https://doi.org/10.1016/j.autcon.2018.02.032

Quantifying cost-effectiveness of subsurface strata exploration in excavation projects through geostatistics and spatial tessellation

Yat Fai Leung*, Wenfei Liu, Yangxue Lei, Shu-Chien Hsu

Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Abstract

A major source of uncertainty in civil engineering projects arises from the geological and geotechnical variability at the project site. This paper presents an approach to quantify such uncertainty, and to rationally incorporate them into estimates of cost contingency. Contrary to the conventional approaches that rely on individual expert's opinion or data from past projects, the proposed approach allows site-specific assessments of the intrinsic geotechnical variability through geostatistical techniques. Such uncertainty can be reduced through geotechnical investigation, and spatial tessellation techniques are proposed to facilitate determination of the optimal locations of new boreholes. Cost-effectiveness of the boreholes can be evaluated based on the corresponding reductions in geotechnical uncertainty and their influence on the budget. The approach is illustrated using a hypothetical excavation scenario, where the project costs are affected by uncertainty in the subsurface strata, particularly the rockhead level across the site. Under the specific site conditions, a Pareto frontier is developed to reveal the relationship between the number of boreholes to be drilled and potential savings in contingency budget. Through this approach, the study promotes better utilization of geotechnical information and rational assessments of project risks associated with their variability, which may lead to improved project planning and resource allocation.

Keywords: Excavation, Contingency budget, Geotechnical variability, Delaunay triangulation, Voronoi cells

*Corresponding author
Email address: andy.yf.leung@polyu.edu.hk (Yat Fai Leung)

1. Introduction

Uncertainty in geotechnical engineering is a well-known, yet inadequately understood topic in the civil engineering profession. Delay and cost overrun of many large-scale infrastructure projects have been attributed to 'unforeseen' and complex geological and geotechnical conditions. According to a survey of 28 construction projects in the United Kingdom [1], more than 40% of the geotechnical problems encountered during construction arise from uncertainties related to the subsurface strata and the geotechnical properties. To reduce such uncertainties, geotechnical investigation (e.g., rock and soil sampling and testing) can provide additional information about the ground conditions at the site. However, there is currently no quantitative approach to relate this to the level of uncertainty across the site, or to elucidate how the project risks may be reduced through the additional information. Consequently, practitioners often rely on their individual experience or intuition when planning the geotechnical investigation programme. The qualitative nature of this practice makes it difficult to assess the cost-effectiveness of the investigation, or its implications on the overall budget and delivery time of the construction project. The problem can be exacerbated in infrastructure mega-projects, where delay in one part of the project often triggers cascading effects to the entire development plan. From a management standpoint, a cost or time contingency is usually included in the project budget or programme, as a common approach to control the risks of delay and cost overruns due to unforeseen conditions. In fact, the contingency budget or the 'float' of a particular task should be decided according to the level of uncertainty associated with the task. For excavation projects, it is therefore beneficial to quantify the geotechnical uncertainty, which then allows rational planning and apportioning of the risks.

Traditionally, cost contingency is incorporated as a simple percentage addition onto the base (cost) estimate, considering specific project features, past experience and historical data [2]. For large and complex projects, more rational estimates may be obtained either through deterministic or probabilistic approaches [3]. Some of the common deterministic approaches include linear regression models, artificial neural networks (ANN) for more complex problems, and Least Squares Support Vector Machine (LS-SVM) in price variation modeling for construction management. For example, Sonmez et al. [4] proposed a linear regression model to predict the bidding contingency amount for contractors, by focusing on the major influential factors of contingency decisions identified from previous projects, while Thal et al. [5] developed a multiple linear regression model for similar purposes. Although the regression method outperforms the practice of assigning an arbitrary percentage, a linear relationship may not be able to best fit the available historical data [6]. Therefore, Artificial neural networks (ANN) were utilized to perform nonlinear regression for more complex problems. These include the work by [7], who built a back propagation general regression neural networks (GRNN) to determine the cost contingency and allocation strategies at the preliminary stage. Lhee et al. [8] further proposed a two-step ANN-based model adding contingency rate as an intermediate output variable. Meanwhile, Cheng et al. [9] established a hybrid system based on Least Squares Support Vector Machine (LS-SVM) for modelling construction price variations, which can be used for decision making in construction management. Although these previous studies have illustrated the potentials of deterministic approaches, a few major criticisms remain regarding their applications. These include the heavy influence by subjectivity of individual experts, deficiency in accurately quantifying the project risks, and the fact that some of these techniques work like a 'black-box' [10, 11].

Probabilistic approaches were advocated to tackle these deficiencies. For example, Khalafallah et al. [12] proposed the Bayesian Belief Network to quantify project risk and uncertainty level, which allows the determination of the appropriate contingency percentage for construction projects. This approach was further developed by Kim et al. [13] to assess the probability of construction project delays based on case studies in developing countries. Meanwhile, other researchers adopted the Monte Carlo Simulation (MCS) to quantify cost contingency at different risk levels [14]. Since the risk factors in construction projects often contain both 'random' and 'fuzzy' variables [15], a Fuzzy MCS framework was established by Sadeghi et al. [16] to evaluate both components in the estimation of contingency range.

These previous approaches mainly rely on historical data or qualitative experts' opinion and experience [17], with little discussion on the intrinsic source of uncertainty. This paper attempts to quantify a major source of uncertainty in excavation projects, arising from the geotechnical and geological variability at the project site. An automated strategy is proposed to quantify geotechnical uncertainty, and to evaluate its changes with additional boreholes in the project site. It accounts for site-specific geologic features based on the available existing information, which may include irregularly-spaced boreholes revealing variations of subsurface strata in different directions. The quantitative approach enables optimization to be performed to determine the number and locations of sampling points that lead to the most cost-effective investigation programme, with respect to the impacts on time and costs of the tasks. The proposed approach will be demonstrated through the scenario of an excavation project, where the major uncertainty arises from the variations of rockhead level across the site. Such variations heavily influence the quantity of rock materials to be excavated, and hence the planning of project budget and delivery time.

2. Methodology

This study utilizes the geostatistical approach discussed by Liu et al. [18] and Liu and Leung [19] to quantify the geotechnical variability associated with subsurface strata. Meanwhile, spatial tessellation techniques are adopted for the derivation of optimal geotechnical sampling strategies. Their cost-effectiveness can be evaluated through the reductions of uncertainty, and the subsequent implications on the budget and time of the excavation project. The three individual components of the proposed approach are described in the following sections.

2.1. Quantification of geotechnical variability

Liu et al. [18] and Liu and Leung [19] presented the details and verification of an integrated framework established to characterize the spatial variability of geological profiles and geotechnical properties. This will be described briefly herein as it forms the basis of the sampling strategy proposed in this study. In general, the spatial variations of the subsurface strata ($\boldsymbol{z}$) can be represented by a linear mixed model consisting of a large-scale trend ($\mathbf{X}\boldsymbol{\beta}$), and the residual effects ($\boldsymbol{\varepsilon}$) that describe the deviations of the actual values from the trend (Fig. 1):

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

where $\mathbf{X}$ is a matrix containing information of the spatial coordinates of sampled points, and $\boldsymbol{\beta}$ represent the trend coefficients. $\boldsymbol{\varepsilon}$ is often observed to be spatially correlated, with greater variations between components ε_i and ε_j associated with larger separation distances between locations i and j. Accordingly, the variance of $\boldsymbol{\varepsilon}$ can be represented by:

$$\mathbf{V} = \sigma_e^2\mathbf{R} + \sigma_n^2\mathbf{I} = (\sigma_e^2 + \sigma_n^2)\left[s\mathbf{R} + (1-s)\mathbf{I}\right] \quad \text{where } \ 0 \leq s = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_n^2} \leq 1 \tag{2}$$

where $\mathbf{I}$ is the identity matrix; σ_n^2 are the random natural effects (white noise effects) which are independent of separation distances; σ_e^2 are the smooth scale variations, or the component of total variance that correlate with separation distance, and such autocorrelation is described by the matrix $\mathbf{R}$. s is referred to as the spatial dependence, and represents the proportion of σ_e^2 within the total variance. Individual components of $\mathbf{R}$ (R_{ij}) describe the correlations between ε_i and ε_j, and the relationship between R and separation distance (h_{ij}) can be modeled by different mathematical functions, such as the exponential, Gaussian (squared exponential), or spherical function, all of which involve a parameter θ that defines the range of correlation. Alternatively, the scale of fluctuation (δ) is another parameter used to define the extent of the correlation [20], and is often taken as the separation distance where the autocorrelation R drops to the value of 0.05. The parameters θ and δ are related to each other according to the adopted correlation function. For example, $\delta \approx \sqrt{\pi}\theta$ for the Gaussian function.

