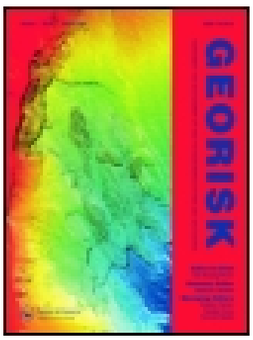

Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards

ISSN: 1749-9518 (Print) 1749-9526 (Online) Journal homepage: http://www.tandfonline.com/loi/ngrk20

Effect of sampling plan and trend removal on residual uncertainty

Gordon A. Fenton, Farzaneh Naghibi & Michael A. Hicks

To cite this article: Gordon A. Fenton, Farzaneh Naghibi & Michael A. Hicks (2018): Effect of sampling plan and trend removal on residual uncertainty, Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, DOI: 10.1080/17499518.2018.1455106

To link to this article: https://doi.org/10.1080/17499518.2018.1455106

Published online: 02 Apr 2018.

GEORISK, 2018
https://doi.org/10.1080/17499518.2018.1455106

Taylor & Francis
Taylor & Francis Group

Effect of sampling plan and trend removal on residual uncertainty

Gordon A. Fenton[a,b], Farzaneh Naghibi[a] and Michael A. Hicks[b]

[a]Department of Engineering Mathematics and Internetworking, Dalhousie University, Halifax, NS, Canada; [b]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands

ABSTRACT

The ground is one of the most highly variable of all engineering materials. As a result, geotechnical designs depend upon a site investigation to estimate the ability of the ground to perform acceptably. For example, when a shallow foundation is being proportioned to avoid a bearing capacity failure under a certain applied load, the frictional and cohesive properties of the ground under the foundation must first be estimated through a site investigation. Questions which arise are: How does the quality and intensity of the site investigation affect the design? Is more investigation cost effective? If the site is sampled at one location and the foundation placed at a different location, how does this mismatch affect the target design and the reliability of the final foundation? By modelling the ground as a spatially variable material, questions such as the above can be investigated through Monte Carlo simulation and sometimes theoretical probabilistic models. Using such tools, this paper looks specifically at how the intensity (frequency and spatial distribution) of a site sampling plan, and how the samples are used, affects the understanding of the ground properties under a foundation. Interestingly, it is found that removing the sample mean outperforms removing the best linear unbiased estimate (BLUE) when the actual field correlation length is small but the BLUE correlation length is assumed equal to the field size. Recommendations are made regarding number of samples and the type of trend to best characterise the field.

Abbreviations: BLUE: best linear unbiased estimate; MCS: Monte Carlo simulation; LAS: local average subdivision

ARTICLE HISTORY
Received 27 December 2017
Accepted 1 March 2018

KEYWORDS
Site characterisation; residual uncertainty; sampling; required number of samples; sampling plans

Symbols

a, b, c	coefficients of the plane equation $a + bx_i + cy_i$
a_j, b_j, c_j	coefficients used to construct a, b, c
A, A_i	matrices used in regression
b_i	vector of covariances using random field correlation length
b_i^k	vector of covariances using BLUE correlation length
b_{ij}^k	elements of vector b_i^k
b_{ij}	covariance between X_i and X_j^o
C_{ij}	elements of covariance matrix
d_{ij}	parameter used in definition of residual variance ($= a_j + b_j x_i + c_j y_i$)
D	edge dimension of the $D \times D$ square random field
e	random field cell area or domain
e_i	ith random field cell or domain
m	number of random field cells in either x or y direction
n	number of random field cells
n_s	number of samples
$G(\mathbf{x})$	stationary Gaussian random field
K	BLUE covariance matrix
S_x, S_y, S_{xy}, S_{xx}, S_{yy}, S^o, S_x^o, S_y^o	sums used in regression
V	theoretical semi-variogram
$\hat{V}$	estimated semi-variogram
$\mathbf{x}_i$	coordinates of the centre of the ith cell$= (x_j, y_j)$
$X(\mathbf{x}_i)$	local average of the random field over the ith cell ($=X_i$)
X_i^o	ith sample observation
$X_r(\mathbf{x})$	residual random field ($= X(\mathbf{x}) - \hat{\mu}(\mathbf{x})$)
$\hat{y}_j$	parameter used in estimate of correlation length
β_i	vector of BLUE coefficients
β_{ij}	elements of vector β_i
γ_e	variance reduction due to averaging over a random field cell
γ_{ij}	average correlation coefficient between ith and jth cells
Δx	distance between field cell centres
$\boldsymbol{\eta}$	dummy variable of integration

CONTACT Gordon A. Fenton gordon.fenton@dal.ca

θ	random field correlation length
$\hat{\theta}$	estimated correlation length
θ_k	BLUE correlation length
μ	mean
$\hat{\mu}$	estimated constant mean
$\hat{\mu}(\mathbf{x})$	estimated mean trend
$\boldsymbol{\xi}$	dummy variable of integration
ρ	correlation coefficient
σ	standard deviation
σ^2_{cell}	variance of a local average over a random field cell
$\hat{\sigma}^2_{\text{cell}}$	estimated value of σ^2_{cell}
σ^2_G	variance of the Gaussian random field
σ^2_r	variance of the residual
$\hat{\sigma}^2_r$	estimated variance of the residual
τ	distance between points on the random field
τ_j	semi-variogram lag distance ($= j\Delta x$)

1. Introduction

Site characterisation is clearly an essential component of any geotechnical design and a great deal of effort has been devoted over recent decades on how to best perform such a characterisation. Questions such as "How many samples should be taken?", "How should these samples be used in the design process?", and "How do these samples affect my level of understanding of the site?" have always been of great concern. This paper looks specifically at the answers to some of these questions.

The ground is one of the most complex of engineering materials due to its high spatial variability and uncertainty about its engineering properties. While the engineering properties of steel, concrete, and wood, for example, have fairly well established and relatively small uncertainties, ground properties can vary by even many orders of magnitude from site to site, and even within a single site.

As a result of the large uncertainty in the ground, all geotechnical designs should start with a geotechnical investigation so that the best "nominal" or "characteristic" ground parameters can be used in the design process. Traditionally, the intensity of the site investigation has not been particularly important, so long as a reasonable estimate of the characteristic design values could be estimated. More specifically, the benefit of an increased intensity of site investigation has not been recognised nor rewarded in most geotechnical design codes. These codes specify a single resistance factor regardless of site investigation effort.

Recent impetus has been towards providing reasonable estimates of the reliability of designed geotechnical systems and in properly reflecting target reliabilities in design codes. In order to economically achieve target reliabilities, the degree of understanding of the ground providing the geotechnical resistance needs to be properly evaluated. To investigate this problem, the accuracy of characteristic value estimates needs to be estimated as a function of site investigation effort.

Jaksa et al. (2005) investigated the effects of site investigation scope on geotechnical risk reduction and specifically found the change in likelihood of over- or under-design as a function of site investigation intensity. Yang et al. (2017) studied slope reliability considering site investigation data by conditional random field simulations. Ching and Phoon (2017) investigated the uncertainties associated with predicting trends in ground properties. While Li, Hicks, and Vardon (2016) investigated the effect of number and location of samples on the reduction in uncertainty in a slope's factor of safety, the research presented here instead focuses on the more general question of how the number of samples taken, as well as how those samples are used to characterise the site, affects the overall residual uncertainty, i.e. the uncertainty that remains after accounting for the sample.

