Fast maximum likelihood estimate of the Kriging correlation range in the frequency domain

Jouke H.S. DE BAAR[1], Richard P. DWIGHT[1], and Hester BIJL[1]

[1] Aerospace Engineering, TU Delft, The Netherlands,
j.h.s.debaar@tudelft.nl

Peer-reviewed IAMG 2011 publication

doi:10.5242/iamg.2011.0268

Abstract

We apply Ordinary Kriging to predict 75,000 terrain survey data from a randomly sampled subset of < 2500 observations. Since such a Kriging prediction requires a considerable amount of CPU time, we aim to reduce its computational cost. In a conventional approach, the cost of the Kriging analysis would be dominated by the optimization routine required to find the maximum likelihood, which provides an estimate of the correlation ranges. We propose to transform the optimization problem to the frequency domain, such that the cost of the optimization is now dominated by that of a single Fourier transform required to find the power spectrum of the observations, as a result of which the computational cost is now virtually independent of the number of optimization steps. For the present application, we find that the proposed approach is as accurate as the conventional approach for a sample size of 100 or more. The CPU time increases with the number of optimization steps for the conventional approach, while it is virtually constant for the proposed approach.

1 Introduction

Kriging is a powerful Geostatistical tool for spatial interpolation (Cressie 1993). It has also been applied in other fields, such as surrogate modeling and uncertainty quantification (Kennedy and O'Hagan 2000, Chung and Alonso 2002, Laurenceau and Sagaut 2008, Dwight and Han 2009, de Baar et al. 2011). Kriging was developed independently by Matheron (1963) and Gandin (1965) in a conventional statistical framework, however we prefer the derivation in a Bayesian framework, which clarifies the role of the observation error (Kitanidis 1986, Handcock and Stein 1993, Wikle and Berliner 2007, de Baar et al. 2011).

Two important problems arise in the Kriging analysis of a large number N of observations in d dimensions: robustness and cost. (1) With respect to the robustness, in (de Baar et al. 2011) we show how the inclusion of an observation error improves the robustness of the Kriging analysis by its effect on the numerical positive definiteness of the $N \times N$ matrix $A = R + HPH^T$. (2) The remaining problem, the high cost of the Kriging analysis, is dominated by the D-dimensional optimization of the likelihood with respect to the correlation range θ, which for each iteration requires the computation of the determinant and inverse of A.

We propose to transform the entire optimization problem from the spatial to the frequency domain. The main benefit of this approach is that in the continuous frequency domain convolutions are replaced with scalar multiplications, which are much cheaper to evaluate. This approach reduces the computational cost of a Kriging analysis with m optimization steps from $\mathrm{O}\left((m+1)N^3\right)$ to only $\mathrm{O}\left(N^3\right)$, such that it is virtually independent of the number of optimization steps.

1.1 The Kriging predictor equations

We would like to predict the set of discrete values $\vec{x}$, given N observations $\vec{y}$, which are a subset of $\vec{x}$ selected by the observation matrix H. We assume that $\vec{x}$ and $\vec{y}$ are normalized, such that $\vec{y}$ has zero mean and variance one. Note that this normalization is straightforward if we approximate the Kriging mean and variance with the conventional mean and variance. The values $\vec{x}$ are situated at $\vec{\xi}$, a vector containing the spatial locations, normalized such that $\left(\xi_{\min}, \xi_{\max}\right) = \left(-1, 1\right)$. The Kriging predictor for the values and variance is given by (Wikle and Berliner 2007):

$$E(\vec{x} \mid \vec{y}) = PH^T (R + HPH^T)^{-1} \vec{y},$$
$$\mathrm{var}(\vec{x} \mid \vec{y}) = \left[I - PH^T (R + HPH^T)^{-1} \right] P,$$

with correlation matrix $P(\theta)$ and observation error matrix R. In a Bayesian framework, $P(\theta)$ is the covariance of the prior, while R is the covariance of the likelihood (Wikle and Berliner 2007). For uncorrelated observation errors of magnitude ε, we have:

$$R = \varepsilon^2 I.$$

The prior covariance matrix is generated by a covariance function, for example:

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

with lag $h_{ij} = | \xi_i - \xi_j |$ and correlation range θ. For multiple dimensions, this equation is replaced with:

$$p_{ij} = \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{h_{d,ij}^2}{\theta_d^2} \right).$$

Note that this Gaussian correlation function is only an example, and that in many cases it is not the most appropriate choice (Stein 1999). The derivation in Section 2 can be made for any parameterized correlation function, as long as one can provide an analytical Fourier transform.

1.2 Bottleneck: cost of estimating the correlation range

In Ordinary Kriging we have to estimate the correlation range θ. With a maximum likelihood estimate (MLE), we maximize the likelihood with respect to θ. This is equivalent to maximizing (Kitanidis 1986):

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

where $A(\theta)$, like $P(\theta)$, is a function of θ:

$$A = R + HPH^T.$$

This optimization dominates the cost of a Kriging analysis, as for each θ we have to find the determinant and the inverse of the $N \times N$ matrix A. For increased speed and accuracy, in the actual implementation we first make a Cholesky factorization $A = U^T U$, after which we find the eigenvalues of U and solve two linear systems containing U and U^T. The computational cost of these operations is $\mathrm{O}(N^3)$.

1.3 Fast approximation in the frequency domain

To maximize $L(\theta)$ in the spatial domain, we would discretize $a(h, \theta)$, the function (or 'semivariogram') that generates the elements:

$$a_{ij} = \varepsilon^2 \delta_{ij} + p_{ij},$$

of $A(\theta)$. As we have seen, we then would have to find the determinant $|A(\theta)|$ and inverse $B(\theta) = A^{-1}(\theta)$ for each value of θ. This approach is illustrated in Table 1. Again, note that in the actual implementation one would rather find the eigenvalues of $A(\theta)$ and solve the linear system $A^{-1}(\theta)\vec{y}$.

Instead of the optimization in the spatial domain, we propose to maximize $L(\theta)$ in the frequency domain, see Table 2. The reason to do so, is that in the frequency domain the convolution $*$ is replaced by scalar multiplication: Consider, for example, the relation:

$$AB = I,$$

which defines B as the inverse of A. This definition is only a short-hand notation for the following summation:

$$\sum_j a_{ij} b_{jk} = \delta_{ik},$$

where a_{ij} and b_{jk} are the elements of A and B. The continuous equivalent of this summation is given by:

$$\{a * b\}(h) = \delta(h),$$

where $*$ denotes the convolution. Here $b(h)$ is clearly not properly defined in the spatial domain, therefore we consider the Fourier transform:

$$\hat{a}(k)\hat{b}(k) = 1,$$

where the convolution now has been replaced with a scalar product. After discretization, we find that in the frequency domain we can approximate the elements:

$$\hat{b}_n = \frac{1}{\hat{a}_n},$$

with scalar inverses of $\hat{a}_n$. The dotted red lines in Table 2 illustrate how we take similar steps to approximate both terms of $L(\theta)$ in the discrete frequency domain, while the continuous red line indicates that this approach also requires the power spectrum $\hat{y}_n^2$ of the observations.

