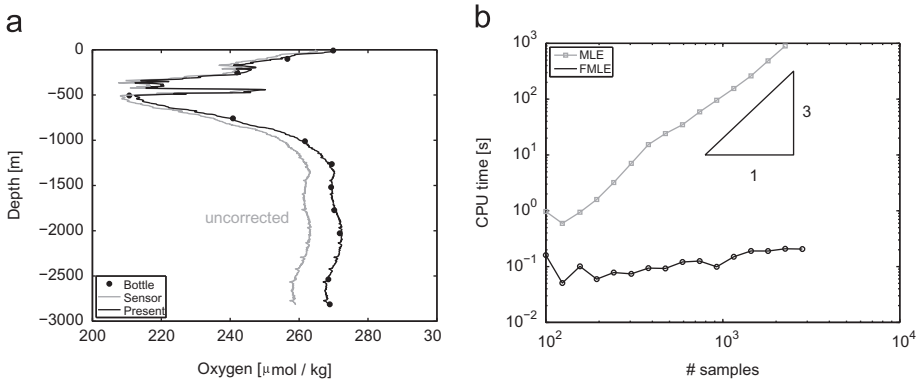

Fig. 7. Dual fidelity oxygen data: (a) high-fidelity data from bottle titration, augmented with low-fidelity sensor data and (b) reduced CPU time when using FSV.

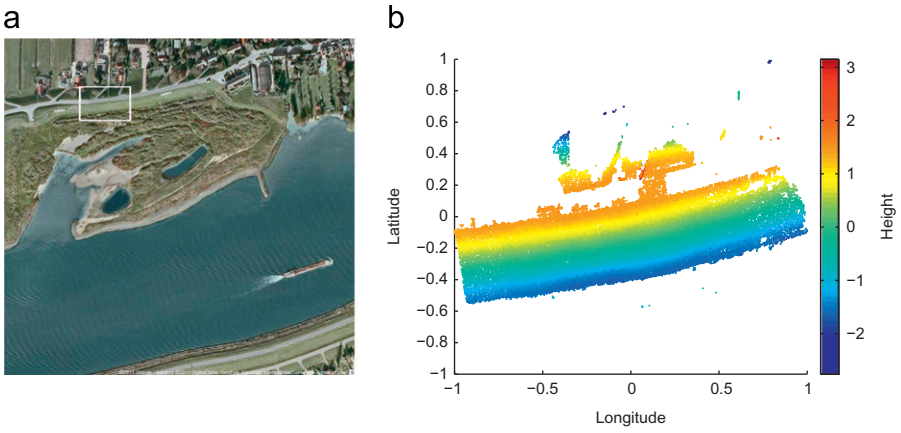

Fig. 8. (a) Arial photograph of a flood barrier, with the area of interest indicated by the white box and (b) the normalized UAV-acquired terrain height data. Image (a) taken from *Google maps*.

The full set of normalized terrain elevation data is shown in Fig. 8. Our task is to find the two correlation ranges. From the complete set of data, we use a NUFFT to find the power spectrum shown in Fig. 9(a). Note that both the flood barrier in Fig. 8(b) and the spectrum in Fig. 9(a) are rotated. We would like to introduce this angle of rotation as a third hyperparameter. A nice feature of FSV is that the rotation angle of the spectrum is equal to that of the data, such that we have to obtain the power spectrum only once, and can continue to fit each of the three hyperparameters. In this case, the FSV spectrum fitting routine is so cheap that we use a brute force minimization, while estimating ϵ_f from the average power at the edge of the spectrum. This leads to the fit of the spectrum shown in Fig. 9(b), from which we find both correlation ranges as well as the rotation angle. For such an increase of the number of hyperparameters the cost of the optimization would make conventional MLE impractical, for FSV the added cost is much smaller due to the reduced cost of the optimization loop.

