TU Delft

MSc thesis in Geo-Engineering

submitted to Delft University of Technology in partial fulfillment of the requirements for Master in Civil Engineering

Development of 2D subsurface site characterization by the fusion of geotechnical and geophysical data

by

Sijun Zeng

July 2023

Committee

Prof. Dr. M.A. (Michael) Hicks (*Chair*)	Geo-Engineering
Dr. ir. A.P. (Bram) van den Eijnden	Geo-Engineering
Dr. G. (Guillaume) Rongier	Applied Geology

Abstract

Site characterization is indispensable in the design phase of geotechnical engineering projects. As a key factor in site characterization, the characterization of soil undrained shear strength (S_u) is always in the spotlight. Various methods, including laboratory and in-situ tests, have been developed to measure S_u. Nevertheless, these measurements are usually sparse at a specific site due to limited time and budget. To enhance S_u characterization, other relevant geotechnical investigation data (e.g., cone penetration test data), can be transformed into S_u through empirical correlations (referred to as transformation models) to provide more information on S_u. Considering this process introduces the transformation uncertainty and a developed transformation model may not be fully applicable to a local site, probabilistic transformation models (PTMs) have been developed to characterize soil parameters in a site-specific way and quantify the uncertainty to augment engineers' judgement.

However, few PTMs incorporate the spatial correlation of soil parameters, especially in the horizontal direction. This limitation hampers the ability to probabilistically characterize S_u in 2D/3D space, which is significant in practice. Moreover, estimating the horizontal spatial correlation from pure geotechnical data is challenging because they are typically sparse. In light of these circumstances, this thesis first proposes a PTM-based scheme to probabilistically characterize S_u in 2D. Then it is proposed to integrate geophysical data into the scheme. Compared to typical geotechnical investigations, geophysical surveys provide abundant 2D/3D measurement data, which are often correlated with geotechnical data. The fusion of these two data sources benefits characterizing geotechnical data including S_u. Particularly the horizontal spatial correlation of S_u 2D domain can be estimated from the abundant geophysical data. To be specific, a well-established PTM, MUSIC-X, by which measured S_u and other relevant soil parameters can be used to preliminarily characterize S_u, is first adopted. In this case, characterization specifically refers to simulating 1D vertical profiles of S_u. It is then combined with the intrinsic collocated co-kriging (ICCK) model, by which primary data (i.e., S_u) in 2D or theoretically 3D space can be estimated through linearly combining the preliminarily characterized S_u from MUSIC-X modelling and observed secondary data (i.e., geophysical data). The secondary parameter considered in this study is interval velocity (V_{int}).

The scheme, to combine the MUSIC-X and ICCK model to estimate S_u in 2D space by the fusion of geotechnical and geophysical data, is applied to a real case study at Hollandse Kust (west) wind farm zone to demonstrate its effectiveness. The results indicate that such a scheme can robustly estimate a 2D cross section of S_u with quantified uncertainty. A comparative analysis is conducted between the proposed scheme and two alternatives, one lacking preliminary S_u characterization (i.e., without MUSIC-X modelling) and one lacking geophysical data, confirming that the proposed scheme has a relatively high accuracy in the estimated cross section. The research reveals it is sensible to combine MUSIC-X and ICCK for 2D S_u characterization and brings a new perspective that integrating geotechnical and geophysical data is promising to characterize soil parameters in higher dimensional space.

Acknowledgements

First and foremost, a special acknowledgement is given to the external assistance received during this Master's thesis project. I am grateful to Prof. Jianye Ching <jyching@gmail.com> for generously providing me with the codes of the MUSIC-X model, which greatly facilitated my understanding of this complex model. In addition, I would like to thank Netherlands enterprise agency (RVO) for offering free access to the site investigation data at Hollandse Kust (west) Wind Farm Zone, and thank Fugro for providing supplementary geophysical data at this site.

Moreover, I would like to express my deepest gratitude to my thesis supervisor, Prof. Dr. M.A. (Michael) Hicks, who inspired me to undertake such an interesting research project, got time to have weekly discussions with me, guided me with the correct way to go and offered me an opportunity to have an additional thesis project. It is also a must to thank the rest members in my thesis committee. Many thanks to Dr. ir. A.P. (Bram) van den Eijnden who was always passionate and patient to answer my questions no matter in this project or the courses he taught. Many thanks to Dr. G. (Guillaume) Rongier who has provided me with valuable suggestions and revised my thesis conscientiously, which can be proved by the 120 comments he gave me. And thanks to Dr. D. (Divya) Varkey. Although she was not involved in the project directly, she used to enthusiastically guide me in the additional thesis project.

Lastly I would like to make genuine acknowledgements to all my friends and family for accompanying and supporting me in the last two years. Special thanks to my parents and Dudu Liu for comforting me during times of stress and anxiety, which is important as TU Delft students can easily get caught up in these two emotions before an exam.

Contents

List of Figures

List of Tables

Acronyms

ACF auto correlation function.

ANN artificial neural network.

BCS Bayesian compressive sampling.

BH borehole.

C characterization.

CART classification and regression tree.

CDF cumulative distribution function.

CDP common depth point.

CI confidence interval.

COV coefficient of variance.

CPT cone penetration test.

