Bayesian supervised learning of 2D subsurface soil stratigraphy using limited cone penetration tests with consideration of uncertainty

Hu, Yue; Wang, Yu

Published in:
Cone Penetration Testing 2022

Published: 01/01/2022

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY-NC

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Bayesian supervised learning of 2D subsurface soil stratigraphy using limited cone penetration tests with consideration of uncertainty

Yue Hu
Department of Civil and Environmental Engineering, National University of Singapore, Singapore

Yu Wang
Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China

ABSTRACT: Cone penetration tests (CPT) have been widely used for soil stratification in geotechnical site investigation for decades. However, due to time and budget limits, the layout of CPT sounding at a specific project site is often sparse, leading to significant interpolation uncertainty in the development of subsurface soil 2D cross-section, particularly at locations without CPT measurements. Such development is often combined with empirical classification criteria, which further introduce model uncertainty to soil stratification. These uncertainties may pose great risks to the geotechnical engineering practice. A Bayesian supervised learning method is presented in this paper for probabilistic soil stratification in a 2D cross-section using limited CPT. The proposed method can not only automatically stratify soils in a 2D cross-section from limited CPT soundings, but also can properly quantify the associated uncertainties. Complete 2D CPT data cross-section is firstly learned from limited number of 1D CPT profiles using Bayesian supervised learning. The associated interpolation uncertainty is modelled numerically using non-parametric random field simulation based on the results of Bayesian supervised learning. Parametric autocorrelation function of CPT data along either vertical or horizontal direction is not needed. A probabilistic model is also developed to account for the model uncertainty of an empirical soil behavior type classification chart. The interpolation uncertainty and soil classification model uncertainty are then evaluated simultaneously in a Monte Carlo simulation framework. A simulated data example is used for illustration. The results suggest that the proposed method performs well.

1 INTRODUCTION

Subsurface soil stratification is an indispensable element in geotechnical engineering as required for geotechnical designs and analyses. However, due to time and budget limits, the layout of site investigation points is often sparse, leading to significant challenge and uncertainty in the development of 2D subsurface soil stratigraphy which is frequently adopted in geotechnical analysis. In engineering practice, such development is often combined with an empirical soil classification system, which inevitably introduce additional model uncertainty to soil stratification. These uncertainties may pose great risks to the geotechnical designs, analyses, and construction process (e.g., Clayton 2001; Mayne 2007). It is therefore necessary to properly evaluate the uncertainties associated with 2D subsurface soil stratification.

Cone penetration tests (CPT) have been widely used for soil stratification in geotechnical site investigation. It can be used to classify subsurface soils, identify stratification, and quantify associated uncertainty in a single sounding (e.g., Robertson 1990; Wang et al. 2013). When interpreting 2D vertical soil cross-section using CPT, interpolation between adjacent CPT profiles or stratigraphy should be performed. However, direct 2D interpolation using conventional geostatistical methods is difficult because of the limitation of CPT soundings number and the spatially varying soil layer boundaries which render the soil properties in the 2D cross-section highly non-stationary (e.g., Wang et al. 2019). In this case, it is even more challenging to reasonably tackle the interpolation uncertainty quantification associated with the 2D subsurface stratigraphy. On the other hand, CPT-based soil classification relies on empirical soil classification charts to transform the continuous CPT measurements to discrete soil behavior type (SBT). Note that these charts rely on a series of deterministic SBT classification boundaries which are developed from site investigation data globally. These classification charts might not provide consistent results at a specific site and introduce model uncertainty when used locally (e.g., Boulanger & Idriss 2014; Maurer et al. 2019). How to incorporate the model uncertainty of SBT classification
boundaries in the CPT-based 2D soil stratification remains unsolved.

To address the abovementioned challenges, a Bayesian supervised learning method is proposed to interpret 2D subsurface soil stratigraphy from limited CPT with explicit evaluation of both interpolation uncertainty and SBT chart model uncertainty. The proposed framework is introduced in the following section and then illustrated using a simulated example.