Site-specific characterization of the spatial features mainly involves determination of the trend

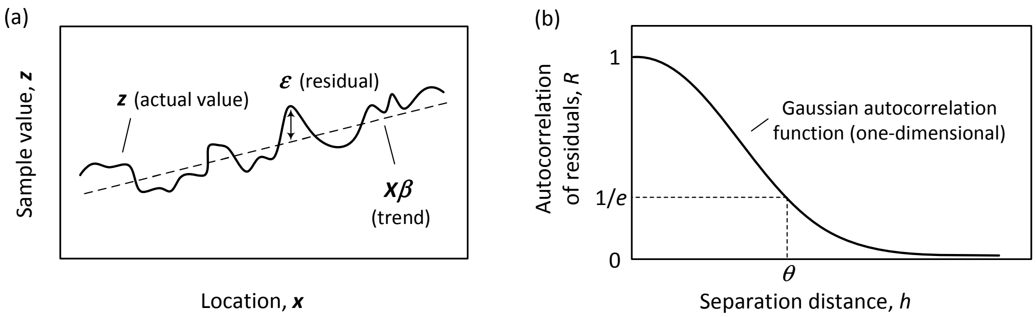

Figure 1: (a) Trend and residuals of spatial variables (after DeGroot and Baecher 1993); and (b) Autocorrelation of residuals

$\boldsymbol{\beta}$, together with correlation parameters s and θ. These parameters can be grouped as a vector $\boldsymbol{\Theta}$, and their values can be obtained through the 'restricted maximum likelihood' (REML) method [21], which involves maximizing the following log-likelihood function:

$$L(\boldsymbol{\Theta}|\boldsymbol{y}) = -\frac{n-m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{W}| - \frac{1}{2}\boldsymbol{y}^T\mathbf{V}^{-1}\mathbf{Q}\boldsymbol{y}$$
$$\text{where } \mathbf{W} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$$
$$\mathbf{Q} = \mathbf{I} - \mathbf{X}\mathbf{W}^{-1}\mathbf{X}^T\mathbf{V}^{-1} \tag{3}$$

In Eq. (3), n is the number of data points (e.g., geotechnical sampling points), and m 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 vector of filtered dataset with the trend components filtered out. The $\mathbf{V}$ matrix depends on the correlation structure (Eq. (2)), which in turn depend on the parameters in $\boldsymbol{\Theta}$.

The determination of $\boldsymbol{\Theta}$ (that leads to maximum $L(\boldsymbol{\Theta}|\boldsymbol{y})$) can be treated as an optimization problem, and various optimization techniques (e.g., gradient-based methods or evolutionary algorithms such as genetic algorithm) can be applied for this purpose. Once $\boldsymbol{\Theta}$ is obtained, predictions can be made for the 'unknown' properties at unsampled locations, through the 'best linear unbiased prediction' approach. More importantly, the level of uncertainty can be quantified through the prediction variance ($\boldsymbol{\sigma}_{\boldsymbol{z}}^2$), which is based on the overall variance across the site and the separation distances between the unsampled ($\boldsymbol{x}_0$) and sampled ($\boldsymbol{x}$) locations. Mathematically, this is represented by:

$$\boldsymbol{\sigma}_{\boldsymbol{z}}^2 = \mathrm{diag}(\mathbf{K}_0 - \mathbf{K}^T\mathbf{V}^{-1}\mathbf{K} + \mathbf{M}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{M}) \tag{4}$$

where $\mathbf{K}$ 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 $\mathbf{M} = \mathbf{X}_0^T - \mathbf{X}^T\mathbf{V}^{-1}\mathbf{K}$, and $\mathbf{X}_0$ is the deterministic component matrix of prediction (unsampled) locations. Details about its derivation and implementation can be found in Liu et al. [18], Liu and Leung [19], and Atkinson et al. [22].

It should be noted that the fundamental assumptions of the REML method require the data to be stationary, which is sometimes shown (or assumed) to be the case in civil or geotechnical engineering applications. Alternatively, the raw data can be transformed, e.g., through log transformation (to become lognormally distributed dataset) or Box-Cox transformation [23, 18]. In these cases, $\boldsymbol{z}$ and $\boldsymbol{\varepsilon}$ in the above equations will represent the transformed data, and the evaluated $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ will be under the transformed space. These can be easily back-transformed to the corresponding values in the original space.

2.2. Sampling strategies to minimize uncertainty

The main purpose of geotechnical investigation is to reduce the uncertainty associated with the geological strata and geotechnical properties at the site. Its effectiveness can be evaluated through the reduction in $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ described by Eq. (4). Therefore, determination of the 'best' sampling strategy can be considered as an optimization problem which, in theory, involves obtaining the optimal configuration of boreholes (with coordinates $\boldsymbol{\zeta}^*$) that minimizes $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ (or the average of its components) throughout the entire site domain:

$$\boldsymbol{\zeta}^* = \arg\min_{\boldsymbol{\zeta}} \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} [\sigma_z(i)]^2 \tag{5}$$

where N_σ represents the number of locations across the site at which the prediction variance is evaluated. A large-scale construction project will entail a large value of N_σ and also a large matrix size for $\mathbf{X}_0$, which renders the optimization in Eq. (5) very time-consuming and sometimes impractical. Therefore, this study adopts the Delaunay triangulation (DT) technique for efficient optimization of the borehole configuration.

DT is often discussed together with the Voronoi diagram due to their dual properties, and the formulations of the two are linked with each other. As shown in Fig. 2, for a domain that consists of n points (or 'seeds') p_i ($i = 1, 2, \ldots, n$), a set of n polyhedral convex regions (r_i, known as Voronoi cells) can be formed. Within each Voronoi cell r_k, any point x is most strongly influenced by (or located closest to) the seed of its cell p_k, compared to any other seeds. Mathematically, this is represented by [24]:

$$\begin{aligned} r_k &= \bigcap_{i \neq k} D(p_k, p_i) \qquad i = 1, 2, \ldots, n \\ \text{where} \qquad D(p_k, p_i) &= \{x | h(p_k, x) \leq h(p_i, x)\} \end{aligned} \tag{6}$$

and h represents the separation distance between the two points. Once the Voronoi diagram is constructed, the DT can be formulated by connecting the seeds that share a common edge in their respective Voronoi cells. The DT possesses several important properties which make it very versatile in computational geometry. One of these is that it maximizes the minimum angle of any triangle [24, 25], which means the triangles formed by DT are the closest to equilateral triangles compared to any other triangulation scheme.

