Delft University of Technology

Predicting subsurface classification in 2D from cone penetration test data

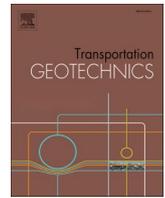Varkey, Divya; Hicks, Michael A.; van den Eijnden, Abraham P.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Predicting subsurface classification in 2D from cone penetration test data

Divya Varkey, Michael A. Hicks [*], Abraham P. van den Eijnden

Geo-Engineering Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands

ABSTRACT

Uncertainty is inevitable in the characterisation of a geotechnical site, especially due to the inherently heterogeneous nature of the ground. In this paper, a method for characterising a subsurface with limited cone penetration test (CPT) data is proposed. The method is based on integrating predictions of CPT parameters with a probabilistic approach for subsoil classification at the CPTs. The predicted stratigraphy is able to capture the spatial variability of soil measured via CPTs and takes account of the uncertainties that arise from transforming CPT measurements into soil units as well as errors in the measurements themselves. The applicability of the proposed method is demonstrated for a site in the Netherlands. The results show that the proposed approach can identify the most likely classification in the domain with good accuracy. Furthermore, the significance of considering the uncertainties in predicting the most likely classification is illustrated via finite element stability analyses of a slope cut-out in the domain.

Introduction

One of the first steps in the analysis of a geotechnical structure is the characterisation of the site, for which field measurements are generally used. The cone penetration test (CPT) is a widely used field investigation technique, which, unlike other traditional investigation techniques, does not involve drilling, is fast, repeatable and gives near-continuous data over its entire depth. However, uncertainty is inevitable in the characterisation and analysis of structures due to the inherently heterogeneous nature of soils. For example, the subsurface is composed of different soil units that were formed due to a combination of various geological, environmental and physicochemical processes [13], resulting in spatial variability of the units and of the properties within the units [12]. Furthermore, the number of CPTs performed at a site is often limited and soil samples are often not collected during the CPTs to assist in soil classification, so that methods for transforming CPT measurements into soil units are required, adding to the already challenging task of site characterisation.

Much research has been done on the influence of spatial variability of soil properties in the analyses of geotechnical structures. For example, Hicks and Samy [4], Griffiths et al. [3], Hicks and Spencer [5], Huang et al. [10], Hicks et al. [6], Hicks and Li [7], Hicks et al. [8] and Varkey et al. [16] have used the random finite element method for investigating the influence of 2D and 3D spatial variability of strength parameters in the reliability-based assessments of slopes. Research on the

characterisation of the subsurface is more limited, although it is now receiving increasing attention. Cao and Wang [1] and Wang et al. [20,21] have proposed methods, based on integrating the Robertson classification chart [15] with a Bayesian approach, for determining the subsurface classification at the test location (i.e. only in the 1D depth direction). Qi et al. [14], Xiao et al. [23] and Zhang et al. [24] have used a coupled Markov chain method for predicting soil units between the measurement locations. In this method, the transitions from one soil unit to another along the vertical and horizontal directions are modelled as Markov processes, which are coupled together to predict soil units at unsampled locations. Varkey et al. [17] have highlighted the drawbacks of this method in estimating the transition probability matrices from a limited dataset and in potentially violating the principles of Markov process stationarity due to the coupling. Wang et al. [22] proposed methods for subsoil classification in a 2D domain by interpolating CPT data in the domain using Bayesian compressive sampling. Hu and Wang [9] combined the interpolated CPT data with the soil behaviour type index (I_c) to classify the subsurface. The uncertainties in the classification were addressed by considering the uncertainty in the I_c classification boundaries, where the latter was based on gauging the I_c data variability in the existing literature.

In this paper, a method for subsoil classification in 2D is proposed by transforming CPT parameter predictions to soil units while accounting for uncertainties in the measurements and transformation. The method is based on integrating the CPT predictions with the method by Wang
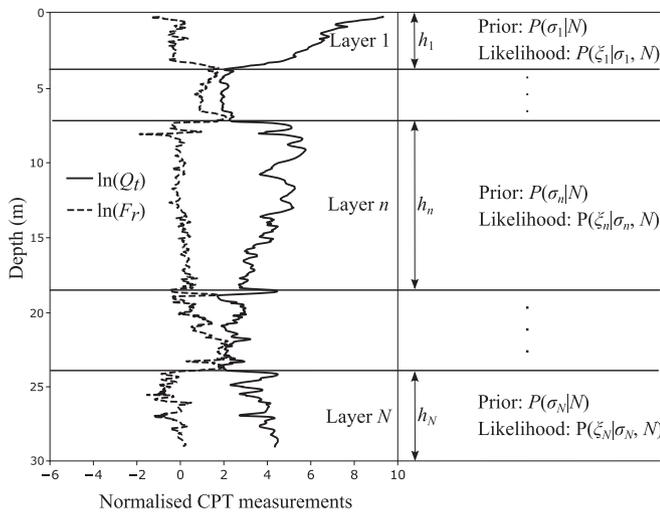
Fig. 1. Illustration of the probabilisitic 1D soil classification based on Wang et al. [20].

Fig. 2. Flowchart of the proposed approach.

et al. [20] within the Bayesian framework, to obtain prior knowledge of the soil distributions in a domain as well as the updated knowledge of soil distributions in the domain. In the following section, a method for predicting CPT parameters between limited measurement locations is discussed and this is followed by a detailed explanation of the proposed probabilistic approach for the transformation to soil units. The subsequent sections discuss the results obtained by applying the proposed approach to a site in the Netherlands.

(a)

(b)

Fig. 4. Vertical scale of fluctuation estimated from (a) $\ln(F_r)$ and (b) $\ln(Q_t)$.
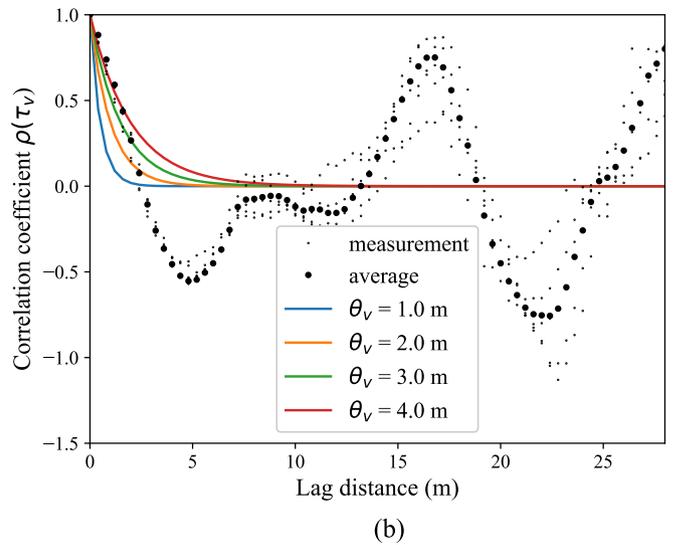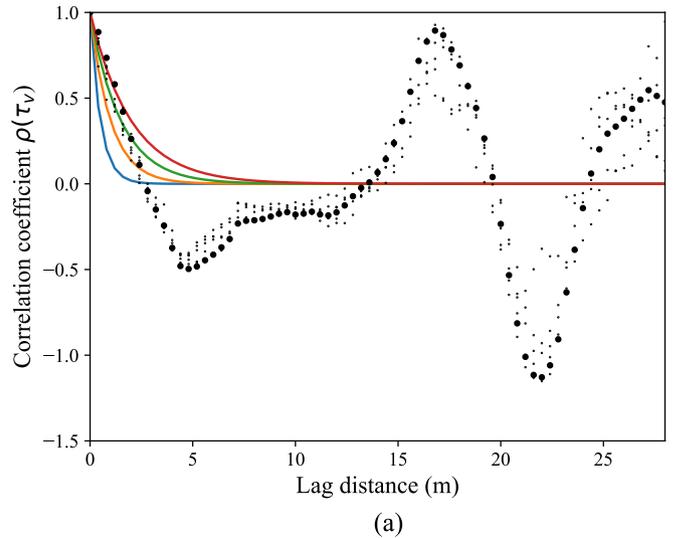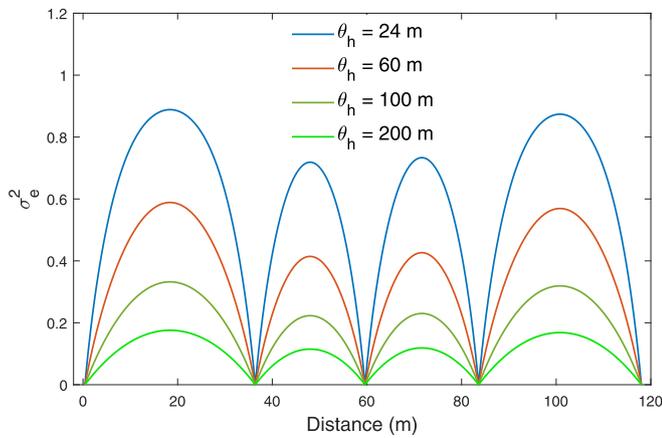