Table 1: MLE in the spatial domain

	Spatial domain	Frequency domain
Discrete	$-\ln \lvert A \rvert$ $-\vec{y}^T B\ \vec{y}$ $AB = I$	$\sum_{0}^{N-1} \ln \hat{a}_n$ $-\frac{1}{N}\sum_{-N/2}^{N/2} \hat{b}_n \hat{y}_n^2$ $\hat{a}_n \hat{b}_n = 1$
Continuous	$\cdots$ $-\int \bar{y}(h)\{b * y\}(h)dh$ $\{a * b\}(h) = \delta(h)$	$\cdots$ $-\int \hat{b}(k)\hat{y}^2(k)dk$ $\hat{a}(k)\hat{b}(k) = 1$

Table 2: proposed MLE in the frequency domain

	Spatial domain	Frequency domain
Discrete	$-\ln \lvert A \rvert$ $-\vec{y}^T B\ \vec{y}$ $AB = I$	$\sum_{0}^{N-1} \ln \hat{a}_n$ $-\frac{1}{N}\sum_{-N/2}^{N/2} \hat{b}_n \hat{y}_n^2$ $\hat{a}_n \hat{b}_n = 1$
Continuous	$\cdots$ $-\int \bar{y}(h)\{b * y\}(h)dh$ $\{a * b\}(h) = \delta(h)$	$\cdots$ $-\int \hat{b}(k)\hat{y}^2(k)dk$ $\hat{a}(k)\hat{b}(k) = 1$

2 Maximum likelihood estimate of the Kriging correlation range in the frequency domain

As we found in the previous section, a fast maximum likelihood estimate of θ in the frequency domain requires the power spectrum $\hat{y}_i^2$ of the observations, as well as the continuous Fourier transform $\hat{a}(k)$ of the matrix generator (see Table 2). First we derive both for the one-dimensional case, and illustrate this with a test function. Then we extend our result to the multi-dimensional case.

2.1 Obtaining the power spectrum from a discrete Fourier transform

In general, the power spectrum can be found from the Discrete Fourier Transform (DFT). For uniform $\vec{\xi}$ it is faster to apply a Fast Fourier Transform (FFT). In the future, we aim to apply a Non-Uniform Fast Fourier Transform (NUFFT) for non-uniform $\vec{\xi}$ (Song et al. 2009). The computational cost of a DFT is $\mathrm{O}\left(N^2\right)$, while that of a FFT or NUFFT is $\mathrm{O}\left(N \log N\right)$.

In the non-uniform case we have to remove some spurious modes from the spectrum. Presently we simply choose a cut-off frequency, above which we reduce the spectrum to zero. In addition, for the terrain survey data spectrum, we apply the simple power spectrum enhancement:

$$\hat{y}^2_{enh,n} = \max\left(\hat{y}^2_n\right)\left[\frac{\hat{y}^2_n}{\max\left(\hat{y}^2_n\right)}\right]^{1.8}.$$

We aim to improve the spectrum enhancement in future work.

2.2 The continuous Fourier transform of the matrix generator of A

The covariance matrix is generated by the continuous function:

$$p(h) = \exp\left(-\frac{h^2}{2\theta^2}\right),$$

which has the Fourier transform:

$$\hat{p}(k) = \theta\sqrt{\frac{Nk_{\max}}{2}}\exp\left(-\frac{k^2\theta^2}{2}\right),$$

with wave number k, where $k_{\max} = \dfrac{\pi(N-1)}{2}$ is the highest wave number represented in the power spectrum. Here we have now taken proper care of the normalization of the Fourier transform, which is chosen such that for intermediate θ we satisfy:

$$\frac{N}{4}\int_{-2}^{2} p(h)dh = \frac{1}{N}\frac{N}{2k_{\max}}\int_{-k_{\max}}^{k_{\max}} \hat{p}(k)dk,$$

such that after uniform sampling of $p(h)$ on the spatial domain $h \in [-2,2]$, and of $\hat{p}(k)$ on the frequency domain $k \in [-k_{\max}, k_{\max}]$, with the present normalization we will approximate the discrete Parseval's Theorem, given by:

$$\sum_i p_i^2 = \frac{1}{N}\sum_n \hat{p}_n^2 .$$

However, this is the result of a continuous Fourier transform, which does not include the effects of aliasing and spectral leakage, which we would have found in the discrete case. It is important to note that, instead of being interested in the continuous functions themselves, we are really interested in simulating this discrete operation, therefore we simulate the effect of aliasing and spectral leakage.

2.2.1 Simulating aliasing

Aliasing is often described as the contribution of the tails of the spectrum folding back on the domain. There are a variety of approaches to simulate this effect. We consider a fairly simple approach, where we add an average contribution in the form of a small constant:

$$\delta_{alias} = \sqrt{\frac{\pi N}{k_{\max}}}\left[1 - erf\left(\frac{k_{\max}\theta}{\sqrt{2}}\right)\right],$$

which can be found by integrating the tails of the Fourier transform of the matrix generator. It can already be seen that this term will only have a significant contribution for very small θ. We therefore disregard the effect of aliasing for intermediate correlation ranges.

As an alternative perspective on this effect, consider that when the correlation range approaches zero, the matrix A is close to the identity matrix, which has $\ln|I| = 0$ and $\vec{y}^T I^{-1}\vec{y} = N\sigma^2 = N$.

2.2.2 Simulating spectral leakage

We propose to simulate spectral leakage by adding to $\hat{p}$ a constant term:

$$\delta_{leak} = \frac{1}{N^2} ,$$

which we aim to derive rigorously in future work.

2.2.3 Fourier transform of the matrix generator of A

The Fourier transform of the observation error term $\varepsilon^2\delta_h$ is given by the constant ε^2. With that last result, the Fourier transform of the full matrix generator of A is now given by:

$$\hat{a}(k) = \varepsilon^2 + \theta\sqrt{\frac{Nk_{\max}}{2}}\exp\left(-\frac{k^2\theta^2}{2}\right) + \delta_{leak} + \delta_{alias} .$$

2.3 Accuracy

At this point we consider three main sources of error in this approach: (1) the approximation of discrete matrix multiplications with convolutions, (2) the approximation of discrete Fourier transforms with continuous Fourier transforms, and (3) numerical noise in the power spectrum.

Firstly, when we approximate discrete matrix operations with convolutions, we effectively approximate a discrete summation with a continuous integral. The error we make here is of the same magnitude as the well-known error estimate of the mid-point rule, which is $\mathrm{O}(N^{-2})$.

The approximation of the discrete Fourier transform with a continuous Fourier transform, which we make when we find the transform of the matrix generator, is in fact also an approximation of a discrete summation with a continuous integral. Again, the error is $\mathrm{O}(N^{-2})$.

For large N we might expect numerical noise in the power spectrum. At this point we do not treat this source of error.