In the current context, the boreholes in a geotechnical investigation programme are equivalent to the 'seeds' of the domain, while the Voronoi cells, and edge lengths of the triangles, are associated with the zones of influence of boreholes in different directions. Through the DT technique, sampling configurations can be designed such that the influence of each borehole will be evenly spread across the domain, thereby enhancing the overall effectiveness of the investigation programme. In other words, when new boreholes are being planned at a site, the 'optimal' sampling strategy can be derived by finding the configuration that minimizes the standard deviation of the edge lengths (t)

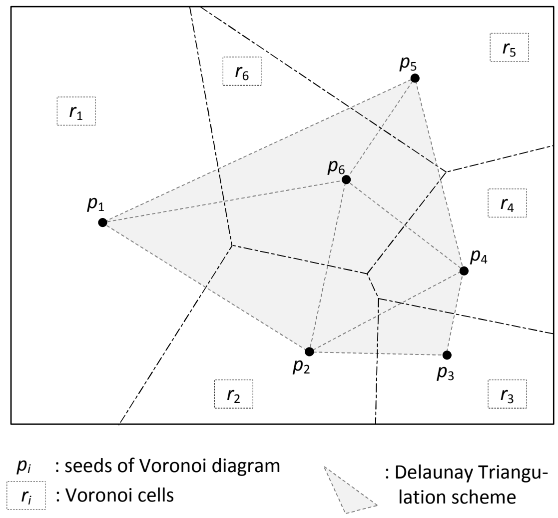

Figure 2: Voronoi diagram and Delaunay Triangulation

from DT, considering locations of both existing and new boreholes:

$$\boldsymbol{\zeta}^* = \arg\min_{\boldsymbol{\zeta}} \sqrt{\frac{1}{N_t - 1}\sum_{j=1}^{N_t}(t_j - \bar{t})^2} \tag{7}$$

which is an alternative way to Eq. (5) for optimal sampling, with N_t representing the total number of edge lengths in the DT scheme, and $\bar{t}$ being the average value of t across the domain. Due to its widespread application, the DT has been programmed efficiently in many common software, making it easier and faster to implement Eq. (7) than Eq (5). Once $\boldsymbol{\zeta}^*$ is determined, the corresponding $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ can be evaluated by Eq. (4).

Fig. 3 demonstrates two simple cases where the optimal sampling strategies obtained through the two different approaches are compared. The domain of the first case is 30 m × 30 m on plan, with eight pre-existing boreholes at the corners and edges. The locations of three new boreholes are to be determined, with the objective of minimizing the uncertainty in rockhead level ($\boldsymbol{z}$). The second case involves a 30 m × 20 m domain, and one more pre-existing borehole is assigned near the center, apart from the eight at the corners and edges. Five new boreholes will be planned across the site for this case. It is further assumed that variance ($\sigma_n^2 + \sigma_e^2$) for the residual of $\boldsymbol{z}$ is

100 m^2, and $s = 1$ for both cases. It is worth noting that the assumed variance corresponds to a standard deviation of 10 m in rockhead variations, while the assumption of $s = 1$ is associated with the measurement error at sampled locations. s approaches 1 when the white noise effects (σ_n^2) are small, and when there are no geologic faults causing abrupt changes in the rock level. The assumed values are close to the range revealed in another study by [18], and they do not actually affect the subsequent analysis results. This is because optimization of new borehole locations is controlled by the spatial distributions of $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$, i.e., their changes with distances, that arise from the existing boreholes. The process will not be affected when $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ is scaled up (or down) uniformly across the site by larger (or smaller) values of the variance. Meanwhile, the spatial correlation between the residuals at locations i and j is modeled by the Gaussian function:

$$R_{ij} = \exp\left(-\frac{h_{ij}^2}{\theta^2}\right) \tag{8}$$

where θ equals 10 m for the first case and 5 m for the second. In addition, to simplify the scenario, a constant mean is adopted for the property so $\mathbf{X}\boldsymbol{\beta}$ becomes a constant vector. This means the assumed trend of the rockhead is a horizontal plane. Again, the trend assumption does not influence the optimization process and the subsequent results, as the uncertainty arises mostly from the residuals, i.e. deviations of properties from the trend. When the framework is applied to actual field data, the trend is removed (or filtered out) to ensure stationarity conditions are satisfied. Subsequently, the residuals have zero mean; and the spatial correlation parameters and $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ are all estimated based on this premise, and the trend coefficients $\boldsymbol{\beta}$ are not involved in Eq. (4). By the same token, any trend assumptions may be adopted for the illustration cases, since $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ is evaluated based on this same principle regardless of the trend.

Optimization of the additional borehole locations is performed by Differential Evolution, which is an evolutionary algorithm developed by Storn and Price [26] and recently adopted in a number of engineering applications. During the optimization process, the potential borehole configurations

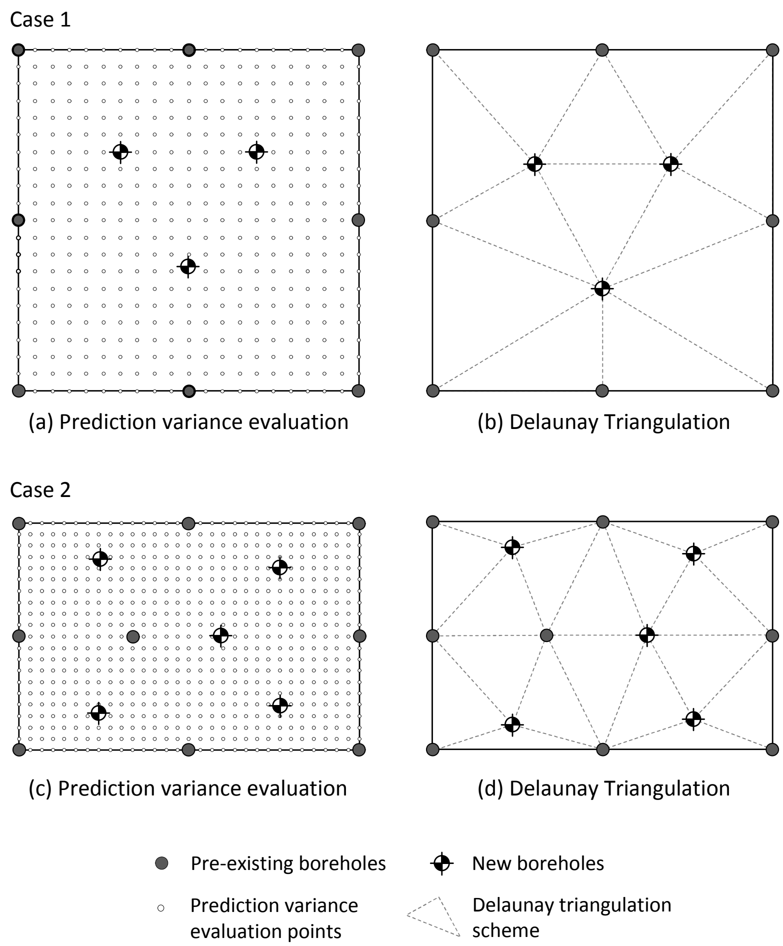

Figure 3: Optimal sampling locations for two illustration cases obtained by prediction variance evaluation and DT scheme

(candidate solutions) are expressed as vectors of variables (e.g., vectors of their coordinates) known as trial vectors. A population of trial vectors is first generated randomly, and the solution space is then explored through linear interpolation/extrapolation and mixing of trial vectors randomly selected from the population. Fitness of different trial vectors are evaluated and compared through an objective function, which determines the survivability of the particular solution: the fitter solutions stay in the population, while the weak ones will be discarded. The procedures are iterated until the population converges to a global optimum, and details of its implementation can be found in Storn and Price [26] or Leung et al. [27].

In the search for the optimal borehole configuration in this study, the objective function can be defined by Eqs. (5) or (7). In the former approach, $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ needs to be evaluated for the entire domain, and discrete evaluation points are assigned with spacing of 1.5 m in both directions for the first case (i.e., $N_\sigma = 441$), and the spacing of these points is 1 m in the second case ($N_\sigma = 651$). Figs. 3 shows that the optimal sampling strategies obtained by the two methods are very similar to each other, even when the two cases involve site domains of different sizes, shapes and pre-existing borehole arrangements. Despite very minor differences in the borehole arrangements, the corresponding discrepancies in the average $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ are smaller than 3% between the two methods. Meanwhile, optimization with the DT scheme (Eq. (7)) is much faster since it is not necessary to evaluate $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ for every potential configuration of new boreholes.