Fig. 3. Relative locations of five CPTs from the Groningen region in the Netherlands.

Fig. 5. Variance of the estimated field generated using different values of θ_h.

Methodology

In this paper, a two-step approach is proposed for the classification of a 2D domain with limited CPT measurements. In the first step, the CPT parameters, i.e. the normalised friction ratio $\ln(F_r)$ and the normalised tip resistance $\ln(Q_t)$ are predicted between the measurement locations by making use of the spatial nature of soil variability. In the second step, the estimated CPT parameters are transformed into soil units. This can be achieved by mapping the estimated parameter values onto, for example, the Robertson chart and thereby classifying each point in the domain into one of the nine soil units. However, this method of classification, based on past observations and experiences, inevitably contains various uncertainties arising from measurement errors and transformation uncertainties, amongst others. Therefore, in order to take account of these uncertainties, a probabilistic subsoil classification in 1D is first carried out at each CPT location. The results of the 1D probabilistic analyses are then integrated with the Kriged CPT predictions to predict updated soil distributions over the 2D domain. An explanation of the various steps followed are given below, along with a detailed implementation procedure. Although the approach has been applied on the Robertson chart, it may be noted that the same approach can easily be adapted for other classification charts as well. This would, for example, allow incorporating additional site-specific knowledge into the classification charts.

Predicting CPT parameters at unsampled locations

Two-dimensional Kriging [11] has been adopted to interpolate CPT parameter fields in the 2D domain by conditioning the fields at the CPT measurements. Kriging is a 'Best Linear Unbiased Estimation' method

that gives the best estimation by minimising the variance with respect to the true field and incorporates the spatial correlation of points into the interpolation procedure. The Kriged interpolation of a parameter Z at some location $\boldsymbol{x}_0\ (=(x_0, y_0))$ is given by

$$\widehat{Z} = \sum_{i=1}^{m} \lambda_i Z_i \tag{1}$$

where $Z_i = Z_1, Z_2, ..., Z_m$ are the m measured values of Z at points $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m$, respectively, and λ_i are the weights at those points. Ordinary Kriging has been used in this paper such that $\sum_{i=1}^{m} \lambda_i = 1$.

The weights in Eq. (1) can be determined by minimising the variance (σ^2_e) of the Kriged predictions with respect to the true field, i.e., $\sigma^2_e = \mathrm{E}\left[(\widehat{Z} - Z)^2\right]$ where $\mathrm{E}[]$ is the expectation operator. To minimise the error variance, the Lagrange method with parameter μ has been used in this paper, which leads to the following equation (see Wackernagel [18] for a detailed derivation):

$$\begin{pmatrix} \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_1) & \cdots & \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_m) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\boldsymbol{x}_m - \boldsymbol{x}_1) & \cdots & \gamma(\boldsymbol{x}_m - \boldsymbol{x}_m) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_0) \\ \vdots \\ \gamma(\boldsymbol{x}_m - \boldsymbol{x}_0) \\ 1 \end{pmatrix} \tag{2}$$

where $\gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)$ is the variogram between the points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Assuming second order stationary data, the variogram can be calculated using

$$\gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) = \sigma^2 \times (1 - \rho(\tau_h, \tau_v)) \tag{3}$$

$$\rho(\tau_h, \tau_v) = \exp\left(-\sqrt{\left(\frac{2\tau_h}{\theta_h}\right)^2 + \left(\frac{2\tau_v}{\theta_v}\right)^2}\right)$$

where σ^2 is the variance of the field, $\rho(\tau_h, \tau_v)$ is the assumed autocorrelation function between the two points, $\tau_h = |x_i - x_j|$ and $\tau_v = |y_i - y_j|$ are the lag distances between the points in the horizontal and vertical directions, and θ_h and θ_v are the scales of fluctuation in the horizontal and vertical directions, respectively.

Transforming CPT parameters into soil units

The Kriged predictions of CPT parameters are transformed into soil units while accounting for uncertainties in the measurements and transformation. This is proposed to be carried out in two sub-steps: using the CPT measurements to predict the most likely 1D classification at the measurement locations, and using the results of the 1D classification to predict the classification over the 2D domain.

