Structural Engineering and Mechanics, *Vol. 54, No. 2 (2015) 363-378*
DOI: http://dx.doi.org/10.12989/sem.2015.54.2.363

Probabilistic real-time updating for geotechnical properties evaluation

Iok-Tong Ng[a], Ka-Veng Yuen[*] and Le Dong[b]

Department of Civil and Environmental Engineering. Faculty of Science and Technology, University of Macau, Macao, China

(Received December 2, 2014, Revised January 14, 2015, Accepted March 12, 2015)

Abstract. Estimation of geotechnical properties is an essential but challenging task since they are major components governing the safety and reliability of the entire structural system. However, due to time and budget constraints, reliable geotechnical properties estimation using traditional site characterization approach is difficult. In view of this, an alternative efficient and cost effective approach to address the overall uncertainty is necessary to facilitate an economical, safe and reliable geotechnical design. In this paper a probabilistic approach is proposed for real-time updating by incorporating new geotechnical information from the underlying project site. The updated model obtained from the proposed method is advantageous because it incorporates information from both existing database and the site of concern. An application using real data from a site in Hong Kong will be presented to demonstrate the proposed method.

Keywords: bayesian inference; empirical correlation; model selection; nonparametric; normally consolidated clays; undrained shear strength

1. Introduction

In the analysis and design of earth structures and foundations, safety, reliability and cost are primary factors for consideration (Kaloop *et al.* 2014, Wang and Ginger 2014, Rezaiee-Pajand and Kazemiyan 2014). The mechanical properties of geomaterials are the major components governing these issues. Since geomaterials are naturally formed materials, their properties vary spatially even within a relatively homogeneous soil stratum (Vanmarcke 1983, Baecher and Christian 2003). In contrast to other construction materials, the properties of geomaterial with considerable inherent spatial variability are difficult to prescribe with prior knowledge but they have to be acquired through geotechnical site characterization. Several researchers have addressed the fundamental levels of uncertainty involved in geotechnical site characterization (e.g., Lumb 1966, Kulhawy 1996). It has been reported that substantial uncertainty were due to equipment and testing error, statistical uncertainty and uncertainty associated with the conversion of test measurements to the design parameters in the evaluation of the geomaterial properties.

∗Corresponding author, Professor, E-mail: kvyuen@umac.mo
[a]Assistant Professor, E-mail: itng@umac.mo
[b]PhD Student, E-mail: dongle0414@gmail.com

Copyright © 2015 Techno-Press, Ltd.
http://www.techno-press.org/?journal=sem&subpage=8 ISSN: 1225-4568 (Print), 1598-6217 (Online)

It is obvious that the aforementioned variability and uncertainty significantly influences the uncertainty of the geomaterial properties estimation and, hence, the analysis and design of the entire structural and foundational system. However, traditional site characterization approaches (e.g., Hvorslev 1949, BS 5930 1981) are largely qualitative and they rely on engineering experience and judgment. On the other hand, the geotechnical information (i.e., the mechanical properties of soils/rocks measured from in situ and laboratory tests) obtained under most of the current site characterization practice is inadequate due to the strict requirement of cost and time duration. As a result, proper characterization of the subsoil conditions may not be achievable (Jaksa *et al.* 2005). In other words, the design value of geotechnical properties may associate with substantial error (Osterberg 1989) and the actual value of a geotechnical property at a specific site can never be obtained with high precision. Therefore, probabilistic treatment is deemed necessary and it is more appropriate to deduce the geotechnical properties statistically based on the appropriate amount of geotechnical information.

A number of researchers proposed probabilistic approaches for systematic treatment of the aforementioned uncertainty in the evaluation of geotechnical properties (Phoon and Kulhawy 1999, Zhang *et al.* 2004, Ching *et al.* 2010, Zhang and Dasaka 2010). However, familiarity with these approaches to evaluate and model the uncertainty is so far not prevalent in practice. Application of their findings in resolving the aforementioned problems has not yet been discovered. On one hand engineers are reluctant to adopt probabilistic analyses because of their lack of proficiency with probability theory (Whitman 1984). On the other hand, it is difficult to isolate and quantify each individual source of uncertainty (Jaksa *et al.* 1997) and it requires considerable budget associated with the quantification of the spatial soil variability. For example, a large amount of data is required to adequately characterize the autocorrelation structure of a soil property for modelling the inherent variability of a soil profile (Asaoka and A-Grivas 1982). One can utilize the published values of the quantified uncertainty level of various types of geotechnical properties for different soils and testing methods (Phoon and Kulhawy 1999). However, the applicability is still limited since there is a wide range of combinations of different geological and geotechnical environments in which the depositional process and stress history may not be similar. As a result, it would be difficult for engineers to select the proper uncertainty level for the site of interest. In this respect, a more objective approach is useful for site characterization.

In geotechnical site characterization practice, it is well recognized that any direct and indirect measurements of the geomaterial properties using laboratory and in-situ testing methods are both expensive and time consuming. For simplicity and cost consideration, a number of empirical correlations have been proposed for the estimation of various important geotechnical properties using easily obtained fundamental soil index properties, such as the normalized undrained shear strength for normally consolidated clays (Skempton 1957, Bjerrum and Simons 1960, Mesri 1975, Larsson 1980, Wroth and Houlsby 1985) and normalized undrained shear strength for over-consolidated clays (Mesri 1975, Ladd *et al.* 1977, Chandler 1988). These correlations, mostly formulated using simple regression analysis, have been widely used in common engineering practice for decades.

Unfortunately, these established empirical correlations provide only the "optimal" estimation but not the associated uncertainty. Therefore, their reliability and applicability are still questionable. It is believed that the shortcomings of these correlations are mainly due to shortage of data and lack of a meticulous method regarding the functional form and the influential soil parameters for developing a probabilistic model with suitable complexity (Mayne 2012, Ng *et al.* 2014). It is anticipated that an empirical model involving more adjustable parameters usually

results in smaller fitting error but a complicated model is more liable to the fitting of the measurement noise and modeling error (Yuen 2010, Yuen *et al.* 2007). Therefore, instead of minimizing the fitting error, a rigorous approach is necessary to determine the suitable complexity of a predictive model. Among different statistical approaches, the Bayesian approach is particular useful for these purposes. The Bayesian probabilistic approach has been successfully applied to many problems in mechanics, structural and geotechnical engineering (e.g., Yuen and Katafygiotis 2005, Zhang *et al.* 2009, Ching *et al.* 2010, Wang *et al.* 2010, Chiu *et al.* 2012a, Kuok and Yuen 2012, Zhang *et al.* 2012, Wang and Cao 2013, Yuen *et al.* 2013, Cao and Wang 2014a, b, Lei *et al*. 2014a, Lei *et al*. 2014b). Recently, the Bayesian model class selection method has been widely used to address various geotechnical issues (Yan *et al.* 2009, Yuen and Mu 2011, Chiu *et al.* 2012b, Cao and Wang 2013, Ng *et al*. 2014). Based on these studies, some of the probabilistic models for the estimation of important geotechnical properties have been developed by use of extensive database to correlate with fundamental soil index properties (e.g., Yan *et al.* 2009, Zhou *et al*. 2013, Ng *et al*. 2014). These models provide an alternative approach for practicing engineers to estimate the geotechnical properties in a more reliable, yet simple and economical manner, compared to direct and indirect measurements.