In this study, the locations of all additional boreholes are determined in a simultaneous manner. It is worth noting that multiple borehole locations may be determined also in a sequential manner: the first borehole is located where the uncertainty is largest, then the variance distribution is re-evaluated to determine location of the next borehole, and the process is repeated until reaching the desired total number of boreholes. The drawback of such approach, however, is that it does not guarantee optimum final arrangement of the multiple boreholes. Using the first case as an example, Fig. 4 compares borehole locations determined by the sequential and the simultaneous approaches, and the corresponding prediction variance contours. With the sequential approach, the first borehole will be assigned at the center of the domain because this is the location furthest away from all existing boreholes, and corresponds to the largest uncertainty in the beginning. The second borehole will then be located midway between the center and one of the corners, with the third one close to another corner. The arrangements will be different from the optimal arrangement obtained by simultaneously considering three new boreholes (Figs. 3a and b). Figs. 4(e) and (f) also show that the sequential approach leads to a worse performance compared to the simultaneous approach: the average $\boldsymbol{\sigma}_{\boldsymbol{z}}^2$ across the domain is about 7% higher with the sequential approach, which also gives rise to localized regions with high uncertainty, and the maximum value of prediction variance is 14% larger than that by the simultaneous approach.

2.3. Cost-effectiveness of sampling strategies in excavation projects

Quantity measurements in geotechnical engineering often depend on the subsurface geological profile at the site, and the variability of such would lead to uncertainties in the cost estimates. The cost-effectiveness of geotechnical sampling can be quantified through evaluating the associated reductions in uncertainty of the project costs. For example, for a building designed to be supported by end-bearing foundation piles, the actual cost of pile material and time of construction would depend on the depth of the competent firm stratum, on which the piles are to be founded. Increasing the number of boreholes would lead to better characterization of the stratum, reduced uncertainty and hence reductions of the contingency budget.

This study applies the concept to projects which involve excavation of soil and rock materials. Excavation of rock materials often costs three to five times more than that of soil materials, and the uncertainty associated with rockhead profile was cited as an important reason leading to the delay and cost overrun of a major underground railway station project in Hong Kong [28]. Following the earlier discussions on uncertainty of rockhead level ($\boldsymbol{\sigma}_{\boldsymbol{z}}$), the total cost of excavation, C, can be

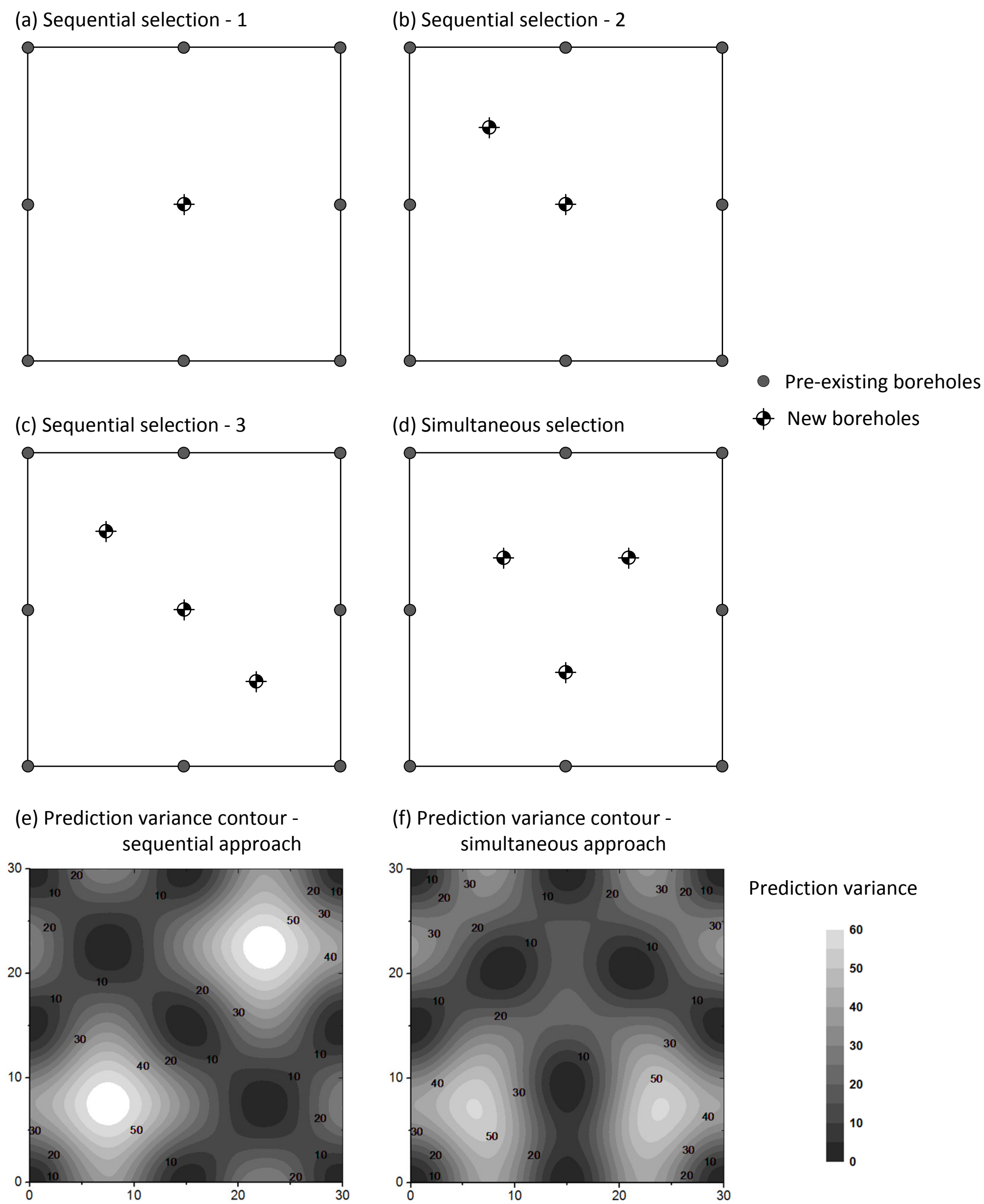

Figure 4: Comparisons between sequential and simultaneous determinations of new borehole locations

estimated for different ‘confidence levels’ of rockhead profile estimates:

$$C = \sum_{i=1}^{N_\sigma} A_i \left[C_r(d_i - d_{r,i}) + C_s(d_{r,i}) \right]$$
$$\text{where} \qquad d_{r,i} = \begin{cases} \hat{z}_i & \text{for 50\% non-exceedance probability} \\ \hat{z}_i - 0.675\,\sigma_{z,i} & \text{for 75\% non-exceedance probability} \\ \hat{z}_i - 1.28\,\sigma_{z,i} & \text{for 90\% non-exceedance probability} \end{cases} \tag{9}$$

and d_i represents the total depth of excavation (soil and rock) at location i, A_i is the area associated with that location, C_s and C_r are the unit rates of excavation in soil and rock, respectively; $d_{r,i}$ is the depth of rockhead with different confidence levels: $d_{r,i} = \hat{z}_i$ represents the best estimate at location i, which can be obtained through simple interpolation or the best linear unbiased prediction technique, while contingency in the budget can be introduced by subtracting from $\hat{z}_i$ multiples of the standard deviation of prediction at that location ($\sigma_{z,i}$) (smaller $d_{r,i}$ represents shallow rockhead, which leads to larger volume of rock to be excavated).

The contingency budget, $\tilde{C}$, can also be expressed by rewriting Eq. (9) as follows:

$$\tilde{C} = \alpha \sum_{i=1}^{N_\sigma} A_i \left(C_r - C_s \right) \sigma_{z,i} \tag{10}$$

where $\alpha = 0.675$ for 75% non-exceedance and 1.28 for 90% non-exceedance probabilities. In cases where the raw data has been transformed (e.g., to the log-space) for REML analyses, $d_{r,i}$, $\hat{z}_i$ and $\sigma_{z,i}$ in Eq. (9) and (10) should be represented in the ‘back-transformed’ original space. In the current practice of project management, the valuation of risks involves much subjectivity. Apart from providing a rational tool for quantification of uncertainty, Eq. (10) is proposed to allow risk perception of different individuals to be incorporated through the parameter α. Different values of α correspond to different confidence levels, or in this context, different non-exceedance probability for rockhead levels. For example, to cater for a high probability of non-exceedance (e.g., 90% chance that the volume of rock excavation is not underestimated), a high value of α (=1.28) is required and that leads to a larger contingency budget. In cases where the project managers are more familiar with the site conditions, or where the risk tolerance is high, it is possible to adopt lower values of non-exceedance probability, and hence lower values of α and contingency budget.