CPTU cone penetration test with pore water pressure measurement.

GANFIS genetic algorithm - adaptive network based Fuzzy inference system.

GS Gibbs sampler.

ICCK intrinsic collocated cokriging.

JD Johnson distribution.

KL Karhunen–Loève.

LMC linear model of coregionalization.

MAE mean absolute error.

MAPE mean absolute percentage error.

MARS multivariate adaptive regression spline.

MASW multichannel analysis of surface wave.

MCMC Markov chain Monte Carlo.

ML machine learning.

MS multi-source.

MSE mean square error.

MUSIC-X multivariate, uncertain and unique, sparse, incomplete and spatial/time dimension. In this thesis it typically refers to a multivariate probabilistic transformation model.

OK ordinary kriging.

PANFIS particle swarm optimization - adaptive network based Fuzzy inference system.

PDF probability distribution function.

PP pocket penetrometer.

PTM probabilistic transformation model.

QExp squared exponential.

RF random forest.

RMSE root mean squared error.

SB bounded system.

SExp single exponential.

SK simple kriging.

SL lognormal system.

SMK second order Markov.

SOF scale of fluctuation.

SPT standard penetration test.

SU unbounded system.

SVM support vector machine.

SVR support vector regression.

UC unconfined compression.

UHR-MCS ultra-high resolution multichannel seismic.

W-M Whittle–Matérn.

XGBoost extreme gradient boosting.

Symbols

δ	Scale of fluctuation
$\gamma(\mathrm{h})$	Variogram
μ_s	Site-specific mean vector
$\mathbf{a}$	Hyperparameter
$\mathbf{C}_\mathrm{s}$	Site-specific covariance matrix
$\mathbf{R}$	Autocorrelation matrix
$\mathbf{X}^\mathrm{o}$	Observed soil parameters
$\mathbf{X}^\mathrm{u}$	Unobserved soil parameters
$\mathbf{X}$	Complete soil parameters
μ	Mean
ν	Smoothness parameter
ϕ'	Effective friction angle
$\rho(\mathrm{h})$	Spatial correlation
σ'_p	Preconsolidation stress
σ'_v	Vertical effective stress
$\sigma_{\mathrm{v}0}$	Overburden stress
σ	Standard deviation
$\mathrm{B_q}$	Pore pressure ratio
b	nugget
c'	Effective cohesion
$\mathrm{C}(\mathrm{h})$	Covariogram (covariance)
C_0	sill

C_c Compression index

C_s Swelling index

Data Observed soil data

E_u Young's modulus

h Lag distance

LI Liquidity index

LL Liquid limit

N_{60} Corrected blowcount for standard penetration test

OCR Over-consolidation ratio

P_a Atmospheric pressure

PI Plasticity index

q_c Cone resistance

q_E Effective cone resistance

q_t Corrected tip resistance

q_{t1} Normalized corrected tip resistance

R^2 Coefficient of determination

S_u Undrained shear strength

$S_u(\text{mob})$ Mobilized undrained shear strength

$S_u(\text{PP})$ Undrained shear strength measured by pocket penetrometer test

$S_u(\text{UC})$ Undrained shear strength measured by unconfined compression test

u_0 In-situ pore water pressure

u_2 Measured pore pressure behind the cone tip

V_{int} Interval velocity

V_{rms} Root mean square velocity

V_s Shear wave velocity

#97, #98, #104 Position 97, Position 98, Position 104

a Range parameter

IC Soil behavior type index

N_k, N_{kt}, N_{ke}, and $N_{\Delta u}$ Empirical cone factors

R Coefficient of correlation

1. Introduction

1.1. Background

Site characterization (i.e., soil stratification, discontinuities, anomalies, physical/mechanical properties and groundwater flow; Phoon et al., 2022) is essential in geotechnical designs and analysis. In this study, the characterization of soil undrained shear strength (S_u) is primarily considered, which is an important design parameter in slope stabilization (Hicks & Samy, 2002), deep excavation (Nguyen & Likitlersuang, 2021), and foundation design (Li et al., 2015). Conventionally sampling and laboratory tests are usually required to measure such a design parameter. However, they are sophisticated and expensive (Ching & Phoon, 2019a). At a specific site, because of time and budget constraints, S_u measurement data are typically sparse. For example, there may be only one borehole (BH) per 300 m^2 to sample soils to a depth of 20 meters and only 10 samples from the BH are applied to laboratory tests (i.e., 10 S_u data points). Actually the sparsity is one of the characteristics of almost all geotechnical data (according to Phoon (2018), the characteristics of geotechnical data can be summarized as MUSIC: multivariate, uncertain and unique, sparse and incomplete). To better characterize S_u, other available geotechnical investigation data are proposed to be transformed into S_u through empirical correlations (referred to as transformation models). These data may be sourced from more cost-effective in-situ tests, such as the cone penetration test (CPT) data, as well as from other accessible laboratory tests such as soil liquid limit (LL). Nevertheless, this transformation process contains a significant amount of transformation uncertainty (Phoon & Kulhawy, 1999). Additionally, the transformation models have the inherent limitation that when applied to a certain site, the soil properties, soil behaviors, and site geology of this site may differ from the data source from where the transformation models are calibrated, directly leading to biases with respect to the actual S_u value (D'Ignazio et al., 2016).