These probabilistic models contain detailed statistical information so they can be utilized to perform reliability analysis of the prediction. However, since geotechnical properties are generally site dependent, the aforementioned probabilistic models should be updated when new information is acquired from the underlying project site. In the context of Bayesian inference, the aforementioned probabilistic models can be treated as prior models and they can be updated whenever new data points are obtained during the site characterization process. This is important to enhance the reliability and applicability of the predictive models. In order to achieve this goal, this paper presents a new probabilistic method to update the prior probabilistic models in a real-time manner using data points from the concerned project site. The proposed approach is computationally very efficient as it updates the model without iteration or nonlinear optimization once a new data point is acquired. Furthermore, the computation involves a minimum amount of data processing. In this study, the implementation of the proposed probabilistic approach is illustrated through an application for a normally consolidated clay site located in Hong Kong. The prior model for the determination of the undrained shear strength for normally consolidated clays is obtained from a comprehensive database and it will be updated using the geotechnical data obtained from the underlying project site. In order to assess the predictive performance of the updated model, a comparative study with the prior model will also be presented. It reconfirms that the real-time updating is important to enhance the accuracy of the predictive model. The results and procedures of the proposed approach highlighted in this paper provide the basis for geotechnical engineers to plan the site characterization programs in a more reliable manner. This approach can be used for the updating of a wide variety of geotechnical properties in geotechnical projects.

2. Multivariate linear model

First, a probabilistic model can be obtained according to the following process. Since simplicity of a predictive model is important for engineers, a linear correlation model is considered

$$Q = \boldsymbol{a}^T \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_N x_N \tag{1}$$

where Q is the target quantity for prediction; $\boldsymbol{a}$ is the column vector for the unknown regression coefficients; $\boldsymbol{x}$ is the input vector including the independent measured variables; and N is the number of terms in the regression model. Considering the measurement noise and modeling error, the measurement of Q is denoted by y and it is modeled as

$$y = \boldsymbol{a}^T \boldsymbol{x} + \epsilon \tag{2}$$

where ϵ is the aggregate of the measurement noise and modeling error and it is modeled as zero-mean Gaussian random variable with variance σ_ϵ^2. Note that the prediction-error variance σ_ϵ^2 is also an uncertain parameter so the total number of uncertain parameters in this model is $N + 1$.

The prediction errors of different records are assumed statistically independent so the likelihood function is given by (Beck and Katafygiotis 1998)

$$p(D|\boldsymbol{\theta}) = (2\pi)^{-N_o/2} \sigma_\epsilon^{-N_o} \exp\left[-\frac{N_o}{2\sigma_\epsilon^2} J_g(\boldsymbol{a})\right] \tag{3}$$

where D is the database including the measurement of $\boldsymbol{x}$ and the corresponding values of y; N_o is the number of records in D; $\boldsymbol{\theta} = [\boldsymbol{a}^T, \sigma_\epsilon^2]^T$ is the vector of uncertain parameters in the model; and $J_g(\boldsymbol{a})$ is the goodness-of-fit function given by

$$J_g(\boldsymbol{a}) = \frac{1}{N_o} \sum_{n=1}^{N_o} [y(n) - \boldsymbol{a}^T \boldsymbol{x}(n)]^2 \tag{4}$$

A smaller value of the goodness-of-fit function implies better fitting to the data. By using a uniform prior distribution of the coefficients, the optimal vector $\hat{\boldsymbol{a}}$ can be obtained by maximizing the likelihood function $p(D|\boldsymbol{\theta})$ or, equivalently, minimizing the goodness-of-fit function $J_g(\boldsymbol{a})$. The latter can be done by solving the following linear equation

$$\frac{\partial J_g(\boldsymbol{a})}{\partial \boldsymbol{a}} = \boldsymbol{0} \tag{5}$$

and the solution is

$$\hat{\boldsymbol{a}} = \left\{\sum_{n=1}^{N_o} [\boldsymbol{x}(n)\boldsymbol{x}(n)^T]\right\}^{-1} \sum_{n=1}^{N_o} [\boldsymbol{x}(n) y(n)] \tag{6}$$

Similarly, the most optimal value of the prediction-error variance $\hat{\sigma}_\epsilon^2$ can be obtained by solving the following equation

$$\left.\frac{\partial p(D|\boldsymbol{\theta})}{\partial \sigma_\epsilon^2}\right|_{\boldsymbol{a}=\hat{\boldsymbol{a}}} \tag{7}$$

and its solution is given by

$$\hat{\sigma}_\epsilon^2 = J_g(\hat{\boldsymbol{a}}) \tag{8}$$

In other words, the optimal value of the prediction-error variance is the goodness-of-fit function evaluated at the optimal regression parameters. In addition to the optimal estimation of these parameters, the Bayesian approach provides also uncertainty quantification. For large number of data points, the posterior probabilistic density function (PDF) can be well approximated as Gaussian distribution. The mean is the optimal parameters vector and the covariance matrix of the

associated estimation uncertainty is equal to the inverse of the Hessian matrix (Yuen 2010): $\Sigma_{\widehat{\boldsymbol{\theta}}} = H_j(\widehat{\boldsymbol{\theta}})^{-1}$. Specifically, the (l, l') component of the Hessian matrix $H(\widehat{\boldsymbol{\theta}})$ can be obtained by

$$H^{(l,l')}(\widehat{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial\theta_l\theta_{l'}} J(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{9}$$

where the objective function is defined as $J(\boldsymbol{\theta}) = -\ln p(D|\boldsymbol{\theta})$ with the likelihood function $p(D|\boldsymbol{\theta})$ given by Eq. (3); and θ_l is the lth component of the parameter vector $\boldsymbol{\theta}$. In this case, the Hessian matrix is readily obtained

$$H(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} (\hat{\sigma}_\epsilon^2)^{-1} \sum_{n=1}^{N_o} \boldsymbol{x}(n)\boldsymbol{x}(n)^T & \mathbf{0}_{N\times 1} \\ \mathbf{0}_{1\times N} & \frac{1}{2} N_o (\hat{\sigma}_\epsilon^2)^{-2} \end{bmatrix} \tag{10}$$

where $\mathbf{0}_{1\times N}$ is the $1 \times N$ zero row vector and $\hat{\sigma}_\epsilon^2$ is given by Eq. (8).