2.4 Example: one-dimensional test function

As an example, we consider the one-dimensional test function:

$$x_0(\xi) = \sin(2\pi\xi) + 0.3\cos(6\pi\xi).$$

We observe this function on the domain $\xi \in [-1,1]$ with an observation error $\varepsilon_0 = 0.001$. The results for uniform ξ are shown in Figure 1, where the results in black are obtained from the conventional approach in the spatial domain, while the results in red are obtained from the proposed approach in the frequency domain. Note that the we have plotted the normalized observations, and that the dotted black line indicates the average spatial separation between the observations. In the uniform case, we compute our results from the raw power spectrum. For non-uniform ξ, the spectrum contains spurious modes, which we remove by setting the power to zero above a cut-off frequency, which in this case corresponds to the 8$^{\text{th}}$ Fourier mode. The results are shown in Figure 2.

Figure 3 shows the effect of aliasing for small correlation ranges. From the spatial domain we obtain the black lines. From the frequency domain, we obtain the red line simulating aliasing, while we obtain the dotted red line without simulating aliasing. Indeed, simulating aliasing improves the estimate, however, this effect is only significant for correlation ranges smaller than the average spatial separation between the observations, as indicated by the dotted black line.

The accuracy of the estimated correlation range for increasing N is given by the relative difference between the value of $\ln(\theta)$ obtained in the spatial and in the frequency domain. Figure 4 shows that, although initially the error is roughly $\mathrm{O}(N^{-2})$, at some point it levels off. This is could either be related to the observation error or to the numerical noise we find in the power spectrum for higher N.