3.3. *Dual fidelity terrain elevation data: satellite data augmented with F-16 acquired terrain elevation data from photogrammetric matching*

The Royal Netherlands Air Force uses a RecceLite pod mounted on a F-16 jet to acquire high resolution terrain images. In the present application, we consider $N_2 = 63$ SRTM satellite terrain

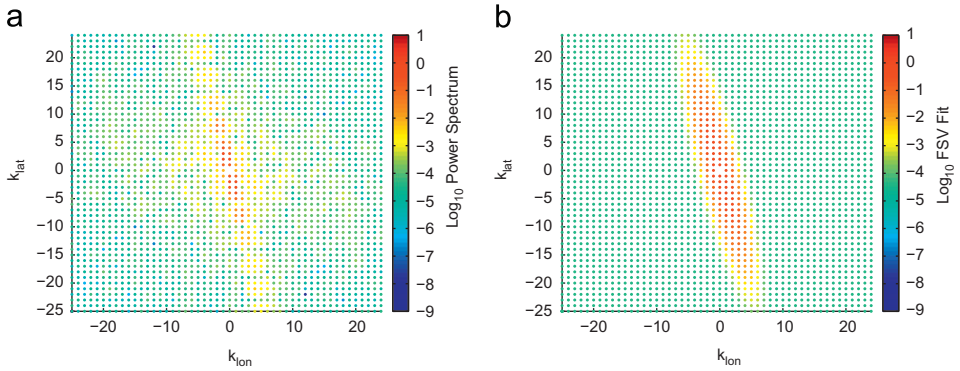

Fig. 9. Fitting the matrix generator (b) to the power spectrum (a).

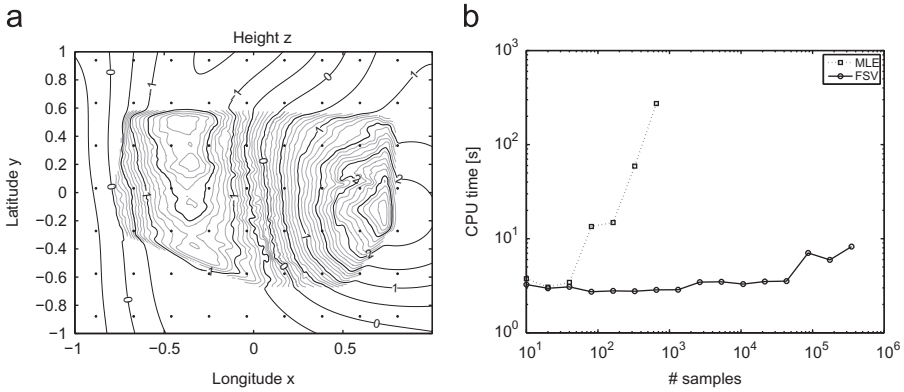

Fig. 10. (a) Normalized contour of the terrain obtained from satellite data, augmented with F-16 acquired terrain elevation data in the central area and (b) reduced CPU time when using FSV.

elevation data, augmented with $N_1 = 35{,}000$ terrain elevations found from photogrammetric matching of overlapping RecceLite images. We use sparse co-Kriging with a screening range of $\theta_{\text{screen}} = 3\theta_{\text{RecceLite}}$ to augment the dual fidelity data (Kennedy and O'Hagan, 2000, see also Section 3.1). We estimate the correlation lengths of the RecceLite data using NUFFT-based FSV. Since the FSV spectrum fitting routine is so cheap, we use a brute force minimization, while ϵ_f is estimated from the edge of the spectrum.

Fig. 10(a) shows a normalized height contour map. The dots indicate the locations where SRTM data is available, while the central area with gray contour lines is the area which has been augmented with the F-16 acquired stereographical images. This approach enables us to combine the accuracy of the SRTM data with the terrain details revealed by the photogrammetric matching. Finding the required hyperparameters using MLE would require an impractical amount of CPU time, making this approach completely infeasible. Fig. 10(b) illustrates the reduction of CPU time when we use FSV. We see that the time required for a conventional MLE increases rapidly with the number of samples and we can only give a rough extrapolation of some 10^6–10^7 s that might be required for the complete dataset which clearly makes this approach impractical, while the time required for the FSV increases only slightly to reach roughly 10 s for the complete dataset.