3. Model class selection

The parametric estimation and the associated uncertainty of a given regression model (i.e., a prescribed functional form on the right hand side of Eq. (1)) can be obtained by the method described in the previous section. Then, Bayesian model class selection method can be used to select the most suitable model class among some possible model class candidates. First, the plausibility of a model class can be obtained by using the Bayes' theorem

$$P(C_j|D, U) = \frac{P(C_j|U)p(D|C_j,U)}{p(D|U)}, \quad j = 1, 2, \dots, N_C \tag{11}$$

where $P(C_j|U)$ is the prior plausibility of the model class C_j; U expresses the user's judgment on the prior plausibility which is taken as uniform in this study, i.e., $P(C_j|U) = 1/N_C$; the denominator $p(D|U) = \sum_{j=1}^{N_C} p(D|C_j, U)P(C_j|U)$ is a normalizing constant that does not depend on the model class; N_C is the number of prescribed model class candidates; and the conditional probability density $p(D|C_j, U)$ is referred to as the evidence of the model class C_j. Note that the user's preference U is irrelevant in $p(D|C_j, U)$ and so it can be dropped from the notation, i.e., $p(D|C_j, U) = p(D|C_j)$. Maximizing $p(D|C_j)$ with respect to j gives the most plausible model class since $P(C_j|U)$ is taken to be uniform. The evidence involves a high dimensional integral with respect to the uncertain parameters in $\boldsymbol{\theta}$ but it can be well approximated with the following asymptotic expansion (Beck and Yuen 2004)

$$p(D|C_j) \approx p(D|\widehat{\boldsymbol{\theta}}, C_j)p(\widehat{\boldsymbol{\theta}}|C_j)(2\pi)^{N_j/2}\left|H_j(\widehat{\boldsymbol{\theta}})\right|^{-1/2}, \quad j = 1,2, \dots, N_C \tag{12}$$

where $\widehat{\boldsymbol{\theta}}$ is the most probable value of model parameters of model class C_j; $p(D|\widehat{\boldsymbol{\theta}}, C_j)$ is the maximum likelihood value; and N_j is the number of uncertain parameters for the model class C_j. The factor $p(\widehat{\boldsymbol{\theta}}|C_j)(2\pi)^{N_j/2}\left|H_j(\widehat{\boldsymbol{\theta}})\right|^{-1/2}$ is the Ockham factor that serves as a measure of

robustness, and it penalizes the model classes that are sensitive to modeling error and measurement noise. For details, please refer to Yuen (2010). By using the Bayesian model class selection method, a regression model with suitable complexity can be chosen among a set of prescribed model class candidates.

4. Real-time updating- based monitoring

The most plausible empirical formula, i.e., the regression model with the most suitable complexity, can be obtained by Bayesian model class selection method. With the parametric identification method presented in Section 2, the regression coefficients and the uncertainty can be quantified. This offline model can be used to predict the target quantity. However, when new data (containing the target quantity and the corresponding input variables for prediction) from the underlying project site becomes available, it can be used to update the predictive model. It is particularly useful because such data is obtained from the underlying project site. Therefore, they are associated with much smaller uncertainty than the data used for the construction of the prior model. The proposed real-time updating approach will update the predictive model once a data point is obtained.

First, the most plausible model class obtained from the Bayesian model class selection method will be fixed for real-time updating. Herein, real-time updating refers to the recursive updating of the regression coefficients once a data point is obtained. Considering *i* data points from the underlying project site, the likelihood function in Eq. (3) will become

$$p(D_i|\boldsymbol{a}) = (2\pi)^{-\frac{N_o+i}{2}}\left(\prod_{n=1}^{N_o+i}\sigma_{\varepsilon_n}^{-1}\right)exp\left\{-\frac{1}{2}\sum_{n=1}^{N_o+i}\frac{[y(n)-\boldsymbol{a}^T\boldsymbol{x}(n)]^2}{\sigma_{\varepsilon_n}^2}\right\} \tag{13}$$

where D_i is the dataset including the training database used for the construction of the prior model and the additional *i* data points from the underlying project site; N_o is the number of data points in the training database; σ_{ε_n} is the standard deviation of fitting-error of *n*th data point; $y(n)$ is the measurement of the target quantity of the *n*th data point; and $\boldsymbol{x}(n)$ is the measurement of the input vector of the *n*th data point.

By using D_i, the optimal coefficient vector $\widehat{\boldsymbol{a}}_i$ can be obtained by maximizing the likelihood function in Eq. (13) or minimizing the following objective function

$$J_i(\boldsymbol{a}) = -\ln p(D_i|\boldsymbol{a}) \tag{14}$$

Furthermore, the fitting-error standard deviation σ_{ε_n} , $n = 1, \dots, N_o$, in the training database can be calculated by

$$\sigma_{\varepsilon_n} = \sigma_0 = \sqrt{\frac{1}{N_o}\sum_{n=1}^{N_o}[y(n) - \widehat{\boldsymbol{a}}_o^T\boldsymbol{x}(n)]^2} \tag{15}$$

where $\widehat{\boldsymbol{a}}_o$ is the optimal coefficient vector obtained using the training database only.

By taking $i = 1$ in Eq. (13), the likelihood function becomes

$$p(D_1|\boldsymbol{a}) = (2\pi)^{-\frac{N_o+1}{2}}\left(\prod_{n=1}^{N_o+1}\sigma_{\varepsilon_n}^{-1}\right)exp\left\{-\frac{1}{2}\sum_{n=1}^{N_o+1}\frac{[y(n)-\boldsymbol{a}^T\boldsymbol{x}(n)]^2}{\sigma_{\varepsilon_n}^2}\right\} \tag{16}$$

and it also can be expanded as follows

$$p(D_1|\boldsymbol{a}) = (2\pi)^{-\frac{N_o}{2}}\sigma_0^{-N_o} exp\left\{-\frac{1}{2\sigma_0^2}\sum_{n=1}^{N_o}[y(n)-\boldsymbol{a}^T\boldsymbol{x}(n)]^2\right\}\times (2\pi)^{-\frac{1}{2}}\sigma_{\varepsilon_{N_o+1}}^{-1} exp\left(-\frac{[y(N_o+1)-\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2}{2\sigma_{\varepsilon_{N_o+1}}^2}\right) \tag{17}$$

Therefore, the likelihood function is proportional to

$$p(D_1|\boldsymbol{a}) \propto exp\left[-\frac{1}{2}(\boldsymbol{a}-\widehat{\boldsymbol{a}}_o)^T A_o(\boldsymbol{a}-\widehat{\boldsymbol{a}}_o)\right]\sigma_{\varepsilon_{N_o+1}}^{-1} exp\left(-\frac{[y(N_o+1)-\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2}{2\sigma_{\varepsilon_{N_o+1}}^2}\right) \tag{18}$$

where A_o is the Hessian matrix of objective function $J_0(\boldsymbol{a}) = -\ln p(D_0|\boldsymbol{a})$, with $\boldsymbol{a}$ evaluated at $\widehat{\boldsymbol{a}}_o$. The (l, l') component of the Hessian matrix A_o is given by

$$A_o{}^{(l,l')} = \frac{\partial^2}{\partial a_l a_{l'}} J(\boldsymbol{a})\Big|_{\boldsymbol{a}=\widehat{\boldsymbol{a}}_o} \tag{19}$$

and the following is readily obtained

$$A_o = \frac{1}{\sigma_0^2}\sum_{n=1}^{N_o}\boldsymbol{x}(n)\boldsymbol{x}(n)^T \tag{20}$$