Considering the limitations of the conventional methods, probabilistic transformation models (PTMs) have emerged and gained traction in the field of soil parameter characterization. PTMs can characterize a target soil parameter based on site-specific sparse data. This ability exactly fits analyzing geotechnical data and mitigates the inherent biases that may arise when employing transformations developed from different sites. In addition, PTMs can quantify the transformation uncertainty (e.g., 95% confidence interval) and thus augment existing physics-based methods and engineers' judgement (Ching & Phoon, 2019a; Phoon, 2018). At the early stage, the PTMs remained at a univariate level (i.e., a single prediction output), solely concerning the pairwise empirical correlation (Feng & Jimenez, 2015; Ng et al., 2015; Wang & Cao, 2013; Yan et al., 2009). However, it has been generally found that a soil parameter is always correlated to more than one soil parameter. Conceptually, the transformation uncertainty can be reduced to the maximum extent by combining multiple input information. Thus an ideal PTM can allow multivariate inputs and predict multivariate outputs (Ching & Phoon, 2019a) and recent studies of PTMs indeed show a trend to be multivariate (Ching & Phoon, 2019a, 2020;

Wang & Akeju, 2016; L. Zhang et al., 2018). Following this trend, the departure point of this study is to characterize S_u based on multivariate PTMs.

1.2. Research questions and objectives

While existing multivariate PTMs mainly focus on utilizing the cross correlation between multiple soil parameters, few of them consider the spatial correlation of soil parameters, especially in the horizontal direction. The incorporation of spatial correlation in vertical and horizontal direction can reduce transformation uncertainty (Ching & Phoon, 2020) and more importantly, endow the models with a more practical significance: probabilistically simulating 2D or even 3D cross sections of a certain soil parameter. Furthermore, even though Ching, Phoon, et al. (2022) investigated incorporating the spatial correlation in 3D space, the estimation of the horizontal scale of fluctuation (SOF), an important parameter reflecting the horizontal spatial correlation, was based on limited CPT profiles in the horizontal direction, which is challenging (Ching et al., 2018). Therefore two research questions are summarized and will be solved in this thesis:

- How can a 2D cross section of S_u be developed by a multivariate PTM based scheme?
- How can the horizontal spatial correlation be inferred for the scheme effectively?

The solution to the first question suggests combining the MUSIC-X model (Ching & Phoon, 2020) with the intrinsic collocated cokriging (ICCK) model (Babak & Deutsch, 2009b). MUSIC-X is a multivariate PTM which can simulate 1D profiles of a geotechnical parameter (e.g. S_u) using the cross correlation between geotechnical parameters and vertical spatial correlation. Specifically, the MUSIC-X model estimates S_u at an unobserved depth conditioned on other geotechnical data (e.g., liquid limit and CPT tip resistance) observed at the same depth and at approximately the same position through the cross correlation between S_u and these geotechnical parameters. Meanwhile, the estimation is also conditioned on S_u observed at nearby depths at the same position through the vertical spatial correlation if it is applicable. Furthermore, it is also conditioned on the other geotechnical data observed at nearby depths at the same position through both the cross correlation and vertical spatial correlation. The vertical spatial correlation can be easily estimated from geotechnical data such as CPT data. For an S_u 1D profile at a BH position at a given resolution, S_u at all unobserved depths can be estimated in the same manner. Thus the S_u 1D profile at this BH position can be simulated. This is the preliminary characterization of S_u. Then the simulated S_u 1D profiles are applied to the ICCK model as the primary data to develop a 2D cross section of S_u. ICCK is a special instance of kriging methods, which can estimate a primary parameter at unknown locations (in 2D or theoretically 3D space) with the assistance of abundant secondary data. Specifically, observed primary data and secondary data are combined linearly with different weights, which are determined through the spatial correlation in both vertical and horizontal direction and the cross correlation between these two parameters, to estimate the primary parameter at target locations. More details of these two models will be discussed in Chapter 3. Note it is not proposed to directly integrate the horizontal spatial correlation to the MUSIC-X model (i.e., an upgraded multivariate PTM) for 2D characterization. Instead it is an alternative approach by combining with the ICCK model. Additionally, note although the ICCK model can be directly adopted to develop the 2D characterization of S_u without the preliminary characterization (i.e., no MUSIC-X modelling, based on measured S_u), it is not recommended. This is because the ICCK model cannot utilize other measured geotechnical data (S_u related) to further inform S_u through their cross correlation. As

mentioned earlier, in a real project, the direct measurement of S_u is sparse. Using the same example, there are only 10 measured S_u data points along a 20 m deep BH. By utilizing other measured geotechnical data near this BH, MUSIC-X may simulate 100 S_u data points (including the measured 10 points). It should be admitted that these simulated data contain uncertainties but 90 uncertain S_u data points are better than 90 unknowns. By inputting better-informed S_u data in the ICCK model, the estimated 2D cross section of S_u is more accurate.