2 THE PROPOSED METHOD

The proposed framework is based on a Monte Carlo simulation (MCS) and comprises of three components. First, SBT index $I_c$ of CPT data is firstly interpolated in the concerned 2D cross-section from limited CPT using a Bayesian supervised learning algorithm in a non-parametric manner. The interpolation uncertainty is quantified automatically during learning process and then modelled by $I_c$ random field simulation based on the learning outcome. Second, a probabilistic SBT chart is developed by adapting an empirical chart to consider the model uncertainty using which random samples of SBT chart with random classification boundaries are drawn. Third, each random field sample of $I_c$ is matched with a random sample of SBT chart to produce a Monte Carlo sample of SBT cross-section with soil stratigraphy. Statistical analysis is then performed on the generated Monte Carlo samples. Both uncertainties are evaluated simultaneously under MCS. The three components are introduced in the following three subsections, respectively.

2.1 Bayesian supervised learning with random field simulation

The full $I_c$ data variability in the concerned 2D cross-section are learned from the profiles of limited CPT. In the context of the Bayesian supervised learning, 2D data matrix $F$ (e.g., $I_c$ data matrix of a 2D cross-section), which is spatially varying along coordinates $x_1$ and $x_2$ (e.g., depth direction and horizontal direction), has a dimension of $N_{x_1} \times N_{x_2}$. Mathematically, $F$ is expressed as a weighted summation of a series of orthonormal 2D basis functions (e.g., Zhao et al. 2018; Hu et al. 2020; Wang et al. 2020&2021):

$$ F = \sum_{t=1}^{N_{x_1} \times N_{x_2}} B_{t}^{2D} \omega_{t}^{2D} \quad (1) $$

in which $B_{t}^{2D}$ is the $t$-th 2D basis function, while $\omega_{t}^{2D}$ is the weight coefficients of $B_{t}^{2D}$. Discrete wavelet transform may be selected to construct $B_{t}^{2D}$ (e.g., Donoho et al. 2006). Note that for compressible images (e.g., spatially correlated CPT data cross-section), most $\omega_{t}^{2D}$ have negligibly small values except for a limited number of non-trivial ones. Therefore, the $F$ may be reconstructed approximately if those non-trivial weight coefficients can be identified and estimated using sparse measurements $Y$ (e.g., CPT profiles data at limited locations), which is a sub-matrix of $F$ with a dimension of $M_{x_1} \times M_{x_2}$ ($M_{x_1} << N_{x_1}$, $M_{x_2} << N_{x_2}$). The relation between $Y$ and $\omega_{t}^{2D}$ is expressed as:

$$ Y = \sum_{t=1}^{N_{x_1} \times N_{x_2}} A_{t}^{2D} \omega_{t}^{2D} \quad (2) $$

in which $A_{t}^{2D}$ is the sub-matrix of $B_{t}^{2D}$ with a dimension of $M_{x_1} \times M_{x_2}$. $A_{t}^{2D}$ just reflects the measured elements in $B_{t}^{2D}$. Equation 2 enables those non-trivial coefficients to be learned through maximum likelihood estimation under a Bayesian framework (e.g., Tipping 2001). The learned weight coefficient vector is denoted as $\omega^{2D}$. After derivation, it is found that the posterior distribution of $\omega^{2D}$ given measurement data follows a multivariate Student’s t distribution, with the mean vector and covariance matrix expressed as (e.g., Zhao et al. 2018; Hu et al. 2020):

$$ \mu_{\omega^{2D}} = HV_{\omega} = (J + D)^{-1} V_{\omega} $$

$$ \text{COV}_{\omega^{2D}} = \frac{d_{n}H}{c_{n} - 1} = \frac{d_{n}(J + D)^{-1}}{c_{n} - 1} $$

in which $J$ is a matrix with element $J_{s,t} = tr[A_{t}^{2D}(A_{s}^{2D})^T]$, $(s,t = 1, 2, ..., N_{x_1} \times N_{x_2})$. “tr” represents trace operation in linear algebra. $D$ is a diagonal matrix with diagonal elements $D_{t,t} = a_t$ ($t = 1, 2, ..., N_{x_1} \times N_{x_2}$) in which $a_t$ are non-negative parameters to be determined by a maximum likelihood algorithm (e.g., Tipping, 2001). $V_{\omega}$ is a $N_{x_1} \times N_{x_2}$ vector with element $V_{\omega,t} = tr[Y(A_{t}^{2D})^T]$. $c_n = M_{x_1} \times M_{x_2}/2 + c$; $d_n = d + ([\|Y\|^2 - \mu_{\omega^{2D}}^T \text{H}^{-1} \mu_{\omega^{2D}}]) / 2$. $c$ and $d$ are non-negative small constants to achieve an uninformative prior in the Bayesian formulation. Due to the compressibility of $F$, only those $N_{a}$ ($N_{a} << N_{x_1} \times N_{x_2}$) non-trivial coefficients need to be estimated, and the rest are zeroed out. Therefore, in Equation 3, the $\mu_{\omega^{2D}}$ is reduced to an $N_{a} \times 1$ vector and $\text{COV}_{\omega^{2D}}$ is reduced to an $N_{a} \times N_{a}$ matrix for simplicity.