By using the relationship

$$(\boldsymbol{a}-\widehat{\boldsymbol{a}}_o)^T A_o(\boldsymbol{a}-\widehat{\boldsymbol{a}}_o) = \boldsymbol{a}^T A_o\boldsymbol{a} - 2\boldsymbol{a}^T A_o\widehat{\boldsymbol{a}}_o + \widehat{\boldsymbol{a}}_o{}^T A_o\,\widehat{\boldsymbol{a}}_o \tag{21}$$

and

$$[y(N_o+1)-\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2 = y(N_o+1)^2 - 2y(N_o+1)\boldsymbol{a}^T\boldsymbol{x}(N_o+1) + [\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2 \tag{22}$$

Eq. (18) can be rewritten as

$$p(D_1|\boldsymbol{a}) \propto \sigma_{\varepsilon_{N_o+1}}^{-1} exp\left(-\frac{1}{2}\boldsymbol{a}^T A_o\boldsymbol{a} + \boldsymbol{a}^T A_o\widehat{\boldsymbol{a}}_o - \frac{y(N_o+1)^2}{2\sigma_{\varepsilon_{N_o+1}}^2} + \frac{y(N_o+1)\boldsymbol{a}^T\boldsymbol{x}(N_o+1)}{\sigma_{\varepsilon_{N_o+1}}^2} - \frac{[\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2}{2\sigma_{\varepsilon_{N_o+1}}^2}\right) \tag{23}$$

According to Eq. (14), the objective function can be expressed as

$$J_1(\boldsymbol{a}) = \ln\sigma_{\varepsilon_{N_o+1}} - \left(-\frac{1}{2}\boldsymbol{a}^T A_o\boldsymbol{a} + \boldsymbol{a}^T A_o\boldsymbol{a}_o - \frac{y(N_o+1)^2}{2\sigma_{\varepsilon_{N_o+1}}^2} + \frac{y(N_o+1)\boldsymbol{a}^T\boldsymbol{x}(N_o+1)}{\sigma_{\varepsilon_{N_o+1}}^2} - \frac{[\boldsymbol{a}^T\boldsymbol{x}(N_o+1)]^2}{2\sigma_{\varepsilon_{N_o+1}}^2}\right) + c_o \tag{24}$$

where c_o is a constant that does not depend on $\boldsymbol{a}$. To obtain the updated optimal coefficient vector, one can solve the equation $\frac{\partial J_1(\boldsymbol{a})}{\partial \boldsymbol{a}} = 0$ and the solution is given by

$$\widehat{\boldsymbol{a}}_1 = \left[A_o + \frac{1}{\sigma_{\varepsilon_{N_o+1}}^2}\boldsymbol{x}(N_o+1)\boldsymbol{x}(N_o+1)^T\right]^{-1}\left[A_o\widehat{\boldsymbol{a}}_o + \frac{y(N_o+1)\boldsymbol{x}(N_o+1)}{\sigma_{\varepsilon_{N_o+1}}^2}\right] \tag{25}$$

Therefore, the parameters can be updated once the first data point is available. In the same fashion, one can update the model parameters after the ith data point is acquired

$$\begin{aligned}\widehat{\boldsymbol{a}}_i &= \left[A_{i-1} + \frac{1}{\sigma^2_{\varepsilon_{N_o+i}}}\boldsymbol{x}(N_o+i)\boldsymbol{x}(N_o+i)^T\right]^{-1}\left[A_{i-1}\widehat{\boldsymbol{a}}_{i-1} + \frac{y(N_o+i)\boldsymbol{x}(N_o+i)}{\sigma^2_{\varepsilon_{N_o+i}}}\right] \\ &= {A_i}^{-1}\left[A_{i-1}\widehat{\boldsymbol{a}}_{i-1} + \frac{y(N_o+i)\boldsymbol{x}(N_o+i)}{\sigma^2_{\varepsilon_{N_o+i}}}\right]\end{aligned} \tag{26}$$

where the Hessian matrix can be obtained by the following recursive formula

$$A_i = A_{i-1} + \frac{1}{\sigma^2_{\varepsilon_{N_o+i}}}\boldsymbol{x}(N_o+i)\boldsymbol{x}(N_o+i)^T \tag{27}$$

However, due to the Woodbury matrix identity, Eq. (26) can be rewritten so that it does not require to compute the matrix inverse

$${A_i}^{-1} = A_{i-1}^{-1} - \frac{1}{\sigma^2_{\varepsilon_{N_o+i}} + \boldsymbol{x}(N_o+i)^T A_{i-1}^{-1}\boldsymbol{x}(N_o+i)}A_{i-1}^{-1}\boldsymbol{x}(N_o+i)\boldsymbol{x}(N_o+i)^T A_{i-1}^{-1} \tag{28}$$

The variable $\sigma_{\varepsilon_{N_0+i}}$ is the standard deviation of the fitting error (including the measurement noise and modeling error) of the ith data point from the underlying project site. Since these data points are acquired from the same site for prediction, it is expected that they are more reliable and possess much lower level of uncertainty than the ones in the training dataset for the construction of the prior model. Therefore, the following reduction relationship is assumed

$$\sigma_{\varepsilon_{N_0+i}} = \frac{\sigma_0}{\gamma}, \qquad i = 1, 2, \ldots \tag{29}$$

where σ_0 is the standard deviation of fitting-error of training data and it is given by Eq. (15); and γ is the reduction factor specified by user. In this study, it is taken as 2. Therefore, the recursive Hessian matrix formula and its inverse in Eqs. (27) and (28) can be expressed by

$$\begin{aligned}A_i &= A_{i-1} + \frac{\gamma^2}{\sigma_0^2}\boldsymbol{x}(N_o+i)\boldsymbol{x}(N_o+i)^T \\ {A_i}^{-1} &= A_{i-1}^{-1} - \frac{1}{\frac{\sigma_0^2}{\gamma^2} + \boldsymbol{x}(N_o+i)^T A_{i-1}^{-1}\boldsymbol{x}(N_o+i)}A_{i-1}^{-1}\boldsymbol{x}(N_o+i)\boldsymbol{x}(N_o+i)^T A_{i-1}^{-1}\end{aligned} \tag{30}$$

It is noted that outlier problem can be a concern and the recently developed Bayesian method (Yuen and Mu 2012, Mu and Yuen 2015) can be considered to screen out the outliers.

