Proceedings of the XVIII ECSMGE 2024
GEOTECHNICAL ENGINEERING CHALLENGES
TO MEET CURRENT AND EMERGING NEEDS OF SOCIETY

ISBN 978-1-032-54816-6
DOI 10.1201/9781003431749-51

ECSMGE 24
XVIII EUROPEAN CONFERENCE ON SOIL MECHANICS AND GEOTECHNICAL ENGINEERING

Integrating cluster-based knowledge for local undrained shear strength prediction

Intégration de connaissances basées sur les clusters pour la prédiction locale de la résistance au cisaillement non drainé

S. Collico
DMT GmbH & Co. KG, Essen, Germany

G. Spagnoli*
Sweco, Essen, Germany

A. Fraccica
Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), Rome, Italy

E. Romero
Universität Politecnica de Catalunya (UPC), Barcelona, Spain

**giovanni.spagnoli@sweco-gmbh.de*

ABSTRACT: The prediction of local (i.e., site-specific) undrained shear strength from soil properties (e.g., water content, Plasticity index, Liquid Limit) is a cost-effective solution to preliminary assess the proper foundation type at a project site. Nevertheless, the local dataset is often incomplete and sparse to infer robust prediction of local strength parameters. To bypass such inconvenience, global or soil-type correlations might be employed to predict undrained shear strength from local soil properties measurements. However, by doing so two main drawbacks can arise: a great uncertainty is introduced for prediction of undrained shear strength due to the large number of soil types the global database is usually composed. Secondly, a problem of pertinence might arise when soil type-specific correlations are employed. In this study the local predictions of undrained shear strength are estimated integrating clustered global information through a Bayesian formulation, which allow to assess the problem of local dataset incompleteness. The methodology is shown by making use of an already compiled and published database.

RÉSUMÉ: La prédiction de la résistance au cisaillement non drainée locale (c'est-à-dire spécifique au site) à partir des propriétés du sol (par exemple, la teneur en eau, l'indice de plasticité, la limite de liquidité) est une solution rentable pour évaluer préliminairement le type de fondation approprié sur un site de projet. Néanmoins, l'ensemble de données locales est souvent incomplet et clairsemé pour déduire une prédiction robuste des paramètres de résistance locaux. Pour éviter de tels inconvénients, des corrélations globales ou par type de sol pourraient être utilisées pour prédire la résistance au cisaillement non drainé à partir de mesures locales des propriétés du sol. Cependant, ce faisant, deux inconvénients principaux peuvent survenir: premièrement, une grande incertitude est introduite pour la prévision de la résistance au cisaillement non drainé en raison du grand nombre de types de sols qui composent habituellement la base de données mondiale. Deuxièmement, un problème de pertinence pourrait surgir lorsque des corrélations spécifiques au type de sol sont utilisées. Dans cette étude, les prédictions locales de la résistance au cisaillement non drainé sont estimées en intégrant des informations globales groupées via une formulation bayésienne, ce qui permet d'évaluer le problème de l'incomplétude des ensembles de données locaux. La méthodologie est présentée en utilisant une base de données déjà compilée et publiée.

Keywords: Undrained shear strength; local dataset; global dataset; Bayesian; cluster-based.

1 INTRODUCTION

At a project site local (i.e., site-specific) geotechnical data are often incomplete and sparse. A typical scenario is the one where the amount of soil properties and Atterberg limits measurements exceed the number of undrained shear strength observations. It is then common practice to compute a weighted average of undrained shear strength measurements with direct and indirect estimates of it. However, the indirect estimates of undrained shear strength might carry large systematic uncertainties when global correlation are used (Collico and Arroyo, 2023; Ching and Phoon, 2019). On the other hand, when soil type-specific correlations are applied a problem of pertinence might arise. To assess such issues, Bayesian methodologies

have been proposed to assess and reduce systematic uncertainties by sequentially updating local information with global and regional measurements (Zhang et al., 2004; Wang and Cao, 2013; Wang et al. 2010). Despite being efficient, these approaches would require the complete set of independent variables to first construct and further validate correlations.