Given the learning results in Equation 3, random samples of $\omega^{2D}$ can be generated through eigen decomposition (e.g., Hu et al. 2019; Hu and Wang 2020):

$$ \omega^{2D} = \mu_{\omega^{2D}} + \sum_{i=1}^{N_{a}} U_{i} \sqrt{\lambda_{i}^{2D}} Z_{i} \quad (4) $$
in which $U_i$ is the $i$-th eigen-vector of the covariance matrix $\text{COV}_{\varphi_{2D}}$; $\varphi_{2D}$ is the $i$-th eigenvalue of $\text{COV}_{\varphi_{2D}}$. $Z_i$ is a set of independently and identically distributed standard Gaussian random variables. Using Equation 4, random vectors $\varphi_{2D}$ can be generated readily through realizations of $Z_i$. After that, extensive RFS of the approximated 2D data $\hat{F}$ (e.g., $I_c$ data 2D cross-section) are subsequently reconstructed by substituting the random vectors $\varphi_{2D}$ into Equation 1 as below:

$$\hat{F} = \sum_{i=1}^{N_x} B_i^{2D} \varphi_i^{2D}$$  \hspace{1cm} (5)

Each generated 2D RFS indicates a possible outcome of Bayesian supervised learning. The ensemble directly reflects the interpolation uncertainty.

### 2.2 Probabilistic SBT chart

The empirical $I_c$ – based SBT chart summarized in Table 1 was developed from a global soil database compiling CPT data obtained predominately within limited depths (e.g., Robertson 1990; Robertson & Wride 1998). It is expected that the SBT chart might not provide accurate classification and introduce model uncertainty when used locally at a specific project site. In other words, those empirical $I_c$ classification boundaries (e.g., see Table 1) can vary from site to site (e.g., Boulanger & Idriss 2014; Maurer et al. 2019). To consider the model uncertainty, a probabilistic $I_c$ – based SBT classification chart is developed. The five SBT classification boundaries (i.e., 1.31, 2.05, 2.6, 2.95, 3.6) listed in Table 1 are denoted as B1 to B5 and modelled as five Gaussian random variables, respectively. The mean values of B1 to B5 are taken as their original values, as summarized in the second column of Table 2. According to literature (e.g., Boulanger & Idriss 2014; Maurer et al. 2019) on $I_c$ data variability, a set of standard deviation (SD) values for B1 to B5 is suggested in this study, as summarized in the third column of Table 2. SD of B1, B2 and B5 are taken as 0.1, while the SD of B3 and B4 are taken as 0.05. The probabilistic SBT chart allows for the varying nature of $I_c$ – based classification criteria at different local sites. Extensive random samples of classification boundaries can be generated by repetitively sampling B1 to B5, leading to random samples of SBT chart. Each random sample is a possible state of the uncertain SBT classification chart and will randomly match with one RFS of $I_c$ data cross-section for classification and stratification purposes.

### 2.3 Statistical analysis of Monte Carlo simulation

By pairing 2D $I_c$ RFS with a random sample of SBT chart, each point in that 2D cross-section can be mapped, leading to a 2D SBT cross-section. After repeating the process $N_B$ times with different combination of 2D $I_c$ data RFS and SBT chart random samples, $N_B$ 2D SBT cross-sections are generated. Each SBT cross-sections serves as a possible state of subsurface soil stratigraphy given limited CPT. Statistical analysis is then performed for those $N_B$ SBT cross-sections. The probability of soil at a given point $(x_1, x_2)$ being mapped to a specific SBT, e.g., SBT=1 (i.e., $t=2, 3, \ldots, 7$), can be calculated as (e.g., Hu and Wang 2020):