The solution to the second question suggests utilizing geophysical data as the secondary data in the ICCK model to estimate the horizontal spatial correlation. In a typical geotechnical project, measured geotechnical parameters (S_u related) are not abundant enough to be the secondary parameter in ICCK estimation for the horizontal spatial correlation. Under such circumstances, geophysical data are considered. Geophysical surveys always provide abundant 2D/3D measurement data. Moreover, geotechnical and geophysical parameters have been usually found to be correlated with each other (e.g., permeability-electrical resistivity, moisture content-electrical resistivity, SPT-electrical resistivity, SPT data-shear wave velocity, CPT data-shear wave velocity, geohydrologic data-electrical conductivity, undrained shear strength-shear wave velocity; Crawford et al., 2018; Hegazy & Mayne, 2006; Hussien & Karray, 2016; Rezaei et al., 2018; Trafford & Long, 2020), so it is reasonable to assume the horizontal spatial correlation of geotechnical data is similar to that of geophysical data. Specifically in this case, it is proposed to use interval velocity (V_{int}) as the secondary parameter. V_{int} refers to the speed at which seismic waves travel through a particular layer, or generally depth interval of rocks or soils in the Earth's subsurface. It is a commonly used geophysical parameter in offshore engineering for site characterization, gas exploration and geohazard assessment. Compared to other alternatives, such as shear wave velocity, the measurement of V_{int} is less technically demanding and more cost-effective, which is attractive given that the offshore geotechnical investigation usually covers a wide scale. Although there are few studies on the correlation between S_u and V_{int}, the seismic wave velocity (including shear and compressional wave velocity) are found to be highly correlated with S_u (Duan et al., 2019). V_{int}, the average of seismic wave velocity within a depth interval, can be reasonably assumed to be also correlated with S_u.

Considering the effect of V_{int} on the characterization of S_u has not been verified by any study, one additional research question arises and serves as an exploratory investigation in the effect of integrating geotechnical and geophysical data:

- What is the effect of utilizing V_{int} to characterize S_u?

This question can be directly addressed through the analysis of results from ICCK modeling. Additionally, integrating V_{int} into the MUSIC-X model and evaluating the impact on the uncertainty of simulated S_u can provide further insights. Actually besides answering the question, this supplementary investigation may offer an extra benefit that the potential uncertainty reduction in simulated S_u profiles from MUSIC-X modelling can make the final S_u cross section more accurate. Therefore it is worthwhile to conduct this investigation.

In summary, the motivation of this thesis is to combine the MUSIC-X model with the ICCK model and utilize geophysical data V_{int} additionally in the combined model to solve the aforementioned research questions. In order to fulfill this motivation, detailed objectives are as follows:

- Integrate V_{int} into the MUSIC-X model.
- Simulate 1D S_u profiles by MUSIC-X and validate the reliability/uncertainty of the simulated profiles for both the V_{int} integrated and not integrated scenario (the original model).

- Compare the simulation performance of the two scenarios and choose a better one as the input of the ICCK model.
- Develop a 2D cross section of S_u by ICCK based on 1D S_u profiles from the MUSIC-X model and V_{int} data and validate the reliability/uncertainty of the estimated S_u cross section.
- Compare the reliability/uncertainty of the estimated 2D S_u cross sections from the scheme proposed in this study (MUSIC-X + ICCK) and the ICCK model and investigate in the computational cost of the scheme.
- Conclude the effect of using V_{int} to characterize S_u.

1.3. Thesis outline

In this thesis, in total 6 chapters are structured in a logical order aiming to answer the research questions clearly.

Chapter 1 comprehensively introduces this study, starting with the importance of characterizing S_u in geotechnical engineering, followed by an overview of mainstream methods to characterize S_u. By analyzing the characteristics of each method and soil parameters, multivariate PTMs are mainly concerned and put forward. Lastly the research questions found in the models are presented, followed by the proposed solutions.

Chapter 2 shows a literature review related to S_u characterization, including conventional methods such as lab tests and in-situ tests and more advanced methods such as PTMs and machine learning. Finally the method to integrate geotechnical and geophysical data is introduced.

Chapter 3 detailly describes the techniques of the MUSIC-X and ICCK model, which are used to solve the research questions. The complicated conditional PDFs derived under a Bayesian framework in MUSIC-X model are provided in Appendix A.

Chapter 4 presents a case study that serves to illustrate the proposed scheme. It begins by providing the basic information about the database under consideration then specifies the data that will be utilized in the study. Subsequently the data are well processed and applied to estimating necessary parameters for the MUSIC-X and ICCK model. Finally, it elaborates the implementation of these models, encompassing the input setting as well as the devised strategy for validating and comparing the output.

Chapter 5 displays the results from both MUSIC-X modelling and ICCK modelling in the case study. Each part is followed by a discussion based on the corresponding validation and comparison strategy.

Chapter 6 summarizes the final conclusions and has an outlook on what can be done in the future to improve the study.

2. Literature review

This chapter first provides an overview of conventional methods (i.e., laboratory and in-situ tests) to estimate soil undrained shear strength (S_u). A more detailed introduction to triaxial tests and cone penetration tests (CPTs), which are the representatives of laboratory and in-situ tests, is presented. Moreover, the broader idea behind using typical in-situ tests to estimate S_u is actually transformation models, so a review on transformation models is given as well. Then a thorough review of applying probabilistic transformation models (PTMs) to soil parameter characterization is provided, including univariate, multivariate, and non-stationary probabilistic methods. Subsequently, considering the trend to apply machine learning (ML) methods to soil parameter prediction and characterization, to make the literature review more comprehensive, ML also has been reviewed. Finally, recent developments to integrate geotechnical and geophysical data to characterize soil parameters are reviewed.