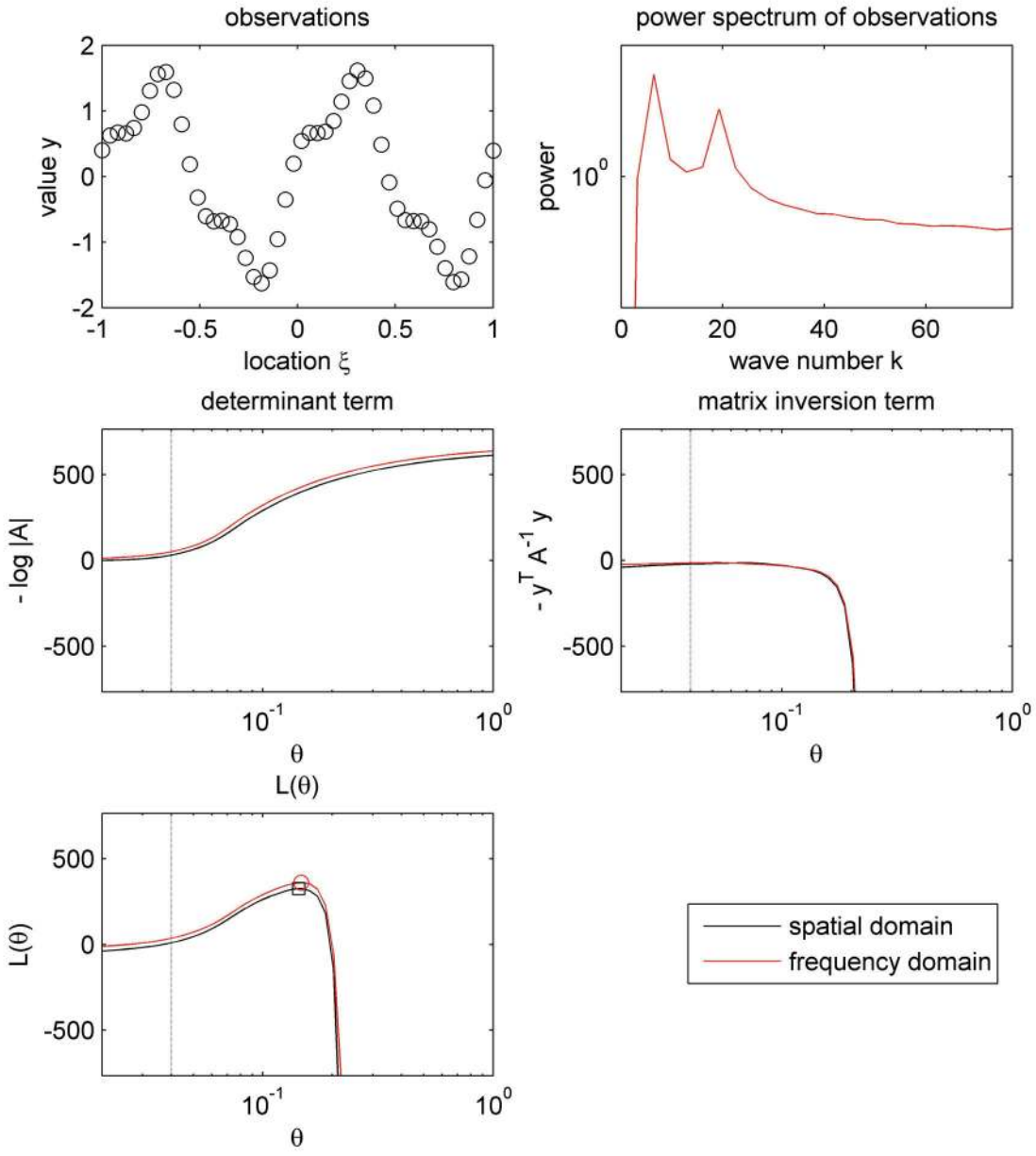

Figure 1: Uniform sampling

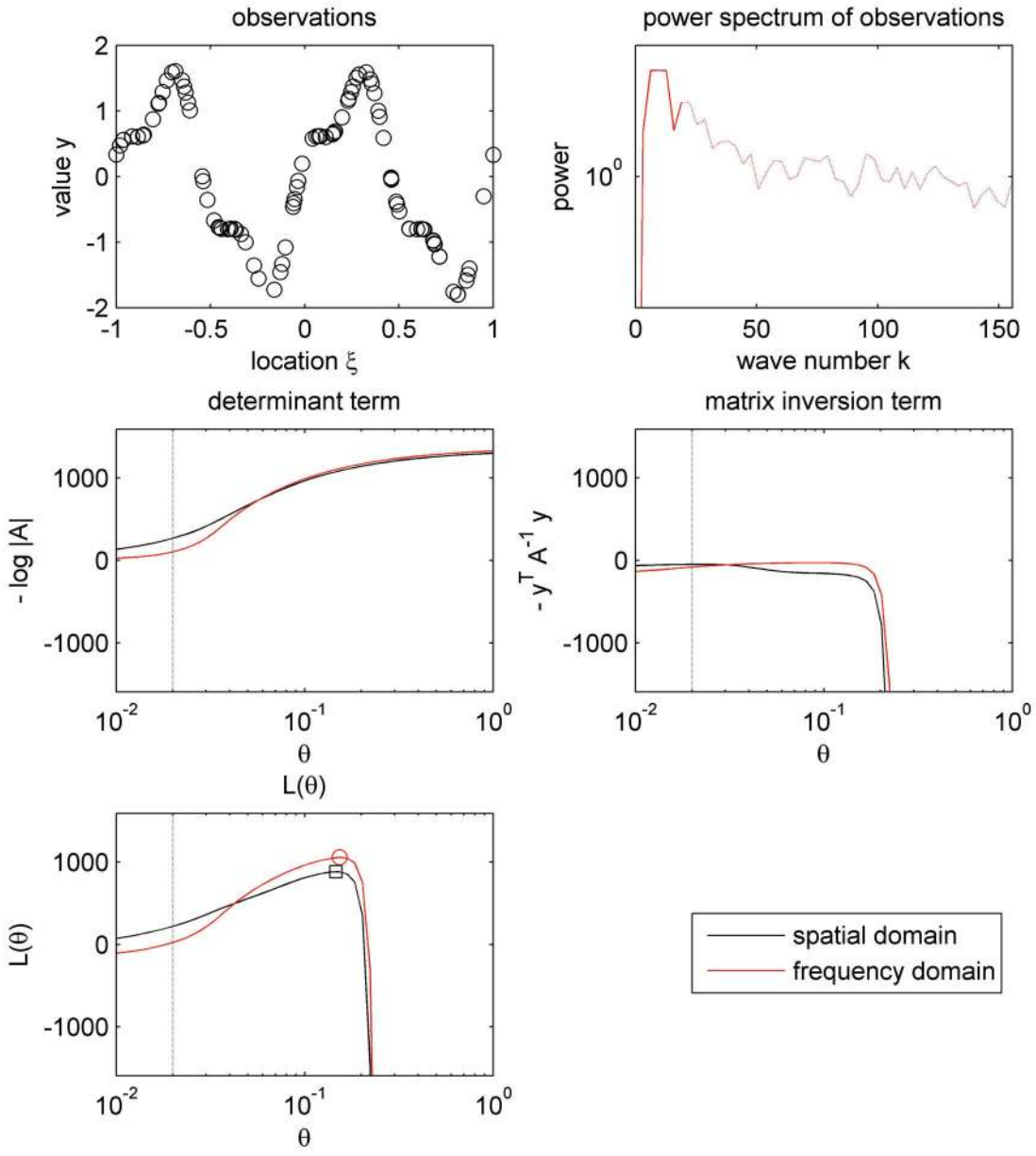

Figure 2: Non-uniform sampling, with cut-off frequency in the power spectrum

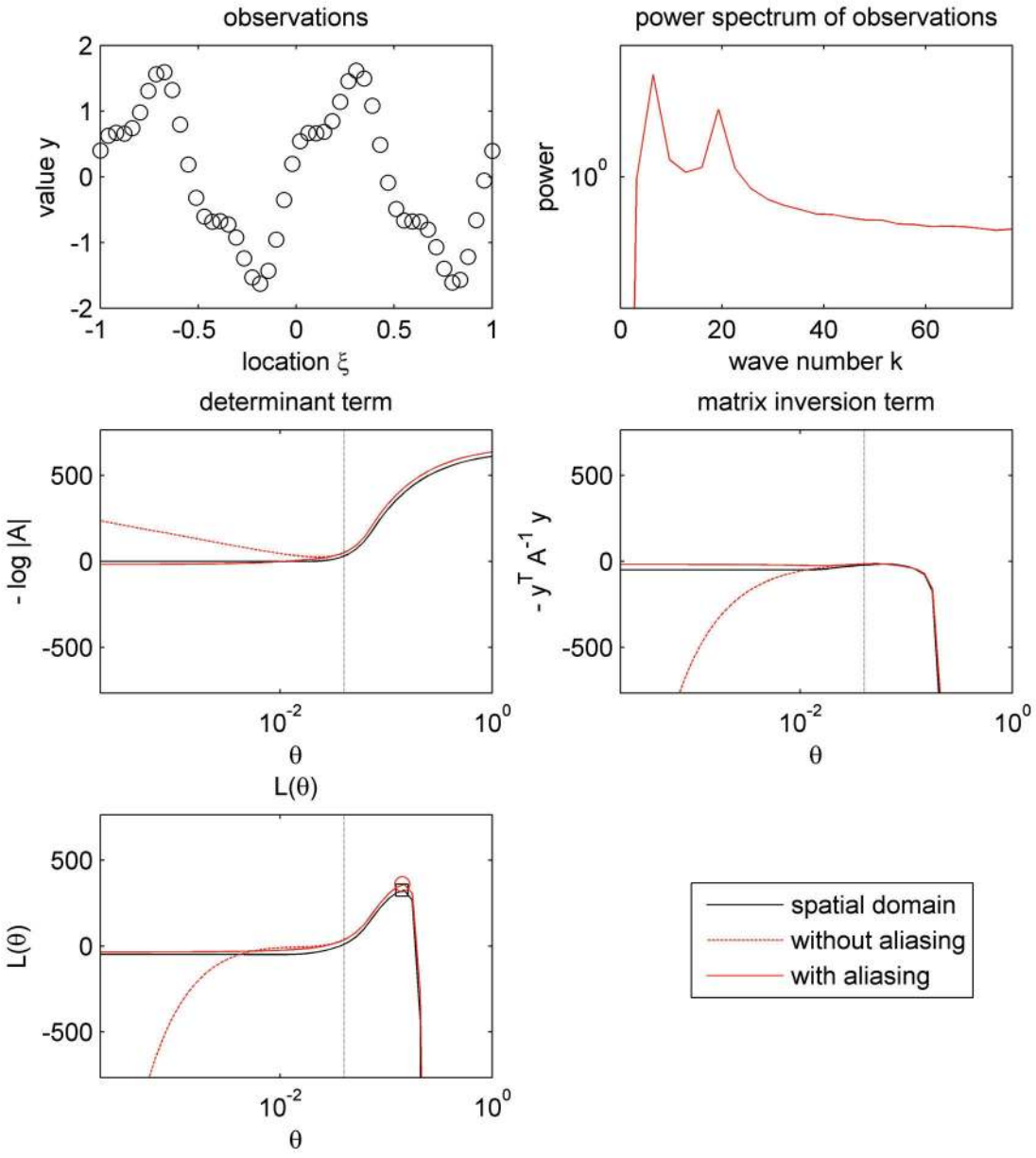

Figure 3: The effect of aliasing for small correlation ranges

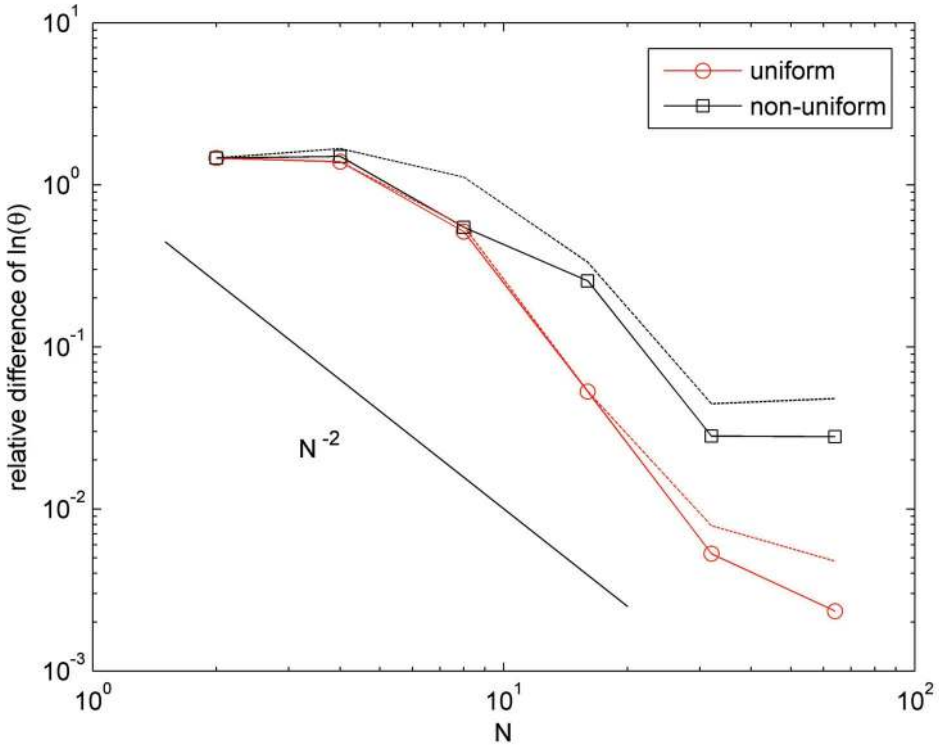

Figure 4: Relative difference of $\ln(\theta)$ obtained in the spatial and in the frequency domain. Average (continuous lines) and standard deviation (dotted lines), taken from 10 runs with updated random sampling and errors

2.5 Generalization to multiple dimensions

In multiple dimensions, the power spectrum is found from the DFT. The number of modes is currently taken to be equal in each dimension, although this can be easily specified differently. For uniform $\vec{\xi}$ we can apply the multi-dimensional FFT, while for non-uniform $\vec{\xi}$ the applicability of the NUFFT depends on the particular implementation, which will be a topic of future research.

The Fourier transform of the D-dimensional matrix generator of A is given by:

$$\hat{a}\left(\vec{k}\right) = \varepsilon^2 + \prod_{d=1}^{D} \theta_d \sqrt{2^{-D} N \prod_{d=1}^{D} k_{d,\max}} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} k_d^2 \theta_d^2 \right) + \delta_{leak} + \delta_{alias},$$

where $\vec{k}$ and $\vec{\theta}$ are now vectors of size D.

The effect of aliasing in multiple dimensions will be the topic of future work, presently we take $\delta_{alias} = 0$, considering that the effect will not be important for intermediate correlation ranges.

2.6 Example: two-dimensional test function

As an example, we consider the two-dimensional test function:

$$x_0(\xi) = \sin(\pi\xi_1) + \sin(4\pi\xi_2)\,.$$

The results for non-uniform sampling with observation error $\varepsilon_0 = 0.05$ are shown in Figure 5. In this figure, the results on the left are from the conventional approach in the spatial domain, while the results on the right are from the proposed approach in the frequency domain. We see that the raw power spectrum contains spurious modes at high frequencies, which results in an underestimation of the correlation ranges.

In Figure 6, we have again applied a simple cut-off frequency in both directions. This improves the resulting estimates, and from the maximum of $L(\theta)$ we find approximately the same correlation ranges as in the conventional approach.