The sampling strategies described earlier aim to reduce $\sigma_{z,i}$ through optimal configuration of additional boreholes, while Eqs. (9) and (10) quantify the effectiveness of drilling these boreholes with respect to the cost commitment of the project. These can be revised to also reflect the programme estimates for the excavation. As will be shown later, when the number of boreholes increases in the site, the return of investment associated with each additional borehole will gradually diminish. The quantification and optimization framework proposed in this study can help engineers make informed decisions regarding the optimal number of boreholes, as well as their best locations, for the specific project conditions.

3. Illustration with hypothetical scenario

The proposed approach is illustrated using a hypothetical excavation scenario. The information on subsurface geological profile, particularly the variations of rockhead level, is based on real data from a project site in Hong Kong, retrieved from the archive maintained by the Civil Engineering Library of the Hong Kong Government. Using the real data of geological variations, the site is hypothetically set up as an excavation project, to investigate the cost-effectiveness of uncertainty reduction by additional boreholes.

3.1. Site descriptions

The data was obtained from a site in the Ngau Tau Kok (NTK) area in Hong Kong, and detailed geostatistical analyses using the REML method have been reported by Liu et al. [18]. The entire site domain is approximately 650 m $\times$ 450 m in area, with 150 irregularly-spaced, existing boreholes already drilled during previous development. The rockhead level is defined as the elevation of moderately decomposed granite, referred to as Grade III material according to GEO [29]. Based

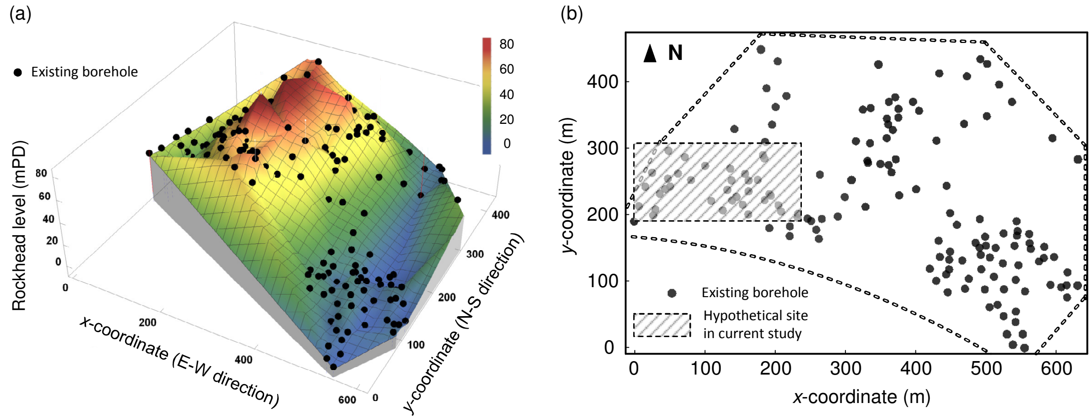

Figure 5: (a) Rockhead variations at NTK; (b) Study site within the NTK domain (modified from Liu et al. 2017)

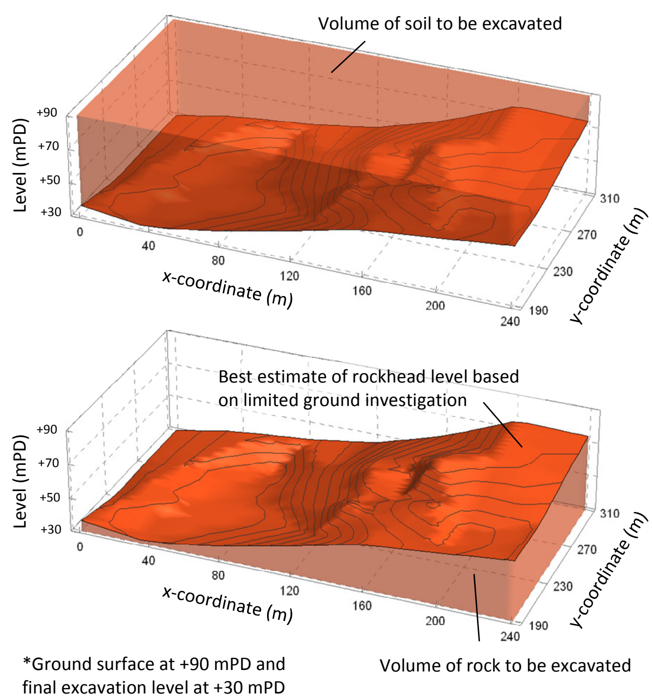

Figure 6: Volumes of soil and rock excavation at hypothetical site

on REML analyses (Eqs. (1) to (3)) of the existing data of rockhead levels, Liu et al. [18] reported the values of spatial dependence (s) and scale of fluctuation (δ) to be 0.75 and 125 m, respectively.

In this study, a portion in the western side of the NTK domain is considered to be the site of a hypothetical excavation project. This hypothetical site is 240 m $\times$ 120 m in area (Fig. 5), where 35 existing boreholes are located. It is further assumed that the ground surface and final excavation levels are at +90 mPD (Principal Datum) and +30 mPD, respectively. Therefore, the excavation depth is 60 m while the amount of rock and soil excavations vary (with uncertainty) across the site according to variations of the rockhead level. A graphical illustration of the project scenario is shown in Fig. 6. In the subsequent analyses, the unit rate of excavation is assumed to be HK$150 per m^3 in soil and HK$500 per m^3 in rock (US$1 $\approx$ HK$7.8).

3.2. Determination of optimal sampling strategy

The uncertain rockhead levels across the site affect the cost estimates associated with excavation of soil and rock materials. The proposed approach is adopted with DT to obtain optimal configurations of additional boreholes to reduce such uncertainty. One major assumption associated with the following analyses is that the data from future additional boreholes will not change the correlation structure of rockhead levels (i.e. s and δ) obtained earlier. This assumption is inevitable during the planning stages of geotechnical investigation, since the data values at future boreholes, and hence their impacts on s and δ, are unknown. In a real project scenario, it is possible to sequentially update s and δ as more information becomes available during the exploration. This is, however, not performed for the hypothetical case in the current study.

Contrary to the scenario depicted in Fig. 3, the site boundaries in this hypothetical case only involve one existing borehole at the southwestern corner. Therefore, a number of auxiliary boundary points need to be assigned to define the boundaries for the triangulation scheme. To determine the spacing and locations of these auxiliary points, a DT analysis is first performed for the 35 existing boreholes (without considering the site boundary), and the median of edge lengths is found to be

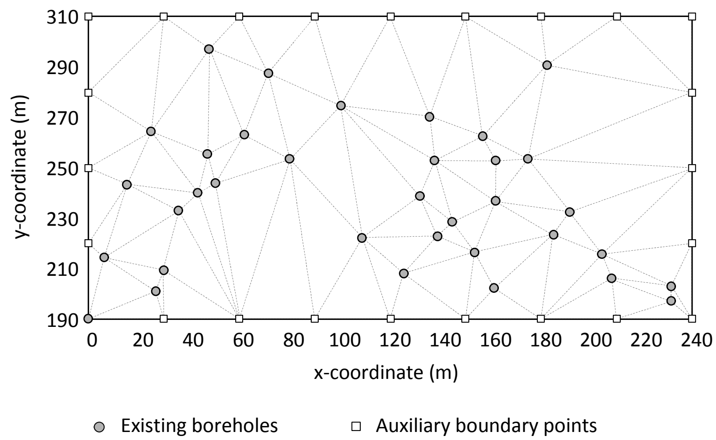

Figure 7: Delaunay triangulation scheme considering existing boreholes and auxiliary boundary points

approximately 30 m. The median is preferred over the mean since the latter appears to be skewed by a few extreme values. This is then adopted as the spacing for auxiliary points along the boundaries of the hypothetical site domain, as shown in Fig. 7. It should be noted that these auxiliary points do not affect the estimates of prediction variance as they are not modeled as boreholes, and they are also omitted in later figures for clarity. However, adopting different spacings for these points will lead to slight influence on the results of triangulation, and hence the optimal sampling strategy. These will be discussed further in a subsequent section.