4. Conclusions

We have applied two methods for fast estimation of the Kriging hyperparameters: the existing FMLE for uniformly sampled data and the new FSV for non-uniformly sampled data. For large N, numerical results show a significant reduction in CPU time, from $\mathcal{O}(N^3)$ to $\mathcal{O}(N \ln N)$, while the accuracy of the Kriging prediction is preserved. During the conventional minimization routine, complications arise as numerical round-off errors influence the positive definiteness of the matrix A, which often leads to a ragged or even incomputable goal function, especially for a larger number of data points N. In the proposed methods, positive definiteness of A ceases to be an issue during the minimization, which results in a smooth goal function.

Acknowledgments

The oxygen data for Section 3.1 was kindly provided by H.M. van Aken of the *Royal NIOZ*. The terrain elevation data for Section 3.2 was kindly provided by P. Wijkstra of *Heering UAS* and G. Vestjens of *Geodelta*. The SRTM data and the RecceLite images for Section 3.3 were kindly provided by P. Hoogeboom of the *NLR* and V.E. Voorneman of the *Royal Netherlands Air Force*. We are very grateful for the financial support by *Technologiestichting STW*, Project number 10113.

Appendix A. The discrete Fourier transform

The *discrete Fourier transform* (DFT) provides the coefficients required to regard a signal as a harmonic superposition. To find these Fourier coefficients $\hat{y}_n$, we apply the following convention of the DFT:

$$\hat{y}_n = \sum_i y_i \exp\left[-\frac{2\pi n i}{N}\right]. \tag{20}$$

The cost of a DFT is $\mathcal{O}(N^2)$, however, the spectrum can be computed more efficiently through a fast Fourier transform (FFT) at cost $\mathcal{O}(N \ln N)$.

The *non-uniform discrete Fourier transform* (NDFT) is given by

$$\hat{y}_n = \sum_i y_i \exp[-\pi n i(\xi_i + 1)], \tag{21}$$

although it is formally not a Fourier transform as it might have non-orthogonal basis functions. This can result in additional noise in the power spectrum of a non-uniformly sampled signal. The cost of a NDFT is $\mathcal{O}(N^2)$, however, the spectrum can be computed more efficiently through a non-uniform fast Fourier transform (NUFFT) at cost $\mathcal{O}(N \ln N)$, see Greengard and Lee (2004).

The following properties are properties of the DFT (strictly they are not necessarily properties of the NDFT):

Parseval's theorem states that the power contained in a signal $\boldsymbol{y}$ is the same as the power contained in the spectrum $\hat{\boldsymbol{y}}$ (Oppenheim et al., 1989)

$$\boldsymbol{y}^2 = \frac{1}{N}\hat{\boldsymbol{y}}^2, \tag{22}$$

where the factor N appears due to the normalization of the DFT in the present convention. This expression can be generalized to the case of two signals $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, which results in

$$\boldsymbol{y}_1^T \boldsymbol{y}_1 = \frac{1}{N}\hat{\boldsymbol{y}}_1^T \hat{\boldsymbol{y}}_2. \tag{23}$$

The *eigenvalue spectrum* contains the eigenvalues of a matrix A. Since A is generated by a correlation function, the spatial eigenvalues can be estimated from the Fourier transform of the correlation function (Whittle, 1954).

The *convolution theorem* is a very powerful property of the Fourier transform. However, we will have to make one assumption: that it is meaningful to think of the observations $\boldsymbol{y}$ as being periodic, such that when we would move beyond the limit of our domain we would find a repeating set of observations. With this assumption, we can apply the circular convolution theorem

$$\hat{\boldsymbol{b}}\hat{\boldsymbol{y}} = \widehat{\boldsymbol{b} * \boldsymbol{y}}_{\text{circ}}. \tag{24}$$

Finally, the *Wiener–Khinchin–Einstein theorem* states that the correlation function is the inverse Fourier transform of the power spectrum (Barkat, 2005). This is equivalent to stating that the Fourier transform of the correlation function is the power spectrum.