2. Effect of sampling intensity

This section looks specifically at how the number of soil samples affects the accuracy of the estimated soil statistics. It is assumed that our samples are error free and are measuring a single soil property. Error free samples are a best case scenario, yielding the greatest uncertainty reduction with increasing sampling effort. Thus, the results of this paper provide a lower bound on the residual uncertainty after sampling.

The samples are used to attempt to estimate the mean, μ, standard deviation, σ, and correlation structure of a site. The correlation structure is characterised by a correlation length, θ. A key question to be answered here is: How does the number of samples affect the accuracy of the estimated statistics? Or, put another way, how many samples are required to achieve a certain desired accuracy in the estimates? The answer is developed by considering a square site and using random field simulation to generate realisations of the soil properties over the site, sampling each realisation, and then comparing the estimated mean, variance, and correlation length to the "true" values. The goal is to investigate the discrepancies between the estimated statistics and the true "local" statistics, with the latter obtained by sampling the field at all locations. Note that the "local" statistics will differ from the population parameters, μ (mean), σ (standard deviation), and θ (correlation

length), which are used by the random field generator, due to the fact that the local statistics are derived from each single realisation. In detail, the soil is represented by a stationary Gaussian random field, $G(\mathbf{x})$, which is discretised into a series of n equal sized cells each having area e. The local average of the ith cell, denoted by $X(\mathbf{x}_i)$ and centred at the spatial position $\mathbf{x}_i$, is defined as

$$X(\mathbf{x}_i) = \frac{1}{e}\int_{e_i} G(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{1}$$

where e_i is the ith cell domain, for $i = 1, \ldots, n$. The resulting discretised field is sampled at n_s locations and the samples are then used to estimate a mean trend, $\hat{\mu}(\mathbf{x})$. The estimated trend can then be compared to the field realisation to assess its ability to represent the actual mean trend. Defining the residual to be

$$X_r(\mathbf{x}) = X(\mathbf{x}) - \hat{\mu}(\mathbf{x}), \tag{2}$$

then $\hat{\mu}(\mathbf{x})$ is a good estimate of the mean trend if X_r is generally small. If the site is sampled at all locations, then $\hat{\mu}(\mathbf{x})$ can be taken to be equal to $X(\mathbf{x})$ in the event that a point-wise trend is assumed for $\hat{\mu}(\mathbf{x})$, in which case $X_r(\mathbf{x}) = 0$ everywhere. Sampling at all locations is the best case since there is then minimum residual uncertainty (zero in the case of a point-wise trend).

Sampling at all locations is, of course, prohibitively expensive and may also change the resulting field properties due to the act of measuring them. In practice, soil properties are estimated from a relatively small number of samples so that $\hat{\mu}(\mathbf{x})$ will at best be an approximation of $X(\mathbf{x})$, with varying degrees of accuracy.

In assessing the ability of $\hat{\mu}(\mathbf{x})$ to represent $X(\mathbf{x})$, it is useful to ask how much residual uncertainty remains? To answer this question, consider the variance of the residual $X_r(\mathbf{x})$ defined by Equation (2),

$$\begin{aligned}\hat{\sigma}_r^2 &= \frac{1}{D\times D}\int_{D\times D} X_r^2(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &\approx \frac{1}{n}\sum_{i=1}^{n}[X(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)]^2,\end{aligned} \tag{3}$$

where D is the edge dimension of the $D \times D$ square random field. Since the domain is broken up into n cells in the simulation, the summation form on the right, in which $\mathbf{x}_i$ is the location of the centre of the ith cell, is used. Each cell has its own random field value, $X_i = X(\mathbf{x}_i)$, and its own estimated mean, $\hat{\mu}(\mathbf{x}_i)$. In both the simulation and theory, X_i is taken to be a local arithmetic average of the Gaussian random field as defined by Equation (1).

The theoretical residual variance, σ_r^2, is calculated as the mean of the sample variance (assuming unbiasedness, i.e. that $E[\hat{\sigma}_r^2] = \sigma_r^2$),

$$\sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \hat{\mu}(\mathbf{x}_i))^2]. \tag{4}$$

Note that, in general, $E[(X_i - \hat{\mu}(\mathbf{x}_i))^2]$ is non-stationary – it depends on where $\mathbf{x}_i$ is relative to sampled locations. For example, when the trend, $\hat{\mu}(\mathbf{x}_i)$, passes through one or more sampled values, the residual values and their variances are both zero at those locations. In addition, the estimate in Equation (4) is the average over the field of mean values, while the estimate in Equation (3) is the average over the field of cell values. There thus might be a sampling difference between the two estimates.

The agreement between $\hat{\mu}(\mathbf{x})$ and $X(\mathbf{x})$ can also be investigated by looking at the residual correlation length, i.e. how does the trend removal affect the perceived correlation length?

Nine sampling schemes are considered, as illustrated in Figure 1, where n_s is the number of samples taken from the field. The paper concentrates on the $n_s = 3$, 5, 7 and 9 sampling schemes for simplicity – the $n_s = 2$, 4, 6 and 8 results are simply intermediate to the presented results. In some cases, a further "maximum" sampling scheme is performed, where every point in the field is sampled, $n_s =$ all, to investigate what the maximum attainable uncertainty reduction is.

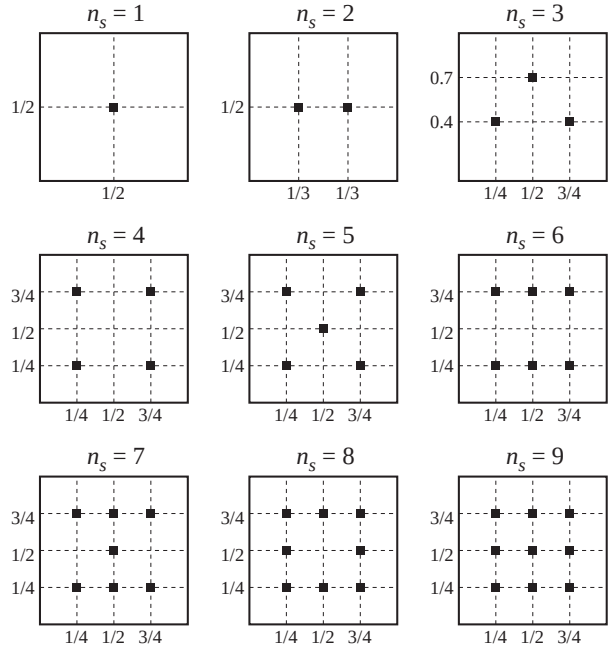

Figure 1. Sampling schemes. Field size is assumed to be 1×1.