In general, drilling more boreholes will always reduce the uncertainty, but that will also introduce additional costs to the investigation programme. The cost-effectiveness of geotechnical investigation at any particular site can be explored by developing the Pareto frontier, which reveals the relationship between the number of additional boreholes (n) and the reduction in contingency budget. For the current hypothetical site, this relationship is obtained through 30 individual optimization analyses, each considering a different number of additional boreholes. The associated percentage reduction in $\tilde{C}$ is plotted in Fig. 8(a), for both 75% and 90% non-exceedance probabilities of rockhead variations. For example, when 10 additional boreholes are drilled at the site, a reduction of 15% in $\tilde{C}$ can be achieved, if $\tilde{C}$ is estimated based on 90% non-exceedance probability of the amount

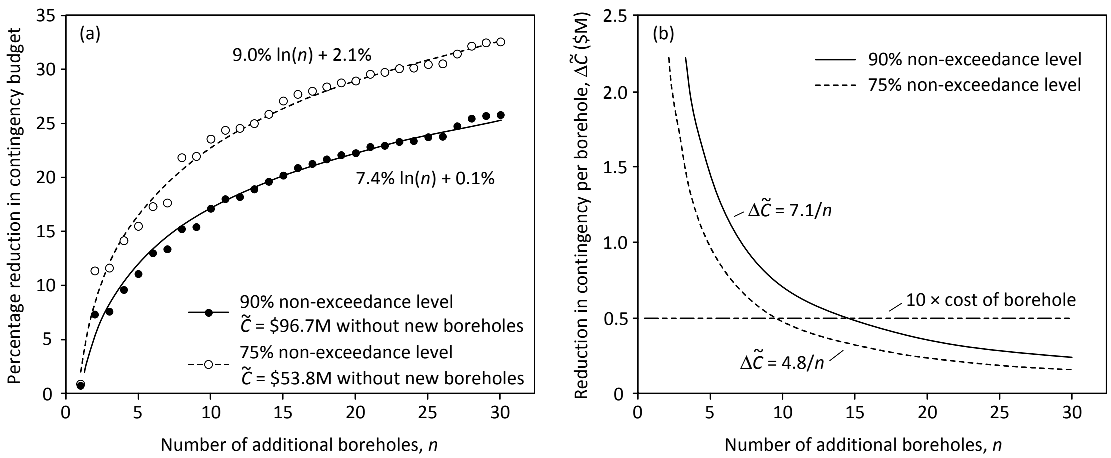

Figure 8: Pareto frontiers for 75% and 90% non-exceedance levels of rockhead variations

of rock materials to be excavated. The same number of boreholes can lead to 23% reduction in $\tilde{C}$ estimated based on 75% non-exceedance probability. These percentage reductions are achievable when the layouts of additional boreholes are optimized using the proposed DT scheme. Under the current site settings, the 10 additional boreholes can lead to potential savings in contingency budget that exceed HK$10 million.

The Pareto frontiers in Fig. 8(a) may be fitted by log curves also shown in the figure. The log relationships help to illustrate the effectiveness of the boreholes, as the slopes of the curves (which are proportional to $1/n$) represent the extra values gained by adding one borehole to the investigation programme. This is denoted as $\Delta\tilde{C}$ in Fig. 8(b). Although $\Delta\tilde{C}$ drops with n, it far exceeds the drilling costs of a borehole under the adopted site settings. For example, assuming an approximate cost of HK$1,250 per meter of drilling, a 40-m borehole would cost HK$50,000. This is 10 times lower than the return of investment (through reducing $\tilde{C}$) by the 15th borehole, based on 90% non-exceedance probability of rockhead level. Even after considering other subsidiary costs involved in sampling or mobilization of rigs, the return of investment is still substantial. The analyses also provide an indication of the balance point between costs and benefits of each additional borehole, and may be used as a quantitative illustration on the rationale behind the optimum number of boreholes. As shown in Fig. 8(b), this optimum number also depends on the risk perceptions and risk tolerance at the project, reflected by the non-exceedance probability and α value (Eq. (10)) adopted in the decision-making process. It is important to note that under different project conditions, the associated drilling costs or contingency budget will differ from this hypothetical scenario, yet the proposed approach allows rational and quantified estimates to be obtained, and informed decisions to be made regarding the geotechnical investigation programme at the excavation site.

Fig. 9 presents two examples of optimal borehole configurations, with 10 and 20 additional boreholes, alongside the original layout of 35 existing boreholes. As mentioned earlier, the optimal arrangements are obtained by minimizing the standard deviations of edge lengths (t) from the DT scheme. The distributions of t in the three configurations are also shown in Fig. 9. With more additional boreholes, both the average and standard deviations of t are reduced, which means the 'coverage area' for each borehole is decreased in a uniform manner, thereby effectively lowering the prediction variance or uncertainty across the site. These coverage areas can be illustrated graphically through the corresponding Voronoi cells, which become more uniform and smaller in size with increasing numbers of boreholes in an optimal arrangement.

The prediction variance contours for these borehole arrangements are shown in Fig. 10. It should be noted that although the color scale in the figures is capped at 50 for clarity, the average values around the edges of Fig. 10(a) (no additional boreholes) and Fig. 10(c) (20 additional boreholes) are approximately 200 and 60, respectively. The substantial reductions in prediction variance lead to the potential savings in contingency budget shown in Fig. 8.

4. Discussions

In the previous section, it is assumed that the new data from additional boreholes do not affect the s and δ parameters that characterize the spatial correlation structure. This is a reasonable assumption for the studied site since the parameters are obtained based on geostatistical analyses