More recently Ching and Phoon (2019) and Wu et al. (2022) developed an efficient and elegant hybrid Bayesian methodology to assess uncertainty of local data and account for global geotechnical information. In this context, it might be pertinent to wonder whether global information or filtered ones according to some criterion should be considered for local undrained shear strength updating. This seems particularly relevant when incomplete and scarce information are available at the project site. A first approach would be to filter global information according to the same soil-types identified at the project site. Nevertheless, by filtering the database according to the same textural soil classification might not represent an optimal criterion. If a soil type classification problem based on some in-situ test is considered, it is thought that the geological textural classification might not be an exhaustive indicator of soil (Collico et al., 2023).

As alternative to filtering, unsupervised data-driven approaches, such as cluster-based solutions can be applied. In particular, cluster-based solution appropriately calibrated by engineers' interpretation and experience might provide an efficient way of data subdivision to allocate soil-similarity (Collico et al. 2024). Aware of that, in this study, the hybrid workflow proposed by Ching and Phoon (2019) is developed and extended to account for global cluster-based information and allocate data site similar for local undrained shear strength updating. The approach is demonstrated making use of already compiled and published databases such as CLAY/10/7490 (Ching et al., 2014) and S-CLAY/7/168 (D'Ignazio et al., 2016).

2 GLOBAL DATABASE DESCRIPTION

Let denoted as Y_g datapoints from global database here represenetd by four geotechnical parameters such as that: Y_g = [Y_{g1}=LL; Y_{g2}=PL; Y_{g3}=PI; Y_{g4}= w; Y_{g5}=Su/σ'_{v0}], with LL=liquid Limit, PL=Plastic Limit, PI=plasticity Index, *w*=water content and Su/σ'_{v0}=normalized undrained shear strength.

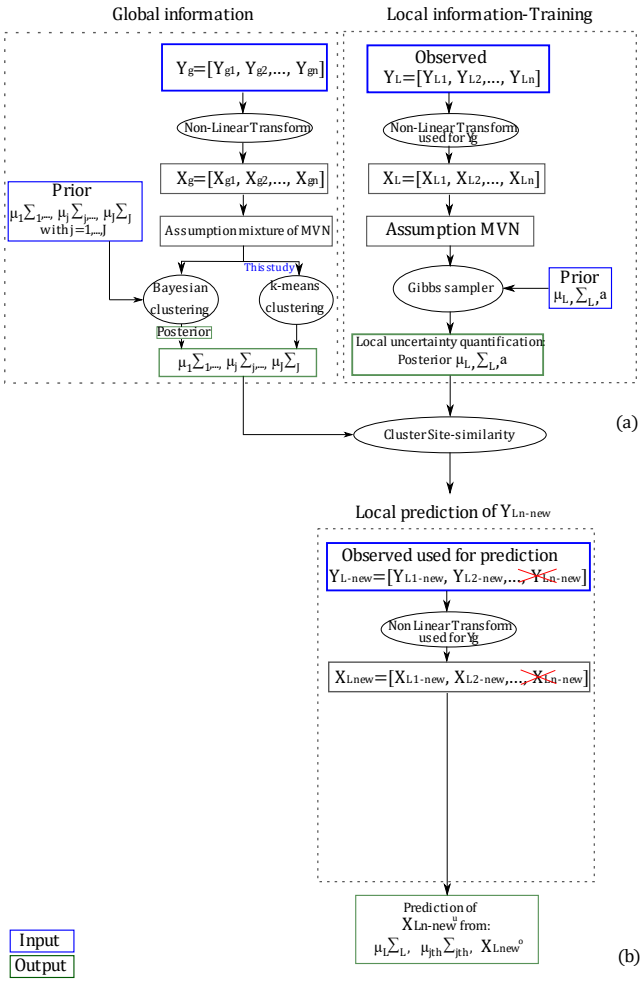

Figure 1. Workflow of methodology employed in this study.

2.1 Probabilistic model for Global information

The proposed method requires the definition of a global (i.e., no-local) multivariate Probability Density Function (PDF). Such no-local PDF is of mainly importance when scarce and incomplete X_L information is available at a project site.