For each sampling scheme, three types of trend removal are performed: (a) removing a constant sample mean estimated from the sample, (b) removing a bilinear trend surface fit to the sample, and (c) removing a best linear unbiased estimate (BLUE) surface based on the sample and an assumed covariance structure. The residual statistics are determined theoretically and then validated by Monte Carlo simulation (MCS) using 2000 realisations for each case. In both the theoretical analysis and the MCS, the field is discretised into 128×128 cells. The MCS random fields are generated using the local average subdivision method (Fenton and Vanmarcke 1990) assuming a Markovian correlation function of the form

$$\rho(\tau) = \exp\left\{\frac{-2|\tau|}{\theta}\right\}, \tag{5}$$

where τ is the distance between two points in the field and θ is the assumed correlation length. This correlation function was selected due to its simplicity, being a function of a single parameter, θ, and, because the paper is not site specific, there is no reason to select any particular alternative correlation function. Indeed, in practice, there are rarely sufficient data to justify the use of other, perhaps more sophisticated, correlation functions. The variance of a local average X_i is defined in terms of the correlation function as

$$\begin{aligned}\mathrm{Var}[X_i] = \sigma^2_{\mathrm{cell}} &= \sigma^2_G \left(\frac{1}{e^2} \int_{e_i} \int_{e_i} \rho(\boldsymbol{\xi} - \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\xi}\, \mathrm{d}\boldsymbol{\eta} \right) \\ &= \sigma^2_G \gamma_e. \end{aligned} \tag{6}$$

The covariance between two local averages, X_i and X_j, is defined as

$$\begin{aligned}\mathrm{Cov}[X_i, X_j] &= \sigma^2_G \left(\frac{1}{e^2} \int_{e_i} \int_{e_j} \rho(\boldsymbol{\xi} - \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\xi}\, \mathrm{d}\boldsymbol{\eta} \right) \\ &= \sigma^2_G \gamma_{ij}. \end{aligned} \tag{7}$$

The random field is assumed to have variance $\sigma^2_G = 1.0$ so that the results presented in this paper are measures of variance reduction relative to the field variance. In other words, the actual value of the field variance is not important to the results of this paper.

3. Theoretical model

In this section, the theoretical model for estimating the residual variance, σ^2_r, is derived for the three types of trend removals mentioned in the previous section.

3.1. Constant sample mean removed

In this case, the mean trend, $\hat{\mu}(\mathbf{x}) = \hat{\mu}$, is assumed to be constant and equal to the sample mean,

$$\hat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} X_i^o, \tag{8}$$

where n_s is the number of samples and X_i^o is the ith sample observation, $i = 1, \ldots, n_s$, as extracted from the field at the locations indicated in Figure 1. The assumption of a constant mean is commonly made in geotechnical site investigations with limited sampling. The mean residual variance is calculated by substituting Equation (8) into Equation (4), giving,

$$\begin{aligned}\sigma^2_r &= \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \hat{\mu})^2] = \frac{1}{n} \sum_{i=1}^{n} E[X_i^2 + 2X_i\hat{\mu} + \hat{\mu}^2] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sigma^2_{\mathrm{cell}} - \frac{2}{n_s} \sum_{j=1}^{n_s} \mathrm{Cov}[X_i, X_j^o] + \frac{1}{n_s^2} \sum_{j=1}^{n_s} \sum_{k=1}^{n_s} \mathrm{Cov}[X_j^o, X_k^o] \right\} \\ &= \sigma^2_G \left\{ \gamma_e - \frac{2}{n n_s} \sum_{i=1}^{n} \sum_{j=1}^{n_s} \gamma_{ij} + \frac{1}{n_s^2} \sum_{j=1}^{n_s} \sum_{k=1}^{n_s} \gamma_{jk} \right\} \end{aligned} \tag{9}$$

The reduction in variability can now be expressed as the dimensionless ratio of standard deviations:

$$\frac{\sigma_r}{\sigma_{\mathrm{cell}}} = \sqrt{1 - \frac{2}{n n_s \gamma_e} \sum_{i=1}^{n} \sum_{j=1}^{n_s} \gamma_{ij} + \frac{1}{n_s^2 \gamma_e} \sum_{j=1}^{n_s} \sum_{k=1}^{n_s} \gamma_{jk}}. \tag{10}$$

3.2. Bilinear trend surface removed

A plane of the form

$$\hat{\mu}(\mathbf{x}_i) = a + bx_i + cy_i, \tag{11}$$

is fitted to the sample using regression. A trend of this form would be normally selected in a geotechnical investigation if the site data display a significant trend. In Equation (11), $\hat{\mu}(\mathbf{x}_i)$ is the estimated mean of the random field value of the ith cell having coordinates $\mathbf{x}_i = (x_i, y_i)$, and

$$a = \sum_{j=1}^{n_s} a_j X_j^o, \quad b = \sum_{j=1}^{n_s} b_j X_j^o, \quad c = \sum_{j=1}^{n_s} c_j X_j^o, \tag{12}$$

are the unknowns to the system of equations

$$\begin{bmatrix} n_s & S_x & S_y \\ S_x & S_{xx} & S_{xy} \\ S_y & S_{xy} & S_{yy} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} S^o \\ S^o_x \\ S^o_y \end{bmatrix}, \tag{13}$$

where

$$S_x = \sum_{j=1}^{n_s} x_j,\ S_y = \sum_{j=1}^{n_s} y_j,\ S_{xy} = \sum_{j=1}^{n_s} x_j y_j,$$
$$S_{xx} = \sum_{j=1}^{n_s} x_j^2,\ S_{yy} = \sum_{j=1}^{n_s} y_j^2,\ S^o = \sum_{j=1}^{n_s} X_j^o,$$
$$S_x^o = \sum_{j=1}^{n_s} x_j X_j^o,\ S_y^o = \sum_{j=1}^{n_s} y_j X_j^o, \tag{14}$$

and where $\mathbf{x}_j = (x_j, y_j)$ are the coordinates of the centre of the jth sampled cell. The unknowns a, b, *and* c in the system of equations shown in Equation (13) can be explicitly solved using Cramer's rule as follows:

$$a = \frac{\det(A_1)}{\det(A)},\quad b = \frac{\det(A_2)}{\det(A)},\quad c = \frac{\det(A_3)}{\det(A)}, \tag{15}$$

where A is the 3×3 matrix of coefficients shown in Equation (13) and $A_i (i = 1,\ 2,\ 3)$ is a 3×3 matrix obtained by replacing the ith column of matrix A with the right-hand side vector in Equation (13). Substituting the corresponding matrix determinants into Equation (15) and extracting the coefficients of X_j^o gives the following components:

$$\begin{aligned} a_j &= \{[S_{xx}S_{yy} - S_{xy}^2] - S_x[S_{yy}x_j - S_{xy}y_j] \\ &\quad + S_y[S_{xy}x_j - S_{xx}y_j]\}/\det(A), \\ b_j &= \{-[S_xS_{yy} - S_{xy}S_y] + n_s[S_{yy}x_j - S_{xy}y_j] \\ &\quad + S_y[S_xy_j - S_yx_j]\}/\det(A), \\ c_j &= \{[S_xS_{xy} - S_{xx}S_y] + n_s[S_{xx}y_j - S_{xy}x_j] \\ &\quad - S_x[S_xy_j - S_yx_j]\}/\det(A), \end{aligned} \tag{16}$$

from which the unknowns a, b, c can be found using Equation (12).