Appendix B. Estimated prediction error

The prior expresses that we expect a Gaussian power spectrum, given by Eq. (11). At first, we expect the prediction error to be related to the number of resolved modes ($N/2$) in our spectrum. Since the unresolved modes are orthogonal, we can expect e_N^2 to be the sum of the unresolved modes. We can approximate this sum with the integral:

$$e_N^2 \approx 4\sigma^2 \left[1 - \text{erf}\left(\frac{N\pi}{2}\frac{\theta}{\sqrt{2}}\right)\right]. \tag{25}$$

We define N_c as the number of modes, for which the above estimate reaches ϵ^2, from which point we would expect the errors in the individual modes to decrease in a Monte Carlo like fashion, such that the overall prediction error

$$e_N^2 \approx 4\sigma^2 \left[1 - \text{erf}\left(\frac{N\pi}{2}\frac{\theta}{\sqrt{2}}\right)\right] + \frac{N_c}{N}\epsilon^2. \tag{26}$$

This result is also obtained by considering Fourier transform of the Kriging variance.

References

van Aken, H., 2011. RV Pelagia Shipboard Report: Cruise 64PE342, Project THOR. Royal NIOZ.

de Baar, J., Dwight, R., Bijl, H., 2011. Fast maximum likelihood estimate of the Kriging correlation range in the frequency domain. In: IAMG Conference Salzburg, 2011.

de Baar, J., Dwight, R., Bijl, H. Improvements to Gradient-Enhanced Kriging using a Bayesian perspective, submitted for publication-a.

de Baar, J., Percin, M., Dwight, R., van Oudheusden, B., Bijl, H. Kriging regression of PIV data using a local error estimate, submitted for publication-b.

Barkat, M., 2005. Signal Detection and Estimation, second ed. Artech House, Boston, MA.

Chiles, J., Delfiner, P., 2012. Geostatistics, Modeling Spatial Uncertainty, second ed. Wiley.

Chung, H.S., Alonso, J.J., 2002. Using Gradients to Construct CoKriging Approximation Models for High-Dimensional Design Optimization Problems. AIAA 40th Aerospace Sciences Meeting and Exhibit.

Cressie, N., 1990. The origins of Kriging. Mathematical Geology 22, 239–252.

Cressie, N., Johannesson, G., 2008. Fixed rank Kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B 70, 209–226.

Dwight, R.P., Han, Z.H., 2009. Efficient uncertainty quantification using gradient-enhanced Kriging. In: 11th AIAA Non-Deterministic Approaches Conference.

Fritz, J., Neuweiler, I., Nowak, W., 2009. Application of fft-based algorithms for large-scale universal Kriging problems. Mathematical Geosciences 41, 509–533.

Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102, 321–331.

Gandin, L., 1965. Objective analysis of meteorological fields: gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad. Translated by Israel Program for Scientific Translations, Jerusalem.

Greengard, L., Lee, J.Y., 2004. Accelerating the nonuniform fast Fourier transform. SIAM Review 46, 443–454.

Kennedy, M.C., O'Hagan, A., 2000a. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B 63, 425–464.

Kennedy, M.C., O'Hagan, A., 2000. Predicting the output from a complex computer code when fast approximations are available. Biometrika 87, 1–13.

Kitanidis, P.K., 1986. Parameter uncertainty in estimation of spatial functions: Bayesian analysis. Water Resources Research 22, 499–507.

Kitanidis, P.K., Lane, R.W., 1985. Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss–Newton method. Journal of Hydrology 79, 53–71.

Mardia, K.V., 1989. Maximum likelihood estimation for spatial models. In: Proceedings of the Symposium on Spatial Statistics: Past, Present, and Future, pp. 203–253.

Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71, 135–146.

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

Oppenheim, A., Schafer, R., Buck, J., 1989. Discrete-time Signal Processing. Prentice Hall.

Percin, M., Eisma, H.E., de Baar, J.H.S., van Oudheusden, B.W., Remes, B., Ruijsink, R., de Wagter, C., 2012. Wake reconstruction of flapping-wing MAV 'DelFly II' in forward flight. International Micro Air Vehicle Conference and Flight Competition, July 2012, Braunschweig Germany.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, second ed. Wiley.

Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–449.

Wikle, C.K., Berliner, L.M., 2007. A Bayesian tutorial for data assimilation. Physica D: Nonlinear Phenomena 230, 1–16.