Integrated framework for characterization of spatial variability of geological profiles

W.F. Liu[1], Y.F. Leung[1*], and M.K. Lo[1]

[1]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong

[*]Corresponding author, email address: yfleung@polyu.edu.hk

ABSTRACT

Despite recent efforts to characterize the uncertainties involved with geological profiles and soil and rock properties, there has been limited study on their spatial correlations and how such features may be included in the engineering decision-making process. This paper presents an integrated framework for geostatisical analyses, which incorporates the Restricted Maximum Likelihood (REML) method with the Matérn autocovariance model. Statistical tests are conducted including those for data normality, constant variance and outliers, which ensure the fundamental assumptions of REML are not violated in the residual analyses of site data, meanwhile offering simple checks for potential errors in the dataset. The proposed approach also allows quantification of uncertainties in the subsurface profiles at the unsampled locations. The approach is illustrated through investigations on spatial correlation features of geological profiles at two project sites in Hong Kong. The numbers of irregularly-spaced boreholes vary from 150 to 350 in the two cases, and the large volume of data enables the variations in rockhead levels to be studied through the proposed framework. In addition, the existence of geological faults in one of the sites is found to significantly affect the spatial variability of rockhead level, as indicated by the reduced scales of fluctuation and spatial dependence, which corresponds to increased uncertainty in areas intersected by faults.

Keywords: Site investigation, Restricted Maximum Likelihood method, Matérn covariance structure, Residual analysis, Rockhead variation

INTRODUCTION

Uncertainties in soil profiles and their properties are often the cause of geotechnical problems encountered during construction. For example, Clayton (2001) conducted a survey of 28 construction projects in the United Kingdom, which revealed that many geotechnical problems encountered during construction stemmed from uncertainties regarding boundaries of the soil strata (22%) and properties of the geo-materials (20%). However, studies on the spatial variability or correlation of soil properties have been hampered by the lack of data. Christian and Baecher (2011) stated that the "unresolved problems in geotechnical risk and reliability" included uncertainties in the variability and spatial correlations of geotechnical properties.

DeGroot (1996) compiled the results from a number of earlier studies, and reported the correlation distances of geotechnical properties including undrained shear strengths and CPT cone tip resistance. Phoon and Kulhawy (1999a,b) also reported the scales of fluctuation of various soil properties, without describing the details of spatial correlation structure. The correlation structure of geotechnical data is sometimes analyzed through the autocorrelation (Vanmarcke 1977; DeGroot and Baecher 1993) or by geostatistics (Matheron 1971), as illustrated in the works by Soulié et al. (1990), Chiasson et al. (1995) and Wang and Chiasson (2006), etc. In these analyses, it is common for researchers to assume certain functional form for the autocorrelation structure (e.g., Gaussian or spherical), and then estimate the parameters for the assumed function (Elkateb et al. 2003; Phoon et al. 2004; Stuedlein et al. 2012; Firouzianbandpey et al. 2014). However, the validity of such assumption have not been discussed in detail. Alternatively, Bayes' Theorem can be applied to determine the most probable correlation function of the spatial data through weighting their posterior probabilities, as illustrated by Cao and Wang (2013, 2014) and Wang et al. (2010, 2013, 2014, 2015, 2016), who adopted the approach in characterizing underground soil stratification and variability of geotechnical properties.

Contrary to the Bayesian approach, this paper presents an integrated framework which ensures the available geotechnical data is best utilized in rigorous statistical analyses. For example, Phoon et al. (2003) stated that stationarity is an important prerequisite for geostatistical analyses, and proposed the use of Modified Bartlett test statistic as a basis to reject the null hypothesis of stationarity. However, many previous studies did not verify stationarity in the data. Also, a constant mean for the residuals is a necessary condition for stationarity, and certain fixed polynomial order is usually assumed in the detrending process (Stuedlein et al. 2012). For example, Liu and Leung (2015) presented the preliminary analyses of the spatial data of geological profiles assuming quadratic and cubic trend structures, but stationarity assumption was not confirmed in the analyses.

The current study proposes a new integrated framework and procedures that incorporate data transform and rigorous residual analyses to ensure stationarity assumptions are satisfied, thereby enhancing the reliability of residual analysis. The framework also enables rational detrending process with the optimal polynomial order, and detection of outliers in the dataset which are not considered in previous attempts to characterize ground variability. The Matérn function (Matérn 1960) is adopted to model the autocorrelation structures, owing to the flexibility of its functional form. Parameters of the function are optimized using a heuristic algorithm, known as the Differential Evolution (Storn and Price 1997), to maximize the log-likelihood value under the Restricted Maximum Likelihood (REML) method. The proposed framework can be used to evaluate the spatial variations of soil and rock strata, obtaining parameters such as the spatial dependence and scale of fluctuation using site-specific information. To illustrate the capabilities of the approach, information on the engineering rockhead level (moderately weathered granite) from irregularly-spaced boreholes at two sites in Hong Kong is analyzed to reveal their spatial variability characteristics. This study also discusses the impacts of the existence of geologic features, such as faults, on the variability of geological profiles.

PROPOSED FRAMEWORK FOR RESIDUAL ANALYSIS

Spatial random variables are often expressed as a combination of fixed effects and random effects, also known as the deterministic trend structure and the residual effects. With $\boldsymbol{x}$ representing the spatial coordinates of sampled points, a general linear mixed regression model for spatial data, $\boldsymbol{z}(\boldsymbol{x})$, can be formulated by:

$$\boldsymbol{z}(\boldsymbol{x}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{X}\boldsymbol{\beta}$ represents the large scale trend, with $\mathbf{X}$ being the deterministic component matrix that contains information on spatial coordinates. $\boldsymbol{\beta}$ is the vector of regression coefficients according to the corresponding trend structure (linear, quadratic, cubic, etc.). The residual, $\boldsymbol{\varepsilon}$, is a combination of the correlation structure (with smooth scale variation of variance σ_e^2) and a white noise process (with variance σ_n^2), since white noise effects are assumed not to correlate with distance. The covariance matrix of $\boldsymbol{\varepsilon}$ is related to the correlation structure, $\mathbf{R}$, by:

$$\begin{aligned}&\mathbf{V} = \mathbf{Var}(\boldsymbol{\varepsilon}) = \sigma_e^2\mathbf{R} + \sigma_n^2\mathbf{I} = \left(\sigma_e^2 + \sigma_n^2\right)\left[s\mathbf{R} + (1-s)\mathbf{I}\right] \\ &\text{where } \ 0 \leq s = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_n^2} \leq 1\end{aligned} \tag{2}$$

In Eq. (2), $\mathbf{I}$ is the identity matrix, and s is the spatial dependence which incorporates the nugget effect (due to white noise) into the covariance model. Previously, the correlation structure, i.e., individual components of $\mathbf{R}$, is often assumed to follow a certain fixed function, such as the Gaussian, exponential or spherical function (DeGroot and Baecher 1993). In the current study, in order to allow flexibility in the functional form of $\mathbf{R}$, the Matérn function is adopted as follows (Matérn 1960):

$$R(h_{ij}) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{h_{ij}}{r}\right)^{\nu} K_{\nu}\left(\frac{h_{ij}}{r}\right) \tag{3}$$

where h_{ij} is the separation distance between points i and j, ν is a smoothness parameter ranging from 0 to infinity, r is the range parameter, Γ is the gamma function, and K_ν

represents the modified Bessel function of the second kind with order ν.

The Matérn function is a generalized function with its shape controlled by the smoothness parameter. For example, it corresponds to the exponential function when $\nu = 0.5$, and is equivalent to the Gaussian function when ν approaches infinity (Minasny and McBratney 2005). The scale of fluctuation, δ, of the Matérn function is determined by both ν and r. In this paper, it is taken as the separation distance where the autocorrelation, $sR(h_{ij})$, equals $0.05s$ (Elkateb et al. 2003; Rue and Held 2005), and Fig. 1 shows the relationship between ν, r and δ accordingly. Also, it should be noted that some researchers proposed a different form for the Matérn function (Stein 1999), but the resulting estimates of scales of fluctuation are essentially the same.

The framework proposed in this study consists of three key components, namely REML analysis with Box-Cox transformation, trend structure determination and statistical tests for residuals. These components, particularly the latter two, have not been considered in previous studies such as Haskard (2007) or existing software such as ArcGIS and geoR. Fig. 2 shows a flowchart of the framework, and the three components are discussed in the following sections.

Restricted Maximum Likelihood (REML)

To ensure stationarity of the spatial data, it is important to estimate and remove the trend (or fixed effects, $\mathbf{X}\boldsymbol{\beta}$), so that the spatial correlation features are not masked by this deterministic component. In some previous studies, the trend component is determined by regression analysis using linear or polynomial functions (Dasaka and Zhang 2012; Lark et al. 2006), and the residuals are then analysed and presented using method of moments or semivariograms. However, the semivariance estimates are not unique when the samples are irregularly spaced, as the semivariance can be affected by subjective decisions on the lag size (bin size). Also, the subsequent semivariograms are estimated through a subjective curve-fitting process.