$\theta_h = 24$ m $\theta_h = 100$ m

(a) $\ln(F_r)$

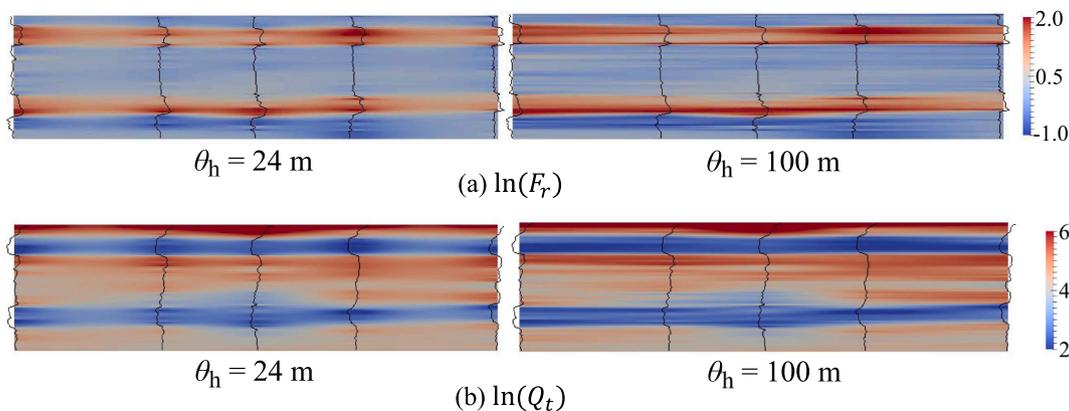$\theta_h = 24$ m $\theta_h = 100$ m

(b) $\ln(Q_t)$

Fig. 6. Kriged predictions of (a) $\ln(F_r)$ and (b) $\ln(Q_t)$ obtained using different values of θ_h. The CPT profiles have been included to indicate the conditioning locations.
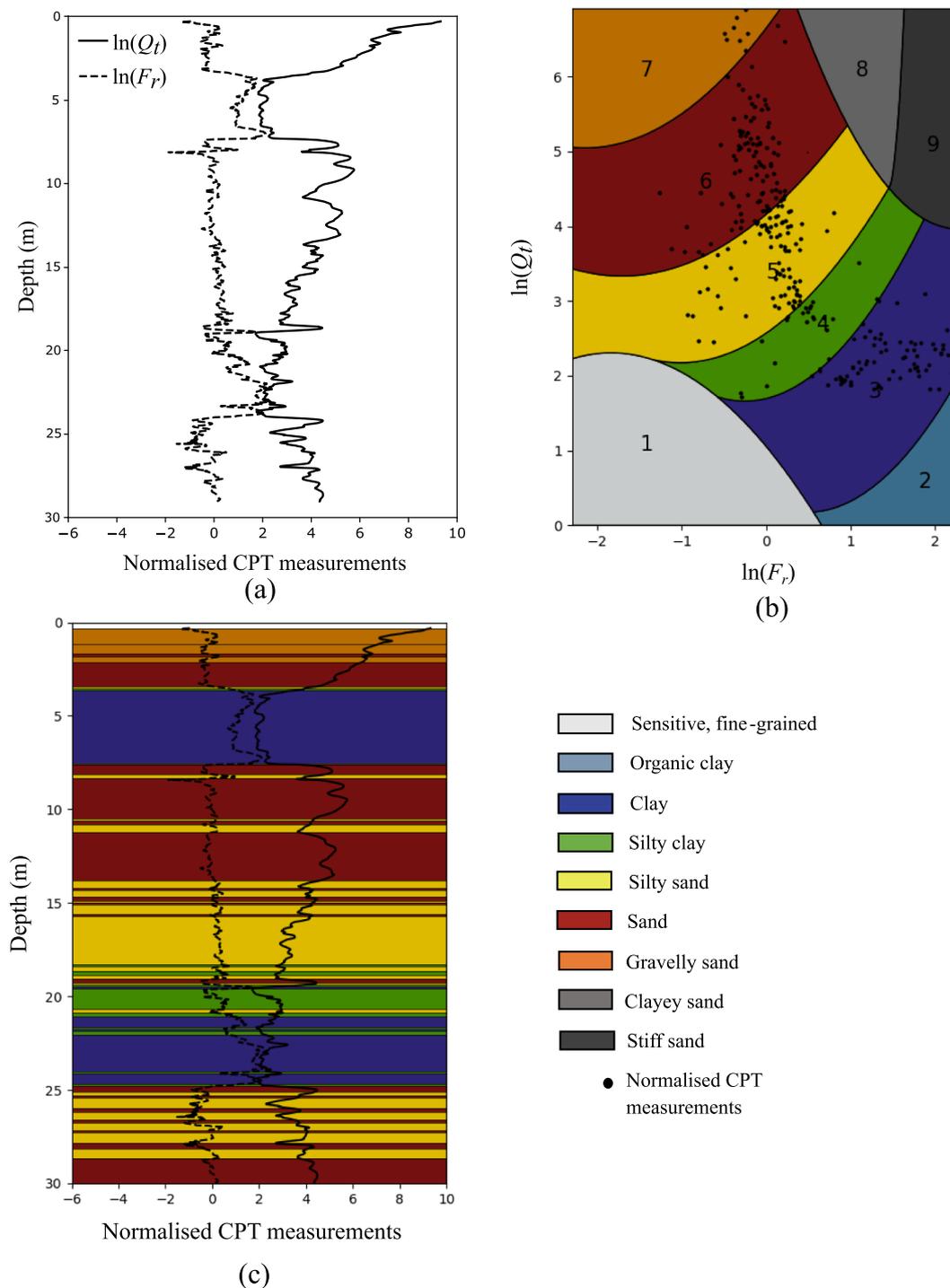
(a)

(b)

(c)

Fig. 7. Conventional method of interpreting CPT data using the Robertson chart: (a) the normalised measurements at CPT3; (b) measurements plotted on the Robertson chart; (c) deterministic classification at CPT3.

Predicting the most likely classification in 1D

The probabilistic approach by Wang et al. [20], following certain modifications, has been used in this paper for predicting the most likely 1D classification of the subsurface at the CPT locations. In this method, uncertainties in the measurements and transformation are accounted for by integrating the Robertson chart with the Bayesian approach. A brief overview of this method is given below and a detailed explanation can be found in Wang et al. [20].

In this method, instead of performing a deterministic classification of the soil at a depth i with normalised CPT measurements ξ^i $(= (\ln(F_r^i), \ln$

$(Q_t^i))$, the probability of the soil being classified as one of the nine units is determined through a Gaussian probability density function (pdf) centred at $(\ln(F_r^i), \ln(Q_t^i))$ in the Robertson chart. The standard deviation of this Gaussian distribution reflects the uncertainties in the measurements and in the Robertson chart (i.e. the transformation uncertainty). The classification is carried out in two steps: (a) for a given number of layers, identifying the most probable soil layer thicknesses, and (b) identifying the most likely number of layers which, together with the most probable thicknesses, maximises the likelihood of observing the CPT measurements.

Table 1
Results of the probabilistic subsoil classification at CPT3.

Case	$\ln P(\xi\|M_N)$	h^*_1 (m)	h^*_2 (m)	h^*_3 (m)	h^*_4 (m)	h^*_5 (m)	h^*_6 (m)	h^*_7 (m)	h^*_8 (m)	h^*_9 (m)
M_1	−536.2	29.6	–	–	–	–	–	–	–	–
M_2	−435.9	2.6	27.0	–	–	–	–	–	–	–
M_3	−381.7	2.8	4.0	22.8	–	–	–	–	–	–
M_4	−336.9	2.6	4.1	7.6	15.3	–	–	–	–	–
M_5	−308.2	2.8	4.0	11.9	5.4	5.5	–	–	–	–
M_6	−287.7	2.8	4.0	7.6	4.2	6.0	5.0	–	–	–
M^*_7	−286.2	2.8	4.0	6.4	7.2	4.4	2.8	2.0	–	–
M_8	−290.9	2.8	4.1	2.6	5.2	4.3	5.4	3.1	2.1	–
M_9	−294.6	2.6	4.1	2.4	5.4	4.5	1.3	4.1	2.6	2.6

As illustrated in Fig. 1, $P(\sigma_n|N)$ and $P(\xi_n|\sigma_n, N)$ are the prior distribution of the model parameters and the likelihood function for the n th layer, respectively, of a subsoil classified into N layers. Here, '|' stands for conditionality, $\sigma_n = (\sigma_{F_r,n}, \sigma_{Q_t,n})$ denotes the set of standard deviations of the joint Gaussian distribution for the n th layer, and σ_{F_r} and σ_{Q_t} are the standard deviations of $\ln(F_r)$ and $\ln(Q_t)$, respectively. Since all the data points in a layer should belong to one soil unit, the likelihood function is given by

$$P(\xi_n|\sigma_n, N) = \sum_{j=1}^{9} \prod_{i=1}^{k_n} P_j(\xi^i_n|\sigma_n, N) \quad (4)$$

where k_n is the total number of data points in the n th layer and $P_j(\xi^i_n|\sigma_n, N)$ is the probability that a data point i belongs to a soil unit j in the Robertson chart. The latter can be computed using a two-dimensional integration of the Gaussian distribution spread over the chart, which in this paper has been performed numerically using a Monte Carlo simulation.