2.1. Conventional methods

Conventionally, S_u is estimated by laboratory and in-situ tests. Among laboratory tests, triaxial tests, simple shear tests, and direct shear tests have been typically used (Mayne, 1985). Generally, laboratory tests can give precise and reliable measurements of S_u, however, they are relatively time-consuming and expensive. Among in-situ tests, cone penetration tests with and without pore pressure measurements (CPTU/CPT), field vane tests, dilatometer tests, fall cone tests, and uniaxial and plane strain tests have been typically used (Mayne, 1985). In-situ tests are more efficient and cost-effective, however, S_u is usually not directly measured but transformed from in-situ data through empirical correlations (i.e., transformation models), which leads to unknown transformation uncertainty. In the following sections, a more detailed review will be given to triaxial tests and CPTs, the most representative tests in laboratory and in situ to measure S_u, followed by the elaboration of transformation models.

2.1.1. Triaxial test

The most reliable laboratory method to assess S_u is the triaxial test (Thakur et al., 2016), which can be categorized as unconsolidated undrained (UU), consolidated drained (CD), and consolidated undrained (CU) compression or extension. The advantages of a triaxial test over other simpler lab tests, such as direct shear, include the ability to control specimen drainage and take measurements of pore water pressure. The schematic diagram of the triaxial test apparatus can be viewed in Fig. 2.1. The descriptions of the apparatus are as follows (Lacasse & Berre, 1988; Nakase & Kamei, 1983): soil samples are enclosed in a rubber membrane and loaded by the piston at the top of the cell; there are valves to control the drainage of the cell, and the cell pressure and pore pressure can be measured; electrical transducers are used for automatic data

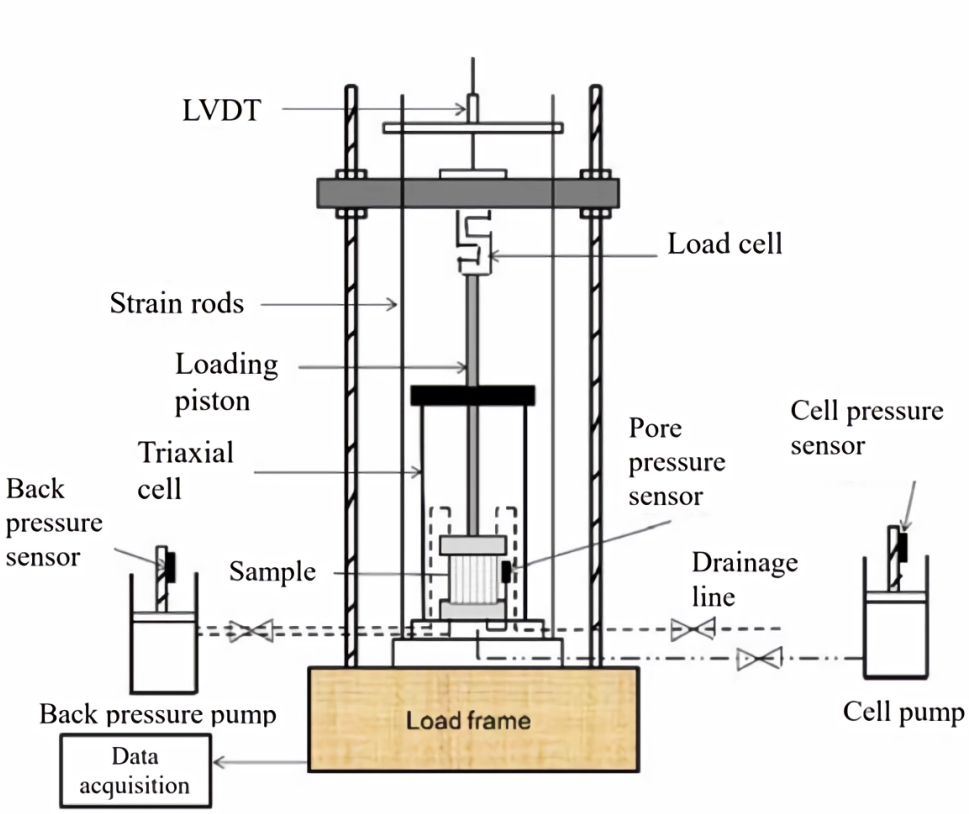


Fig. 2.1. Schematic diagram of triaxial test apparatus (Omar & Sadrekarimi, 2014).

logging. Detailed procedures for conducting a triaxial test were proposed by Berre (1982). The performance and reliability of triaxial tests are influenced by the testing configuration, typically like the slenderness of the soil specimen and the roughness of the platens (Peri et al., 2019).