In the current study, the Restricted Maximum Likelihood (REML) method is applied to simultaneously determine the trend coefficients and estimate the autocorrelation properties of residuals. The method does not require decisions on the bin size so it can be applied to irregularly-spaced sampling points. An important assumption in the development of REML methods is that the data follows a normal distribution, and that the variance of residuals is constant throughout the domain. These assumptions are often made, but rarely verified, in most previous studies. In the current work, the Box-Cox transformation (Box and Cox 1964) is performed on the raw dataset, $\boldsymbol{z}^*$, to ensure these assumptions are satisfied. The transformed dataset, $\boldsymbol{z}$, can be represented by the following equation:

$$z_i = \begin{cases} \dfrac{(z_i^* + \lambda_2)^{\lambda_1} - 1}{\lambda_1 (gm(\boldsymbol{z}^* + \boldsymbol{\lambda}))^{(\lambda_1 - 1)}} & \text{if } \lambda_1 \neq 0 \\ (gm(\boldsymbol{z}^* + \boldsymbol{\lambda})) \log{(z_i^* + \lambda_2)} & \text{if } \lambda_1 = 0 \end{cases} \tag{4}$$

where $\boldsymbol{\lambda}$ is a vector with all the terms equal to λ_2, which is a parameter used to ensure $z_i^* + \lambda_2 > 0$. λ_1 is estimated by minimizing the residual sum of squares (RSS) of $\boldsymbol{z}$, and $gm(\cdot)$ denotes the geometric mean of the vector $\boldsymbol{z}^* + \boldsymbol{\lambda}$.

Details of the REML approach have been described in Cressie and Lahiri (1996) and Lark and Cullis (2004). In short, the autocorrelation structure can be obtained by maximizing the following log-likelihood function with respect to $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{n-p}{2}\log{(2\pi)} - \frac{1}{2}\log{|\mathbf{V}|} - \frac{1}{2}\log{|\mathbf{W}|} - \frac{1}{2}\boldsymbol{y}^T\mathbf{V}^{-1}\mathbf{Q}\boldsymbol{y} \tag{5}$$

where $\mathbf{W} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$ and $\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{W}^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. $\boldsymbol{\theta}$ represents the unknown quantities in the autocorrelation structure, i.e., s, ν and r in Eqs. (2) and (3). In Eq. (5), n is the number of data points, and p is the number of coefficients in the trend structure. $\boldsymbol{y} = (\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\boldsymbol{z}$, which is the matrix of filtered dataset with the trend components filtered out. Therefore, the covariance estimates of REML are independent of the trend estimates.

The determination of $\boldsymbol{\theta}$ can be treated as an optimization problem, aiming to obtain the set of $\{s, \nu, r\}$ parameters that maximize the log-likelihood function. In the current work, this is achieved using the Differential Evolution algorithm (Storn and Price 1997). This is conceptually similar to other evolutionary algorithms, which is not prone to converging at local maxima, and has recently been applied in a number of engineering problems.

Once the covariance structure is determined, the trend coefficients, and subsequently the predicted residuals, can be estimated using generalized least squares (GLS):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{V}^{-1}\boldsymbol{z} \tag{6}$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{z} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{7}$$

The predictions at unsampled locations, $\hat{\boldsymbol{z}}(\boldsymbol{x}_0)$, and the corresponding prediction variance, $\boldsymbol{\sigma}^2_{\boldsymbol{z}}(\boldsymbol{x}_0)$, can be estimated based on the Best Linear Unbiased Prediction (BLUP) technique (Atkinson et al. 2008; Santra et al. 2012):

$$\hat{\boldsymbol{z}}(\boldsymbol{x}_0) = \mathbf{X}_0{}^T\hat{\boldsymbol{\beta}} + \mathbf{K}^T\mathbf{V}^{-1}\hat{\boldsymbol{\varepsilon}} \tag{8}$$

$$\boldsymbol{\sigma}^2_{\boldsymbol{z}}(\boldsymbol{x}_0) = \mathrm{diag}(\mathbf{K}_0 - \mathbf{K}^T\mathbf{V}^{-1}\mathbf{K} + \mathbf{M}^T(\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^T)^{-1}\mathbf{M}) \tag{9}$$

where $\mathbf{X}_0$ is the deterministic component matrix of prediction. $\mathbf{K}$ represents the covariance matrix between observations and predictions, i.e., $\mathbf{K} = \mathrm{cov}\{\boldsymbol{z}(\boldsymbol{x}), \boldsymbol{z}(\boldsymbol{x}_0)\}$, $\mathbf{K}_0 = \mathrm{cov}\{\boldsymbol{z}(\boldsymbol{x}_0), \boldsymbol{z}(\boldsymbol{x}_0)^T\}$ and $\mathbf{M} = \mathbf{X}_0 - \mathbf{X}\mathbf{V}^{-1}\mathbf{K}$.

It should be noted that the predictions and associated variance evaluated by Eqs. (8) and (9) correspond to values in the ‘transformed’ space, under Box-Cox transformation. While $\hat{\boldsymbol{z}}(\boldsymbol{x}_0)$ can be back-transformed to the original space through Eq. (4), back-transformation of the prediction variance can be approximated by multiplying $\boldsymbol{\sigma}^2_{\boldsymbol{z}}(\boldsymbol{x}_0)$ with a factor Φ (Cressie 1993). For unsampled location i:

$$\Phi_i = (\lambda_1 gm(z)^{\lambda_1 - 1}(\lambda_1 gm(z)^{\lambda_1 - 1}(\mathbf{X}_0\hat{\boldsymbol{\beta}})_i + 1)^{\frac{1}{\lambda_1} - 1})^2 \tag{10}$$

As will be shown in the following case study, the distributions of prediction variances in the transformed and back-transformed (original) spaces are broadly similar.

Regression diagnostics for REML

Normality

The Box-Cox transformation (Eq. (4)) minimizes the recovered residuals but does not guarantee normality – the approach assumes that the transformed data has the highest likelihood to be normally distributed when RSS value of $\boldsymbol{z}$ is minimized. It is therefore necessary to perform diagnostics for normality to ensure the assumption of REML is not violated.

Traditional diagnostics for normal errors in regression typically utilize ordinary residuals, based on uncorrelated linear mixed model. The residuals after GLS process, however, are correlated spatially, according to the autocovariance model effects. Therefore, the GLS residuals need to be converted to recovered residuals before executing the normality diagnostics. The current study applied the Kolomogrov-Smirnov (KS) test to diagnose the normality of residuals (Smirnov 1939; Jensen and Ramirez 1999). The KS test evaluates the maximum deviation between the cumulative distribution of the recovered residuals and that of a theoretical normal distribution. The P_N value is defined as the tail probability for this deviation to be small enough for the data to be considered normally distributed. In the current study, a P_N value exceeding 0.05 is considered to satisfy the normality assumption.

Constant variance

After removing the deterministic trend component ($\mathbf{X}\boldsymbol{\beta}$), the residuals ($\boldsymbol{\varepsilon}$) are assumed to be stationary in the REML formulation, which implies a constant variance across the domain. However, the validity of this assumption has rarely been verified in previous applications of geostatistical methods.

The Breusch-Pagan Test (Breusch and Pagan 1979) is frequently used in statistics to verify the constant variance assumption. During the test, a regression is conducted on the squared residuals with the explanatory variables (i.e. across the spatial coordinates in this case), and it checks whether a trend exists in the variance. The null hypothesis, H_0, is defined by all regression coefficients of the variances being zero. The P_C value is defined as the tail probability for the regression coefficients to be considered insignificant. In the current study, a P_C value exceeding 0.05 corresponds to acceptance of H_0, and the data is considered to satisfy the constant variance assumption.

The original Breusch-Pagan test is based on residuals obtained from ordinary least squares (OLS), where spatial correlation is absent. To apply this test to the general linear mixed model, the spatial correlation effect has to be removed from the residuals. This is achieved in the current study by multiplying the negative square root of the covariance matrix to the GLS residuals:

$$\boldsymbol{\varepsilon}^* = \mathbf{V}^{-\frac{1}{2}}\hat{\boldsymbol{\varepsilon}} = \mathbf{P}\,\mathbf{O}^{-\frac{1}{2}}\mathbf{P}^T\hat{\boldsymbol{\varepsilon}} \tag{11}$$

where $\boldsymbol{\varepsilon}^*$ is Pearson residuals for constant variance test, $\hat{\boldsymbol{\varepsilon}}$ is predicted residuals calculated by Eq. (7), $\mathbf{P}$ is the square matrix containing the eigenvectors of covariance matrix $\mathbf{V}$, and $\mathbf{O}$ is the diagonal matrix containing the eigenvalues of $\mathbf{V}$. The Breusch-Pagan test is then applied to $\boldsymbol{\varepsilon}^*$ of the uncorrelated model.