5. Application

In this section, an application of the undrained shear strength estimation is presented to demonstrate the proposed real-time updating method. Consider the following full model for undrained shear strength for normally consolidated clay (Ng *et al.* 2014)

$$S_u = a_1\sigma_v' + a_2 W_n + a_3 LI + a_4 W_n \cdot LI + a_5 LL \cdot LI + a_6 \tag{31}$$

Table 1 Statistics of the training database

Geographical Regions	Asia (n=93)		Americas (n=42)		Europe (n=89)		Oceania (n=38)	
Statistics	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
S_u (kPa)	30.53	17.03	27.16	14.86	28.69	16.93	17.95	5.06
σ_v' (kPa)	97.99	60.94	98.70	53.94	95.17	69.01	60.07	22.71
PL (%)	34	11	28	8	25	8	32	12
W_n (%)	82	33	58	20	61	25	82	27
LL (%)	83	32	58	22	69	22	80	19

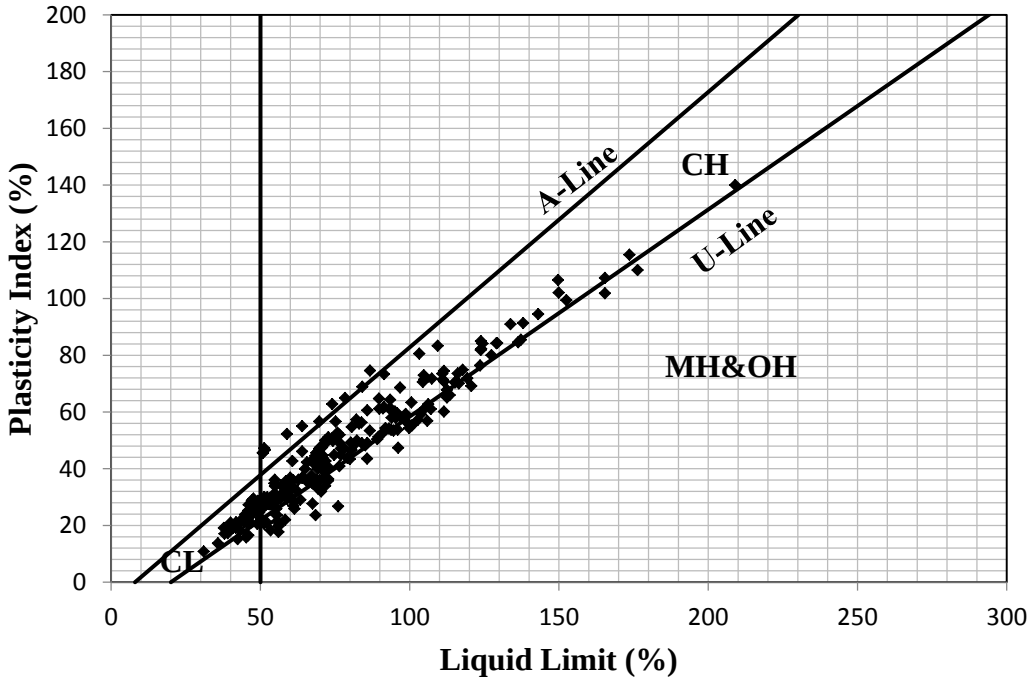

Fig. 1 A-Line chart for classification of the clay types in the training database

where S_u is the undrained shear strength; σ_v' is the effective overburden pressure; W_n is the water content of soil; LI is the liquidity index; and LL is the liquid limit. The constant term is enforced to be included in the model class candidates and models with subsets of terms of this full model are considered as potential model class candidates except the one with the constant a_6 only. There are $2^5-1=31$ model class candidates in this case. A comprehensive training database with 262 data points was compiled (Ng *et al.* 2014) and the information of training database is shown in Table 1. Furthermore, Fig. 1 shows the A-line chart for the records of this training database. According to the Unified Soil Classification System (USCS), the clay samples in this database are classified into high plasticity clay (CH).

Using the Bayesian model class selection method, the most plausible model is shown as follows (Ng *et al.* 2014)

$$S_u = 0.2335\sigma_v' - 2.6915W_n \cdot LI + 8.9657 \quad (32)$$

Therefore, the initial optimal parameter vector is given by: $\hat{\boldsymbol{a}}_o = [0.2335, 2.6915, 8.9657]^T$. Furthermore, the fitting error of the prior model was given by: $\sigma_{\varepsilon_{N_0}} = 6.2994$. For more details, please refer to Ng *et al*. (2014).

To demonstrate the proposed real-time updating method, an additional independent testing database of normally consolidated clay in Hong Kong with 51 soil records has been extracted from Lumb and Holt (1968). In the same fashion, Fig. 2 shows the A-line chart for the records from the underlying project site. Again, the clay samples in this database are classified into high plasticity clay (CH) but it can be clearly seen that the distribution is different from the training database to certain extent. In view of all these, prediction using the prior model in Eq. (32) provides satisfactory estimation but there is certainly significant room for improvement with real-time updating. This will be further elaborated in the followings.

By using the proposed method, the prior predictive model can be updated in the real-time manner. Since the identification results depends on the sequence of the data points used in the updating process, the order of the data points is randomly selected and 1000 independent runs are conducted. Fig. 3 shows the parameter estimation time history in a typical run. It is not surprising to observe minor fluctuation in the time histories but a main trend can be clearly observed. This reconfirms the importance of the real-time updating using data from the underlying project site. The final updated model after using all the 51 data points is

$$S_u = 0.2501\sigma_v' - 1.9860W_n \cdot LI + 6.1663 \quad (33)$$

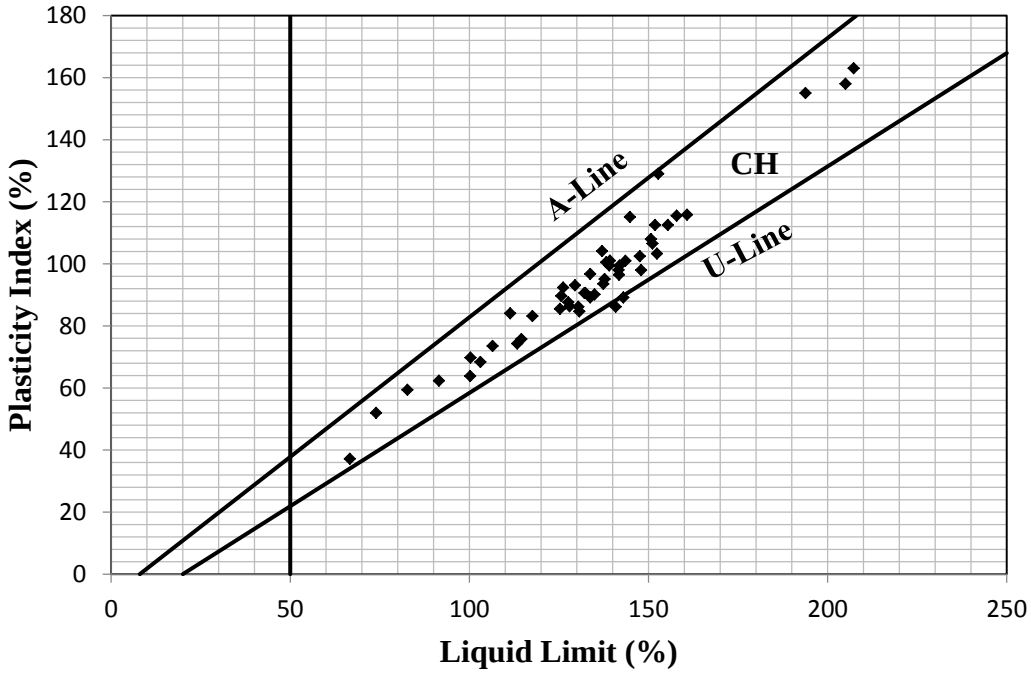

Fig. 2 A-Line chart for the clay samples for real-time updating in the site of concern