2.1.2. Cone penetration test

The CPT has been used to estimate S_u for a few decades. It is cost-effective and convenient compared to laboratory tests because CPTs can be conducted quickly in field and no sampling is needed. Moreover they can offer continuous measurements (i.e., 1D vertical profiles) instead of point measurements. There are several empirical correlations between CPT data and S_u, and in an actual project, the correlation to apply is decided on a case-by-case basis. Rémai (2013) summarized four forms of empirical correlations to derive S_u, which use different empirical cone factors (i.e., the N terms in the following equations, with different subscripts to distinguish different forms). The first form is the most typical, derived from many theoretical studies: $S_u = (q_c - \sigma_{v0})/N_k$, where q_c is the cone resistance, σ_{v0} is the total overburden stress; the second form is similar to the first one but cone resistance is replaced with tip resistance, which is the corrected value by pore pressure: $S_u = (q_t - \sigma_{v0})/N_{kt}$, where q_t is the tip resistance; the third form is to utilize effective cone resistance: $S_u = q_E/N_{ke} = (q_t - u_2)/N_{ke}$, where q_E is the effective cone resistance, u_2 is the measured pore pressure behind the cone tip; the fourth form is developed for soft clays based on excess pore water pressure: $S_u = \Delta u/N_{\Delta u} = (u_2 - u_0)/N_{\Delta u}$, where u_0 is the in-situ pore water pressure.

Based on the aforementioned four forms of empirical correlations, it can be observed that

the accuracy of transformed S_u relies on the determination of cone factors (i.e., N_k, N_{kt}, N_{ke}, and $N_{\Delta u}$), which vary from different sites, soil types, and soil characteristics (i.e., unsaturated or saturated, normally consolidated or overconsolidated). Previous studies have proposed to use site-specific soil parameters such as plastic index, pore water pressure ratio, and over consolidation ratio to determine the cone factor value (or the range of the value), contributing to a more reliable empirical correlation. Otoko et al. (2016) estimated the correlations between S_u and CPT data of soils from the Sombreiro-Warri deltaic plain of Niger Delta, Nigeria. They found the cone factor N_k was proportional to the plastic index and the best N_k was proposed to be 50. Zein (2017) found that N_k seemed to depend on the moisture condition and degree of stiffness based on the Sudanese fine-grained soils, and the value of N_k was found to be in the range of 37.5 to 44.1. Hong et al. (2010) conducted a more comprehensive analysis for N_{kt}, N_{ke} and $N_{\Delta u}$ of Busan clay, and it was found that N_{kt} and N_{ke} were reversely proportional to the pore water pressure ratio and $N_{\Delta u}$ increased as the plasticity index increased. The values of N_{kt}, N_{ke}, and $N_{\Delta u}$ were respectively in the range of 7-20, 3-18, and 4-9. It is evident that there is always a relatively extensive range for those cone factors. Using empirical correlations to derive S_u exists a large amount of uncertainty.

2.1.3. Transformation models

In geotechnical engineering projects, there is a common condition that it is not always feasible to measure a relevant target parameter (e.g., S_u) directly, so estimates of such a parameter will have to be made from other available data, such as the results of laboratory tests and in-situ tests (Kulhawy & Mayne, 1990). So a model used to transform the test measurements to an appropriate design parameter (target parameter) is called a transformation model (Phoon & Kulhawy, 2001). The development of such a model is based on empirical or semi-empirical relations between the target data and other data obtained by regression fitting to a calibration dataset. Typically for S_u, several transformation models have been developed in the last several decades (Hansbo, 1957; JAMIOLKOWSKI et al., 1985; Karlsrud & Hernandez-Martinez, 2013; Ladd & Foott, 1974; Mesri, 1975). Of course, the aforementioned empirical correlations between CPT data and S_u are also transformation models. However, such models have some severe limitations. The first limitation is that when applied to a certain site, the soil properties, soil behaviors, and site geology of this site may differ from the data source from where the transformation models are calibrated, directly leading to bias with respect to the actual S_u (D'Ignazio et al., 2016). One improvement scheme is to calibrate the models with a greater number of data to make it "global" (Ching & Phoon, 2012, 2014a; D'Ignazio et al., 2016). Benefitting from regression fitting to a larger dataset covering several different sites and soil types, the global transformation models can be less biased. However, there is a consequent drawback: the global transformation models generate a significant amount of uncertainty when applied to a specific site as it accommodates a wide range of soil types and site conditions. The second limitation is that the transformation uncertainty remains overlooked instead of being able to be quantified. Experienced engineers may estimate error bounds based on their expertise but this is not a general situation. So it is preferable for a transformation model to quantify the uncertainty (e.g., 95% confidence interval), which can augment existing physics-based methods and engineer's judgment (Ching & Phoon, 2019a; Phoon, 2018).

2.2. Probabilistic transformation models

Considering the limitations in conventional methods, PTMs have been widely adopted to characterize S_u, or any other soil parameter in a broad sense, in the last two decades. PTMs can be constructed in a site-specific (or quasi-site-specific) way, in the absence of big data and quantify the uncertainty. In this section, a thorough review on univariate, multivariate, and non-stationary PTMs is presented.