To construct a valid multivariate normal distribution, the variables Y_g (non-normal distributed) (Figure 1) needs to be transformed into normally distributed ones X_g. This is accomplished through a non-linear transform (see Ching and Phoon, 2014). It is then assumed a linear dependence among X_g variables such that their probability density can be described by a $n-th$ dimensional Gaussian Mixture Model which reads:

$$f(X_g|\Theta) = \sum_{j=1}^{J} \pi_j f(X_g|\mu_j, \sum_j) \qquad (1)$$

with $f(X_g|\mu_j, \sum_j)$ denoting the j-th multivariate normal component (i.e., j-th cluster) with mean μ_j, covariance $\sum_j$ and weight π_j. J is the number of Gaussian mixture components and Θ vector containing

all Gaussian Mixture parameters (i.e., means, Covariances and weights). For J=1 eq. (1) recovers the definition of multivariate normal distribution. In this study the global datapoints are modelled by eq. (1) through the k-Means algorithm. Nevertheless, a Bayesian Mixture formulation could be also applied instead of eq. (1) allowing to infer prior knowledge about clusters depending on local dataset information.

2.2 Probabilistic model for Local dataset

Let consider the local dataset reported in Table 1 for a the Norwegian site of Drammen (D'Ignazio et al., 2016) as example. Table 1 represents an optimistic scenario at a project-site. A more realistic situation is the one where 60% of datapoints in Table 1 are missing. Let denote as $X_L{}^u$ the missing entries and as $X_L{}^o$ the observed datapoints of X_L. In such a case, great uncertainty would characterize local PDF parameters, making Bayesian inference necessary.

To construct a valid local PDF, the datapoints Y_L are transformed into X_L by considering the same non-linear transform coefficients employed for Y_g (Figure 1a). After transformation, the local datapoints X_L are assumed to be distributed according to a multivariate normal distribution as:

$$f(X_L|\mu_L, \sum_L) = mvn(X_L|\mu_L, \sum_L) \quad (2)$$

with $\mu_L, \sum_L$ mean and covariance matrix of local PDF.

2.3 Bayesian formulation for incomplete datasets

To assess the local data incompleteness the Bayesian formulation proposed by Ching and Phoon, (2019) is employed. We here recall only the main steps of the methodology.

As a first step, the prior knowledge of the unknown random variables has to be defined (Figure 1a). Therefore, the prior knowledge of μ_L, $\sum_L$ and hyperparameters a for $\sum_L$ parametrization, need to be elicited. Having denote as $f(\mu_L)$ and $f(\sum_L)$ the prior density distribution of μ_L and $\sum_L$, their conjugate form is assumed since enable exact sampling. The prior knowledge of μ_L is taken as a normal multivariate distribution, while $\sum_L$ is modelled as the inverse Wishart distribution such as:

$$f(\mu_L) = mvn\,(\mu_L|\mu_0, C_0) \quad (3)$$

$$f(\sum_L) = IW\,(\sum_L|\sum_0, v) \quad (4)$$

Table 1. Local observation for a Norway Site (Drammen).

Depth	LL (%) (Y_{L1})	PL (Y_{L2})	PI (Y_{L3})	*w* (%) (Y_{L4})	*Su/σ'v0* (Y_{L5})
Training Y_L					
5.21	58.72	10.34	0.80	65.47	0.189
6.16	65.58	18.82	0.81	65.58	0.164
7.14	75.2	20.56	0.78	56.15	0.148
7.45	88.51	18.73	0.73	65.25	0.141
7.50	88	18	0.73	65	0.245
7.80	60	29	0.89	52	0.322
8.48	75.84	15.3	0.76	61.65	0.210
9.04	78.22	19.1	0.76	58.22	0.133
9.25	33	23	1.12	32	0.157
9.42	92.23	20.35	0.72	64.08	0.186
11.93	40.16	8.75	0.89	27.60	0.0872
13.00	25	3	0.96	26	0.178
Testing Y_L					
4.01	39.28	9.74	0.90	30.68	0.181
13.04	25	3	0.96	25.7	0.119
17.39	23.3	2.73	0.97	17.2	0.122