Now the residual variance can be calculated by substituting Equation (11) into Equation (4),

$$\begin{aligned} \sigma_r^2 &= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \hat{\mu}(\mathbf{x}_i))^2] = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - a - bx_i - cy_i)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[\left(X_i - \sum_{j=1}^{n_s} a_j X_j^o - \left(\sum_{j=1}^{n_s} b_j X_j^o\right)x_i - \left(\sum_{j=1}^{n_s} c_j X_j^o\right)y_i\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[\left(X_i - \sum_{j=1}^{n_s} d_{ij} X_j^o\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\sum_{j=1}^{n_s} d_{ij}X_j^o + \left(\sum_{j=1}^{n_s} d_{ij}X_j^o\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma_G^2\gamma_e - 2\sum_{j=1}^{n_s} d_{ij}\mathrm{Cov}[X_i, X_j^o] + \sum_{j=1}^{n_s}\sum_{k=1}^{n_s} d_{ij}d_{ik}\mathrm{Cov}[X_j^o, X_k^o]\right\} \\ &= \sigma_G^2\left\{\gamma_e - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s} d_{ij}\gamma_{ij} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{k=1}^{n_s} d_{ij}d_{ik}\gamma_{jk}\right\}, \end{aligned} \tag{17}$$

where

$$d_{ij} = a_j + b_j x_i + c_j y_i, \tag{18}$$

and a_j, b_j, c_j are obtained via Equation (16). As above, the reduction in variability can now be expressed as the dimensionless ratio of standard deviations:

$$\frac{\sigma_r}{\sigma_{\text{cell}}} = \sqrt{1 - \frac{2}{n\gamma_e}\sum_{i=1}^{n}\sum_{j=1}^{n_s} d_{ij}\gamma_{ij} + \frac{1}{n\gamma_e}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{k=1}^{n_s} d_{ij}d_{ik}\gamma_{jk}}. \tag{19}$$

3.3. Best linear unbiased surface removed

BLUE surfaces are commonly used in the mining industry in the form of "Kriging" and often are a reasonable means of estimating geotechnical properties given a set of observations. However, BLUE requires an a-priori knowledge of the site's covariance structure. Various options regarding covariance structure will be considered here. Assuming a zero mean for simplicity, the BLUE of the field at the spatial location $\mathbf{x}_i$ is defined as

$$\hat{\mu}(\mathbf{x}_i) = \sum_{j=1}^{n_s} \beta_{ij} X_j^o, \tag{20}$$

where n_s is the number of samples and X_j^o is the random field value of the jth sample, $j = 1,\ \ldots,\ n_s$. The BLUE coefficients, β_{ij}, are obtained from

$$\beta_{ij} = K_{ij}^{-1} b_i^k, \tag{21}$$

or in matrix form

$$\underset{\sim}{\beta}^i = \underset{\approx}{K}^{-1}\ \underset{\sim}{b}_i^k, \tag{22}$$

where the matrix $\underset{\approx}{K}$, having elements

$$K_{ij} = \mathrm{Cov}[X_i^o, X_j^o] = \sigma_G^2\gamma_{ij},\quad i, j = 1,\ \ldots,\ n_s, \tag{23}$$

is the covariance matrix between samples using the BLUE correlation length, θ_k, in Equations (5)–(7). Note that BLUE uses covariances to estimate the field values at unobserved locations, which requires that a correlation length be specified. Since it is unlikely that the actual correlation length of the field is known on the basis of just the sample, the correlation length used for the BLUE estimation, θ_k, may be different from the actual correlation length, θ.

The vector $\underset{\sim}{b}_i^k$, having elements

$$b_{ij}^k = \mathrm{Cov}[X_i, X_j^o] = \sigma_G^2\gamma_{ij},\quad j = 1,\ \ldots,\ n_s, \tag{24}$$

is the covariance vector between samples and the ith element, again using the BLUE correlation length, θ_k, in Equations (5)–(7).

The residual variance is then calculated by substituting Equation (20) into Equation (4) yielding the following result,

$$\sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \hat{X}_i)^2] = \frac{1}{n}\sum_{i=1}^{n}\{E[X_i^2] - 2E[X_i\hat{X}_i] + E[\hat{X}_i^2]\}$$
$$= \sigma_{\text{cell}}^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\beta_{ij}b_{ij} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{l=1}^{n_s}\beta_{ij}\beta_{il}C_{jl}$$
$$= \sigma_G^2\left\{\gamma_e - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\beta_{ij}\gamma_{ij} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{l=1}^{n_s}\beta_{ij}\beta_{il}\gamma_{jl}\right\}, \quad (25)$$

where C_{ij} is defined in the same fashion as K_{ij},

$$C_{ij} = \text{Cov}[X_i^o, X_j^o] = \sigma_G^2\gamma_{ij}, \; i,j = 1,\ldots,n_s, \quad (26)$$

except that this is the covariance between samples using the actual field correlation length, θ, in Equations (5)–(7). Similarly,

$$b_{ij} = \text{Cov}[X_i, X_j^o] = \sigma_G^2\gamma_{ij}, \quad j = 1, \ldots, n_s, \quad (27)$$

is the covariance vector between samples and the ith element using the actual correlation length, θ, in Equations (5)–(7). Again, the reduction in variability can be expressed as the dimensionless ratio:

$$\frac{\sigma_r}{\sigma_{\text{cell}}} = \sqrt{1 - \frac{2}{n\gamma_e}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\beta_{ij}\gamma_{ij} + \frac{1}{n\gamma_e}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\sum_{l=1}^{n_s}\beta_{ij}\beta_{il}\gamma_{jl}}. \quad (28)$$

If the actual and the BLUE correlation lengths are equal, i.e. if $\theta = \theta_k$, then Equation (25) simplifies to

$$\sigma_r^2 = \sigma_{\text{cell}}^2 - \frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\beta}_i^{\text{T}}\,\underset{\sim}{b}_i$$
$$= \sigma_G^2\left\{\gamma_e - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_s}\beta_{ij}\gamma_{ij}\right\}. \quad (29)$$

The case where $\theta = \theta_k$ gives the best estimate of the field and will be considered to be one of the options for the BLUE surface removed residual field. Note also, that if the entire field is sampled, n_s = all, then the BLUE surface becomes identical to the random field, $X(\mathbf{x}_i)$, and as such the residual field becomes zero everywhere. In other words, for BLUE, the n_s = all case is a perfect representation of the random field and will not be considered further in this paper.

4. Estimation of correlation length

Once $\hat{\mu}(\mathbf{x})$ has been established using the soil samples, the correlation length is estimated as follows:

(1) for each direction through the soil domain, $i = 1, 2$,
(2) estimate the semi-variogram along all lines through the domain in direction i using the entire $X_r(\mathbf{x})$ field. If in any direction, there are m cells, then the semi-variogram is estimated according to

$$\hat{V}(\tau_j) = \frac{1}{2(m-j)}\sum_{i=1}^{m-j}(X_{i+j} - X_i)^2, \quad j = 0, 1, \ldots, m-1, \quad (30)$$

where $\tau_j = j\Delta x$ with Δx being the distance between field cell centres.
(3) The theoretical semi-variogram is defined as

$$V(\tau_j) = \frac{1}{2}E[(X_{i+j} - X_i)^2] = \sigma_{\text{cell}}^2(1 - \rho(\tau_j)). \quad (31)$$

Assuming a Markovian correlation function (see Equation (5)), the theoretical semi-variogram can be written in terms of the correlation length as

$$V(\tau_j) = \sigma_{\text{cell}}^2(1 - e^{(-2|\tau|)/\theta}), \quad (32)$$

(4) fit the theoretical semi-variogram in Equation (32), having parameter θ (correlation length), to the semi-variogram estimated in step 2, also defined in Equation (30), by minimising the sum of squared errors (i.e. regression),

$$E = \sum_{j=1}^{0.9m}(\hat{V}(\tau_j) - V(\tau_j))^2$$
$$= \sum_{j=1}^{0.9m}(\hat{V}(\tau_j) - \sigma_{\text{cell}}^2(1 - e^{(-2|\tau|)/\theta}))^2, \quad (33)$$

with respect to θ and solving for the estimated correlation length $\hat{\theta}$:

$$\hat{\theta} = \frac{\sum_{j=1}^{0.9m}\tau_j^2}{\sum_{j=1}^{0.9m}\tau_j\hat{y}_j}, \quad (34)$$

where

$$\hat{y}_j = -\frac{1}{2}\ln\left(1 - \frac{\hat{V}(\tau_j)}{\hat{\sigma}_{\text{cell}}^2}\right). \quad (35)$$

Note that the sums do not include the estimated semi-variograms at the longest lags, since these have large sampling uncertainty.