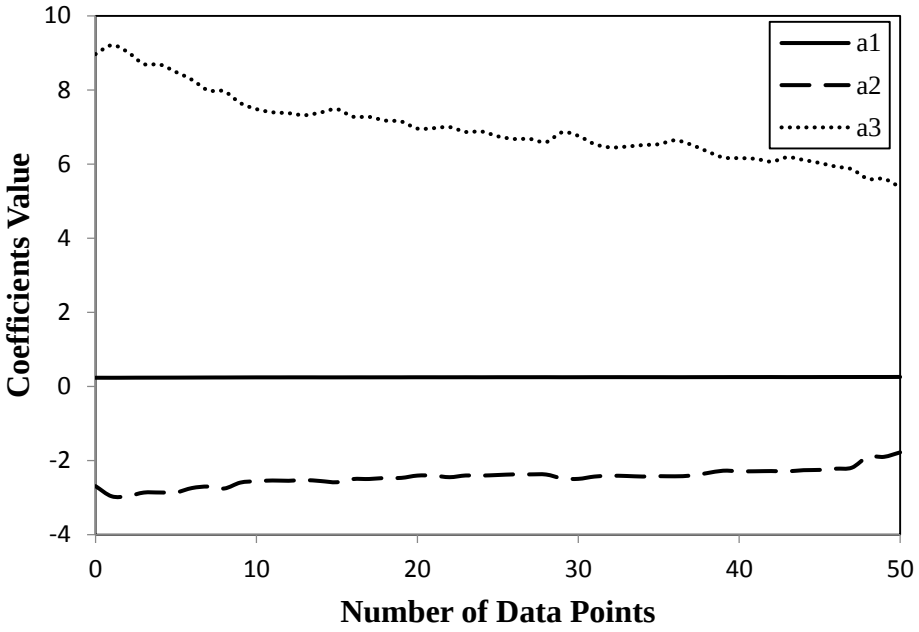

Fig. 3 A typical updated history of the regression coefficients

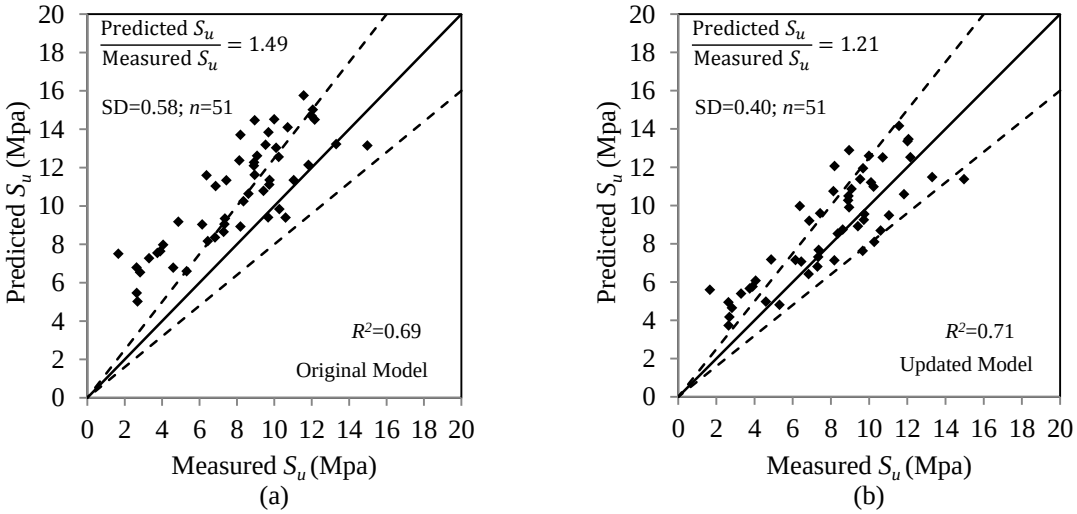

Fig. 4 Comparative predict results of original model and updated model

To justify the necessity of the real-time updating of the predictive model, a comparison study is performed using the prior model (Fig. 4(a)) in Eq. (32) and the final updated model (Fig. 4(b)) in Eq. (33) in predicting the undrained shear strength. The two dashed lines in each subplot are ±20% from the 1:1 perfect agreement line (solid line) and they are used to facilitate the visual judgment of the model performance. The average of the ratio between the predicted and measured values is shown on the left top corner in each subplot, and SD indicates the standard deviation of the

predicted-measured ratio. Both average ratios are larger than 1 and this implies that both models overestimate the undrained shear strength. However, the updated model corrects itself by adopting the new information from the data points of the underlying project site so this bias is substantially reduced. Furthermore, the standard deviation of the updated model is significantly smaller than the prior model, indicating improved prediction results.

Once a data point is obtained, the predictive model (i.e., the coefficient vector $\widehat{\boldsymbol{a}}$ and the Hessian matrix) can be updated and it can be used to predict the rest of the data points. Specifically, after the updating with the ith data point, the model will be used to predict the (i+1) th, (i+2) th,…, data points and the associated prediction errors are given by

$$\varepsilon(n) = y(n) - \widehat{\boldsymbol{a}}_i^T \boldsymbol{x}(n) \qquad n = i+1, i+2, \ldots \tag{34}$$

Two statistical indicators are used for the assessment of the prediction performance and they are the mean absolute error (MAE)

$$\mathrm{MAE}(i) = \frac{1}{51-i}\sum_{n=i+1}^{51}|\varepsilon(n)| \qquad i = 0,1,2,\ldots,50 \tag{35}$$

and the root-mean-squares error (RMSE)

$$\mathrm{RMSE}(i) = \sqrt{\frac{1}{51-i}\sum_{n=i+1}^{51}\varepsilon(n)^2} \qquad i = 0,1,2,\ldots,50 \tag{36}$$

Then, the averages of the MAE and RMSE over the 1000 aforementioned independent runs are computed and they are shown in Fig. 5. From this figure, it can be seen that the MAE and RMSE are reduced by 40% and 50%, respectively, by using the in-situ data for real-time updating. Both curves exhibit rapid decrease at the beginning but the rate slows down gradually. This is because