The mean μ_0 is taken as zero mean vector of $nx1$ dimension, while the prior covariance matrix C_0 (nxn) is taken to be a diagonal matrix with very large diagonal elements (i.e., 104), such that $f(\mu_L)$ tend to be uninformative. Concerning $f(\sum_L)$ the inverse-Wishart distribution is assumed with scale matrix $\sum_0$ and degree of freedom $v = n + 1$. To make $f(\sum_L)$ noninformative a heirarchical form can be applied (Huang and Wand 2013; Ching and Phoon 2019):

$$f(\sum_L|a) = IW\,(\sum_L|\sum_0, v) \quad (5)$$

$$f(a_i) = IG(a_i|\alpha, \beta) \quad (6)$$

where $(a = [a_1, \ldots, a_n])$ are the random hyperparameters required for $\sum_L$ parametrization. The prior of each hyperparameters $f(a_i)$ is assumed an inverse gamma distribution with scale and shape parameters α, β. In this study α, β are taken as 0.5 and 10^{-4} respectively such that a non-informative linear dependence and variance is assumed (Wu et al., 2022).

2.4 Local posterior inference

The Posterior inference of the local unknown random variables (μ_L, $\sum_L, a$) is derived through the Gibbs sampler (GS) (Geman and Geman, 1984), which decompose the unknown random parameters into groups to randomly sample each group conditional on the remaining ones. This allows to sequentially draw T times the unknown random variables, to final generate (T-t_b) random samples of μ_L, $\sum_L$ and a, which are asymptotically distributed as the target posterior distribution. t_b is defined as burning period (i.e., the

number of generated samples to be discard for posterior inference). It worth noticing that, when missing information characterize the local dataset, by partitioning local mean and covariance matrix accordingly, the Bayesian formulation allow to draw T samples of missing datapoints X_{L^u} conditional to the observed X_{L^o} ones (see Ching and Phoon, 2019 and Wu et al., 2022).

2.5 Local prediction integrating cluster-based information

When the local training dataset is scarce and incomplete, the statistical uncertainty of local PDF parameters $(\mu_L, \sum_L, a, X_{L^u})$ can be significant, propagating into local soil parameters prediction, here denoted as X_{L-new}. In this case, it is reasonable to rely on global information, which are statistically more robust. Nevertheless, global databases might be composed of too different soil-types leading to a too generic prediction of local soil parameters. In this context, the clustering approach previously described in section 2.2 might come in help since enable to filter global information "similar" to the local site of interest. The data similarity between each t-th local samples, generated during the Gibbs sampler, and the j-th global cluster is quantified as follows:

$$p(\mu_{L,t}|\mu_{gj}, \sum_{gj}) = \frac{\pi_j f(\mu_{L,t}|\mu_{gj}, \sum_{gj})}{\sum_{j=1}^{J} \pi_j f(\mu_{L,t}|\mu_{gj}, \sum_{gj})} \quad (7)$$

with:

$$f(\mu_{L,t}|\mu_{gj}, \sum_{gj}) = \left|\sum_{gj}\right|^{-1/2} \cdot (2\pi)^{-n/2} \cdot \exp\left(\left[-0.5(\mu_L - \mu_{gj})^T \sum_{gj}^{-1} (\mu_L - \mu_{gj})\right]\right. \quad (8)$$

and π_j weigth of the j-th Gaussian mixture component.

Once site similarity of the t-th sample is assessed, prediction of X_{L-new} can be inferred according to the hybrid method proposed by Ching and Phoon (2019) as:

$$f(X_{L-new-t}|hb) = f\left(X_{L-new}|\mu_{gj}, \sum_{gj}\right) \cdot f(X_{L-new}|\mu_L, \sum_L) \quad (8a)$$

$$f(X_{L-new-t}|hb) = \sum_{t=tb+1}^{T} w_t \cdot f\left(X_{L-new-t}|\mu_{hb,t}, \sum_{hb,t}\right) \quad (8b)$$

with $\mu_{gj}, \sum_{gj}$, mean and covariance matrix of the most similar j-th global cluster, $\mu_{hb,t}, \sum_{hb,t}$ mean and covariance matrix of the hybrid multivariate distribution given by product of $f(X_{L-new}|\mu_{gj}, \sum_{gj})$ and $f(X_{L-new}|\mu_L, \sum_L)$ and w_t is the weight of each condition density distribution $f(X_{L-new-t}|hb)$.