Assuming independence between the layers, the prior distribution of the model parameters and the likelihood function are

$$P(\Omega_N|N) = \prod_{n=1}^{N} P(\sigma_n|N) \quad (5)$$

$$P(\xi|\Omega_N, N) = \prod_{n=1}^{N} P(\xi_n|\sigma_n, N) \quad (6)$$

where $\Omega_N = \sigma_1, \sigma_2, \cdots, \sigma_N$ is the set of model parameters for the N layers. Within the Bayesian framework, the posterior distribution of the model parameters is given by

$$P(\Omega_N|\xi, N) = K \times P(\xi|\Omega_N, N) \times P(\Omega_N|N) \quad (7)$$

where K is a normalising constant. The most probable layer thicknesses (h^*_N) are identified by approximating this posterior distribution by a Gaussian distribution. Under this assumption, maximising $P(\Omega_N|\xi, N)$ leads to the most likely values of the model parameters and the layer thicknesses.

If M_N denotes a scenario with N layers, the most likely scenario (M^*_N) with the maximum occurrence probability amongst all possible scenarios is the one which gives the maximum value of the following:

$$P(M_N|\xi) = P(\xi|M_N) \times P(M_N)/P(\xi) \quad (8)$$

where $P(\xi)$ is the pdf of ξ, and $P(M_N)$ is the prior probability of M_N. Under the assumption of no prior information, M^*_N can be found by maximising the conditional probability $P(\xi|M_N)$, which is approximated by (see Wang et al. [20] for details)

$$P(\xi|M_N) \approx \frac{1}{H_t^{N-1}} \int P(\xi|\Omega_N, M_N) P(\Omega_N|M_N) d\Omega_N \quad (9)$$

where H_t is the total thickness of the soil strata. The term $1/H_t^{N-1}$ re-flects the occurrence probability of the most probable layer thicknesses and represents the penalty against choosing a higher value of N.

As the number of layers increases, the computational time required to identify the scenarios with the most probable thickness configuration increases significantly. Therefore, an optimisation technique (basin-hopping algorithm [19]) has been used in this paper. Furthermore, the above procedures (see Eq. (4)) do not consider the fact that the resulting configuration can have adjacent layers classified by the same soil unit, representing an unrealistic scenario. In this paper, all such configurations are therefore discarded before carrying out the optimisations.

Predicting classifications in 2D

The soil classifications at the CPTs together with the Kriged predictions of CPT parameters have been used in this paper to predict soil classifications in the 2D domain. In the absence of any additional information, the soil compositions at the CPT locations reflect a prior knowledge about the soil distributions in the 2D domain. Within the Bayesian framework, the updated knowledge of the soil distributions at some location in the 2D domain, with Kriged predictions $\widehat{F_r}$ and $\widehat{Q_t}$ of $\ln(F_r)$ and $\ln(Q_t)$, respectively, can be expressed as

$$P(j|\widehat{F_r}, \widehat{Q_t}) = K \times P(\widehat{F_r}, \widehat{Q_t}|j) \times P(j) \quad (10)$$

where K is a normalising constant, $P(j)$ is the prior distribution of soil unit j and $P(\widehat{F_r}, \widehat{Q_t}|j)$ is the likelihood of observing the CPT predictions given the soil unit.

The likelihood function in Eq. (10) can be computed using the results of the most likely 1D soil classifications as follows. The results of the 1D classification can be grouped together to generate distributions of the CPT parameters conditional to the soil units. Using the conditional distributions for a given soil unit, the likelihood of observing a certain CPT value can be obtained by computing the joint pdf using

$$P(\widehat{F_r}, \widehat{Q_t}|j) = P(\widehat{F_r}|\mu_{F_r,j}, \sigma_{F_r,j}) \times P(\widehat{Q_t}|\mu_{Q_t,j}, \sigma_{Q_t,j}) \quad (11)$$

where $P(\widehat{F_r}|\mu_{F_r,j}, \sigma_{F_r,j})$ and $P(\widehat{Q_t}|\mu_{Q_t,j}, \sigma_{Q_t,j})$ are the pdfs of the conditional distributions of $\ln(F_r)$ and $\ln(Q_t)$ with means $\mu_{F_r,j}$ and $\mu_{Q_t,j}$ and standard deviations $\sigma_{F_r,j}$ and $\sigma_{Q_t,j}$, respectively.

Subsequently, the realisation of soil units at a location in the domain can be generated using a Monte Carlo simulation by

$$\sum_{j=1}^{s-1} P(j|\widehat{F_r}, \widehat{Q_t}) < u \leq \sum_{j=1}^{s} P(j|\widehat{F_r}, \widehat{Q_t}) \quad (12)$$

where s is the soil unit realised at the location and u is a standard uniform random number sampled in that realisation over the entire domain.

Implementation procedure

A flowchart of the proposed approach is shown in Fig. 2 and a summary of the various steps followed is given below:
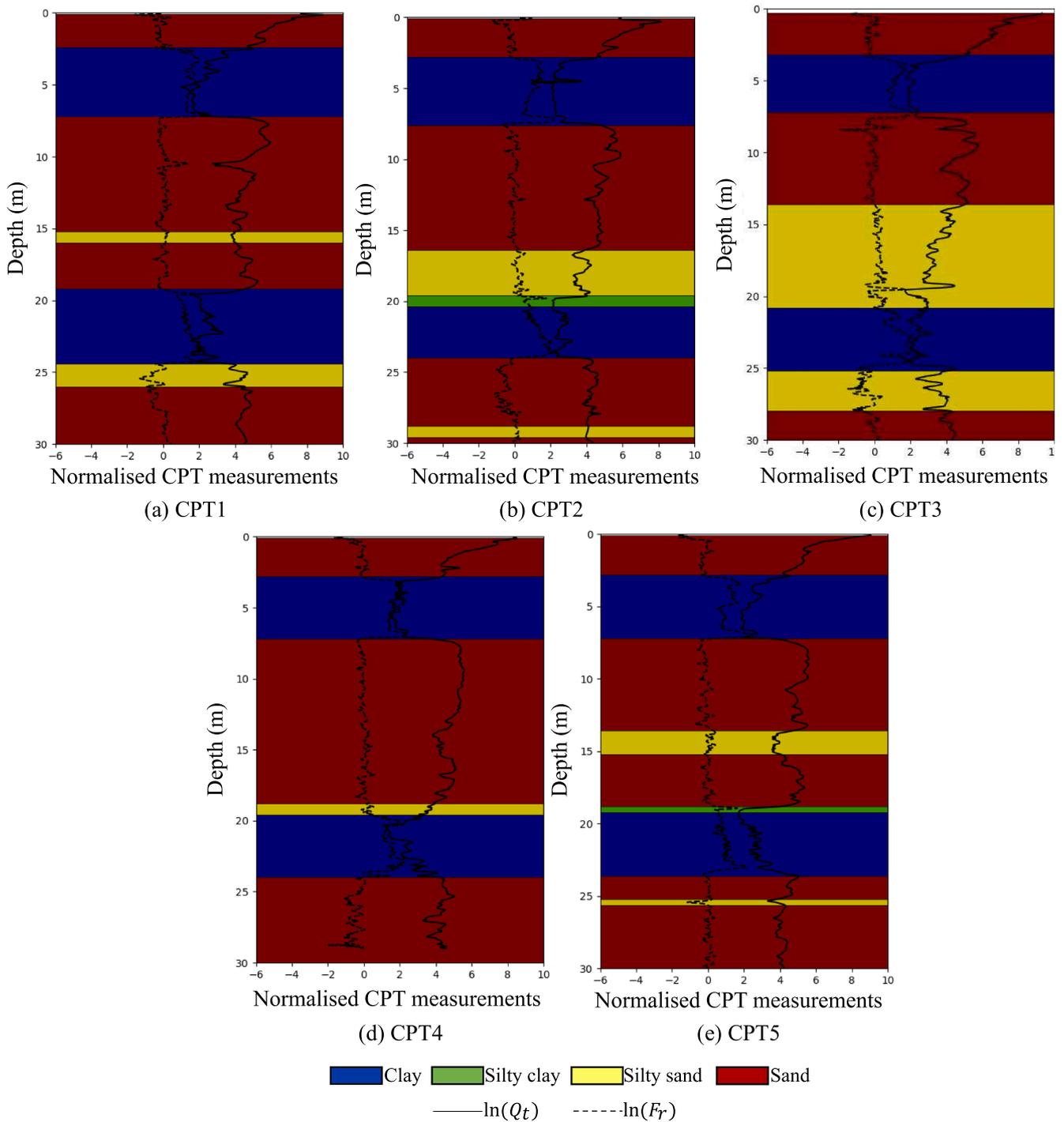
(a) CPT1   (b) CPT2   (c) CPT3