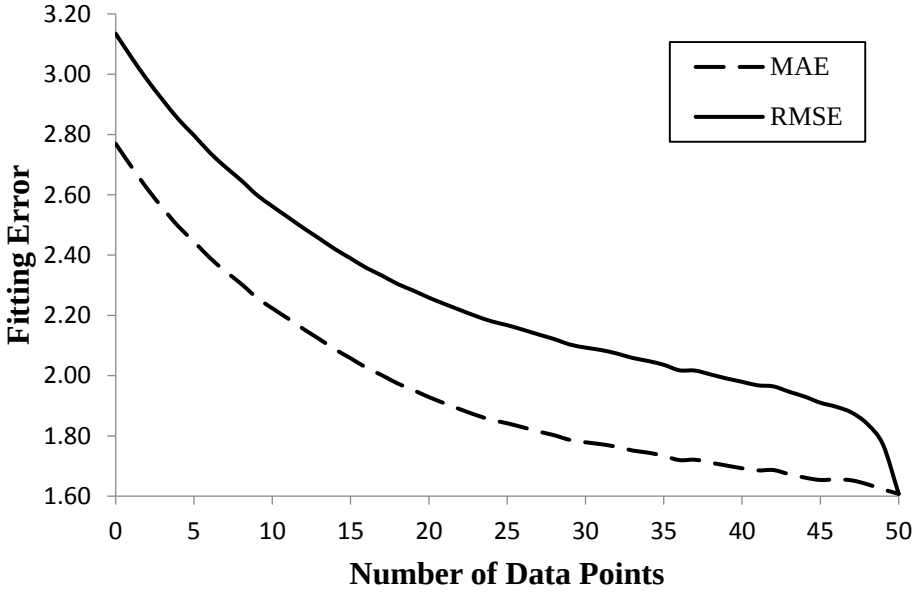

Fig. 5 Prediction error statistics by the real-time updated model

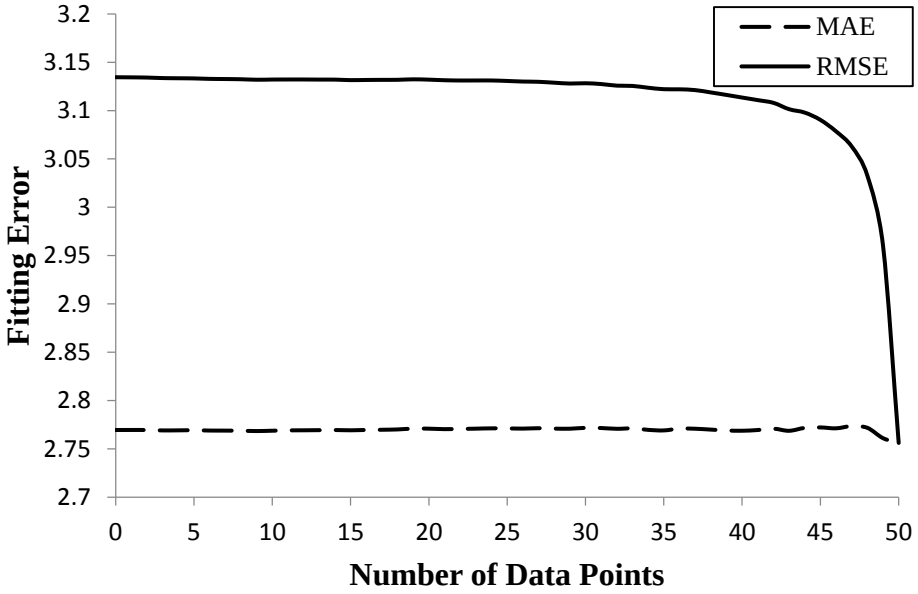

Fig. 6 Prediction error statistics by the prior model

the posterior uncertainty in terms of variance is inversely proportional to the number of data points. On the other hand, the curves converge to the same point because MAE and RMSE are identical when there is only one last data point for prediction (i=50). The error analysis is repeated for the prior model without updating and the results are shown in Fig. 6. It can be clearly seen that the prediction errors are much larger than those in Fig. 5 so it reconfirms the advantage of the proposed real-time updating.

6. Conclusions

Due to high level of complexity and uncertainty in the geomaterial properties, it is challenging to obtain reliable and representative estimation in site characterization practice. In this paper, a probabilistic real-time updating approach is proposed for this purpose. It uses a prior model constructed by use of a comprehensive database and this model can be updated once a new data point is obtained from the underlying project site. An application is presented to demonstrate the effectiveness and efficiency of the proposed method. In this example, a recently developed empirical model for the undrained shear strength of normally consolidated clays was adopted and it was updated using the data from a site located in Hong Kong. Compared with the prior model, it reconfirms that the updated model enhances the predictive performance so the proposed method is useful for reliability analysis and reliability-based design of geotechnical projects.

Acknowledgments

This work was supported by the Science and Technology Development Fund of the Macau

SAR government under Research Grant No. 012/2013/A1. This generous support is gratefully acknowledged.

References

Asaoka, A. and A-Grivas, D. (1982), "Spatial variability of the undrained strength of clays", *J. Geotech. Eng.*, ASCE, **108**(5), 743-756.

Baecher, G.B. and Christian, J.T. (2003), *Reliability and Statistics in Geotechnical Engineering*, John Wiley & Sons, Hoboken, New Jersey.

Beck, J.L. and Katafygiotis, L.S. (1998), "Updating models and their uncertainties. I: Bayesian statistical framework", *J. Eng. Mech*., **124**(4), 455-461.

Beck, J.L. and Yuen, K.V. (2004), "Model selection using response measurements: Bayesian probabilistic approach", *J. Eng. Mech*., **130**(2), 192-203.

Bjerrum, L. and Simons, N.E. (1960), "Comparison of shear strength characteristics of normally consolidated clays", *Proceedings of the 1st ASCE Specialty Conference on Shear Strength of Cohesive Soils*, Boulder, Colorado, 711-726.

Cao, Z. and Wang, Y. (2013), "Bayesian approach for probabilistic site characterization using cone penetration tests", *J. Geotech. Geoenviron. Eng*., **139**(2), 267-276.

Cao, Z. and Wang Y. (2014a), "Bayesian model comparison and characterization of undrained shear strength", *J. Geotech. Geoenviron. Eng*., **140**(6), Article number 04014018.

Cao, Z. and Wang Y. (2014b), "Bayesian model comparison and selection of spatial correlation functions for soil parameters", *Struct. Saf*., **49**, 10-17.

Chandler, R.J. (1988), "The in-situ measurement of the undrained shear strength of clays using the field vane", Vane Shear Strength Testing in Soils: Field and Laboratory Studies, 13-44.

Ching, J., Phoon, K.K. and Chen, Y.C. (2010), "Reducing shear strength uncertainties in clays by multivariate correlations", *Can. Geotech. J*., **47**(1), 16-33.

Chiu, C.F., Yan, W.M. and Yuen, K.V. (2012a), "Estimation of water retention curve of granular soils from particle-size distribution-a Bayesian probabilistic approach", *Can. Geotech. J*., **49**(9), 1024-1035.

Chiu, C.F., Yan, W.M. and Yuen, K.V. (2012b), "Reliability analysis of soil-water characteristics curve and its application to slope stability analysis", *Eng. Geol*., **135**, 83-91.

Hvorslev, M.J. (1949), "Subsurface exploration and sampling of soils for civil engineering purposes", Waterways Experiment Station, Vicksburg.

Jaksa, M.B., Brooker, P.I. and Kaggwa, W.S. (1997), "Inaccuracies associated with estimating random measurement errors", *J. Geotech. Geoenviron. Eng*., **123**(5), 393-401.