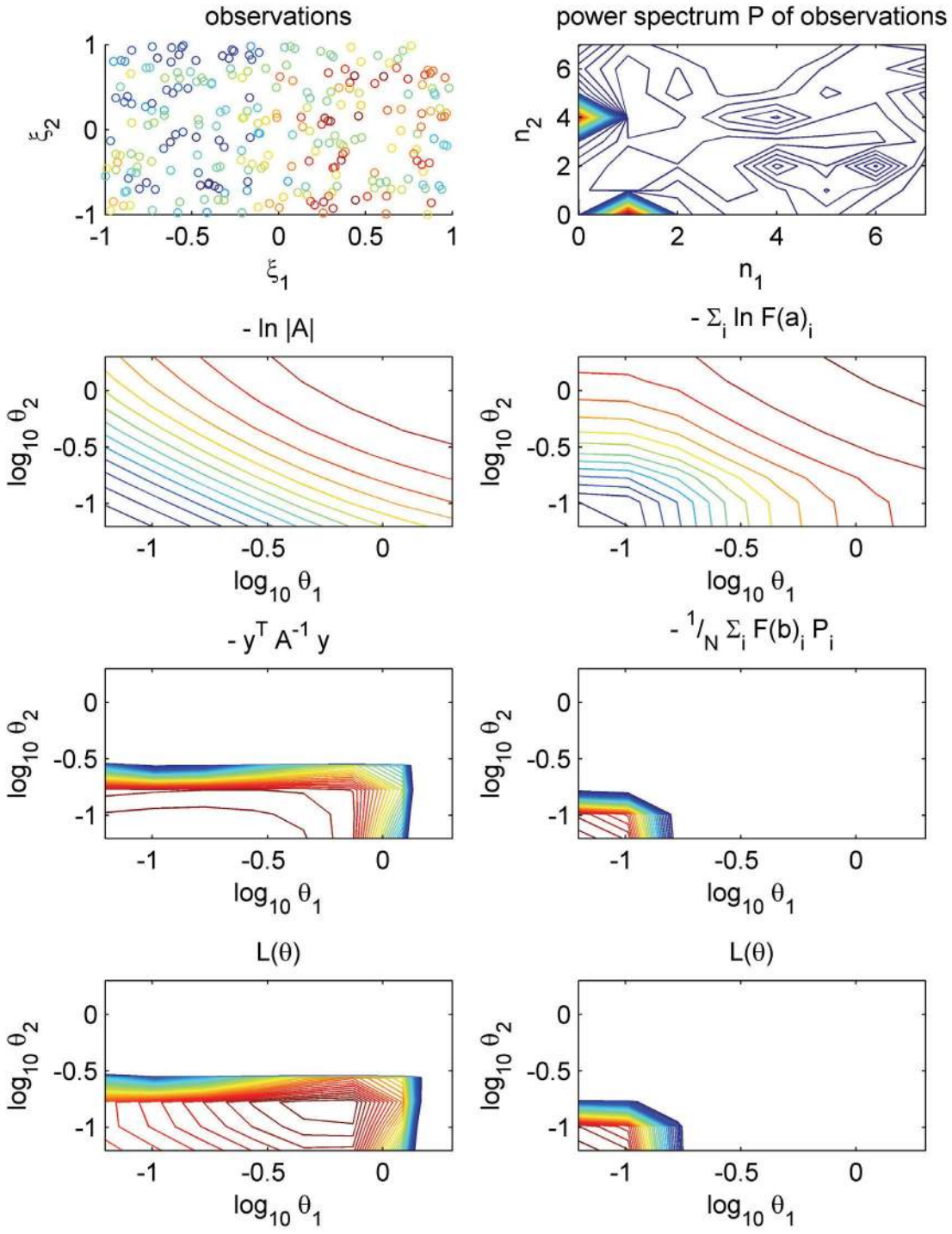

Figure 5: Two-dimensional non-uniform sampling, raw spectrum

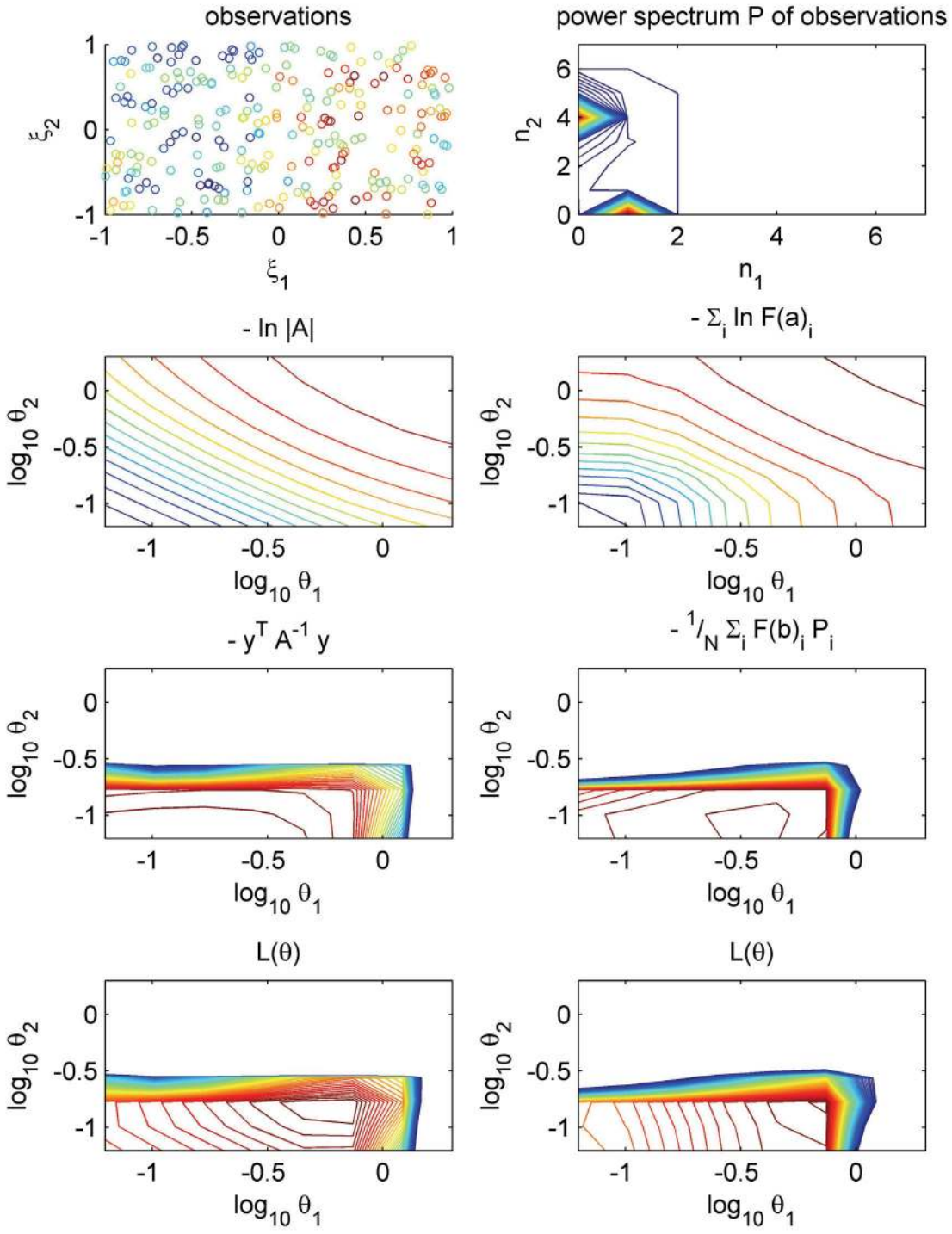

Figure 6: Two-dimensional non-uniform sampling, with cut-off frequencies

3 Application to terrain survey data, obtained by a UAV

Efficient interpolation is an important issue for terrain survey. Here we consider 75,000 height measurements of a segment of a flood protection dike, obtained with an Unmanned Aerial Vehicle (UAV) by *Heering UAS*. High resolution photographs taken by the UAV are combined with GPS recordings to survey the terrain height. The accuracy of the data is improved by the fact that, contrary to a piloted aircraft, the UAV is permitted to fly at low altitude.

Figure 7 shows a photograph of the surrounding terrain, while Figure 8 shows the normalized 75,000 height measurements of the dike segment. In our analysis, this set of 75,000 data acts as $\vec{x}$. We would like to predict $\vec{x}$ from a much smaller sub-set of observations $\vec{y}$ of size N. We randomly sample $\vec{y}$ from $\vec{x}$, an example of such a sample of size $N = 64$ is shown by the black squares in Figure 8. Considering the apparently random variations in neighboring data, we estimate a normalized observation error of $\varepsilon = 0.05$.

We would like to emphasize that we are currently not using the complete set of 75,000 data as observations in our Kriging analysis. We only use a small subset of the data as observations, and from this subset wet attempt to predict the complete set of data.

Figure 7: Aerial photograph of the dike segment and surrounding area (© 2011 Google - Imagery © 2011 DigitalGlobe, GeoEye, Aerodata International Surveys)

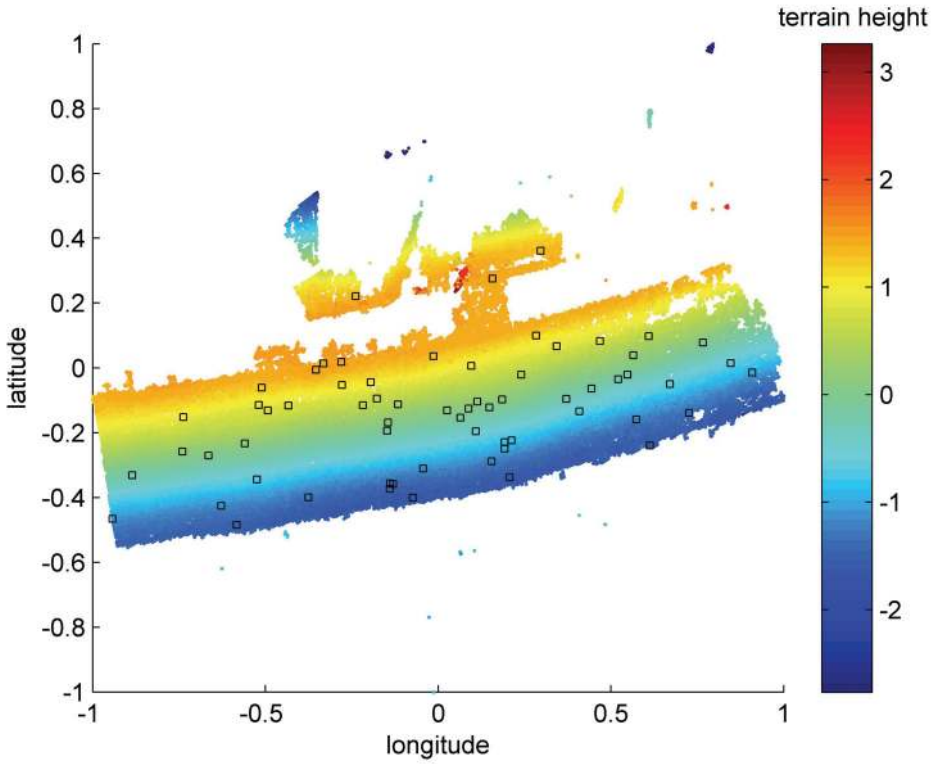

Figure 8: Normalized terrain height of a dike segment, data obtained by a UAV

From $\vec{y}$ we estimate the correlation lengths by brute force optimization, these estimates are shown in Figure 9. With these correlation lengths we can make the prediction $E(\vec{x}\,|\,\vec{y})$, and compare this prediction to the observed $\vec{x}$. Figure 10 shows the standard deviation of the prediction error $E(\vec{x}\,|\,\vec{y})-\vec{x}$, which is a measure of the accuracy of the prediction. We see that for small N the proposed approach is not as accurate as the conventional approach, however for increasing $N>100$ both approaches show the same accuracy.