Detection of potential outliers

Outliers are data points that vary significantly from the neighboring points in the spatial context, and can be indications of peculiarities or errors in the dataset which influence geostatistical analyses. For example, Lark (2000) stated that the maximum likelihood method was susceptible to asymmetry caused by outliers. There are multiple methods for detection of outliers, and this is achieved by two approaches in the current study. The first approach examines the distribution of residuals ($\hat{\boldsymbol{\varepsilon}}$) in the spatial domain, and the data points with residuals exceeding ±1.96 times their standard deviation ($\sqrt{\sigma_e^2 + \sigma_n^2}$) are identified as potential outliers. This represents the 95% inter-percentile range assuming the residuals follow a normal distribution, and those outside this range may be considered ‘extreme’ values of deviations from the trend.

The second approach evaluates the Cook's distance (Cook 1977) of each data point, which is an indicator of its influence to regression results. The Cook's distance is based on the difference between regression coefficients estimated with all the observations, i.e., $\hat{\boldsymbol{\beta}}$; and coefficients estimated without a particular observation i, i.e., $\hat{\boldsymbol{\beta}}(i)$ (Haslett and Hayes 1998). This difference is often termed **DFBETA**, and is referred to as $\boldsymbol{D}$ herein for simplicity. A large value of $\boldsymbol{D}_{ji}$ suggests that the $i\,^{th}$ data point is influential in determining the $j\,^{th}$ regression coefficient, which may indicate an outlying data point. $\boldsymbol{D}_{ji}$ can be estimated by:

$$\begin{aligned} \boldsymbol{D}_{ji} &= \frac{\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j(i)}{\sqrt{\mathbf{Var}(\hat{\boldsymbol{\beta}})_{jj}}} = \frac{\mathbf{B}_{ji}\;\widetilde{\boldsymbol{\varepsilon}}_i}{\sqrt{\mathbf{Var}(\hat{\boldsymbol{\beta}})_{jj}}} \\ \text{where}\;\; \mathbf{B} &= \mathbf{W}^{-1}\mathbf{X}^T\mathbf{V}^{-1} \\ \widetilde{\boldsymbol{\varepsilon}} &= \left[\mathbf{I}\;\mathrm{diag}(\mathbf{E})\right]^{-1}\mathbf{E}\,\boldsymbol{z} \\ \mathbf{E} &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\,\mathbf{W}^{-1}\mathbf{X}^T\mathbf{V}^{-1} \end{aligned} \tag{12}$$

where diag($\mathbf{E}$) is a vector consisting of the diagonal components of $\mathbf{E}$. $\boldsymbol{D}_{ji}$ measures the change of the $j\,^{th}$ individual regression coefficient when the $i\,^{th}$ observation is deleted, scaled by the variance of $\hat{\boldsymbol{\beta}}$, which is $\mathbf{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{W}^{-1}$, and the subscript jj indicates the $j\,^{th}$ diagonal element of $\mathbf{Var}(\hat{\boldsymbol{\beta}})$. Therefore, the $\boldsymbol{D}_i$ vector summarizes the changes in all regression coefficients resulting from deleting the i^{th} observation.

The Cook's distance is defined as the average of the squared $\boldsymbol{D}_{ji}$ components, which is proportional to the squared length of the $\boldsymbol{D}_i$ vector. For example, the Cook's distance for the $i\,^{th}$ observation point can be expressed as:

$$C_i = \frac{1}{p}\sum_{j=1}^{p}\boldsymbol{D}_{ji}^2 = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j(i)}{\sqrt{\mathbf{Var}(\hat{\boldsymbol{\beta}})_{jj}}}\right]^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^T(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))}{p\;\mathbf{Var}(\hat{\boldsymbol{\beta}})_{jj}} \tag{13}$$

where p is the number of regression coefficients. Belsley et al. (2005) suggested $2/\sqrt{n}$ as the cutoff value for $\boldsymbol{D}_{ji}$ for outlier diagnostics. Correspondingly, the cutoff value for C_i can be taken as $(2/\sqrt{n})^2 = 4/n$ (Nieuwenhuis et al. 2012). In other words, an observation point i is classified as an outlier if $C_i > 4/n$.

The necessity and implementation of the two approaches for outlier detection will be illustrated through the NTK case study described in later sections.

Determination of polynomial order of trend structure

The REML approach, together with GLS, allow determination of the trend coefficients $\hat{\boldsymbol{\beta}}$. However, the polynomial order of the trend structure (i.e. the size of $\hat{\boldsymbol{\beta}}$ vector) is often a subjective decision of the analyst. A higher order trend will fit the data better and hence reduce the residuals and their variance. On the other hand, an ever-increasing trend flexibility may lead to overfitting of the data, which means random noise and errors are included in the statistical model, sacrificing its predictive power. Previous researchers have proposed the Akaike's information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (Schwarz 1978) as a means for model selection, both of which compare the changes in log-likelihood value with respect to the number of associated model parameters. More recently, Beck (2010) proposed the Bayesian system identification approach which evaluates the expected information gain for individual model class. In the current study, the proposed framework incorporates objective criteria to determine the optimal polynomial order for the trend structure, balancing the needs for regression diagnostics, the significance of the high order coefficients and the predictive power of the model.

Significance of trend coefficients

The current work adopts statistical hypothesis testing to assess the significance of trend coefficients, in order to avoid overfitting of data. The basic idea is to test whether the highest order trend coefficients are statistically different from zero, by calculating the Wald statistics which are tested against the F-distribution. Two competing hypotheses are defined as: null hypothesis H_0 – all highest order regression coefficients are zero; and alternative hypothesis H_1 – at least one regression coefficient is nonzero, so that H_1 is the complement of H_0. A P_F value can be computed by comparing the Wald statistics to a $F(m,n\text{-}p)$ distribution, where m is the number of the highest order coefficients. The highest order coefficients are statistically significant with the acceptance of H_1, indicated by a P_F value smaller than 0.05. Before executing statistical hypothesis testing, it is important to ensure the data follow a normal distribution, hence the necessity of the above-mentioned regression diagnostics.

Leave-one-out cross validation

The predictive power of a model can be evaluated through assessing the accuracy of its estimates by the “leave-one-out cross validation” method, which is performed by removing one observation at one time from the dataset, and then predicting its value using the remaining data (Haslett and Hayes 1998; Haslett 1999). During this process, the trend coefficients ($\boldsymbol{\beta}$) and variance parameters ($\sigma_e^2 + \sigma_n^2$) may change with removal of each data point. The spatial correlation model, however, is assumed to remain constant, so there is no attempt to re-evaluate the correlation structure using REML. In this study, the cross validation scores, S_{cv}, is formulated based on the stacked vector for prediction errors, $\widetilde{\boldsymbol{\varepsilon}}$, as below:

$$S_{cv} = \frac{(\widetilde{\boldsymbol{\varepsilon}})^T \widetilde{\boldsymbol{\varepsilon}}}{n} \tag{14}$$

S_{cv} defined herein can be interpreted as the average of squared prediction error at each borehole location under leave-one-out-cross validation. The advantage of this approach is that it does not require data partitioning, and hence minimizes perturbation to the data. It therefore provides an asymptotically unbiased estimate of the prediction errors, and is attractive for the purposes of model selection (Cawley and Talbot 2003).

IMPLEMENTATION OF THE INTEGRATED FRAMEWORK

Incorporating the components discussed in previous sections, the proposed framework ensures that the assumptions of REML are satisfied in geostatistical analyses of the data. Meanwhile, it allows objective determination of the order of trend polynomial and identification of potential outliers in the dataset. The predictive power of the model is also assessed through the leave-one-out cross validation method. While Fig. 2 shows the flowchart outlining the framework, its implementation can be summarised as follows:

1. The analysis starts with a linear trend structure (order $i = 1$). The autocovariance structure ($\mathbf{V}$) and trend coefficients ($\hat{\boldsymbol{\beta}}$) are estimated by REML (with Box-Cox transformation) and GLS. Normality and constant variance checks are performed on the residuals.
2. A polynomial order i is rejected if either the normality or constant variance conditions is violated. The analysis is then repeated with a $(i+1)^{th}$ order polynomial for the trend structure ($i = i+1$).
3. If both normality and constant variance conditions are satisfied for the residuals, i becomes a potential candidate. The significance of trend coefficients, $P_F(i)$, and cross validation scores, $S_{cv}(i)$, will be evaluated for polynomial order i.
4. Meanwhile, the same analyses will be performed also on $(i+1)^{th}$ order polynomial to obtain $P_F(i+1)$ and $S_{cv}(i+1)$. If $P_F(i+1)$ indicates non-significant polynomial order or $S_{cv}(i) < S_{cv}(i+1)$, then i is the optimal order for the trend structure.
5. If $P_F(i+1)$ indicates a significant polynomial order, with $S_{cv}(i) > S_{cv}(i+1)$ and normality and constant variance conditions are satisfied, then order $i+1$ replaces i as the potential candidate ($i = i+1$), and Step (4) onwards will be repeated.
6. Once the optimal polynomial order is determined for the trend structure, the $\mathbf{V}$ and $\hat{\boldsymbol{\beta}}$ estimates become final. Outliers in the dataset are also determined.