(d) CPT4   (e) CPT5

Fig. 8. Most likely classifications at the five CPT locations.

(a) Collect the available CPTs as well as any prior site knowledge.

(b) Generate Kriged predictions of the normalised CPT parameters using Eqs. (1)–(3).

(c) Perform probabilistic classification at the CPTs and identify the most likely classifications:

    i. For a given number of soil layers, identify the most likely layer thicknesses using Eqs. (4)–(7).

    ii. Identify the most likely number of layers with the most likely thicknesses using Eqs. (8) and (9)).

(d) Combine (a)–(c) to predict soil classifications in the 2D domain:

    i. Using the results of the most likely 1D classifications, generate distributions of CPT parameters conditional on the soil units.

    ii. Compute the likelihood of observing the Kriged CPT predictions for a given soil unit using Eq. (11).

    iii. Compute the posterior probability of each soil unit using Eq. (10) by using the likelihood function and prior knowledge of the soil distributions.

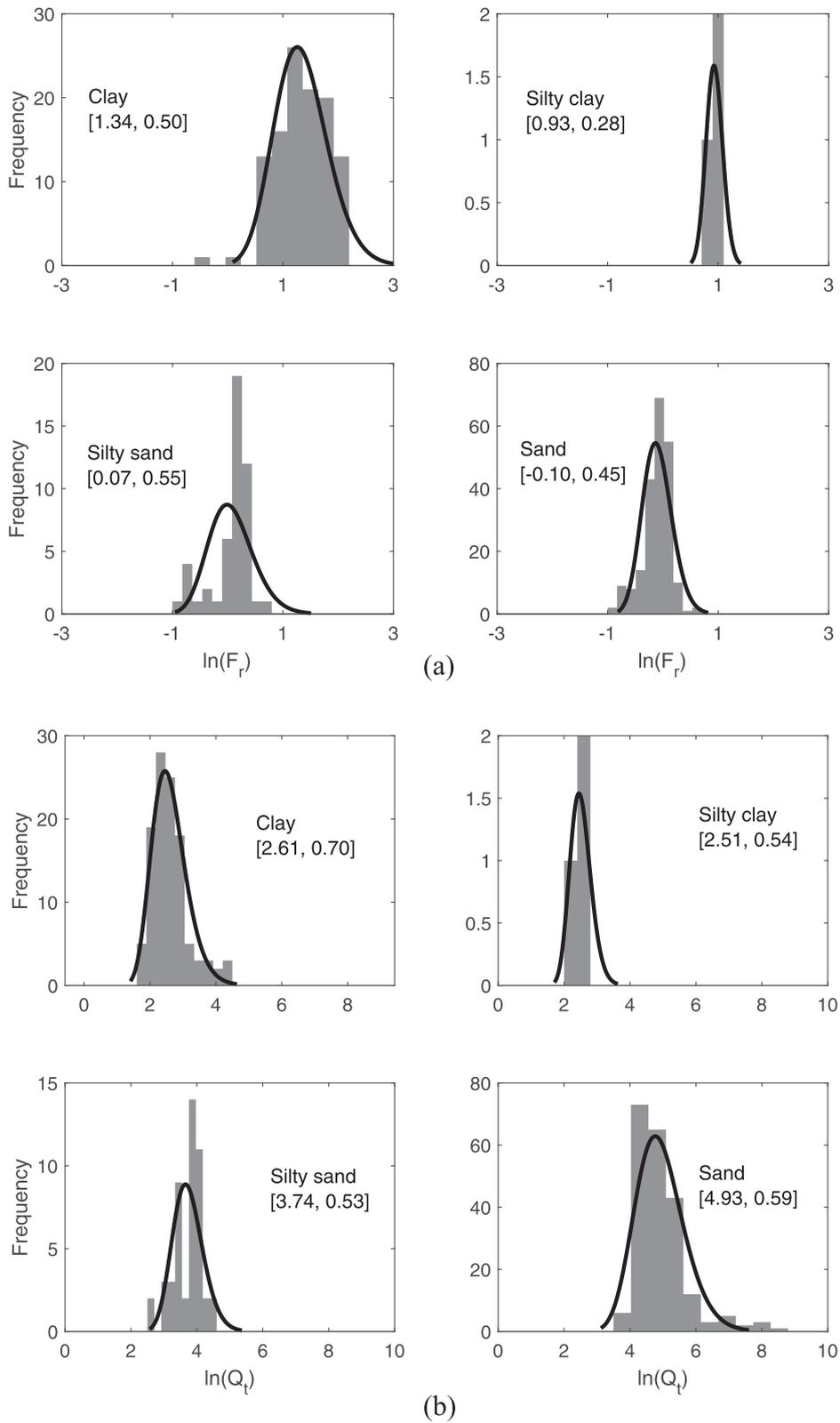   iv. Generate different possible subsoil configurations using Monte Carlo simulation (see Eq. (12)).