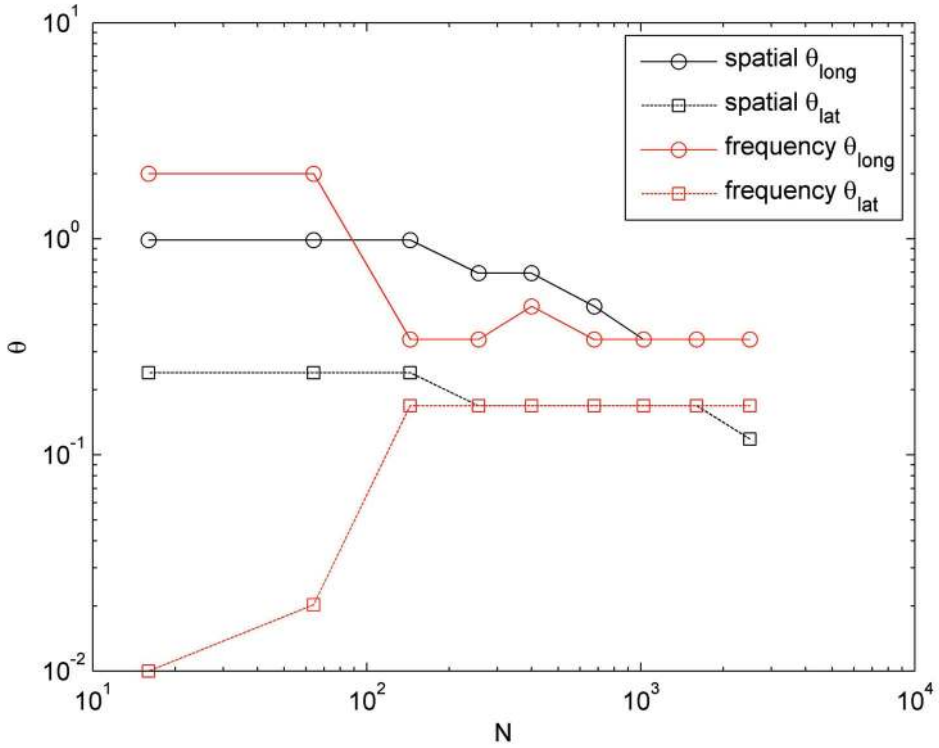

Figure 9: Estimated correlation ranges

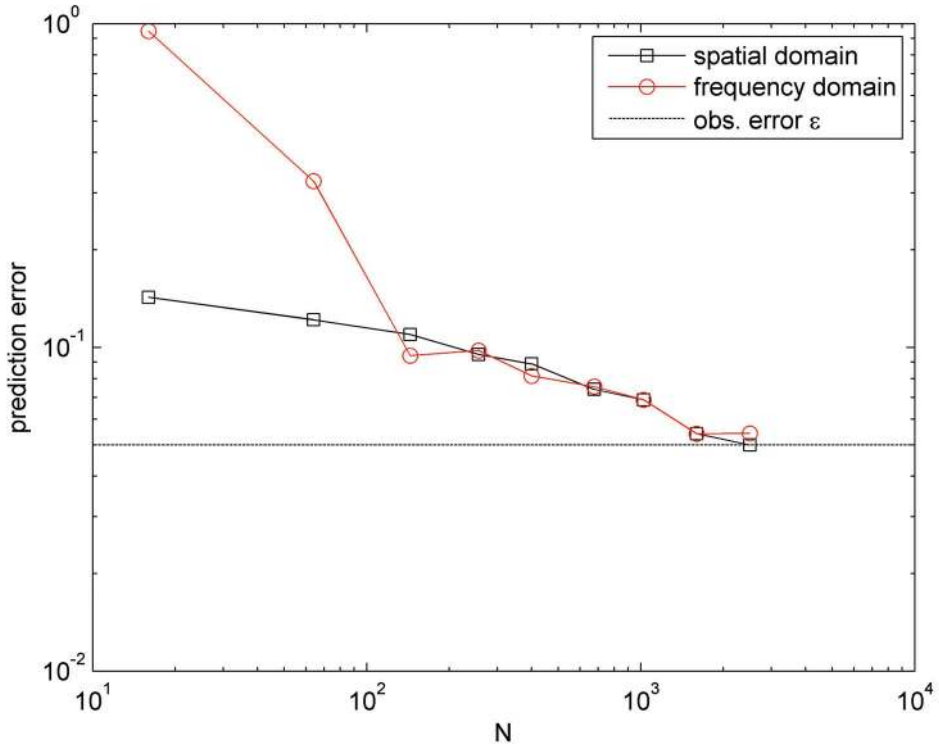

Figure 10: Prediction error $std\left(E(\vec{x}\mid\vec{y})-\vec{x}\right)$ of the terrain height