CASE STUDIES

Study regions and site descriptions

The proposed integrated framework is applied to analyze the spatial variability of engineering rockhead levels at two project sites in Hong Kong, namely the Ngau Tau Kok (NTK) site and the Cheung Wang Estate (CWE) site. Borehole information at the two sites was obtained from geotechnical investigation reports of previous government projects in the areas, which were archived in the Civil Engineering Library maintained by the Civil Engineering and Development Department of the Hong Kong Government. For both cases, the boreholes are irregularly spaced across the site, and the focus of the analysis is on the level of moderately decomposed granite, referred to as Grade III material (GEO 1988) and commonly taken as the rockhead level in the local practice.

Table 1 summarizes the sampling information for the two cases, including the areas of the project sites and the sample sizes (i.e., number of boreholes). In addition, previous geotechnical investigation had revealed the existence of a geological fault across the site of CWE. To understand the effects of faults on spatial variability features, two sub-regional blocks were extracted from CWE, with the fault crossing Block 1 but not Block 2. Details of the analyses will be presented in the following sections where the benefits of the proposed framework are also illustrated. It should be noted that the proposed approach will not replace conventional geotechnical investigation techniques in identifying rockhead level or existence of fault zones. However, it provides 'added value' to existing borehole information by revealing the spatial characteristics and uncertainties regarding these geological features.

NTK study site

The NTK site is located in the eastern part of Hong Kong, where data from 150 boreholes have been collected over an area of 650 m × 450 m. Spatial variations on the level of engineering bedrock, i.e., Grade III moderately weathered granite at the site, are analyzed using the proposed framework outlined in previous sections. To illustrate the importance of residual diagnostics, and to elucidate the effects of assumptions on trend structures, Table 2 compares the analyses with different polynomial orders for the trend, with and without the Box-Cox transformation. It should be noted that the proposed framework already incorporates the selection criteria without the need to individually examine and compare each separate analysis. The main purpose of Table 2 is to shed insights through the comparisons and allow meaningful discussions on the significance of the proposed framework.

For illustration purposes, two series of analyses were performed, the first with Box-Cox transformation of the raw data of rockhead level, and the second without. For each series, the order of trend structure was varied from $i = 1$ (linear trend) to $i = 4$ (quartic trend) in the REML analyses. Normality and constant variance tests were then conducted based on the recovered residuals and Pearson residuals, respectively, with P_N and P_C values exceeding 0.05 indicating satisfaction of these conditions, as described earlier.

Table 2 shows that normality condition is in fact satisfied in all cases. However, without Box-Cox transformation of the data, the condition of constant variance is not satisfied with any trend structure. In other words, the REML (or other geostatistical) analyses are not representative in those cases as the variance $(\sigma_e^2 + \sigma_n^2)$ changes across various locations of the domain. This issue will be discussed again in later sections on trend structure selection.

According to the proposed framework, the cubic trend structure $(i = 3)$ was adopted since its residuals satisfied the normality and constant variance tests, and it produced better prediction than the 4^{th} order polynomial based on the leave-one-out cross validation scores. With this optimal trend structure, the autocovariance structure for rockhead variations is estimated by REML and shown in Fig. 3(b). The estimates by method of moments are also provided for comparison purposes, and the two methods produce similar results of autocovariance structures.

Using Eqs. (8) and (9), predictions can be made at unsampled locations and the corresponding prediction variances (uncertainties) can be quantified, as shown in Fig. 3(c). The prediction variance $(\boldsymbol{\sigma}_{\boldsymbol{z}}^2)$ may be interpreted as the confidence level in the estimated rockhead levels at unsampled locations, which varies spatially across the site according to the autocorrelation structure and locations of existing boreholes. It contains contributions from both uncertainties in deterministic trend structure and the corresponding residual effects. In general, the prediction variance is low near sampled locations and increases with distance away from boreholes. Such contour can provide guidance to determine the locations of additional sampling points (if necessary), in order to achieve a specific level of confidence in the predictions.

To illustrate the validity of $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ estimates, the leave-one-out cross validation method is again applied, where the prediction error $(\widehat{\varepsilon})$ at location x is normalized by $\sigma_z(x)$, estimated with $z(x)$ removed from the dataset. Fig. 3(d) shows the histogram of such normalized prediction errors for all observation points, which broadly follows the standard normal distribution curve. This implies that $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ provides reasonable estimates on the prediction uncertainties at unsampled locations. While Figs. 3(c) and (d) show the results in the transformed space, similar patterns are observed for the back-transformed prediction variance by Eq. (10), and the associated normalized prediction errors, as presented in Figs. 3(e) and (f).

Influence of trend structure determination

Table 2 shows that the autocorrelation structure is highly influenced by the polynomial order of the trend structure. In the current study, the optimal trend is selected with considerations on the residual diagnostics, significance of trend coefficients (P_F), and leave-one-out cross validation scores (S_{cv}), which shows whether over-fitting of the data has occurred.

In general, a high order trend tends to match the existing observation points ($\boldsymbol{z}$) more closely, therefore reducing the magnitudes of residuals ($\boldsymbol{\varepsilon}$) and their variances. Closer examination on Table 2 also reveals that higher polynomial orders are associated with smaller scales of fluctuation (δ) and spatial dependence values (s). This is because with an increasing polynomial order i, the trend structure becomes more flexible and the effective range of residuals becomes shorter. Also, with increasing i, the smooth scale variation (σ_e^2) is reduced due to better 'fitting' of the existing data. Meanwhile, the white noise effects (σ_n^2) are also reduced since high order polynomials tend to 'absorb' some of the random noise in measurements. As σ_e^2 decreases at a greater rate than σ_n^2, the s value also reduces with increasing i.

To further illustrate the importance of residual diagnostics in trend order determination, Fig. 4 compares the diagnostics of recovered residuals under linear and cubic trends. The histograms of recovered residuals show that normality conditions are satisfied for both trend orders. With a sufficient sample size, the complexity of trend structure does not seem to affect the normality of residuals.

On the contrary, constant variance tests based on the two trend structures produce different results, with the cubic trend–but not the linear trend–satisfying constant variance conditions. Plots of Pearson residuals along the north-south (N-S) and east-west (E-W) directions are shown in Fig. 4. With a linear trend structure, the magnitudes of residuals gradually diverge along both the N-S and E-W directions, implying that a potential trend is still hidden in the residuals which masks the correlation features of the data, even after filtering the linear trend component. On the other hand, the Pearson residuals under cubic trend structure uniformly distribute around the zero axis along the two directions, indicating that no significant trend exists among the residuals. The REML analyses are therefore representative since $\sigma_e^2 + \sigma_n^2$ can be considered as constant throughout the study domain or project site.

Potential outliers at NTK site

In the current study, outliers are defined as the data points with large deviations from the trend (based on their residuals), and those that significantly influence the trend structure (based on their Cook's distances). By implementing the proposed integrated framework, eight outliers are identified with the optimal (cubic) trend structure.

Figs. 5(a) and (b) show the locations of outliers identified by the two approaches. Four outliers (No. $1 - 4$) are identified based on the Cook's distances, and Fig. 5(c) shows the magnitudes of the residuals for two of them, compared to their neighboring points. When examining the residuals, significant differences can be observed between the outliers and their neighbors, which explain their substantial influence on the regression coefficients and hence large values of Cook's distances. However, such influence may not be obvious by only examining their corresponding raw data values. This shows that although outliers are associated with significantly different residuals, they may not be easily detected when the sample size is large, such as in the NTK site with 150 borehole records.

A potential deficiency of this approach is that the Cook's distances can be affected by the leverage effect: a data point near the edge of the sampling domain tends to demonstrate a higher influence on the regression coefficients than those near the central region. To supplement the Cook's distance method, the residual values are also examined in the current study, and those exceeding ±1.96 times their standard deviation are also considered as potential outliers (Fig. 5(d)), as they represent 'extreme' values outside the 95% inter-percentile range (assuming normal distribution). It should be noted that among those data points, two clusters are identified at NTK, where groups of boreholes spatially close to each other have similar values of residuals. They may be manifestations of local rockhead variations that are not captured by the large-scale trend, instead of results of measurement errors. The proposed framework thereby allows such details to be revealed so that engineers and geologists can focus on a small number of potential outliers to ensure accuracy and consistency of the dataset.

In addition, Table 2 shows that the number of potential outliers are affected by the adopted trend structure. In general, with a higher order polynomial, the trend involves greater flexibility and hence a larger number of 'influential' data points may be identified as outliers. In many cases, the data points identified as outliers using the low order trend are also outliers under higher order trend structure. The current approach is established to automatically identify these statistical influential or extreme points, so they can be reviewed again by engineers or geologists to determine whether they indeed contain measurement errors or mistakes.

CWE study site and effects of faults

The second study site is located at the Cheung Wang Estate (CWE) on the Tsing Yi Island in Hong Kong. A total number of 321 borehole records were obtained within an area of 800 m × 500 m, and the variations of Grade III moderately weathered granite was studied. At the CWE site, a geological fault has been reported from previous geotechnical investigation. Effects of the fault on the spatial variability of rockhead level are also evaluated in the current study.

Geological faults often form discontinuities in the rockhead level, which may have significant implications on the design and construction of an engineering project. As shown in Fig. 6(a) and (c), one NW-SE fault cuts through the western part of the CWE site. To understand the influence of the fault, two sub-regional blocks were extracted from the CWE site, with the same sample domain size and similar sampling densities. The borehole locations and partition scheme of the blocks are also illustrated in Fig. 6(a). Block 1 is designed to be intersected by the fault, while Block 2 is deemed to be free of its influence.