2.2.1. Univariate probabilistic transformation models

At the early stage, PTMs were univariate (i.e., a single prediction output). Wang and Cao (2013) proposed a PTM based on the Bayesian framework which integrated the prior knowledge (i.e., maps and surveys, local experience, engineering judgment, visual observations, and published reports and studies) and site-specific sparse data to characterize undrained Young's modulus (E_u) using standard penetration test (SPT) N values. The uncertainty of the model is first considered, and two sources are involved: E_u inherent variability and transformation model uncertainty. Then by using the theorem of total probability and Bayes' theorem, the posterior probability distribution function (PDF) of E_u can be derived based on prior knowledge (minimal and maximal values of mean and standard deviation of E_u; assumption on joint uniform distribution of mean and standard deviation of E_u, non-informative) and site-specific data (i.e., SPT-N values). Due to the complexity of the derived PDF of E_u, it is challenging to be analytically or explicitly expressed. So Markov chain Monte Carlo (MCMC) simulation method is adopted to generate samples based on the PDF. After enough E_u samples are collected based on MCMC, conventional statistical methods (e.g., mean and standard deviation) are used to characterize E_u. As these samples are distributed under the PDF, containing equivalent information to the original PDF, this approach is referred to as the equivalent sample approach in this study. The proposed model was applied to probabilistically characterize E_u using only 5 SPT-N values obtained from the clay site of the NGES at Texas A&M University. The cumulative distribution function (CDF) of E_u at this site is shown in Fig. 2.2 to show the quantified uncertainty.

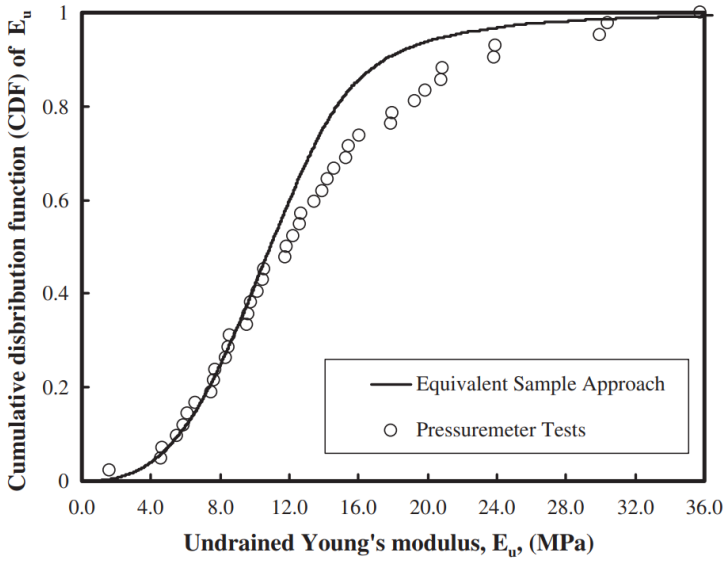


Fig. 2.2. Validation of the probability distribution for undrained Young's modulus estimated from equivalent samples (Wang & Cao, 2013).

Yan et al. (2009) proposed to predict the compression index based on a Bayesian probabilistic approach. Similar to S_u, there are also many transformation models for compression index with uncertain reliability and predictability. The Bayesian probabilistic approach is used for model selection, by which the model with good fitting performance can be found. Notice in this study, the model is also selected by criteria that it should offer an acceptable degree of robustness to measurement noise and modelling error in order to avoid an unnecessarily complicated transformation model. Prior PDFs of model parameters are assumed to be uniform distribution (non-informative) with prior knowledge of minimal and maximal values, and site-specific data (not sparse in this case) are also used.

Univariate PTM studies have also been developed to predict S_u of soils (Cao & Wang, 2014), uniaxial compressive strength of rocks (Ng et al., 2015), and deformation modulus of rocks (Feng & Jimenez, 2015).

2.2.2. Multivariate probabilistic transformation models

The univariate PTMs have been subsequently developed into multivariate probabilistic models as multivariate can exploit the cross-correlations between soil parameters. Wang and Akeju (2016) extended to probabilistically modelling two cross-correlated parameters, effective cohesion (c') and effective friction angle (ϕ'). c' and ϕ' are first assumed to be a bivariate normal distribution. The prior knowledge (i.e., possible ranges of $\mu_{c'}$, $\sigma_{c'}$, $\mu_{\phi'}$, $\sigma_{\phi'}$, ρ and assumption on joint uniform distribution of these parameters; ρ is the correlation coefficient between c' and ϕ') as well as the site-specific data (c' and ϕ') are both used in this study. Then based on the theorem of total probability and Bayes' theorem, the posterior joint PDF of c' and ϕ' can be calculated. Subsequently, MCMC simulation is employed to generate sample pairs of c' and ϕ' based on the joint PDF, and conventional statistical methods are finally applied to characterize these two parameters. The proposed model is tested by different numbers of data pairs (5, 10, 20, 30) measured from alluvial fine-grained soils at the Paglia River alluvial plain in Central Italy and as a result, not only the marginal CDF of c' and ϕ' but also their probabilistic correlation coefficients are obtained.