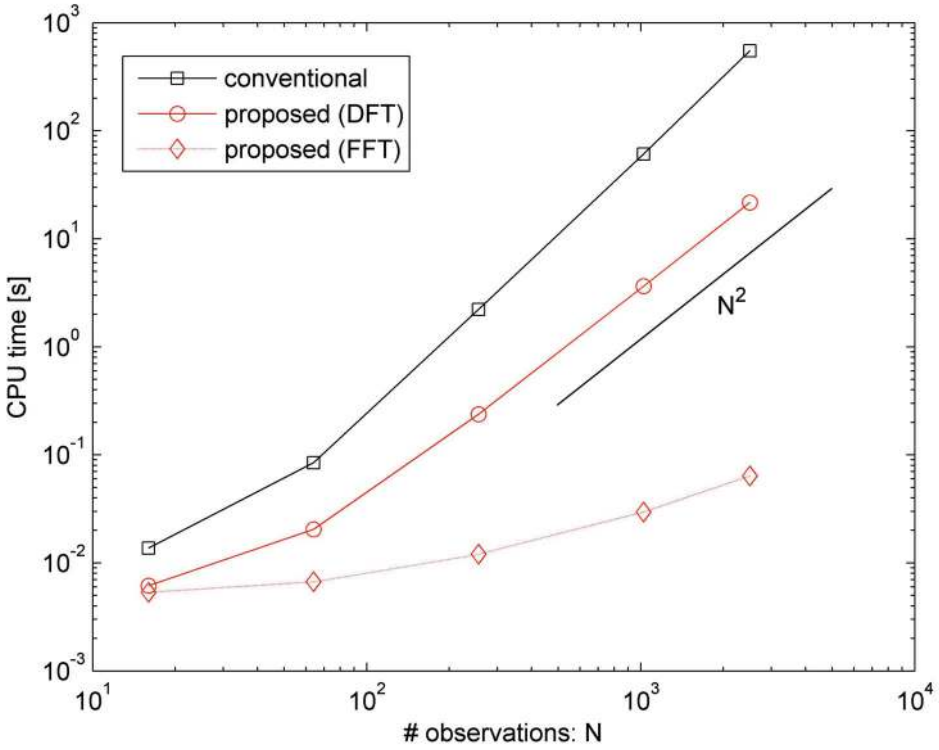

Figure 11: Cost of the maximum likelihood estimates for different N

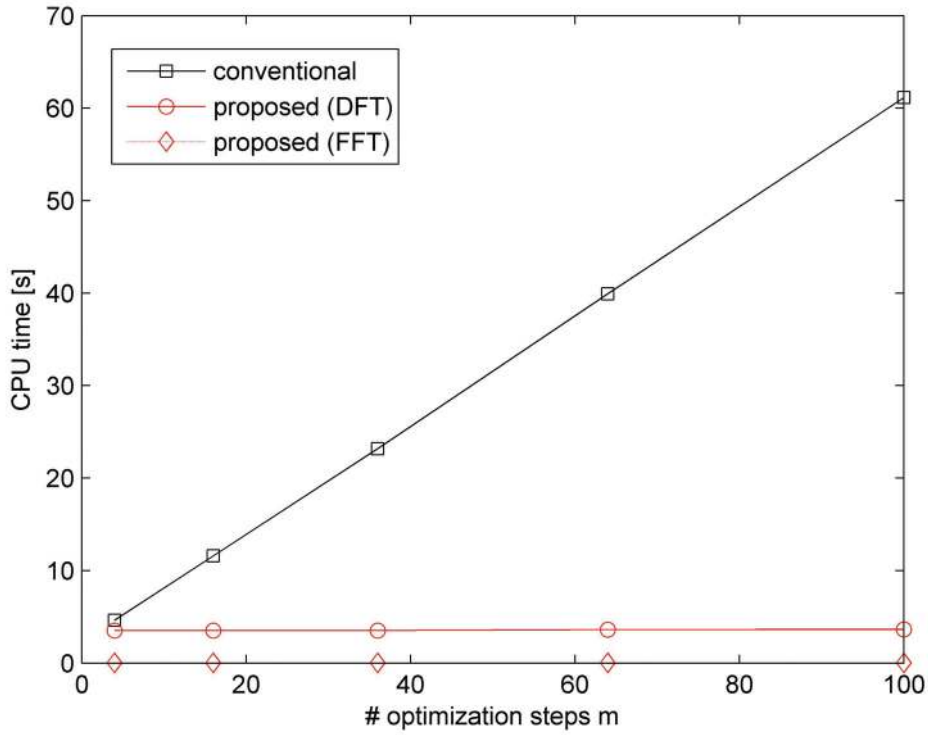

Figure 12: Cost of the maximum likelihood estimates for different m

Since the objective is to have a faster estimate of the correlation ranges, we make a rough comparison of CPU time. First we consider the CPU time required for the maximum likelihood estimates using $m = 100$ iteration steps, for various numbers N of observations of the terrain height. Figure 11 illustrates that the cost of the proposed optimization (red line) is proportional to N^2, while the cost of the conventional optimization is significantly higher (black line). However, in this comparison we use highly optimized Matlab functions to find the Cholesky factors, to find the eigenvalues and to solve the linear systems in the conventional approach, while we use a non-optimized DFT to obtain the power spectrum in the proposed approach. Although the use of a FFT instead of the DFT does not give a correct power spectrum, it does provide a rough indication of the CPU time to be expected from an optimized NUFFT in future implementation (dotted red line with diamonds).

Finally, we consider the CPU time required for the maximum likelihood estimates based on $N = 1024$ observations of the terrain height, for various numbers m of optimization steps. Figure 12 illustrates that the cost of the optimization is proportional to m in the conventional approach (black line), while it is virtually independent of m for the proposed approach (red line). Again, the use of a FFT provides a rough indication of the CPU time to be expected from an optimized NUFFT in future implementation (dotted red line with diamonds).

4 Discussion and outlook

We have proposed a fast maximum likelihood estimate of the Kriging correlation range in the frequency domain. The aim of this approach is to reduce the cost of a Kriging analysis.

We have shown the convergence of the estimated correlation ranges and of the Kriging prediction for randomly sampled terrain survey data. The CPU time required for the optimization is proportional to the number of optimization steps in the conventional approach, while it is virtually constant in the proposed approach. However, we observe that due to high frequencies in the raw power spectrum, enhancement of the spectrum is necessary for non-uniform sampling as well as for large data sets. The present spectrum enhancement, mainly through high frequency cut-off, is clearly insufficient.

Future work will focus on implementation of a NUFFT, on a more thorough derivation of the simulation of spectral leakage and aliasing, and on the enhancement of the power spectrum for non-uniform sampling and large data sets.

Acknowledgements

We would like to express our gratitude to P. Wijkstra of *Heering UAS* and to G. Vestjens of *Geodelta*, who kindly provided the Unmanned Aerial Vehicle (UAV) terrain survey data.

References

BAAR, J.H.S. DE; DWIGHT, R.P.; BIJL, H.: Improvements to gradient-enhanced Kriging using a Bayesian interpretation. *Submitted to a Wiley Journal, May 2011*.

CRESSIE, N. (1993): *Statistics for spatial data*. 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* .

DWIGHT, R. P.; HAN, Z.-H. (2009): Efficient uncertainty quantification using gradient enhanced Kriging. *11th AIAA Non-Deterministic Approaches Conference*.

GANDIN, L. (1965): *Objective analysis of meteorological fields: Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad*. Translated by Israel Programfor Scientific Translations, Jerusalem.

HANDCOCK, M. S.; STEIN, M. L. (1993): A Bayesian analysis of Kriging. *Technometrics 35*(4), pp. 403–410.

MATHERON, G. (1963): Principles of Geostatistics. *Economic Geol*. 58, 1246–1266.

KENNEDY, M. C.; O'HAGAN, A. (2000): Bayesian calibration of computer models. *J. R. Statist. Soc. B 63*, 425–464.

KITANIDIS, P. K. (1986): Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research 22*(4), 499–507.

LAURENCEAU, J.; SAGAUT, P. (2008): Building efficient response surfaces of aerodynamic functions with Kriging and Cokriging. *AIAA Journal 46*(2), 498–507.

SONG, J.; LIU, Q.H.; GEWALT, S.L.; COFER, G.; JOHNSON, G.A. (2009): Least Square NUFFT Methods Applied to 2D and 3D Radially Encoded MR Image Reconstruction. *IEEE Trans Biomed Eng. 56(4)*, 1134–1142.

STEIN, M. L. (1999): *Interpolation of spatial data, some theory for Kriging*. Springer.

WIKLE, C. K.; BERLINER, L. M. (2007): A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena 230* (1-2), 1 – 16.