$$p(SBT_{x_1, x_2} = i) = \frac{N_{x_1, x_2}^i}{N_B} \times 100\%$$ \hspace{1cm} (6)

in which $N_{x_1, x_2}^i$ is the number of SBT values at point $(x_1, x_2)$ that equal to $i$. Equation 6 quantitatively measures how likely the soil at a point $(x_1, x_2)$ is classified as any one of the six SBT. The SBT with the highest probability is taken as the most likely SBT at the point $(x_1, x_2)$. Similarly, the most likely SBT cross-section is obtained. In addition, the classification uncertainty is quantified by the SD of SBT samples at a given point. The higher the SD, the higher the uncertainty. Similarly, an SD cross-section can be calculated which reflects the pattern of uncertain region in the 2D cross-section. The SD cross-section directly assesses the reliability of soil stratigraphy.

### 3 ILLUSTRATIVE EXAMPLE

To illustrate the proposed framework, a simulated geological cross-section example is provided in this section. As shown in the Figure 1a, a 2D vertical cross-section is simulated. Four soil types exist in this cross-section, i.e., clay, silt mixtures, sand mixtures and sand, which correspond to SBT values of 3 to 6, respectively. Spatial variability of $I_c$ data in each soil layer is generated using assumed random field model. The random field parameters used (e.g., mean $\mu_{Ic}$, standard deviation $\sigma_{Ic}$, correlation lengths along horizontal direction $\lambda_h$ and vertical direction $\lambda_v$) are summarized in Table 3. An exponential correlation structure is adopted in this simulation. The simulated 2D $I_c$ data cross-section is shown by colormap in the Figure 1b. The $I_c$ data cross-section is a 128×256 matrix with resolution of 0.1m for both directions. It is regarded as a geological setting at a specific site. Note that the complete $I_c$ data cross-section is usually not available in engineering practice. This example is just for illustration and validation purposes. Suppose that six (i.e., $M=6$) CPT soundings are conducted within the cross-section, as denoted by black dash lines (e.g., M1-M6) in Figure 1. $I_c$ data profiles of M1-M6 are shown in Figure 2. These six 1D $I_c$ data
profiles are used as input to interpret the soil stratigraphy under the proposed framework. After constructing the Y matrix in Equation 2 from data profiles of M1-M6, the Bayesian supervised learning is implemented. Then \( N_B = 500 \) 2D RFs of \( I_c \) data are generated from the learning results. Four examples of \( I_c \) cross-section are shown by colormap in Figure 3. Each plot in Figure 3 is a possible interpolation of \( I_c \) data in this 2D cross-section from six CPT soundings. No parametric correlation structure is needed for the learning process. Next, \( N_B = 500 \) random SBT classification charts are generated from the probabilistic SBT classification boundary model. Note that a Gaussian random variable ranges from negative infinity to positive infinity. To mitigate the overlapping problem, the Gaussian probability density functions (PDF) for B1-B5 are truncated respectively to a range of mean ± three standard deviation, as summarized in the fourth column of Table 2.

By randomly pairing a RFS of \( I_c \) data cross-section with one random sample of SBT chart, examples of SBT cross-sections are shown in Figures 4a-4d. These four 2D SBT cross-sections correspond respectively to the four RFS of 2D \( I_c \) data shown in Figure 3 mapped with random sample of SBT charts shown above each subplot of Figure 4. Note that each of these SBT cross-section indicates a possible soil stratigraphy.

Using \( N_B = 500 \) Monte Carlo samples of 2D SBT cross-section, statistical analysis is performed using Equation 6. The most likely SBT cross-section is shown in Figure 5a. In the most likely SBT cross-section, four SBT (i.e., SBT3-6) are presented, which is consistent with the underlying true stratigraphy. The original soil zone boundaries are shown by black solid lines for comparison. Note that the most likely SBT cross-section is generally comparable to the underlying true one, although the soil zone boundaries are not perfectly learned due to interpolation uncertainty and model uncertainty in SBT chart. The uncertainties can be evaluated simultaneously through SD of \( N_B = 500 \) SBT cross-sections, as shown in Figure 5b. It is found that the bright areas with high uncertainty are generally consistent with the underlying true boundaries (i.e., the black solid lines). The results in Figure 5b suggest that the underlying true soil zone boundaries can be approximated based on the SD of \( N_B \) SBT cross-sections. The proposed framework performs reasonably well in the interpretation of 2D soil stratigraphy and uncertainty quantification, given only six CPT soundings.