Fig. 9. Histograms of (a) $\ln(F_r)$ and (b) $\ln(Q_t)$ for four soil units obtained from the most likely soil classifications identified at the CPT locations. The values inside square brackets are the means and standard deviations of lognormal distribution fits to the histograms.
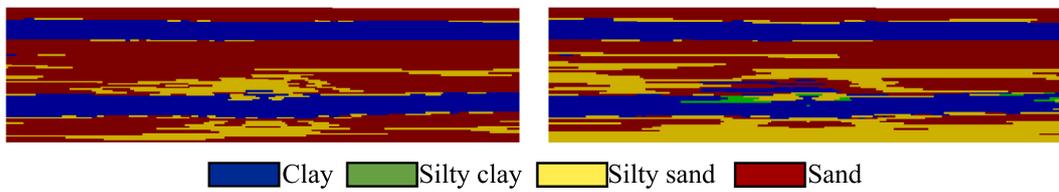
Fig. 10. Typical realisations of soil classification of the domain, obtained using the proposed approach with $\theta_h = 100$ m.
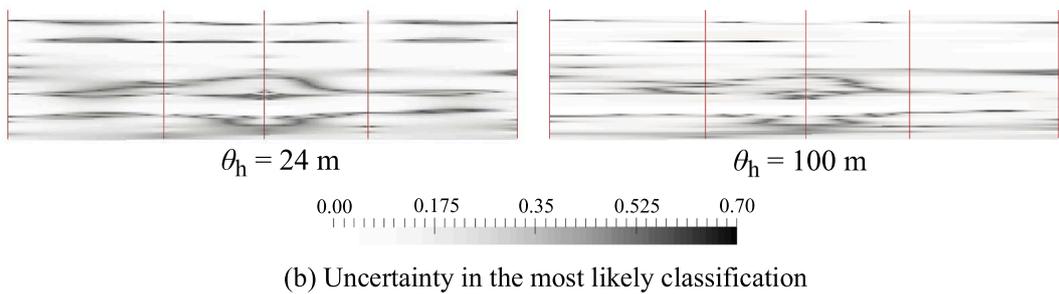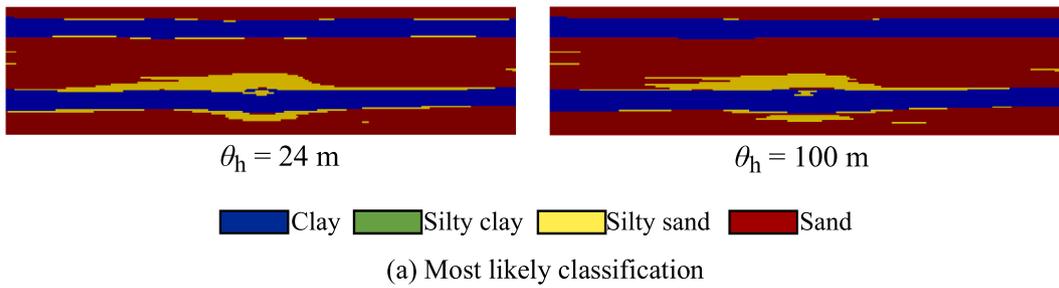

$\theta_h = 24$ m $\qquad\qquad$ $\theta_h = 100$ m

(a) Most likely classification

$\theta_h = 24$ m $\qquad\qquad$ $\theta_h = 100$ m

(b) Uncertainty in the most likely classification

Fig. 11. Most likely classification of the domain obtained using the proposed approach and its uncertainty.

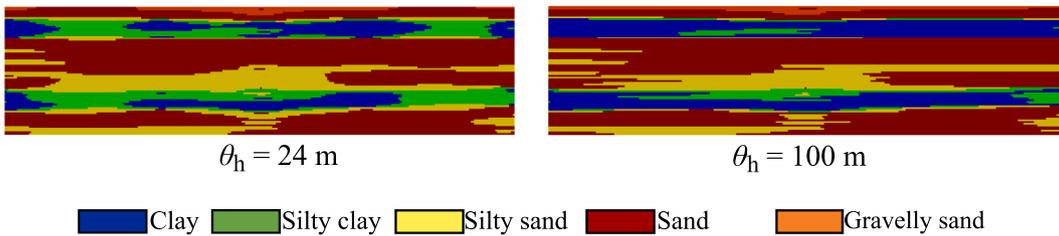$\theta_h = 24$ m $\qquad\qquad$ $\theta_h = 100$ m

Fig. 12. Deterministic classification of the domain obtained by ignoring the uncertainties.

Table 2
Percentages of most likely soil units predicted in the 2D domain using the proposed approach and those at the CPT locations (1D) based on Wang et al. [20].

Soil unit	2D		1D
	$\theta_h = 24$ m	$\theta_h = 100$ m	
Clay	27.9 %	29.7 %	29.6 %
Silty clay	0.0 %	0.0 %	0.8 %
Silty sand	9.1 %	6.8 %	12.8 %
Sand	63.0 %	63.5 %	56.8 %

Fig. 13. Confidence intervals of predictions at the CPTs obtained using the verification procedure.

Results and discussions

The applicability of the proposed approach has been demonstrated for a domain comprising five CPTs (CPT1–CPT5) from the Groningen region in the Netherlands. The relative locations of the CPTs are shown in Fig. 3. The CPTs are distributed over a length of 118 m and are
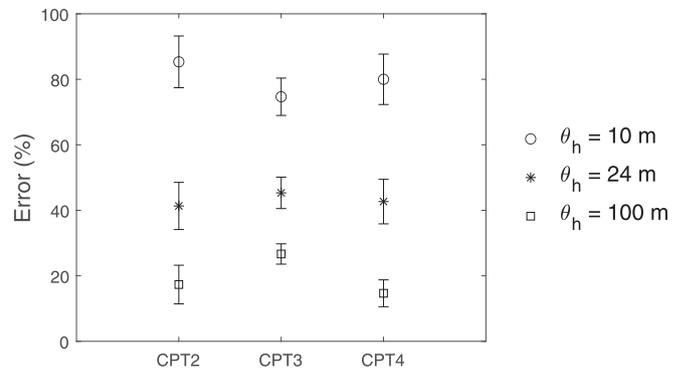
Table 3

Shear strength parameter values used in the slope stability analyses.

Soil unit	Cohesion (kPa)	Friction angle (°)
Clay	50	0
Silty clay	40	10
Silty sand	10	25
Sand	2	32
Gravelly sand	0	36

conducted to a depth of 30 m below the ground surface. The dashed line shown in Fig. 3 indicates the location of the representative cross-section, along which the classification has been carried out by discretising the domain into square cells of dimension 0.4 m each. At each CPT location, the measured values of sleeve friction (f_s) and tip resistance (q_c) have been normalised to obtain $\ln(F_r)$ and $\ln(Q_t)$, respectively, using the following equations:

$$F_r = \frac{100f_s}{q_t - \sigma_v}$$

$$Q_t = \frac{q_t - \sigma_v}{\sigma'_v} \tag{13}$$

where $q_t = q_c + u_2(1 - \alpha)$ is the corrected cone tip resistance, u_2 is the pore pressure measured behind the cone, α is the shape correction factor (which has here been assumed to equal 0.8), and σ_v and σ'_v are the total and effective vertical stresses, respectively.

The CPT parameters between the measurement locations have been predicted using the Kriging approach outlined earlier (see Eqs. (1)–(3)).

The standard deviation of each parameter has been taken to be the average of the standard deviations of all five normalised CPT profiles. A vertical scale of fluctuation of 2 m has been used for each parameter and is based on approximately fitting the auto-correlation function in Eq. (3) to the measurements over the entire depth of each CPT, see Fig. 4. (It may be noted that the negative values appearing in the figure are a result of correlating the measurements over their entire depth range, based on a limited number of CPTs, rather than within individual soil layers as is more usual.) On the other hand, estimating the horizontal scale of fluctuation is more difficult. This is because it requires several closely spaced CPTs to be placed strategically next to each other, for example, see de Gast et al. [2]. Because of the difficulties in accurately estimating θ_h using the very few CPTs available in this case, various values of θ_h have been considered and the Kriged fields of CPT parameters were generated for each combination of θ_v and θ_h. The Kriging variances of the predicted fields (in standard normal space) obtained using the various values are plotted in Fig. 5. This figure shows that the variance decreases significantly as the value of θ_h increases. The Kriged predictions of $\ln(F_r)$ and $\ln(Q_t)$ generated using $\theta_h = 24$ m (the minimum spacing between the CPTs) and $\theta_h = 100$ m are shown in Fig. 6.