Jaksa, M.B., Goldsworthy, J.S., Fenton, G.A., Kaggwa, W.S., Griffiths, D.V., Kuo, Y.L. and Poulos, H.G. (2005), "Towards reliable and effective site investigations", *Geotechnique*, **55**(2), 109-121.

Kaloop, M.R., Sayed, M.A., Kim, D. and Kim, E. (2014), "Movement identification model of port container crane based on structural health monitoring system", *Struct. Eng. Mech*., **50**, 105-119.

Kulhawy, F.H. and Trautmann, C.H. (1996), "Estimation of in-situ test uncertainty", *Uncert. Geol. Envir., F. Theory Pract*., Geotechnical Special Publication, **58**(1), 269-286.

Kuok, S.C. and Yuen, K.V. (2012), "Structural health monitoring of Canton tower using Bayesian framework", *Smart Struct. Syst*., **10**(4-5), 375-391.

Ladd, C.C., Foote, R., Ishihara, K., Schlosser, F. and Poulos, H.G. (1977), "Stress-deformation and strength characteristics", *Proceedings of 9th International Conference on Soil Mechanics and Foundation Engineering*, **2**, Tokyo.

Larsson, R. (1980), "Undrained shear strength in stability calculation of embankments and foundations on clays", *Can. Geotech. J*., **17**(4), 591-602.

Lei, Y., Lai, Z., Zhu, S. and Zhang, X. (2014a), "Experimental study on impact induced damage detection using an improved extended Kalman filter", *Int. J. Struct. Stab. Dyn*., **14**(5), 1440007.

Lei, Y., Chen, F. and Zhou, H. (2014b), "An algorithm based on two-step Kalman filter for intelligent structural damage detection", *Struct. Control Hlth. Monit*., **22**, 694-706.
Lumb , P. (1966), "The variability of natural soils", *Can. Geotech. J*., **3**(2), 74-79.
Lumb, P. and Holt, J.K. (1968), "The undrained shear strength of a soft marine clay from Hong Kong", *Geotechnique*, **18**(1), 25-36.
Mayne, P.W. (2012), "Quandary in geomaterial characterization: new versus the old", Shaking the Foundations of Geo-engineering Education, 15-26.
Mesri, G. (1975), "Discussion on new design procedures for stability of soft clays", *J. Geotech. Eng*., ASCE, **101**(4), 409-412.
Mu, H.Q. and Yuen, K.V. (2015), "Novel outlier-resistant extended Kalman filter for robust online structural identification", *J. Eng. Mech*., ASCE, **141**(1), 04014100.
Ng, I.T., Yuen, K.V. and Dong, L. (2014), "Nonparametric estimation of undrained shear strength for normally consolidated clays", *Marine Geores. Geotech.*, DOI:10.1080/1064119X.2014.970305.
Osterberg, J.O. (1989), "Necessary redundancy in geotechnical engineering", *J. Geotech. Eng.*, ASCE, **115**(11), 1513-1531.
Phoon, K.K. and Kulhawy, F.H. (1999), "Characterization of geotechnical variability", *Can. Geotech. J*., **36**(4), 612-624.
Rezaiee-Pajand, M. and Kazemiyan, M.S. (2014), "Damage identification of 2D and 3D trusses by using complete and incomplete noisy measurements", *Struct. Eng. Mech*., **52**, 149-172.
Skempton, A.W. (1957), "Discussion of the planning and design of the new Hong Kong Air Port", *Proceedings of the Institution of Civil Engineers*, London, **7**(2), 305-307.
Standard, B. (1981), *Code of Practice for Site Investigations*, British Standards Institution, London.
Vanmarcke, E.H. (1983), *Random Fields*, MIT Press, Cambridge.
Wang, V.Z. and Ginger, J.D. (2014), "Maximum a posteriori estimation based wind fragility analysis with application to existing linear or hysteretic shear frames", *Struct. Eng. Mech*., **50**, 653-664.
Wang, Y., Au, S.K. and Cao, Z. (2010), "Bayesian approach for probabilistic characterization of sand friction angles", *Eng. Geol*., **114**(3-4), 354-363.
Wang, Y. and Cao, Z. (2013), "Probabilistic characterization of Young's modulus of soil using equivalent samples", *Eng. Geol*., **159**, 106-118.
Whitman, R.V. (1984), "Evaluating calculated risk in geotechnical engineering", *J. Geotech. Eng.*, ASCE, **110**(2), 145-188.
Wroth, C.P. and Houlsby, G.T. (1985), "Soil mechanics-property characterization and analysis procedures", *Proceedings of the 11th International Conference on Soil Mechanics and Foundation Engineering*, **1**, San Francisco.
Yan, W.M., Yuen, K.V. and Yoon, G.L. (2009), "Bayesian probabilistic approach for the correlations of compressibility index for marine clays", *J. Geotech. Geoenviron. Eng*., **135**(12), 1932-1940.
Yuen, K.V. (2010), "Recent developments of Bayesian model class selection and applications in civil engineering", *Struct. Saf*., **32**(5), 338-346.
Yuen, K.V., Hoi, K.I. and Mok, K.M. (2007), "Selection of noise parameters for Kalman filter", *Earthq. Eng. Eng. Vib*., **6**(1), 49-56.
Yuen, K.V. and Katafygiotis, L.S. (2005), "Model updating using noisy response measurements without knowledge of the input spectrum", *Earthq. Eng. Struct. Dyn*., **34**(2), 167-187.
Yuen, K.V., Liang, P.F. and Kuok, S.C. (2013), "Online estimation of noise parameters for Kalman filter", *Struct. Eng. Mech*., **47**(3), 361-381.
Yuen, K.V. and Mu, H.Q. (2011), "Peak ground acceleration estimation by linear and nonlinear models with reduced order Monte Carlo simulation", *Comput. Aid. Civil Infrastr. Eng*., **26**(1), 30-47.
Yuen, K.V. and Mu, H.Q. (2012), "Novel probabilistic method for robust parametric identification and outlier detection", *Probab. Eng. Mech*., **30**, 48-59.
Zhang, L. M., Tang, W. H., Zhang, L.L. and Zheng, J. G. (2004), "Reducing uncertainty of prediction from empirical correlations", *J. Geotech. Geoenviron. Eng*., **130**(5), 526-534.
Zhang, L.M. and Dasaka, S.M. (2010), "Uncertainties in geologic profiles versus variability in pile founding

depth", *J. Geotech. Geoenviron. Eng*., **136**(11), 1475-1488.
Zhang, J., Zhang, L.M. and Tang, W.H. (2009), "Bayesian framework for characterizing geotechnical model uncertainty", *J. Geotech. Geoenviron. Eng*., **135**(7), 932-940.
Zhang, J., Tang, W.H., Zhang, L.M. and Huang, H.W. (2012), "Characterising geotechnical model uncertainty by hybrid Markov Chain Monte Carlo simulation", *Comput. Geotech*., **43**, 26-36.
Zhou, W.H., Yuen, K.V. and Tan, F. (2013), "Estimation of maximum pullout shear stress of grouted soil nails using Bayesian probabilistic approach", *Int. J. Geomech*., **13**(5), 659-664.