5. Results

Consider first the normalised standard deviation of the residual, $X_r(\mathbf{x})$, given by Equations (10), (19), and (28),

and as estimated by simulation. This measure of the remaining uncertainty should decrease as the trend estimate improves. Figure 2 illustrates the effect of actual correlation length and number of samples taken on the normalised residual standard deviation. Plot (c) in Figure 2 is generated

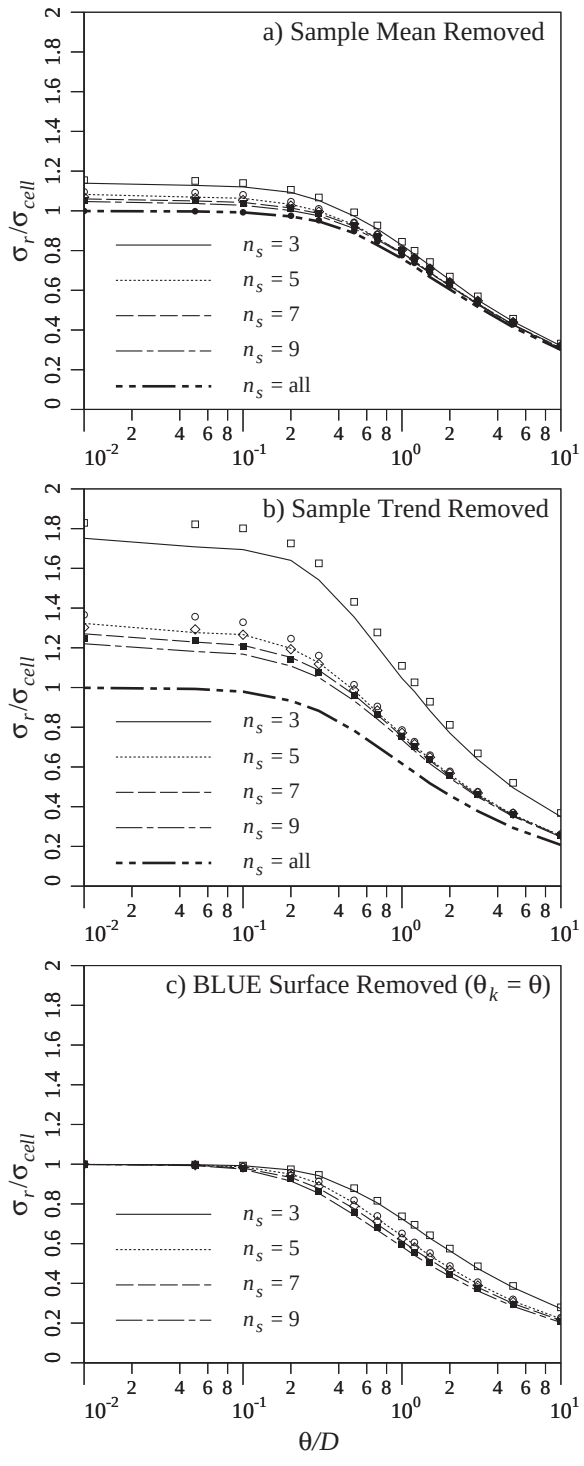

Figure 2. Normalised standard deviation of the residual versus normalised correlation length for (a) sample mean removed, (b) sample bilinear trend removed, and (c) BLUE surface removed. Theoretical results (Equations (10), (19), (28)) are shown as points while simulation results are shown using lines.

for the case where the actual and the BLUE correlation lengths are equal, i.e. $\theta_k = \theta$, which is the best case scenario for the BLUE surface removed since it leads to the most accurate prediction of the random field.

It is apparent from Figure 2 that the agreement between theory and simulation is excellent for the constant sample mean and BLUE methods, although, theory somewhat overestimates simulation in the bilinear trend method for smaller values of θ/D and particularly when $n_s = 3$. This discrepancy may be due to sampling error or a possible bias in the estimator $\hat{\sigma}_r^2$. However, the agreement improves as $\theta/D \to \infty$. In all cases, the error is less than about 5%, so that the theory is sufficiently accurate to replace the simulation for all three methods considered.

Figure 2 shows that the ability of $\hat{\mu}(\mathbf{x})$ to represent $X(\mathbf{x})$ improves as the actual correlation length increases. In the limit, as $\theta/D \to \infty$, all random fields become uniform (under the assumed finite variance correlation structure); i.e. random from realisation to realisation, but constant within each realisation. In this limiting case, the sample perfectly predicts the uniform field, and the residual becomes zero everywhere so that $\sigma_r = 0$. It is apparent in Figure 2 that all curves are heading towards 0, as $\theta/D \to \infty$.

Figure 3 illustrates the effect of the type of trend removed on the residual uncertainty for $n_s = 3$ and $n_s = 9$. The BLUE surface is obtained in two ways: first by assuming a fixed correlation length, $\theta_k/D = 1.0$, and second by assuming that the BLUE correlation length is equal to the actual field correlation length, $\theta_k = \theta$, as in Figure 2(c), which is a best case scenario.

One of the perhaps surprising results of Figure 3 is that the removal of a bilinear trend is not nearly as good as the removal of the constant sample mean and BLUE surface at smaller correlation lengths, and especially at a lower number of samples. The reason for this becomes apparent when, for example, the case where $n_s = 3$ is considered. If the correlation length is small, then the three samples will be largely independent, and the resulting fitted bilinear plane could (and often does) end up with quite an unrepresentative slope, leading to a high variability in the residual. Even when $n_s = 9$ the residual variability is higher at low correlation lengths than seen using the constant sample mean. The performance of the bilinear trend might be improved by choosing different sampling locations, perhaps more spread out, but this idea was not investigated in this paper. At larger correlation lengths, for example, above about $\theta/D = 1$, the bilinear trend does start to show slightly better performance than the constant sample mean at higher numbers of samples, but the difference is slight.

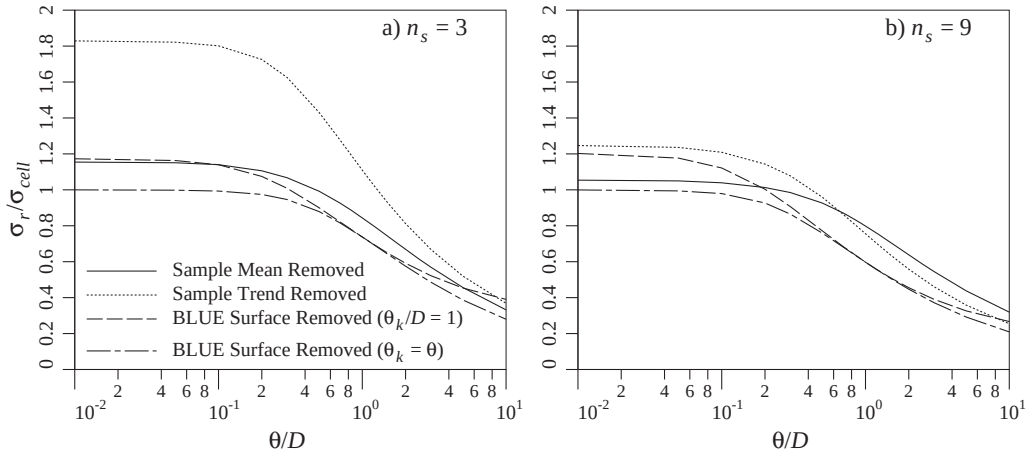

Figure 3. Normalised standard deviation of the residual versus normalised correlation length using theory (Equations (10), (19), (28)) for (a) $n_s = 3$ and (b) $n_s = 9$.