Table 3 compares the spatial variation features of the two blocks. At the CWE study site, the fault is associated with reductions in scale of fluctuation (δ) (about 50%) and in spatial dependence (s) (about 20%) , which imply higher levels of uncertainties in the rockhead levels. The differences in the two autocorrelation structures and prediction variances are also shown in Figs. 6(b) and (d), respectively. Intuitively, the existence of geological faults or other discontinuities at the site will increase the uncertainty in the subsurface profiles. Analyses by the proposed framework provide a quantitative evaluation of such effects, which may then be coupled with risk analyses by reliability methods.

DISCUSSIONS

The framework proposed in the current study ensures that spatial correlation analyses performed on geotechnical data satisfy the fundamental assumptions of REML and are statistical sound. A key feature of the framework is the methodological and objective determination of the optimal trend structure. As shown in Table 2 and discussed by Lark and Webster (2006), residual analyses can be substantially affected by the choice of trend structure. The proposed framework involves holistic considerations on the significance of trend coefficients and variance distributions of the residuals, which lead to rational decisions on the trend component in the analyses.

The proposed framework also offers simple and automatic detection of potential outliers or errors in the dataset, through the Cook's distances of the data points and distribution of their residuals. Automatic detection of outliers is especially beneficial in the case of a large dataset, where anomalies may not be easily identified manually. The potential outliers identified using the current approach can be reviewed again by engineers or geologists, who can then determine whether they indeed involve measurement errors or human mistakes.

Using the BLUP technique (Eqs. (8)) and (9)), contours of the prediction variances can be produced to quantify the level of confidence in predictions at unsampled locations. This can form a useful guidance to determine necessity and/or locations of additional sampling. In addition, the predicted properties and prediction variance from BLUP can be used to construct a conditional random field (Li et al. 2016; Lo and Leung 2016), to be adopted for probabilistic geotechnical models by Random Finite Element Method (RFEM)(Fenton and Griffiths 2003; Griffiths et al. 2009). By quantifying the spatial uncertainty around the observed data points, predictions of the probabilistic models can be more precise than those using an unconditional random field.

It should be noted that the presented case studies involve large numbers of boreholes (150 to 350), and the current study aims to fully utilize such information to demonstrate the proposed framework and reveal spatial correlation features of the rockhead profile. This, however, does not imply that the approach is only applicable to such sample sizes. While any statistical analysis will improve with a large sample size, the proposed method will also produce more robust results in smaller dataset than traditional method of moments or maximum likelihood methods, due to the rigorous consideration of stationarity requirements, detrending and detection of outliers.

To illustrate the robustness of the proposed framework, Block 1 of the CWE case is taken as an example where 100 subsets are extracted, each containing 50% of data points randomly chosen from the original dataset. These 100 subsets are analysed using both the proposed framework and the approach in Liu and Leung (2015), which consists of REML but not the other key features of this study such as data transformation, regression diagnostics, trend order determination and outlier detection. Fig. 7 compares the statistics of the two series of analyses, and shows that the proposed framework produces closer estimates of spatial dependence and scale of fluctuation compared to results from the complete dataset, and are associated with smaller variances which indicate more robust analyses. In addition, the cross validation scores are generally lower under the proposed framework. Fig. 7(d) also shows an analysis on one subset using the method of moments with different lag sizes. Traditional method of moments does not include simultaneous determination of the large scale trend, so in this case the trend is adopted from REML analysis. Even so, estimates by the method of moments are shown to be dependent on subjective decisions on lag size and curve-fitting for the spatial correlation parameters.

CONCLUSION

This paper presents an integrated framework for geostatistical analyses, incorporating the REML method with the Matérn autocovariance model, to estimate the spatial correlation features of rockhead levels. The approach is a robust technique which includes efficient determination of optimal trend structure and identification of spatial outliers, meanwhile ensuring the basic premises of REML, including assumptions on normality and constant variance of residuals, are satisfied across the study region.

The framework is demonstrated through analyses on the spatial variations of Grade III rockhead levels using borehole data from two sites in Hong Kong. As illustrated in the CWE case, geological faults can have significant influences on the spatial variability features of rockhead levels. In particular, the scale of fluctuation and spatial dependence reduce with existence of faults, which corresponds to higher spatial uncertainty. It is recommended that sub-regional analyses be performed separately whenever local geological features are identified at the project site.

ACKNOWLEDGEMENTS

The work presented in this paper is financially supported by the Research Grants Council of the Hong Kong Special Administrative Region (Project No. 25201214). Also, the authors would like to acknowledge the permission of the Civil Engineering and Development Department, the Government of Hong Kong Special Administrative Region, to present analyses of data obtained from the Civil Engineering Library. Readers may contact the corresponding author for details of the dataset used in the study.

REFERENCES

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Atkinson, P. M., Pardo-Iguzquiza, E., and Chica-Olmo, M. 2008. "Downscaling cokriging for super-resolution mapping of continua in remotely sensed images." *Geoscience and Remote Sensing, IEEE Transactions on*, 46(2), 573–580.

Beck, J. L. 2010. "Bayesian system identification based on probability logic." *Structural Control and Health Monitoring*, 17(7), 825–847.

Belsley, D. A., Kuh, E., and Welsch, R. E. 2005. *Regression diagnostics: Identifying influential data and sources of collinearity*, Vol. 571. John Wiley & Sons.

Box, G. E. P. and Cox, D. R. 1964. "An analysis of transformations." *Journal of the Royal Statistical Society, Series B (Methodological)*, 26(2), 211–252.

Breusch, T. S. and Pagan, A. R. 1979. "A simple test for heteroscedasticity and random coefficient variation." *Econometrica: Journal of the Econometric Society*, 1287–1294.

Cao, Z. and Wang, Y. 2013. "Bayesian approach for probabilistic site characterization using cone penetration tests." *Journal of Geotechnical and Geoenvironmental Engineering*, 139(2), 267–276.

Cao, Z. and Wang, Y. 2014. "Bayesian model comparison and selection of spatial correlation functions for soil parameters." *Structural Safety*, 49, 10 – 17 Special Issue In Honor of Professor Wilson H. Tang.

Cawley, G. C. and Talbot, N. L. 2003. "Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers." *Pattern Recognition*, 36(11), 2585–2592.

Chiasson, P., Lafleur, J., Soulié, M., and Law, K. T. 1995. "Characterizing spatial variability of a clay by geostatistics." *Canadian Geotechnical Journal*, 32(1), 1–10.

Christian, J. T. and Baecher, G. B. 2011. "Unresolved problems in geotechnical risk and reliability." *ASCE GSP 224: Geo-Risk 2011: Risk Assessment and Management*, 50–63.

Clayton, C. R. I. 2001. *Managing geotechnical risk: improving productivity in UK building and construction*. Thomas Telford.

Cook, R. D. 1977. “Detection of influential observation in linear regression.” *Technometrics*, 19(1), 15–18.

Cressie, N. and Lahiri, S. N. 1996. “Asymptotics for reml estimation of spatial covariance parameters.” *Journal of Statistical Planning and Inference*, 50(3), 327–341.

Cressie, N. A. C. 1993. *Statistics for Spatial Data (Revised Edition)*. John Wiley & Sons.

Dasaka, S. M. and Zhang, L. M. 2012. “Spatial variability of in situ weathered soil.” *Géotechnique*, 62(5), 375–384.

DeGroot, D. J. 1996. “Analyzing spatial variability of in situ soil properties.” *Uncertainty in the Geologic Environment: From Theory to Practice*, Vol. 1, 210–238.

DeGroot, D. J. and Baecher, G. B. 1993. “Estimating autocovariance of in-situ soil properties.” *Journal of Geotechnical Engineering*, 119(1), 147–166.

Elkateb, T., Chalaturnyk, R., and Robertson, P. K. 2003. “An overview of soil heterogeneity: quantification and implications on geotechnical field problems.” *Canadian Geotechnical Journal*, 40(1), 1–15.

Fenton, G. A. and Griffiths, D. V. 2003. “Bearing-capacity prediction of spatially random c-ϕ soils.” *Canadian Geotechnical Journal*, 40(1), 54–65.

Firouzianbandpey, S., Griffiths, D. V., Ibsen, L. B., and Andersen, L. V. 2014. “Spatial correlation length of normalized cone data in sand: case study in the north of denmark.” *Canadian Geotechnical Journal*, 51(8), 844–857.

GEO 1988. *GEOGUIDE 3: Guide to Rock and Soil Descriptions*. Geotechnical Engineering Office, Civil Engineering Department, Government of Hong Kong.

Griffiths, D., Huang, J., and Fenton, G. 2009. “Influence of spatial variability on slope reliability using 2-d random fields.” *Journal of Geotechnical and Geoenvironmental Engineering*, 135(10), 1367–1378.

Haskard, K. A. 2007. *An anisotropic Matérn spatial covariance model: REML estimation and properties*. Ph.D. thesis, The University of Adelaide.