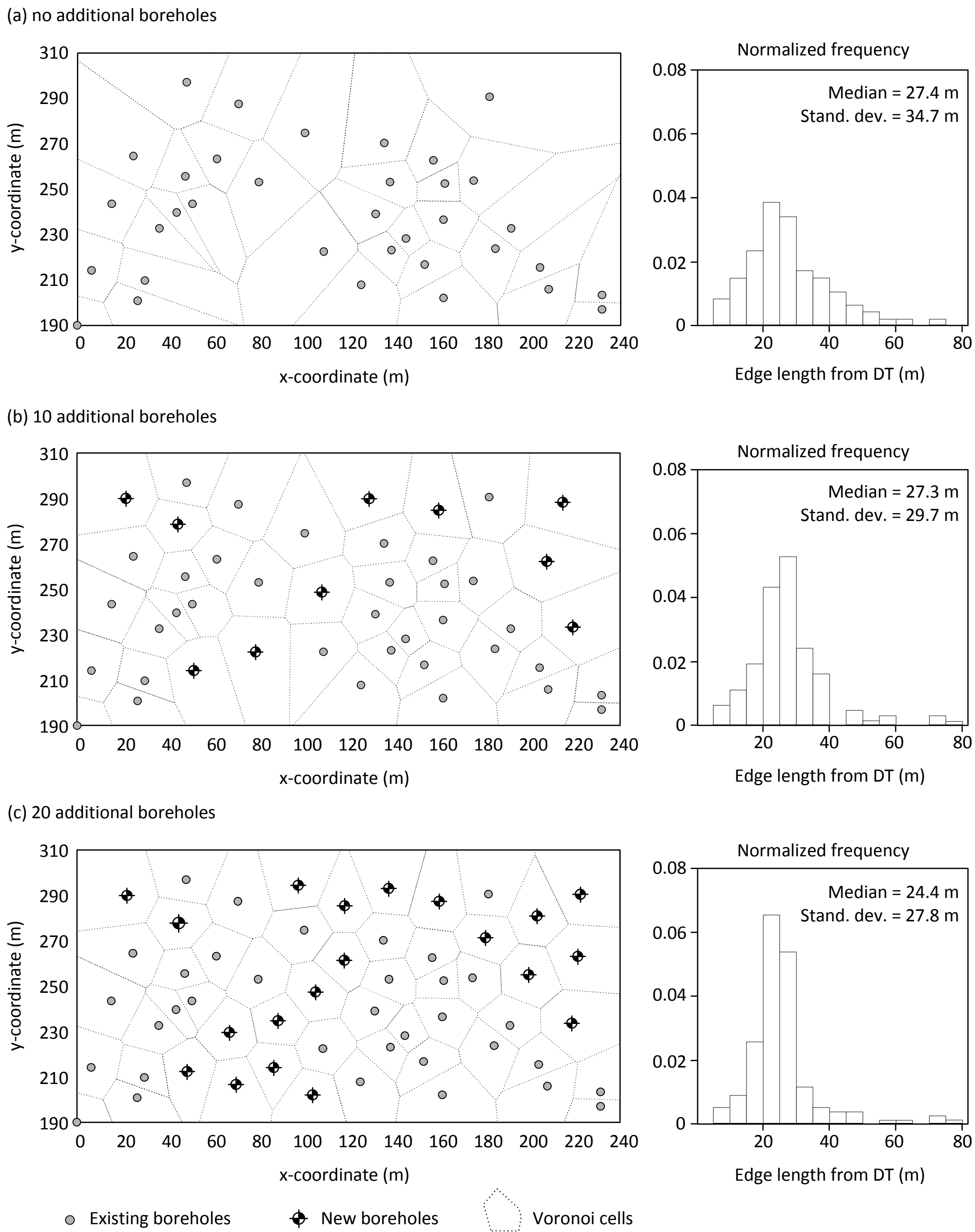

Figure 9: Optimal layout for (a) zero and (b) 10 and (c) 20 additional boreholes, with frequency distributions of edge lengths from Delaunay Triangulation

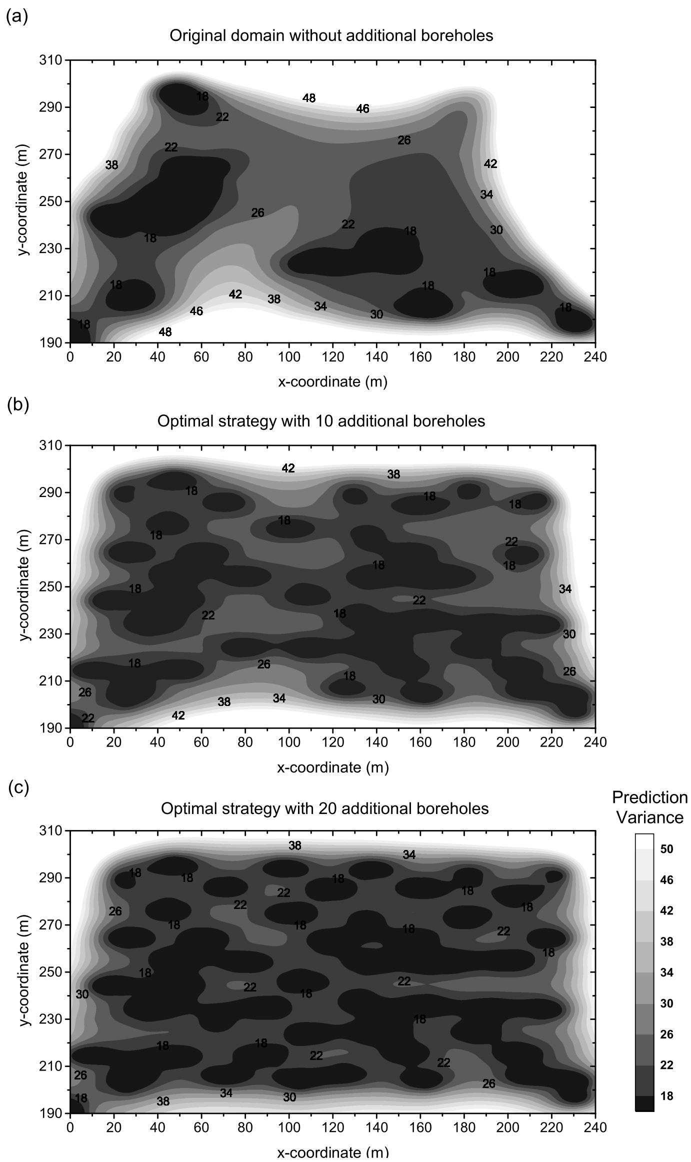

Figure 10: Prediction variance contours for (a) no additional boreholes; (b) 10 additional boreholes and (c) 20 additional boreholes

Table 1: Influence of auxiliary boundary points on Delaunay triangulation and prediction variance estimates

	10 additional boreholes			20 additional boreholes		
Spacing of boundary points (m)	20	30	40	20	30	40
Median of t from DT (m)	26.8	27.3	27.7	24.7	24.4	25.2
Standard deviation of t from DT (m)	29.5	29.7	29.0	25.9	27.8	23.8
Average prediction variance	38.7	37.4	37.6	33.9	31.6	29.7
Cost contingency ($M)(90% non-exceedance)	80.5	80.2	79.8	76.1	75.2	73.5
(75% non-exceedance)	41.5	41.1	41.2	39.1	38.2	37.3

of 150 boreholes in the NTK domain, leading to relatively robust estimates. Similar analyses can be applied to other projects using information from previous developments around the region, or pre-tender investigation data. In cases where prior knowledge of the site is more limited, it would be advisable to progressively update the correlation parameters using newly available data, and reassess the cost-effectiveness of boreholes as the geotechnical investigation is being performed. Also, there may be situations where no prior information is available in the project area. In these cases, the first estimates of spatial correlation parameters may be reasonably assumed based on understanding of the local geology, or published information from the literature [18, 20, 30].

In preceding analyses of the hypothetical case, auxiliary boundary points are assigned along the site boundaries to facilitate the triangulation scheme, with the spacing of 30 m decided based on the median of edge lengths from DT of 35 existing boreholes. To explore the influence of the adopted spacing of these auxiliary points, additional analyses are performed, with the spacing set to be 20 m and 40 m, respectively. Table 1 compares these results with optimization of 10 and 20 additional boreholes. In general, sparsely spaced auxiliary points lead to the additional boreholes being slightly closer to the boundaries of the domain. However, the overall impact on prediction variance, and hence the reductions in uncertainty or contingency, is not significant.

In the current study, the focus is placed on the impact of geological variability on the cost uncertainty of the bulk excavation, i.e., removal of soil and rock materials. There are a number of other factors that influence the total costs of an excavation project, most notably the choice and installation of retaining wall and shoring systems. Although these factors are not evaluated explicitly in the current study, it is expected that they will also be influenced by data from additional boreholes. For example, reduced uncertainty in rockhead levels can lead to more accurate estimates on the depths of retaining walls. In fact, Eqs. (9) and (10) may be revised to assess the influence of reduced uncertainty on the potentials of project delays and the ensuing liquidated damages. The cost-effectiveness of geotechnical investigation would be even more significant if these benefits are also considered.

Apart from excavation projects, the proposed approach may be applicable to other civil and geotechnical project scenarios, where the quantity estimates and/or design assumptions are heavily influenced by uncertainty in subsurface strata. For example, the designs of both onshore and offshore foundation systems rely on proper understanding of the underground geological conditions and depths of firm strata supporting the foundations. The designs and cost estimates will benefit from rational decisions on the optimal planning of geotechnical investigation.

It should be noted that the current study adopts the isotropic autocovariance structure for variations in geological profiles, with the same spatial variability properties in both directions. This, however, may be inadequate when modelling certain geotechnical properties (e.g., shear strength and stiffness) that display anisotropic spatial correlation features in the three-dimensional subsurface domain. A potential extension of this study may focus on addressing this challenge in the future.

5. Conclusion

This study illustrates how the uncertainty in subsurface geological strata can be quantified through geostatistical analyses of the existing borehole information, and the benefits of incorporating such considerations in the cost estimates of excavation projects. To maximize the cost-effectiveness of geotechnical investigation, the locations of boreholes can be optimized to cater for the specific conditions of the project. This is achieved in the current study through coupling an optimization algorithm, the Differential Evolution, with spatial tessellation techniques including generation of Voronoi diagrams and Delaunay triangulation.