3 CASE STUDY

To capture the benefit of cluster-based information on local prediction, the Norway site, Drammen (Table 1), extracted from D'Ignazio et al. (2016) is considered. The local dataset is subdivided into training (to derive posterior estimated of local PDF parameters) and testing. Four case scenarios are exploited:

- Only 40% of datapoints of Table 1 is considered available for training, while the remaining 60% is considered missing. We denote such scenario as Incomplete case;
- All the datapoints of Table 1 are considered for training. Such case is denotated as Complete scenario.
- For both Complete and Incomplete scenarios, global information employed in eq. (8) is subdivided in one (no-cluster scenario) and two clusters (cluster-based scenario) for a total of four different cases (i.e., Incomplete no-cluster, Incomplete cluster-based, Complete no-cluster, Complete cluster-based). It is worth noticing that for illustration purposed optimal data subdivision (criterion such as BIC) is not considered.

For local X_{L-new} prediction, we will assume that all information (LL, PL, PI, w= Y_{L-new^o}) will be available except for Su/σ'_{v0}, which is here taken as variable of interest to be predicted (i.e., $Su/\sigma'_{v0} = Y_{L-new^u}$).

4 RESULTS

As a first step both local and global datapoints are transformed from Y into X domain (Figure 2a and Figure 2b). As a second step global information are subdivided according to the number of clusters selected (i.e., one and two in this study). The k-means algorithm is used as illustrative example. Figure 2c reports the global database subdivision re-transformed in the original Y space when two clusters are selected.

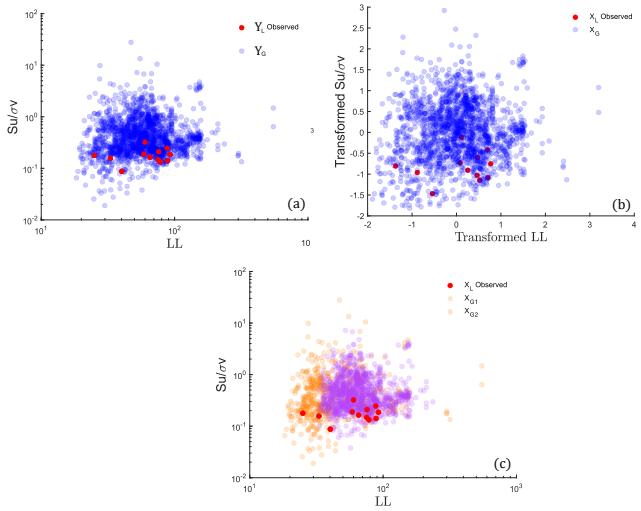

Figure 2. Global and local Normalised undrained shear strength and Liquid Limit datapoints: a) in the original Y space. b) in the X space. c) subdivided according to two clusters after k-NN algorithm.

As a third step the Gibbs sampler is run to compute posterior estimates of local PDF parameters (μ_L, $\sum_L$, a) and (μ_L, $\sum_L$, a, $X_L{}^u$) for the Complete and Incomplete case scenarios respectively. These posterior estimates are then employed to further generate X_{L-new} to be converted into the original Y space. It is worth noticing that at this stage no testing datapoints have been used. The generated samples X_{L-new} would then represent soil parameter prediction if no additional information at the site is acquired.

As an example, Figure 3 reports the generated (T-t_b) samples of local undrained shear strength and LL according to the four scenarios considered. It can be observed that by considering the Complete scenario, no major difference subsists between generated samples of no-cluster (Figure 3a) and cluster-based (Figure 3b) cases. This does not come as surprise since a not clean separation among clusters subsists. When 40% of datapoints are considered for training (Incomplete scenario), generated samples start to converge to global datapoints scatter independently whether one or two cluster are considered (Figure 3c, Figure 3d respectively).