Haslett, J. 1999. "A simple derivation of deletion diagnostic results for the general linear model with correlated errors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 603–609.

Haslett, J. and Hayes, K. 1998. "Residuals for the linear model with general covariance structure." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 201–215.

Jensen, D. R. and Ramirez, D. E. 1999. "Recovered errors and normal diagnostics in regression." *Metrika*, 49(2), 107–119.

Lark, R. 2000. "Estimating variograms of soil properties by the method-of-moments and maximum likelihood." *European Journal of Soil Science*, 51(4), 717–728.

Lark, R., Cullis, B., and Welham, S. 2006. "On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (e-blup) with reml." *European Journal of Soil Science*, 57(6), 787–799.

Lark, R. M. and Cullis, B. R. 2004. "Model-based analysis using reml for inference from systematically sampled data on soil." *European Journal of Soil Science*, 55(4), 799–813.

Lark, R. M. and Webster, R. 2006. "Geostatistical mapping of geomorphic variables in the presence of trend." *Earth Surface Processes and Landforms*, 31(7), 862–874.

Li, X., Zhang, L., and Li, J. 2016. "Using conditioned random field to characterize the variability of geologic profiles." *Journal of Geotechnical and Geoenvironmental Engineering*, 142(4), 04015096.

Liu, W. F. and Leung, Y. F. 2015. "Analyzing spatial variability of geologic profiles for four sites in Hong Kong." *The 5th International Symposium on Geotechnical Safety and Risk, Rotterdam, The Netherlands*, 157–162.

Lo, M. K. and Leung, Y. F. 2016. "Bayesian updating of subsurface spatial correlation through monitoring of infrastructure and building developments." *International Conference on Smart Infrastructure and Construction, Cambridge, United Kingdom*, (Accepted).

Matérn, B. 1960. "Stochastic models and their application to some problems in forest surveys and other sampling investigations." *Meddelanden från Statens Skogsforskningsinstitut*, 49(5), 1–144.

Matheron, G. 1971. *The theory of regionalized variables and its application.* Fontainebleau, France.

Minasny, B. and McBratney, A. B. 2005. "The Matérn function as a general model for soil variograms." *Geoderma*, 128(3-4), 192–207.

Nieuwenhuis, R., te Grotenhuis, H., and Pelzer, B. 2012. "Influence. me: tools for detecting influential data in mixed effects models." *The R Journal*, 38–47.

Phoon, K. K. and Kulhawy, F. H. 1999a. "Characterization of geotechnical variability." *Canadian Geotechnical Journal*, 36(4), 612–624.

Phoon, K. K. and Kulhawy, F. H. 1999b. "Evaluation of geotechnical property variability." *Canadian Geotechnical Journal*, 36(4), 625–639.

Phoon, K.-K., Quek, S.-T., and An, P. 2003. "Identification of statistically homogeneous soil layers using modified bartlett statistics." *Journal of Geotechnical and Geoenvironmental Engineering*, 129(7), 649–659.

Phoon, K.-K., Quek, S.-T., and An, P. 2004. "Geostatistical analysis of cone penetration test (cpt) sounding using the modified bartlett test." *Canadian Geotechnical Journal*, 41(2), 356–365.

Rue, H. and Held, L. 2005. *Gaussian Markov Random Fields: Theory and Applications.* CRC Press, Taylor and Francis.

Santra, P., Das, B. S., and Chakravarty, D. 2012. "Spatial prediction of soil properties in a watershed scale through maximum likelihood approach." *Environmental Earth Sciences*, 65(7), 2051–2061.

Schwarz, G. 1978. "Estimating the dimension of a model." *Ann. Statist.*, 6(2), 461–464.

Smirnov, N. 1939. “On the estimation of the discrepancy between empirical curves of distribution for two independent samples.” *Bulletin Mathématique de l’Université de Moscou*, 2, 3–14.

Soulié, M., Montes, P., and Silvestri, V. 1990. “Modeling spatial variability of soil parameters.” *Canadian Geotechnical Journal*, 27(5), 617–630.

Stein, M. L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer.

Storn, R. and Price, K. 1997. “Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces.” *Journal of Global Optimization*, 11(4), 341–359.

Stuedlein, A. W., Kramer, S. L., Arduino, P., and Holtz, R. D. 2012. “Geotechnical characterization and random field modeling of desiccated clay.” *Journal of Geotechnical and Geoenvironmental Engineering*, 138(11), 1301–1313.

Vanmarcke, E. H. 1977. “Probabilistic modeling of soil profiles.” *ASCE Journal of Geotechnical Engineering Division*, 103(GT11), 1227–1246.

Wang, Y., Au, S.-K., and Cao, Z. 2010. “Bayesian approach for probabilistic characterization of sand friction angles.” *Engineering Geology*, 114(34), 354 – 363.

Wang, Y., Cao, Z., and Li, D. 2016. “Bayesian perspective on geotechnical variability and site characterization.” *Engineering Geology*, 203, 117 – 125 Special Issue on Probabilistic and Soft Computing Methods for Engineering Geology.

Wang, Y., Huang, K., and Cao, Z. 2013. “Probabilistic identification of underground soil stratification using cone penetration tests.” *Canadian Geotechnical Journal*, 50(7), 766–776.

Wang, Y., Huang, K., and Cao, Z. 2014. “Bayesian identification of soil strata in london clay.” *Géotechnique*, 64(3), 239–246.

Wang, Y., Zhao, T., and Cao, Z. 2015. “Site-specific probability distribution of geotechnical properties.” *Computers and Geotechnics*, 70, 159 – 168.

Wang, Y.-J. and Chiasson, P. 2006. “Stochastic stability analysis of a test excavation involving spatially variable subsoil.” *Canadian Geotechnical Journal*, 43(10), 1074–1087.

List of Tables

TABLE 1. Domain scales and sample sizes for two cases

Case study		Area of domain	Sample size
NTK		650 m × 450 m	150
CWE	Block 1 (with fault)	360 m × 400 m	172
	Block 2 (no fault)	360 m × 400 m	149

TABLE 2. Comparisons of spatial correlation analyses for NTK site

Case	Trend order (i)	Spatial dependence (s)	Scale of fluctuation (δ)	Constant variance test (P_C)	Normality test (P_N)	Trend coefficients test (P_F)	No. of potential outliers	Cross validation score (S_{cv})
NTK (transformed)	1	0.93	308 m	0.0010(N)	0.3041(Y)	0.0183(Y)	1	27.49
	2	0.89	227 m	0.0102(N)	0.7123(Y)	0.2226(N)	4	28.42
	3	0.75	125 m	0.2700(Y)	0.7250(Y)	2.71×10^{-6}(Y)	8	31.95
	4	0.58	70 m	0.5995(Y)	0.3459(Y)	3.30×10^{-10}(Y)	10	33.50
NTK (raw data)	1	0.94	273 m	0.0005(N)	0.5540(Y)	0.0088(Y)	5	28.48
	2	0.90	215 m	0.0008(N)	0.5908(Y)	0.2384 (N)	8	29.29
	3	0.74	128 m	0.00001(N)	0.3063(Y)	8.17×10^{-6}(Y)	10	32.99
	4	0.65	96 m	0.0004(N)	0.2752(Y)	2.11×10^{-6}(Y)	15	35.44

TABLE 3. Effects of geological faults on spatial correlation features

CWE sub- regional block	Optimal trend order	Scale of fluctuation (δ)	Spatial dependence (s)
Block 1 (with fault)	2 (quadratic)	74 m	0.53
Block 2 (no fault)	2 (quadratic)	143 m	0.64

List of Figures

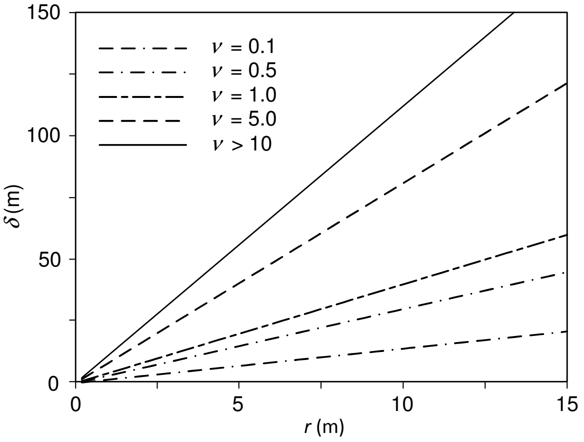


FIG. 1. Relationship between scale of fluctuation, δ, and Matérn function parameters, ν and r