The proposed approach is demonstrated through analyses of a hypothetical excavation site consisting of real data of subsurface geological variations. It is shown that the contingency budget for excavation of soil and rock materials can be rationally and substantially reduced by considering the influence of additional boreholes located in optimal configurations. In addition, Pareto frontiers are developed to reveal the relationship between the number of additional boreholes and the cost-effectiveness of geotechnical investigation. In general, the benefits of geotechnical investigation to the finances and programme of the project will vary according to the specific site conditions. Nonetheless, the proposed approach can be a versatile tool applicable to various civil engineering projects, as it provides a platform where the assumptions and risks associated with subsurface profiles can be properly communicated, allowing various options of site investigation to be compared quantitatively in an evidence-based decision making process.

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.

References

[1] C. R. I. Clayton, Managing geotechnical risk: improving productivity in UK building and construction, Thomas Telford, ISBN: 9780727729675, 2001.

[2] R. B. Lorance, R. V. Wendling, Basic techniques for analyzing and presenting cost risk analysis, AACE International Transactions (1999) K11, ISBN: 9781885517166.

[3] M. M. Elbarkouky, A. R. Fayek, N. B. Siraj, N. Sadeghi, Fuzzy arithmetic risk analysis approach to determine construction project contingency, Journal of Construction Engineering and Management 142 (12) (2016) 04016070, doi.org/10.1061/(asce)co.1943–7862.0001191.

[4] R. Sonmez, A. Ergin, M. T. Birgonul, Quantitative methodology for determination of cost contingency in international projects, Journal of Management in Engineering 23 (1) (2007) 35–39, doi.org/10.1061/(asce)0742–597x(2007)23:1(35).

[5] A. E. Thal Jr, J. J. Cook, E. D. White III, Estimation of cost contingency for air force construction projects, Journal of Construction Engineering and Management 136 (11) (2010) 1181–1188, doi.org/10.1061/(asce)co.1943–7862.0000227.

[6] C. Wilmot, B. Mei, Neural network modelling of highway construction costs, Journal of Construction Engineering and Management 131 (7) (2005) 765771, doi.org/10.1061/(asce)0733–9364(2005)131:7(765).

[7] D. Chen, F. T. Hartman, A neural network approach to risk assessment and contingency allocation, AACE International Transactions (2000) RI7A, ISBN: 9781885517258.

[8] S. C. Lhee, I. Flood, R. R. Issa, Development of a two-step neural network-based model to predict construction cost contingency, Journal of Information Technology in Construction (ITcon) 19 (24) (2014) 399–411, ISSN 1874–4753.

[9] M. Cheng, N. Hoang, Y. Wu, Hybrid intelligence approach based on LS-SVM and differential evolution for construction cost index estimation: A Taiwan case study, Automation in Construction 35 (2013) 306–313, doi.org/10.1016/j.autcon.2013.05.018.

[10] M. Choi, G. Lee, Decision tree for selecting retaining wall systems based on logistic regression analysis, Automation in Construction 19 (7) (2010) 917–928, doi.org/10.1016/j.autcon.2010.06.005.

[11] S. Mak, D. Picken, Using risk analysis to determine construction project contingencies, Journal of Construction Engineering and Management 126 (2) (2000) 130–136, doi.org/10.1061/(asce)0733–9364(2000)126:2(130).

[12] A. Khalafallah, M. Taha, M. El-Said, Estimating residential projects cost contingencies using a belief network, In Proceedings of Project Management: Vision for Better Future Conference (2005) Cairo, Egypt, 21–22 November 2005, 1–19, Available `http://cee.ucf.edu/people/khalafal/home/publications_files/EstimatingContingencies.pdf`, Accessed date 29 May 2017.

[13] S. Y. Kim, N. Van Tuan, S. O. Ogunlana, Quantifying schedule risk in construction projects using bayesian belief networks, International Journal of Project Management 27 (1) (2009) 39–50, doi.org/10.1016/j.ijproman.2008.03.003.

[14] G. A. Barraza, R. A. Bueno, Cost contingency management, Journal of Management in Engineering 23 (3) (2007) 140–146, doi.org/10.1061/(ASCE)0742–597X(2007)23:3(140).

[15] D. Guyonnet, B. Bourgigne, D. Dubois, H. Fargier, B. Côme, J. Chilès, Hybrid approach for addressing uncertainty in risk assessments, Journal of Environmental Engineering 129 (1) (2003) 6878, doi.org/10.1061/(asce)0733–9372(2003)129:1(68).

[16] N. Sadeghi, A. R. Fayek, W. Pedrycz, Fuzzy Monte Carlo simulation and risk assessment

in construction, Computer-Aided Civil and Infrastructure Engineering 25 (4) (2010) 238252, doi.org/10.1111/j.1467–8667.2009.00632.x.

[17] A. Salah, O. Moselhi, Contingency modelling for construction projects using fuzzy set theory, Engineering, Construction and Architectural Management 22 (2) (2015) 214241, doi.org/10.1108/ecam–03–2014–0039.

[18] W. F. Liu, Y. F. Leung, M. K. Lo, Integrated framework for characterization of spatial variability of geological profiles, Canadian Geotechnical Journal 54 (1) (2017) 47–58, doi.org/10.1139/cgj–2016–0189.

[19] W. F. Liu, Y. F. Leung, Characterising three-dimensional anisotropic spatial correlation of soil properties through *in situ* test results, Géotechnique (2017) 1–15, doi.org/10.1680/jgeot.16.P.336.

[20] K.-K. Phoon, F. H. Kulhawy, Characterization of geotechnical variability, Canadian Geotechnical Journal 36 (4) (1999) 612–624, doi.org/10.1139/cgj–36–4–612.

[21] R. Lark, B. Cullis, S. Welham, 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) (2006) 787–799, doi.org/10.1111/j.1365–2389.2005.00768.x.

[22] P. M. Atkinson, E. Pardo-Iguzquiza, M. Chica-Olmo, Downscaling cokriging for super-resolution mapping of continua in remotely sensed images, IEEE Transactions on Geoscience and Remote Sensing 46 (2) (2008) 573–580, doi.org/10.1109/tgrs.2007.909952.

[23] G. E. P. Box, D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society, Series B (Methodological) 26 (2) (1964) 211–252, http://www.jstor.org/stable/2984418.

[24] F. Aurenhammer, Voronoi diagrams – a survey of a fundamental geometric data structure, ACM Computing Surveys 23 (3) (1991) 345–406, doi.org/10.1145/116873.116880.

[25] S. J. Fortune, Voronoi diagrams and delaunay triangulations, in: Computing in Euclidean Geometry, 2nd ed., 1995, pp. 225–265, doi.org/10.1142/9789812831699_0007.

[26] R. Storn, K. Price, Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization 11 (4) (1997) 341–359, doi.org/10.1023/A:1008202s821328.

[27] Y. F. Leung, A. Klar, K. Soga, N. A. Hoult, Superstructure-foundation interaction in multi-objective pile group optimization considering settlement response, Canadian Geotechnical Journal 54 (10) (2017) 1408–1420, doi.org/10.1139/cgj–2016–0498.

[28] MTRC, Second Report by the Independent Board Committee on the Express Rail Link Project, MTR Corporation Limited. 2014. Available `http://www.expressraillink.hk/pdf/en/report/2nd%20Report_ENG(Full).pdf`, Accessed date: 29 May 2017.

[29] GEO, GEOGUIDE 3: Guide to Rock and Soil Descriptions, Geotechnical Engineering Office, Civil Engineering Department, Government of Hong Kong, 1988. Available `http://www.cedd.gov.hk/eng/publications/geo/doc/eg3.pdf`, Accessed date: 22 May 2017.

[30] D. J. DeGroot, Analyzing spatial variability of in situ soil properties, in: Uncertainty in the Geologic Environment: From Theory to Practice, Vol. 1, 1996, pp. 210–238, ISBN: 9780784401880.