The fields of $\ln(F_r)$ and $\ln(Q_t)$ are transformed into soil units using the two-step approach explained earlier, starting with the 1D classification at each CPT. For example, Fig. 7 shows the normalised measurements at CPT3 and the deterministic classification at the CPT based on mapping the measurements onto the Robertson chart. On considering uncertainties, i.e. by following the probabilistic approach using Eqs. (4)–(7), the most probable thickness configurations (h^*_N) obtained for the various possible cases with N layers for CPT3 are listed in Table 1. The conditional probabilities ($P(\xi|M_N)$) calculated using Eq. (8) for each
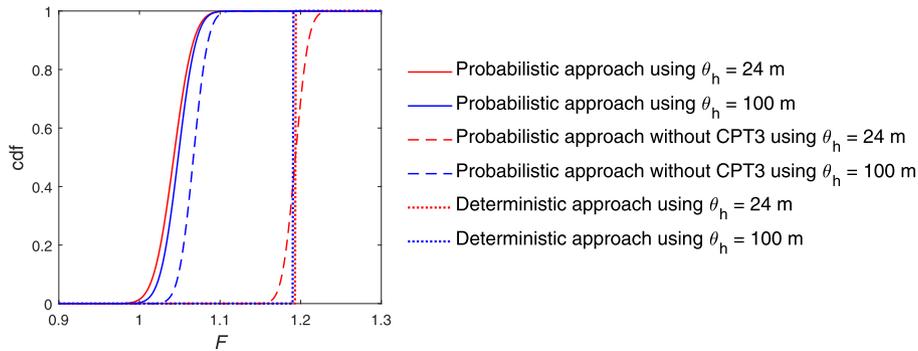

Fig. 14. Cumulative distribution functions of factor of safety obtained for slopes cut-out in the domain stratified either using the proposed probabilistic approach or using the deterministic approach.
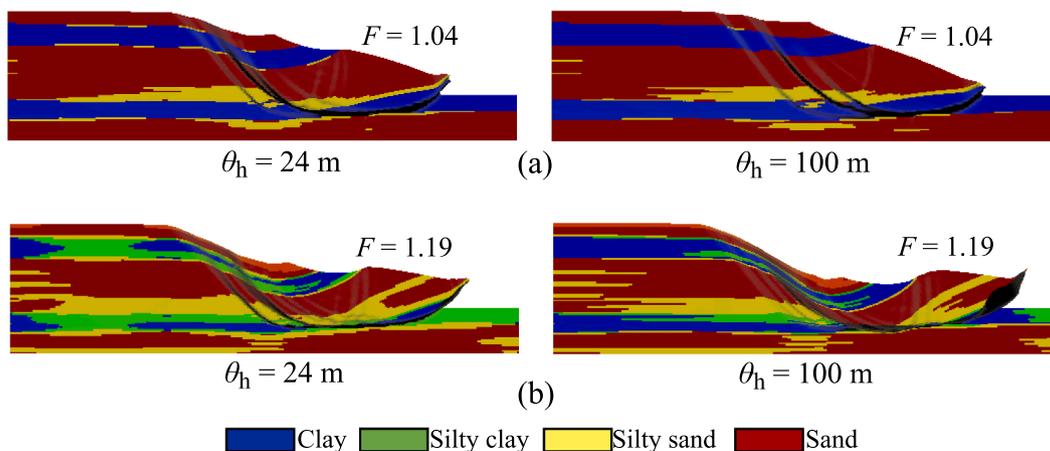

Fig. 15. Factors of safety and failure mechanisms obtained for slopes cut-out in the domain classified as in (a) Fig. 11 and (b) Fig. 12 (deformations have been exaggerated for better visualisation).

case are also listed in the table. As can be seen from these values, the most likely scenario (i.e. with the highest value of $P(\xi|M_N)$) for CPT3 is the case with seven layers. The subsoil classifications corresponding to this most likely scenario are illustrated in Fig. 8c). The same procedure was applied to the other four CPTs and the resulting most probable classifications at these CPT locations are also shown in Fig. 8. Due to the lack of high quality prior information about the measurements and the site, values of 1.0 and 1.5 were assumed for the standard deviations of the Gaussian distributions of $\ln(F_r)$ and $\ln(Q_t)$, respectively, and are approximately 75 % of their maximum possible values [20]. As can be seen from the classification at the CPTs in Fig. 8, the domain consists of clay (29.6 %), silty clay (0.8 %), silty sand (12.8 %) and sand (56.8 %).

The above results are used to predict soil distributions in the domain by using Eqs. (10)–(12). As shown in Fig. 9, the results of the most likely 1D soil classifications are used to generate conditional distributions by generating histograms of the CPT parameters that resulted in a certain soil unit. Also shown in the figure are the parameters of the lognormal distribution fit to each histogram. These distributions, together with the Kriged CPT predictions at a location, are used to predict the likelihood function and posterior probabilities of the different soil units at the location using Eqs. (10) and (11). Although, in this example, only the most likely 1D classification at the CPTs have been used as prior knowledge, any additional information on the domain from other sources, for example borehole data, could also be integrated into the framework.

Following the calculation of the probabilities, a Monte Carlo simulation has been carried out to predict the different possible scenarios. Typical realisations of subsoil classifications generated for the domain are shown in Fig. 10. Fig. 11 shows the most likely classification of the 2D domain and its uncertainty, which was obtained by comparing the most likely classification with those obtained in 500 realisations of the Monte Carlo simulation. In this figure, the most likely soil unit at a location is chosen as the one which gives the maximum value of the following:

$$O(s) = \sum_{R=1}^{500} \mathrm{I}\left(P\left(j^R\middle|\widehat{F_r}, \widehat{Q_t}\right) = s\right)\middle/500 \qquad (14)$$

where $O(s)$ is the occurrence probability of a soil unit s at the location, j^R is the soil unit simulated at the location in a realisation R using Eq. (12), and I() is an indicator function which is equal to 1 if the condition within the brackets is true and equal to 0 otherwise.

For comparison, Fig. 12 shows the deterministic classification obtained by simply mapping the predicted values of the CPT parameters, obtained using the two values of θ_h, onto the Robertson classification chart. The classifications in Figs. 11 and 12 demonstrate that ignoring the uncertainties in the measurements and transformation results in a clear difference in the composition of the domain and in the locations of the soil units.

The percentages of various soil units in the most likely classification predicted using the proposed approach, for the two values of θ_h, are listed in Table 2. As a result of the very small fraction of silty clay (=0.8 %) obtained in the most likely 1D classifications, the classification of the 2D domain is most likely composed of only clay, silty sand and sand, whose percentages agree well with those obtained in the most likely 1D classifications. Note that, as a result of using the proposed approach with the limited number of CPTs, the soil units predicted in the domain at the CPT locations are not exactly the same as their most likely 1D classifications (Fig. 8). Nevertheless, it has been observed that the soil units at these locations are predicted with an accuracy of 80–97 % compared to those in their most likely 1D classifications.

The accuracy of the proposed approach can be verified by investigating its predictive capabilities at known locations. This has here been carried out by: (a) removing a CPT from the input and generating the Kriged predictions of $\ln(F_r)$ and $\ln(Q_t)$ conditioned on the remaining four CPTs, (b) using the predicted fields together with the distributions in

Fig. 9 to predict the classification in the 2D domain, and (c) comparing the soil classification at the location of the missing CPT with its most likely 1D classification (shown in Fig. 8). Fig. 13 shows the results of this verification procedure, in terms of the means and standard deviations of the errors, that have been obtained in the Monte Carlo simulation using various values of θ_h. As can be seen from the figure, on using $\theta_h = 100$ m the proposed approach can predict the soil classifications at the CPTs with good accuracy (>72 %). However, the predictions are very poor on using a significantly smaller value of θ_h. This is mainly due to a larger Kriged variance of the predicted CPT parameters as a result of the larger distance between the adjacent conditioning CPTs relative to the value of θ_h.