When additional local information becomes available (i.e., testing datapoints), the local undrained shear strength $Y_{L-new}{}^u$ can then be predicted from the observed testing datapoints $Y_{L-new}{}^o$ and trained PDF parameters μ_L, $\sum_L$, a, for the four scenarios. It can be observed that when the training is performed using the complete local dataset a better prediction of undrained shear strength in terms of accuracy and precision is obtained.

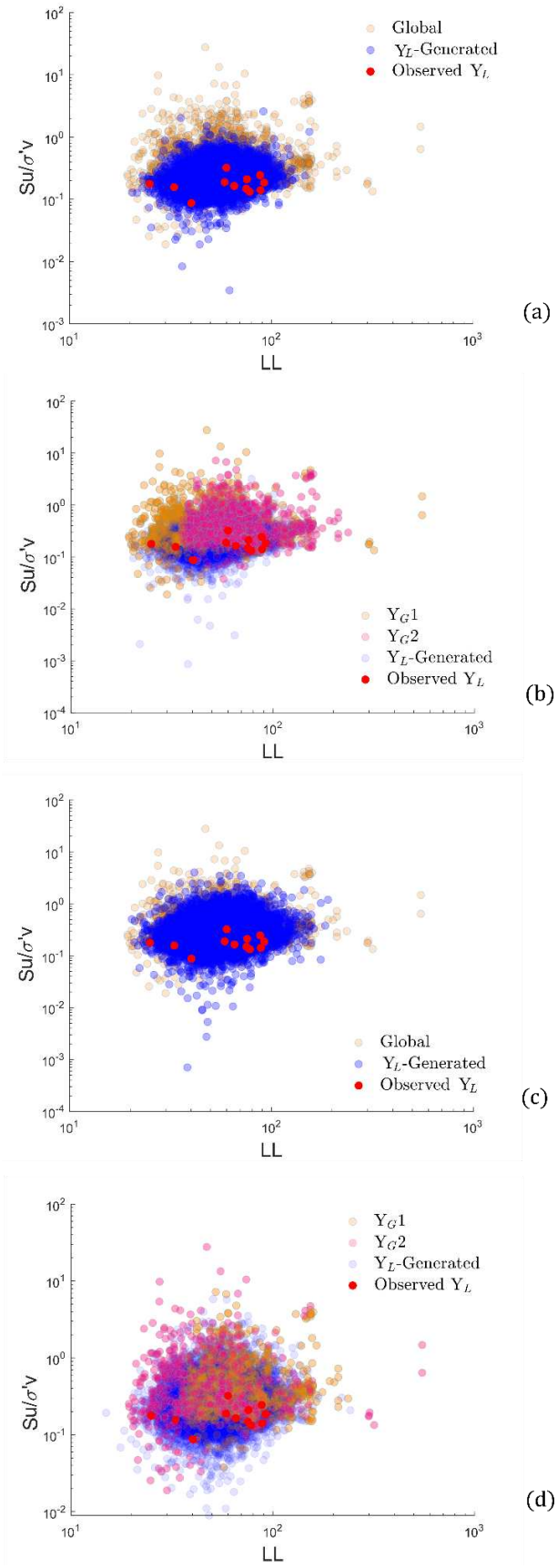

Figure 3. Generated $Y_{L\text{-}new}$ samples for: a) Complete no-cluster scenario. b) Complete cluster-based scenario. c) Incomplete no-cluster scenario. d) Incomplete cluster-based scanrio.

The accuracy seems to slightly increase for both Complete and Incomplete scenarios when clustering approach is employed. Nevertheless, such difference is negligible since the two clusters introduced according to the k-means approach are not characterized by major difference in terms of Pearson correlation and Covariance.