The BLUE surface outperforms the other two methods when the BLUE correlation length is set equal to the actual field length (ideal case). However, the difference between the BLUE surface and the constant sample mean approaches is not large even at small correlation lengths, where more of a difference might have been expected.

In real situations where only a handful of samples are taken from a site, the actual field correlation length is unlikely to be known. In this case, the BLUE correlation length must either be estimated from the sample or assumed, perhaps from the literature about similar sites, and almost certainly the actual and the BLUE correlation lengths will be different. Figure 4 shows the normalised residual standard deviation for fixed BLUE correlation lengths of $\theta_k/D = 0.2$, 1.0, and 2.0. It is apparent from Figure 4 that the agreement between theory and simulation is excellent for $\theta_k/D = 1.0$, and 2.0. When $\theta_k/D = 0.2$ the agreement degrades somewhat, with theory being larger than simulation, when θ/D becomes large. The largest error seen is less than 15%, however, so that theory can be used reasonably accurately in place of simulation.

When $\theta/D \leq 0.2$, the lowest normalised standard deviation of about 1 is obtained when $\theta_k/D = 0.2$. The lowest normalised standard deviation ranges from 0.7 to 0.9 when $\theta_k/D = 1.0$ and when $0.2 < \theta/D \leq 1.0$, and is as low as 0.2 when $\theta_k/D = 2.0$ and $\theta/D > 1.0$. In other words, when the actual correlation length is small, the smallest normalised residual standard deviation is achieved if the BLUE correlation length is also selected to be small, which makes sense. At the other end of the spectrum, when the actual correlation length is large, the lowest normalised residual standard deviation overall is achieved when the BLUE correlation length is also selected to be large. A reasonable compromise appears to be to select $\theta_k/D = 1.0$ since this choice gives reasonably low normalised residual standard deviation over the whole range of actual correlation lengths considered.

Figure 5 shows how the assumed value of the BLUE correlation length affects the residual variability. Three cases are considered; $\theta_k/D = 0.2$, $\theta_k/D = 2.0$, and the ideal case where $\theta_k = \theta$. Clearly, the ideal case gives the lowest residual variability over the whole range of actual correlation lengths. As implied previously, when the actual correlation length is small, the $\theta_k/D = 0.2$ and the ideal case are similar. For large actual correlation lengths, the $\theta_k/D = 2.0$ and the ideal case are similar. The effect of the number of samples is relatively minor at small correlation lengths but larger n_s leads to lower residual variability at larger correlation lengths (in agreement with the findings of Lloret-Cabot, Hicks, and Van Den Eijnden 2012). In other words, using BLUE, there is little advantage to increasing the sample size if the actual correlation length is small.

A second measure of the quality of the trend type considered in this paper is how well the estimated correlation length agrees with the actual correlation length, as shown in Figure 6. Figure 6 is based solely on simulation results and the BLUE results do not include n_s = all, since that case leads to the ideal residual of zero everywhere whose correlation length is undefined.

The correlation length estimated from the residual, θ_r, will agree with the actual correlation length used in the simulation, θ, when the ratio $\theta_r/\theta \approx 1$ and the curve approaches a diagonal line. In Figure 6(a,b), which are the sample mean and sample trend removed cases, it can be seen that this agreement only occurs when the entire field is sampled and the correlation length is relatively small (i.e. significantly less than D). In other words, when the entire field is sampled (n_s = all), the estimated correlation length becomes approximately equal to the

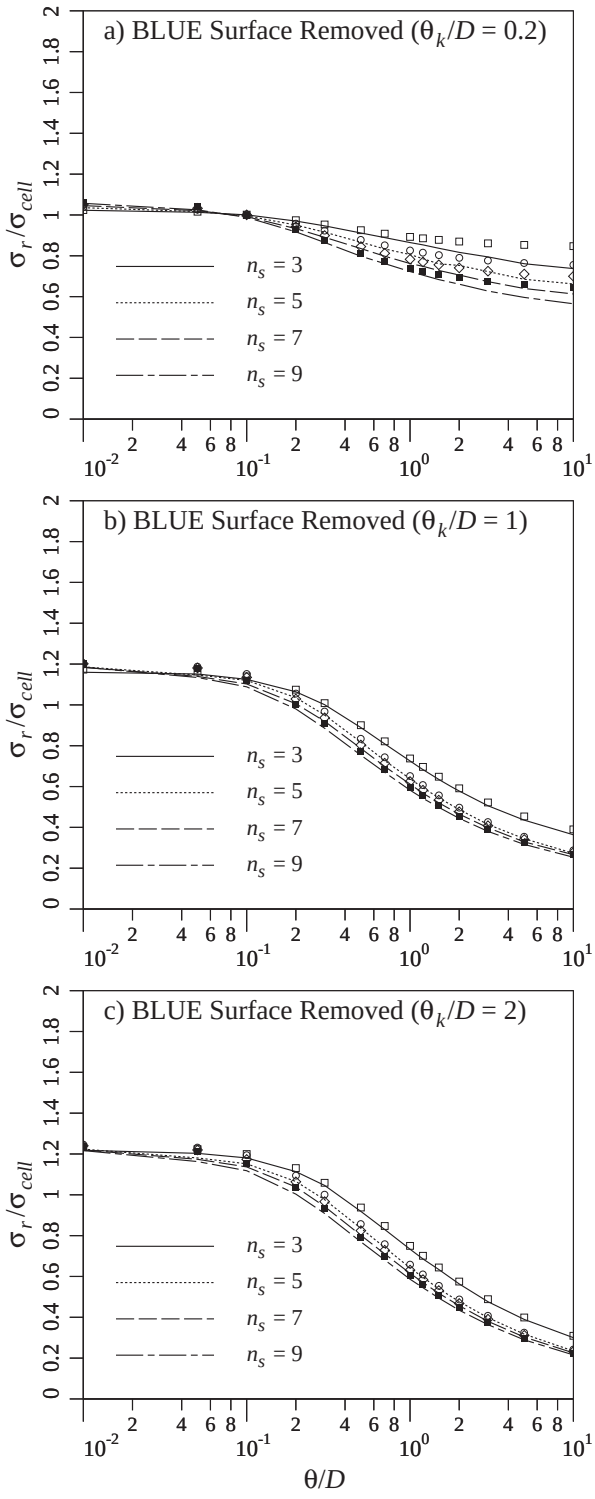

Figure 4. Normalised standard deviation of the residual versus normalised correlation length for BLUE surface removed using fixed $\theta_k/D = 0.2$ in (a), 1.0 in (b) and 2.0 in (c). Theoretical results (Equation (28)) are shown as points while simulation results are shown using lines.

actual correlation length when samples are relatively independent (small θ).