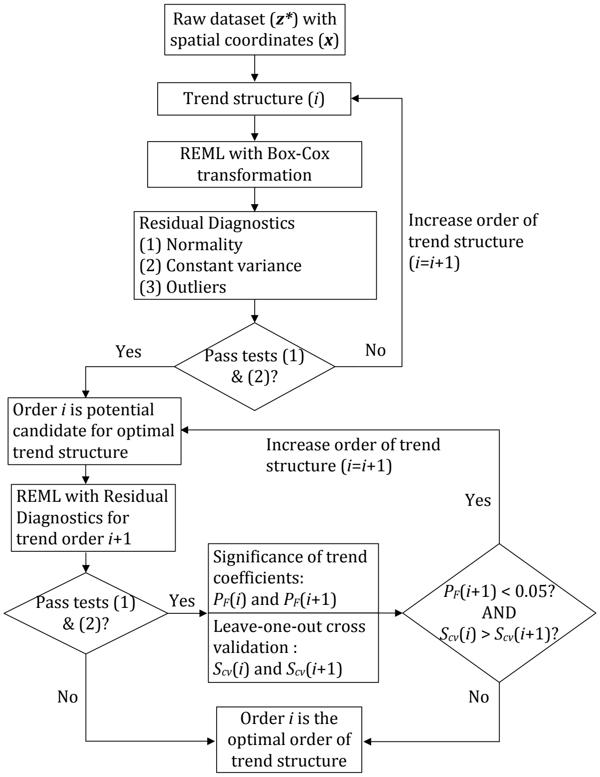

Raw dataset (z*) with spatial coordinates (x)
Trend structure (i)
REML with Box-Cox transformation
Residual Diagnostics
(1) Normality
(2) Constant variance
(3) Outliers
Increase order of trend structure (i=i+1)
Yes
Pass tests (1) & (2)?
No
Order i is potential candidate for optimal trend structure
Increase order of trend structure (i=i+1)
Yes
REML with Residual Diagnostics for trend order i+1
Significance of trend coefficients: P_F(i) and P_F(i+1)
Leave-one-out cross validation : S_cv(i) and S_cv(i+1)
Pass tests (1) & (2)?
Yes
P_F(i+1) < 0.05? AND S_cv(i) > S_cv(i+1)?
No
No
Order i is the optimal order of trend structure

FIG. 2. Flowchart of integrated residual analysis framework

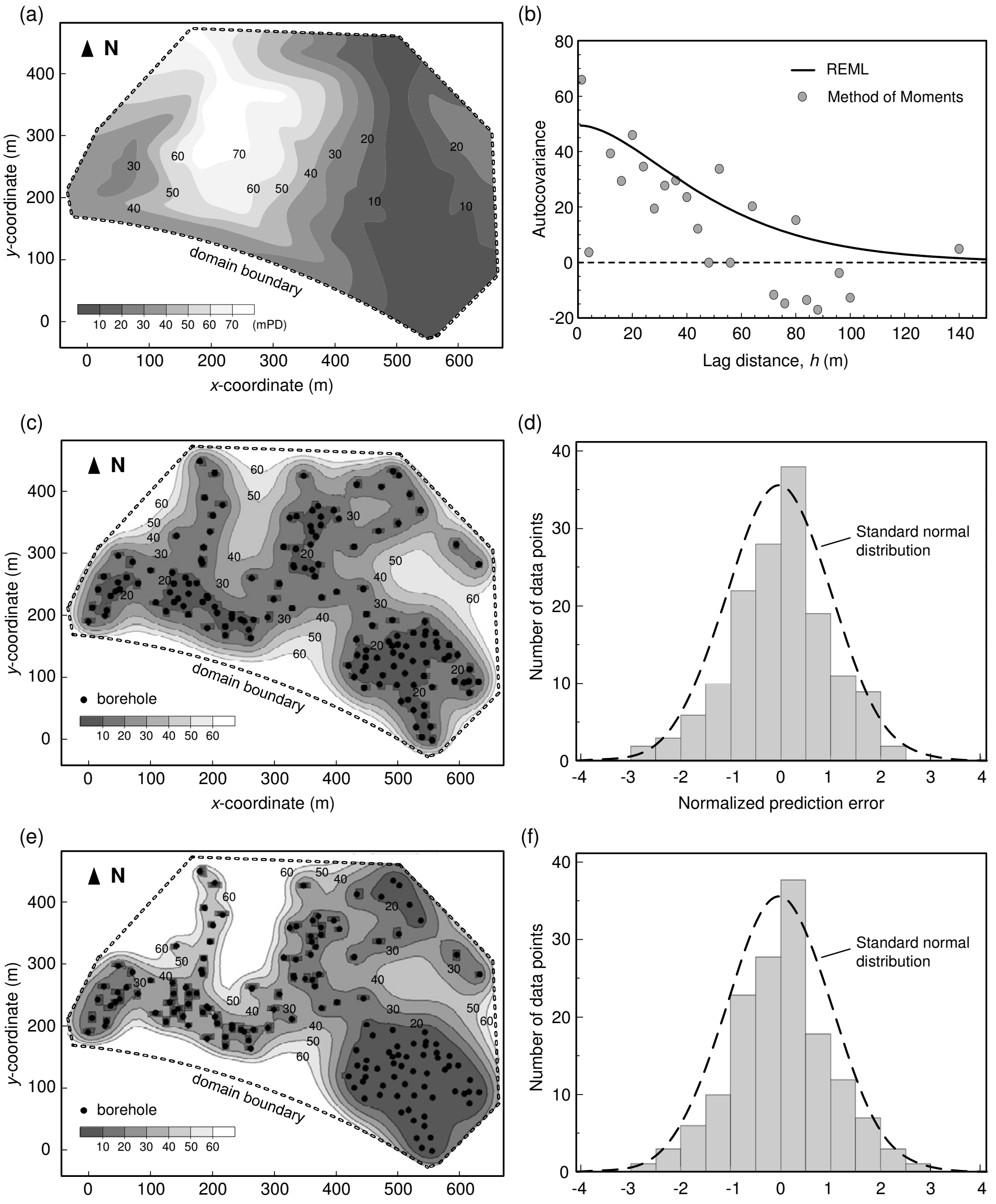


FIG. 3. (a) Rockhead level; (b) Autocovariance structure; (c,d) Prediction variance contour and normalized prediction errors in transformed space; (e,f) Prediction variance contour and normalized prediction errors in back-transformed original space

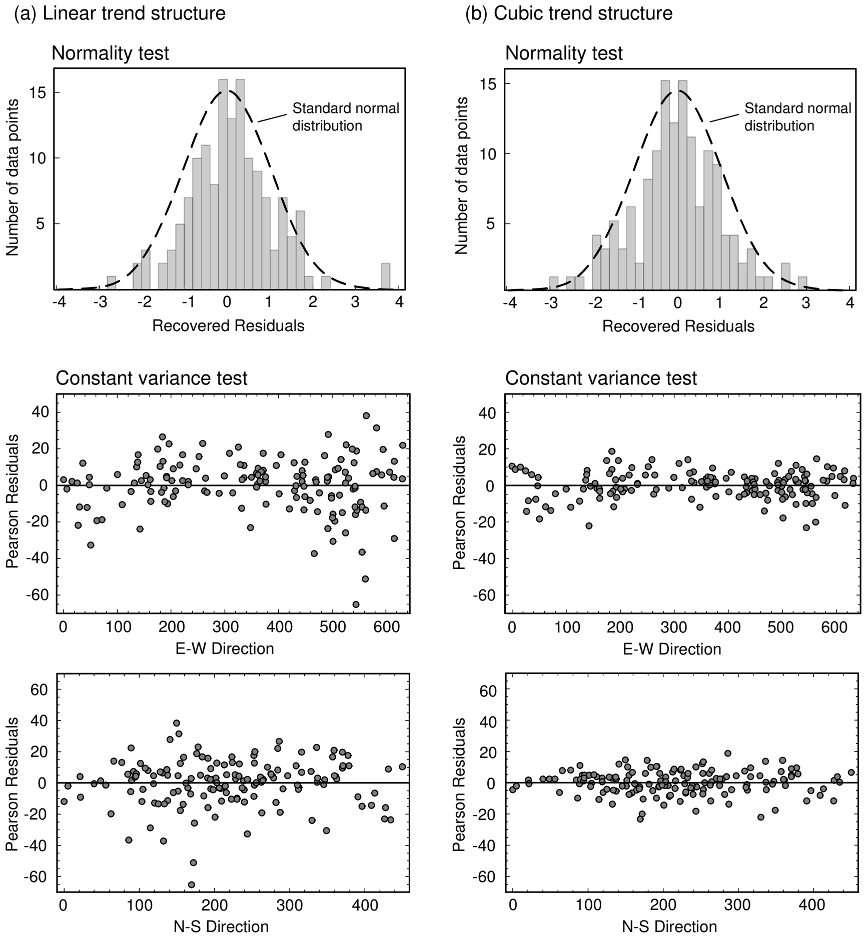
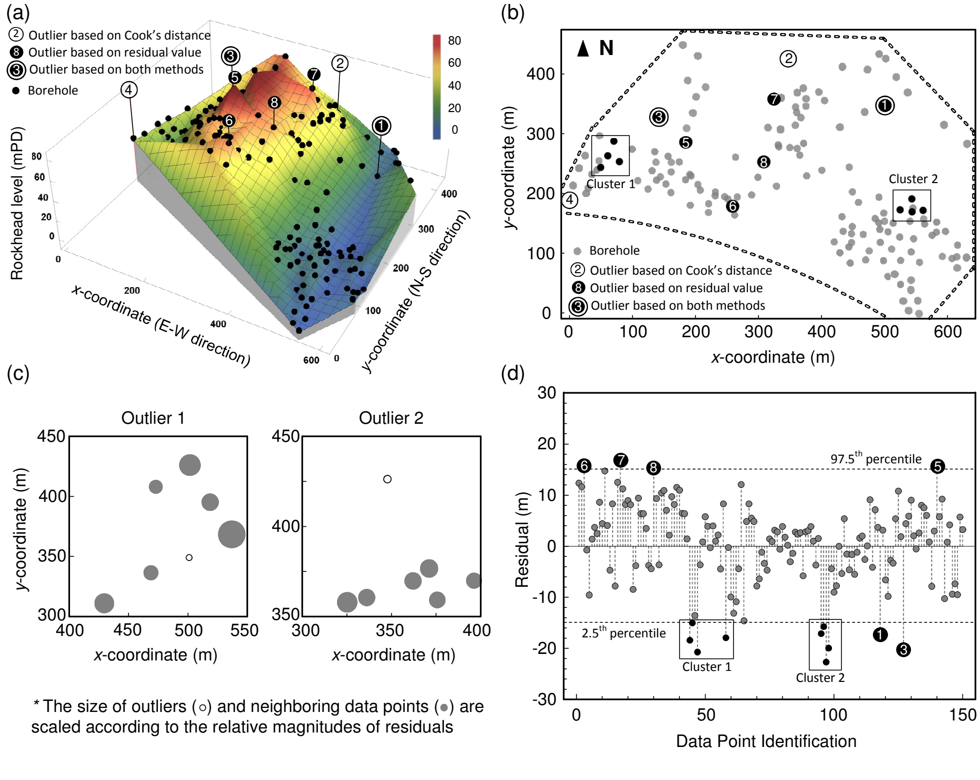