Table 1. \( I_c \) – based SBT classification chart (after Robertson 1998).

<table>
<thead>
<tr>
<th>Range of SBT index ( I_c )</th>
<th>SBT ID</th>
<th>SBT description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I_c &lt; 1.31 )</td>
<td>7</td>
<td>Gravelly sand to dense sand</td>
</tr>
<tr>
<td>( 1.31 &lt; I_c &lt; 2.05 )</td>
<td>6</td>
<td>Sands: clean sand to silty sand</td>
</tr>
<tr>
<td>( 2.05 &lt; I_c &lt; 2.60 )</td>
<td>5</td>
<td>Sand mixtures: silty sand to sandy silt</td>
</tr>
<tr>
<td>( 2.60 &lt; I_c &lt; 2.95 )</td>
<td>4</td>
<td>Silt mixtures: clayey silt to silty clay</td>
</tr>
<tr>
<td>( 2.95 &lt; I_c &lt; 3.60 )</td>
<td>3</td>
<td>Clays: silty clay to clay</td>
</tr>
<tr>
<td>( I_c &gt; 3.60 )</td>
<td>2</td>
<td>Organic soil: peats</td>
</tr>
</tbody>
</table>

Table 2. Probabilistic model of the \( I_c \) – based SBT classification boundaries.

<table>
<thead>
<tr>
<th>Boundaries</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>1.31</td>
<td>0.1</td>
<td>[1.01, 1.61]</td>
</tr>
<tr>
<td>B2</td>
<td>2.05</td>
<td>0.1</td>
<td>[1.75, 2.35]</td>
</tr>
<tr>
<td>B3</td>
<td>2.6</td>
<td>0.05</td>
<td>[2.45, 2.75]</td>
</tr>
<tr>
<td>B4</td>
<td>2.95</td>
<td>0.05</td>
<td>[2.8, 3.1]</td>
</tr>
<tr>
<td>B5</td>
<td>3.6</td>
<td>0.1</td>
<td>[3.3, 3.9]</td>
</tr>
</tbody>
</table>
4 EFFECT OF SOUNDED NUMBER

To investigate the effect of CPT soundings number \( M \) on the proposed framework, an additional scenario with \( M=25 \) CPT soundings is discussed. Those 25 CPT soundings are performed with equal space in the 2D cross-section of an illustrative example (i.e., see Figure 1). Following the same procedures as described above, \( N_M=500 \) Monte Carlo samples of 2D SBT cross-section are obtained. The most likely SBT cross-section in this scenario is shown in Figure 6a. The approximated soil stratigraphy in this scenario becomes more accurate. The quantified uncertainty also shrinks significantly, as shown in the Figure 6b. Those thin bright areas with high uncertainty are in good agreement with boundary locations. The proposed method is data-driven and features fine scale spatial variability to the interpolation result when \( M \) increases.

5 CONCLUSIONS

A novel Monte Carlo simulation (MCS) – based framework was proposed in this paper for interpreting soil stratigraphy in a 2D cross-section from

![Figure 3. Four examples of 2D \( I_c \) data RFS.](image)

![Figure 4. SBT cross-sections for the four \( I_c \) samples in Figure 3.](image)
Bayesian supervised learning. The associated interpolation uncertainty was modelled by non-parametric random field simulation based on the learned results. A probabilistic soil behavior type (SBT) chart was developed for incorporating the model uncertainty in the empirical chart. The interpolation uncertainty of \( I_c \) data and model uncertainty in SBT chart were considered simultaneously under MCS. Key equations and detailed implementation procedures were provided. Numerical example was illustrated and showed that the proposed method performed reasonably well. Sensitivity study suggested that the proposed method was data-driven. As the number of CPT soundings increased, the most likely SBT cross-section became accurate and associated interpolation uncertainty reduced significantly.

ACKNOWLEDGEMENTS

The work described in this paper was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CityU 11213119 and CityU 11202121). The financial supports are gratefully acknowledged.

REFERENCES