Illustrative example

The influence of accounting for the uncertainties considered in this paper has been illustrated via the stability analysis of a slope cut-out in the domain, with a slope angle of 26.6° and a height of 20 m. The shear strength parameters assumed for the various soil units are listed in Table 3. The finite element model for the slope is meshed by 8-noded quadrilateral (mainly square) elements of size 0.8 m each and uses 2 × 2 Gaussian integration. The mesh is fixed at the base, whereas rollers are applied on the two ends to prevent movement horizontally. The soil units identified at each interval of 0.4 m in the 2D domain are mapped to the integration points of the elements and the slope is analysed using the strength reduction method. In this method, gravity loading is applied to generate the initial in-situ stresses. The resulting shear stresses are checked against the Mohr–Coulomb failure criterion and the excess stresses are iteratively redistributed throughout the model following a viscoplastic stress integration scheme [25]. The lowest strength reduction factor at which equilibrium cannot be achieved within 500 iterations is the factor of safety (F) of the slope.

Fig. 14 shows the cumulative distributions of F for the slope cut-out in: (a) generated realisations of the domain that have been classified using the proposed approach, and (b) the domain classified using the deterministic approach of ignoring uncertainties in the measurements and transformation. To investigate the sensitivity of the results to θ_h, Fig. 14 also shows the distributions of F obtained by removing CPT3 from the input and following the proposed approach to generate realisations of subsoil classification. The results show that F is less sensitive to the chosen value of θ_h if the distance between the CPTs is smaller than that value of θ_h. Fig. 15 shows a comparison of the results obtained using the deterministic approach with those obtained in the most likely classification using the proposed approach. Both figures highlight the significant difference in computed responses due to differences in the composition of the domain and the location of the soil units obtained as a result of considering or ignoring the uncertainties. Note that these results are based on considering deterministic strength properties for the soil units and that it may be worth investigating the differences when the spatial variability of properties within each soil unit is also taken into consideration.

Conclusions

A methodology for predicting subsurface classification using limited CPT data has been proposed. The predicted classification is able to capture the inherent spatial variability of soil measured via CPTs and takes account of the uncertainties that arise from transforming CPT measurements into soil units as well as measurement errors. In order to do this, a two-step framework has been proposed. In the first step, the CPT parameters between the measurement locations are predicted using Kriging. In the second step, these parameters are transformed into soil units while accounting for uncertainties in the measurements and transformation. The latter has been carried out by first implementing and updating an existing method of integrating the Robertson chart with the Bayesian approach for predicting the most likely 1D classification at

the CPTs. These results are integrated with the CPT predictions within the Bayesian framework to predict posterior distributions of soil units in the domain.

The applicability of the proposed approach has been illustrated for a 2D domain consisting of five CPTs in the Netherlands. The results show that the proposed approach, when used with a larger horizontal correlation length of the CPT parameters than the distance between the CPTs, can predict the stratigraphy in the domain with good accuracy. Furthermore, the significance of considering the above mentioned factors in carrying out classification has been demonstrated via finite element stability analyses of slopes cut-out in the domain. It was observed that considering the uncertainties resulted in lower safety factors and different failure geometries than the unconservative solution obtained by ignoring the uncertainties.

CRediT authorship contribution statement

Divya Varkey: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **Michael A. Hicks:** Funding acquisition, Conceptualization, Supervision, Writing – review & editing. **Abraham P. van den Eijnden:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The CPT data source will be made available on request.

Acknowledgement

References

[1] Cao Z, Wang Y. Bayesian approach for probabilistic site characterization using cone penetration tests. J Geotech Geoenviron Eng 2013;139(2):267–76.

[2] de Gast T, Vardon PJ, Hicks MA. Assessment of soil spatial variability for linear infrastructure using cone penetration tests. Géotechnique 2021;71(11):999–1013.

[3] Griffiths DV, Huang J, Fenton GA. Influence of spatial variability on slope reliability using 2-D random fields. J Geotech Geoenviron Eng 2009;135(10):1367–78.

[4] Hicks MA, Samy K. Influence of heterogeneity on undrained clay slope stability. Q J Eng Geol Hydrogeol 2002;35(1):41–9.

[5] Hicks MA, Spencer WA. Influence of heterogeneity on the reliability and failure of a long 3D slope. Comput Geotech 2010;37:948–55.

[6] Hicks MA, Nuttall JD, Chen J. Influence of heterogeneity on 3D slope reliability and failure consequence. Comput Geotech 2014;61:198–208.

[7] Hicks MA, Li Y. Influence of length effect on embankment slope reliability in 3D. Int J Numer Anal Meth Geomech 2018;42:891–915.

[8] Hicks MA, Varkey D, van den Eijnden AP, de Gast T, Vardon PJ. On characteristic values and the reliability-based assessment of dykes. Georisk: Assess Manag Risk Eng Syst Geohazards 2019;13:313–9.

[9] Hu Y, Wang Y. Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation. Comput Geotech 2020;124:103634.

[10] Huang J, Griffiths DV, Fenton GA. System reliability of slopes by RFEM. Soils Found 2010;50(3):343–53.

[11] Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. J Chem Metall Min Soc S Afr 1951;52(6):119–39.

[12] Mayne PW, Christopher BR, DeJong J. Subsurface investigations – geotechnical site characterisation. Technical report, United States Federal Highway Administration, FHWA-NHI-01-031, Washington DC; 2002.

[13] Phoon KK, Kulhawy FH. Characterization of geotechnical variability. Can Geotech J 1999;36(4):612–24.

[14] Qi XH, Li DQ, Phoon KK, Ciao ZJ, Tang XS. Simulation of geologic uncertainty using coupled Markov chain. Eng Geol 2016;207:129–40.

[15] Robertson PK. Soil classification using the cone penetration test. Can Geotech J 1990;27(1):151–8.

[16] Varkey D, Hicks MA, Vardon PJ. Effect of uncertainties in geometry, inter-layer boundary and shear strength properties on the probabilistic stability of a 3D embankment slope. Georisk: Assess Manag Risk Eng Syst Geohazards 2023;17(2):262–76.

[17] Varkey D, van den Eijnden AP, Hicks MA. Predicting subsurface stratigraphy using an improved coupled Markov chain method. In: Proceedings of the 14th International Conference on Applications of Statistics and Probability in Civil Engineering, Dublin, Ireland, 2023.

[18] Wackernagel H. Multivariate geostatistics: An introduction with applications. New York: Springer; 2003.

[19] Wales DJ, Doye JPK. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A 1997;101(28):5111–6.

[20] Wang Y, Huang K, Cao Z. Probabilistic identification of underground soil stratification using cone penetration tests. Can Geotech J 2013;50:766–76.

[21] Wang Y, Huang K, Cao Z. Bayesian identification of soil strata in London clay. Géotechnique 2014;64(3):239–46.

[22] Wang Y, Hu Y, Zhao T. Cone penetration test (CPT)-based subsurface soil classification and zonation in two-dimensional vertical cross section using Bayesian compressive sampling. Can Geotech J 2020;57(7):947–58.

[23] Xiao T, Zhang LM, Li XY, Li DQ. Probabilistic stratification modelling in geotechnical site characterization. ASCE-ASME J Risk Uncert Eng Syst Part A: Civil Eng 2017;3(4):04017019.

[24] Zhang JZ, Liu ZQ, Zhang DM, Huang HW, Phoon KK, Xue YD. Improved coupled Markov chain method for simulating geological uncertainty. Eng Geol 2022;298:106539.

[25] Zienkiewicz OC, Cormeau IC. Visco-plasticity–plasticity and creep in elastic solids – A unified numerical solution approach. Int J Numer Methods Eng 1974;8(4):821–45.