FIG. 4. Residual analyses for NTK site under linear and cubic trend assumptions



* The size of outliers (○) and neighboring data points (●) are scaled according to the relative magnitudes of residuals

FIG. 5. (a,b) Locations of potential outliers identified by (c) Cook's distances and (d) residual analysis

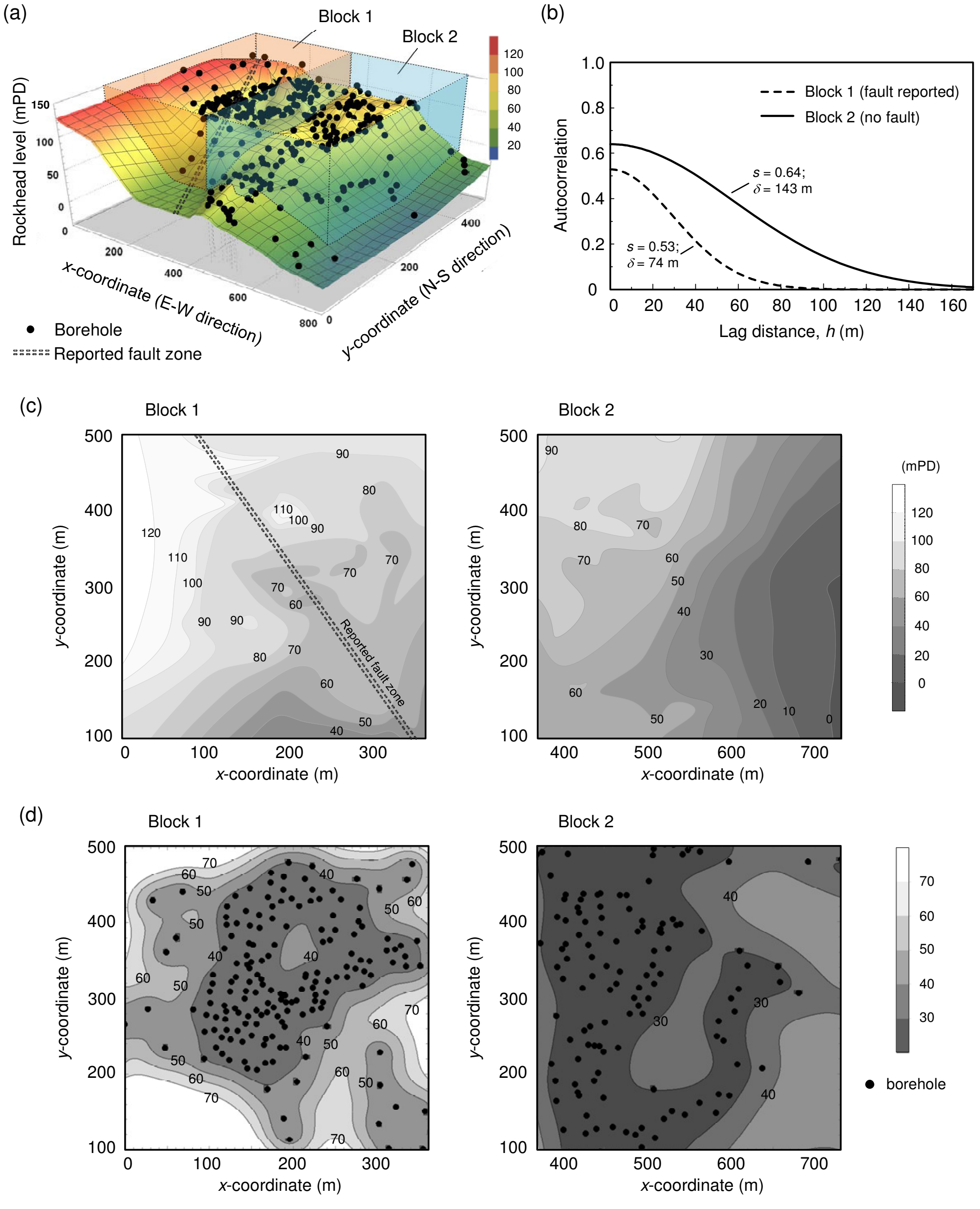


FIG. 6. (a) Two sub-regional blocks at CWE site; (b) Autocorrelation structures for two blocks; (c) Variations in rockhead levels; (d) Prediction variance contours for two blocks

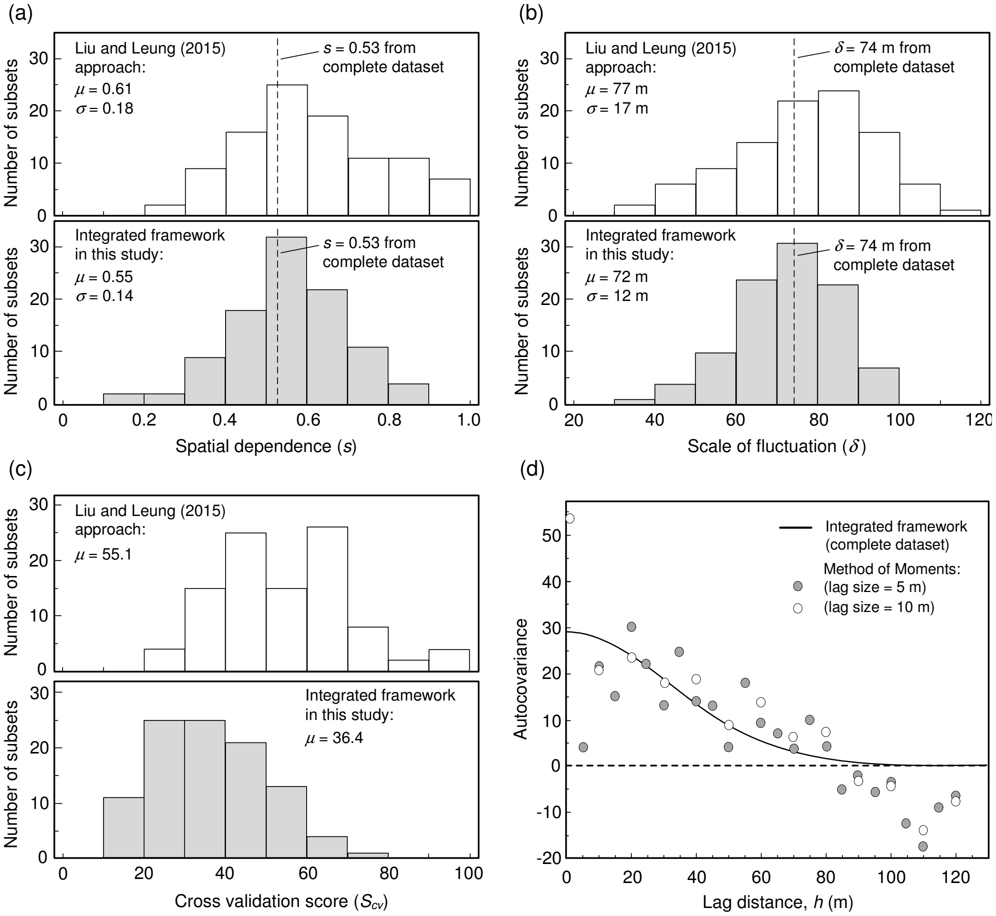


FIG. 7. Analyses of subsets by the proposed framework, the REML approach adopted in Liu and Leung (2015) and method of moments