Note that the correlation length estimated from the simulated residual, θ_r, is actually the correlation length acting between local averages, while θ is the point-wise correlation length of the Gaussian random field, G. The correlation length acting between local averages is expected to be somewhat higher than the point-wise correlation length because correlation between averages is generally higher than correlation between points, especially at smaller correlation lengths. This can be seen in Figure 6(a–c) for the best cases (n_s = all and $\theta_k = \theta$) at very small correlation lengths, where the estimate is reasonably accurate and unbiased, by the fact that θ_r remains somewhat higher than θ.

In general, when $n_s <$ all, θ_k is fixed, and $\theta/D < 1$, the correlation length is overestimated, and often considerably overestimated, especially when the actual correlation length is very small. This occurs because errors between the estimated trend and actual field trend are perceived in the estimation process to be caused by a strong lingering correlation – hence a longer correlation length is estimated to account for the apparent residual trend. For example, the constant mean estimated from $n_s = 3$ samples will almost certainly not be equal to the actual field average. The resulting residual field will have a non-zero average suggesting a longer correlation length – field values tend to be either all above or all below the assumed field mean of zero. As it turns out, the estimated correlation length seems to approach the distance between the samples when $\theta/D < 1$. The reason for this is under investigation.

For larger correlation lengths, i.e. when $\theta/D > 1$, the estimated correlation length is generally less than the actual correlation length due to estimator bias (Fenton and Griffiths 2008). Interestingly, when $\theta_k/D = 0.2$ in the BLUE case, the longer correlation lengths are actually more accurately estimated. Why this is so is also under investigation.

The effect of number of samples on the correlation length estimate is minor in the case of the BLUE approach and not particularly large in the sample mean and bilinear trend removed approaches. In other words, over the range of the number of samples considered in Figure 1, there is no particular sampling scheme which results in a significant improvement in the accuracy of the estimated correlation length.

In summary, the best practical approach to estimating the correlation length seems to be using BLUE with a small value of θ_k/D.

Figure 7 shows in more detail the effect of the number of samples and sampling methodology on the estimated correlation length. The general effect of increasing the number of samples taken is to reduce the estimated correlation length (see Figure 7(a,b)). This makes sense, since an increase in the number of samples improves the trend estimate, resulting in a more "independent"

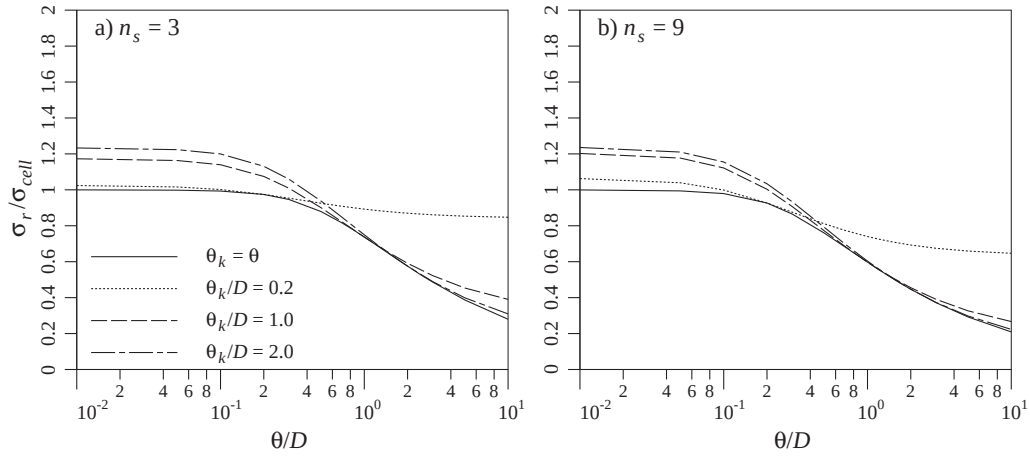

Figure 5. Normalised standard deviation of the residual versus normalised correlation length for BLUE surface removed (Equation (28)) for $n_s = 3$ in (a) and $n_s = 9$ in (b).

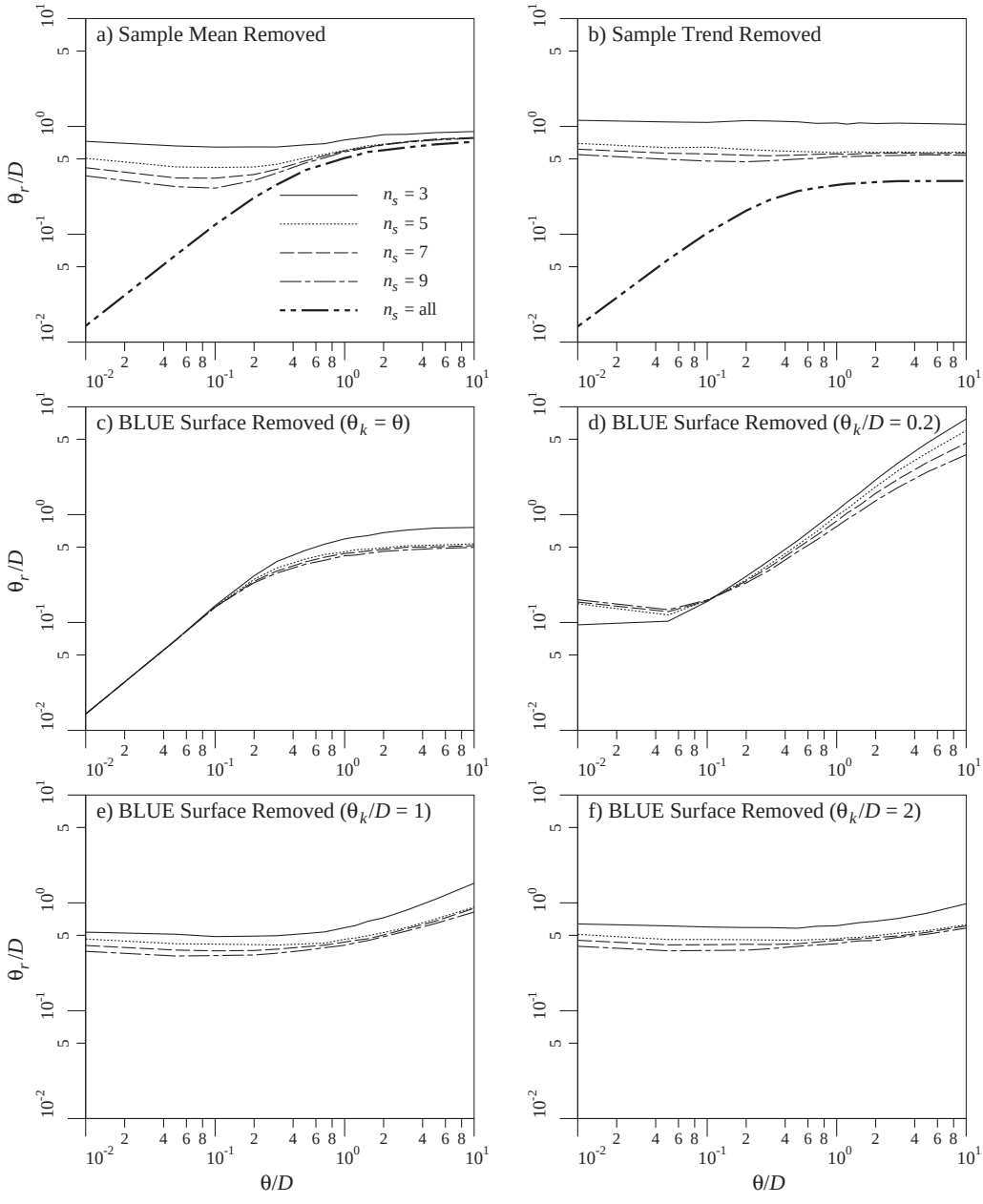

Figure 6. Simulation-based estimated correlation length of the residual versus actual random field correlation length for each method.