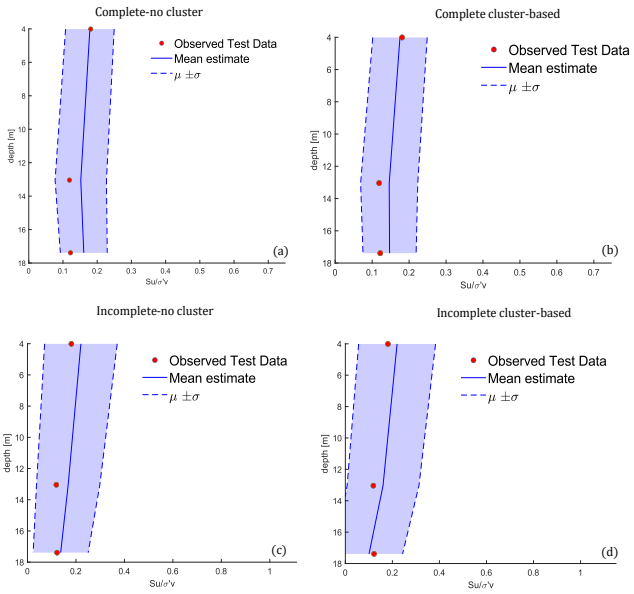

Figure 4. Predicted Su/σ'_{v0} for: a) Complete no-cluster scenario. b) Complete cluster-based scenario. c) Incomplete no-cluster scenario. d) Incomplete cluster-based scanrio.

5 CONCLUSIONS

This paper propose a Bayesian formulation (implementing and extending the formulation by Ching and Phoon (2019)) to integrate cluster-based global knowledge for local undrained shear strength prediction, and assess the challenge of local database incompletness. The work hihglight how cluster-based knowledge might slighlty increase accuracy of local undrained shear strength prediction. Neverthless, the work hihglights how cluster-based solution should be integrated with reason within the described Bayesian formulation. To do so, the approach proposed could provide more significance improvement to local undrained shear strength prediction when a Bayesian Mixture formulation is employed for global dataset manipulation. This would allow to account for more insightful information based on local dataset structure and enegineering experience.

REFERENCES

Ching, J. and Phoon, K.K. (2019) Constructing site-specific multivariate probability distribution model using Bayesian machine learning. *Journal of Engineering Mechanics*, 145 (1). https://doi.org/10.1061/(asce)em.1943-7889.0001537.

Ching, J., Phoon, K.K. and Chen, C.H. (2014) Modeling piezocone cone penetration (CPTU) parameters of clays as a multivariate normal distribution. *Canadian Geotechnical Journal*, 51(1), https://doi.org/10.1139/cgj-2012-0259.

Collico, S. and Arroyo, M. (2023) Bayesian mixture analysis of a global database to improve unit weight prediction from CPTu. *Engineering Geology*, 327, pp. 107353. https://doi.org/10.1016/j.enggeo.2023.107353.

Collico, S., Arroyo, M. and Devincenzi, M. (2024) A simple approach to probabilistic CPTu-based geotechnical stratigraphic profiling. *Computers and Geotechnics*, 165, pp. 105905. 10.1016/j.compgeo.2023.105905.

Tan, S.A. and Länsivaara, T.T. (2016) Correlations for undrained shear strength of Finnish soft clays. *Canadian Geotechnical Journal*, 53 (10), pp. 1628–45. https://doi.org/10.1139/cgj-2016-0037.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 721–41.

Huang, A. and Wand, M. P. (2013) Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8 (2), pp. 439–52. https://doi.org/10.1214/13-BA815.

Wang, T. Au, S.K. and Cao, Z. (2010) Bayesian approach for probabilistic characterization of sand friction angles. *Engineering Geology*, 114 (3–4), pp. 354–363, https://doi.org/10.1016/j.enggeo.2010.05.013.

Wang, Y. and Cao, Z. (2013) Probabilistic characterization of Young's modulus of soil using equivalent samples. *Engineering Geology*, 159, pp. 106–118. https://doi.org/10.1016/j.enggeo.2013.03.017.

Wu, S., Ching, J. and Phoon, K.K. (2022) Quasi-site-specific soil property prediction using a cluster-based hierarchical Bayesian model. *Structural Safety*, 99, pp. 102253. https://doi.org/10.1016/j.strusafe.2022.102253.

Zhang, L., Tang, W.H., Zhang, L. and Zheng, J. (2004) Reducing uncertainty of prediction from empirical correlations. *Journal of Geotechnical and Geoenvironmental Engineering*, 130 (5), https://doi.org/10.1061/(ASCE)1090-0241(2004)130:5(526).