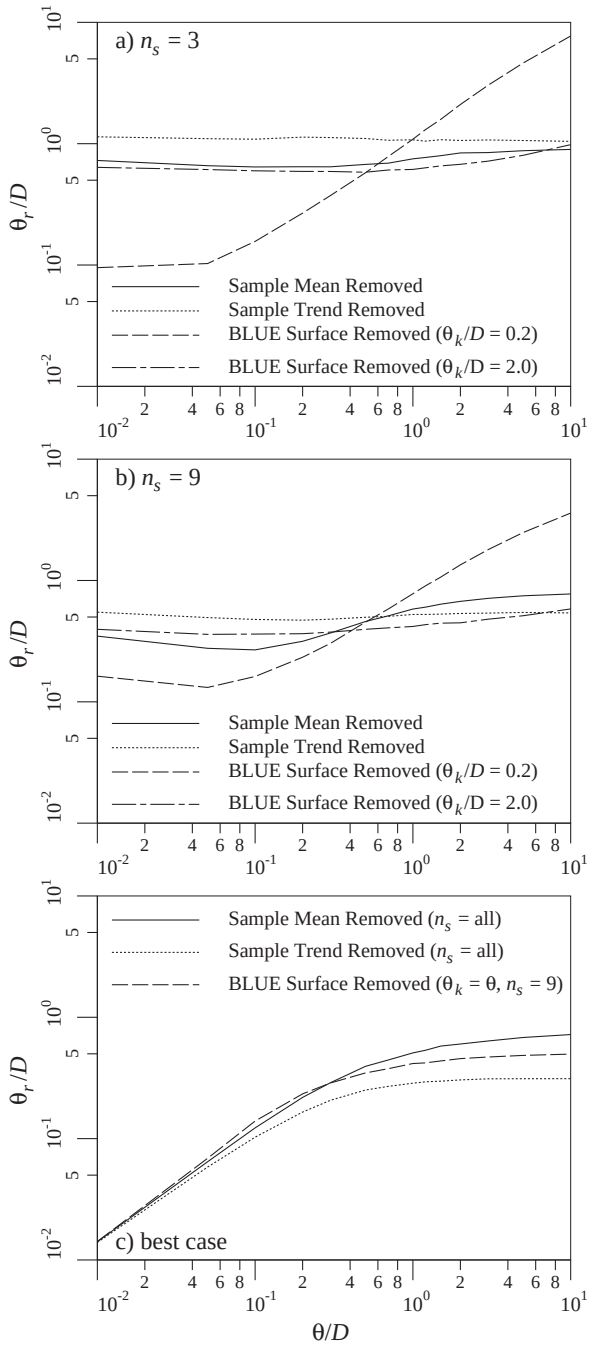

Figure 7. Simulation-based estimated correlation length of the residual versus actual random field correlation length for n_s= 3, 9, and best case.

residual field. The difference between the various methods is not particularly large, except for the case of BLUE with $\theta_k/D = 0.2$, which produces the best estimate of the correlation length. Figure 7(c) illustrates the accuracy of the correlation length estimate in the best cases (n_s = all for sample mean and trend removed methods and $\theta_k = \theta$ for BLUE). This plot again reflects the fact that the correlation length can only be accurately estimated when $\theta/D << 1$.

6. Conclusions

The paper investigates the effect of number of samples and type of trend removal on residual uncertainty. The basic goal is to study how our uncertainty in a geotechnical site investigation is best reduced, taking into account the investigation intensity.

Probably, the most important measure of the value of site investigation is the magnitude of the residual field standard deviation after a trend estimated from the site sample has been removed. Figure 3 suggests that more samples reduces uncertainty when the field correlation length is small, but does not have much impact when the field correlation length is large. The BLUE trend removed approach under the ideal conditions where the assumed correlation length equals the actual correlation length ($\theta_k = \theta$) outperforms all of the other methods considered. However, it is unrealistic to expect that the assumed correlation length will equal the true correlation length. The best compromise appears to be taking $\theta_k/D = 1$, which can be seen from Figure 5 to lead to good variance reduction at high correlation lengths and intermediate variance reduction at small correlation length. The BLUE approach using $\theta_k/D = 1$ matches the sample mean removed case for small correlation lengths and $n_s = 3$, although, the sample mean removed case is better than BLUE when $n_s = 9$, as seen in Figure 3. In general, this last observation suggests that for small correlation lengths, the optimum trend is the sample mean for any reasonable number of samples. At longer correlation lengths, all considered methods are similar, with the BLUE approach using $\theta_k/D = 1$ slightly in the lead for larger numbers of samples. This observation suggests that BLUE is better to use if the actual correlation length is large relative to the domain size.

As mentioned above, a larger number of samples leads to lower residual variability at smaller correlation lengths, which can be seen in Figures 2 –4, but does not seem to make that much difference at larger correlation lengths. In other words, if the actual correlation length is large, there is no particular advantage to increasing the number of samples above, say, $n_s = 3$.

Figure 6 shows that correlation length is not well estimated except under ideal cases where n_s = all and/or $\theta_k = \theta$, which are not very realistic, nor practical. Figure 7 provides the rather surprising observation that the overall best correlation length estimate is obtained when the BLUE trend with $\theta_k/D = 0.2$ is removed from the random field.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors are thankful for the support provided by the Natural Sciences and Engineering Research Council of Canada and by Delft University of Technology.

References

Ching, J., and K. K. Phoon. 2017. "Characterizing Uncertain Site-specific Trend Function by Sparse Bayesian Learning." *Journal of Engineering Mechanics* 143 (7): 04017028.

Fenton, G. A., and D. V. Griffiths. 2008. *Risk Assessment in Geotechnical Engineering*. Hoboken, NJ: John Wiley & Sons.

Fenton, G. A., and E. H. Vanmarcke. 1990. "Simulation of Random Fields via Local Average Subdivision." *Journal of Engineering Mechanics* 116 (8): 1733–1749.

Jaksa, M. B., J. S. Goldsworthy, G. A. Fenton, W. S. Kaggwa, D. V. Griffiths, Y. L. Kuo, and H. G. Poulos. 2005. "Towards Reliable and Effective Site Investigations." *Geotechnique* 55 (2): 109–121.

Li, Y. J., M. A. Hicks, and P. J. Vardon. 2016. "Uncertainty Reduction and Sampling Efficiency in Slope Designs Using 3D Conditional Random Fields." *Computers and Geotechnics*, 79: 159–172.

Lloret-Cabot, M., M. A. Hicks, and A. P. Van Den Eijnden. 2012. "Investigation of the Reduction in Uncertainty Due to Soil Variability When Conditioning a Random Field Using Kriging." *Géotechnique Letters* 2: 123–127.

Yang, R., J. Huang, D. V. Griffiths, and D. Sheng. 2017. "Probabilistic Stability Analysis of Slopes by Conditional Random Fields". Georisk 2017, Denver, Colorado, 4–7 June.