L. Zhang et al. (2018) utilized the copula approach to model the posterior joint probability distribution of c' and ϕ'. A copula function refers to a function that couples a multivariate distribution to its one-dimensional marginal distribution. Simply speaking, it functions to connect multiple marginal distributions to a multivariate distribution. There are two main steps in this study: 1) identification of the best-fit marginal distribution (the distribution of c' and ϕ' are not deterministically assumed to be normal; four candidate distributions: normal, lognormal, Gumbel and Gamma) and copula (four bivariate copula candidates: Gaussian, Plackett, Frank and No.16); 2) estimation of distribution parameters of the best-fit marginal distributions (i.e., mean and standard deviation) and copula parameter of the best-fit copula (i.e., Kendall rank correlation coefficient tau). In the first step, there are 4 (marginal distributions for c') × 4 (marginal distributions for ϕ') × 4 (copulas) = 64 bivariate distribution models. Based on a Bayesian framework, the occurrence probabilities of each model are calculated and the most probable model is selected. After determining the most probable model, it is adopted to characterize the joint PDF of c' and ϕ' under a Bayesian framework. Similar to Wang and Akeju (2016), MCMC is adopted to sample from the PDF. And finally conventional statistical methods are adopted to analyze the generated samples. The proposed method is tested using 64 data pairs measured from clays in the core wall of Xiaolangdi rockfill dam in China and as a result, the

marginal distribution of c′ and ϕ' are obtained and their correlation coefficient can be calculated probabilistically.

Ching and Phoon (2019a) constructed a site-specific multivariate PTM based on a novel Bayesian framework. Unlike the normal Bayesian framework shown above, which calculates the multivariate PDF integrally (i.e., all components in the PDF are calculated as a combination), the novel Bayesian framework separately calculates each parameter of the multivariate distribution (in this study, mean vector, covariance matrix, hyperparameter, and multivariate). Each parameter is conditioned on the rest of them and after being sampled by Gibbs sampler, a special instance of MCMC, based on its conditional PDF, it is updated and becomes the condition for the rest of the parameters. Such a process goes through all the parameters and iterates until enough samples are collected to carry on traditional statistical analysis. Based on such a framework, the model makes the exact sampling more efficient. More importantly, this model outperforms other multivariate models in that it can accommodate incomplete site-specific data and it can switch to depending on site-specific data when they are "abundant" or depending on generic databases when site-specific data are extremely sparse. For the former advantage, to elaborate incomplete data better, an exemplary input dataset for this model is shown in Table 2.1. It can be evidently seen that some soil parameters are not available at some depths (denoted

Table 2.1. Site investigation for a silty clay layer at a Taipei, Taiwan, site (Ou & Liao, 1987).

Depth (m)	S_u (kPa)		S_u(mob) (kPa)	Test results (training data)						
				LL (Y_1)	PI (Y_2)	LI (Y_3)	σ'_v/P_a (Y_4)	σ'_p/P_a (Y_5)	$S_u(mob)/\sigma'_v$ (Y_6)	q_{t1} (Y_9)
12.8	UU	55.2	46.9	30.1	9.1	1.2	1.26	1.71	0.37	5.17
14.8	VST	50.7	52.9	32.8	12.8	1.43	1.43	N/A	0.36	4.22
16.1	UU	61.9	51.7	36.4	14.5	1.24	1.54	N/A	0.33	4.12
17.8	UU	54.2	42.8	41.9	18.9	0.9	1.68	1.79	0.25	4.03
18.3	VST	59.5	59.3	N/A	N/A	N/A	1.72	N/A	0.34	5.27
20.2	UU	73.1	60.5	38.1	17.3	0.7	1.88	N/A	0.32	4.53
22.7	VST	63.3	64.4	37	16	0.58	2.08	N/A	0.31	4.76
24	UU	82.2	67.5	38	16.2	0.75	2.19	2.19	0.3	5.12
26.6	UU	98.1	82.1	34.8	13.8	0.8	2.41	N/A	0.34	5.32

as N/A). Incomplete geotechnical data are always met in actual projects, so this ability is quite attractive. For the latter advantage, in case the site-specific data are too sparse leading to significant statistical uncertainty, a scheme to hybridize site-specific data and generic data, which can offer prior knowledge is further proposed in the study. The hybridization is straightforward as the hybrid multivariate PDF is proportional to the product between generic and site-specific multivariate PDFs as:

$$f(\mathbf{x}_{new}|hb) \propto f(\mathbf{x}_{new}|\mu_g, \mathbf{C}_g) \cdot f(\mathbf{x}_{new}|\mathbf{X}^o) \tag{2.1}$$

where $\mathbf{x}_{new}$ is a vector for multivariate soil parameters at a new (predicted) depth; μ_g and $\mathbf{C}_g$ are the mean vector and covariance matrix (prior knowledge) obtained from generic databases like CLAY/10/7490; $\mathbf{X}^o$ is the site-specific dataset. Fig. 2.3 explains how it works: when $\mathbf{X}^o$ is sparse (a), $f(\mathbf{x}_{new}|hb) \propto f(\mathbf{x}_{new}|\mu_g, \mathbf{C}_g) \times$ (a relatively flat PDF) $\approx f(\mathbf{x}_{new}|\mu_g, \mathbf{C}_g)$; whereas when $\mathbf{X}^o$ is abundant (b), $f(\mathbf{x}_{new}|hb) \propto$ (a relatively flat PDF) $\times f(\mathbf{x}_{new}|\mathbf{X}^o) \approx f(\mathbf{x}_{new}|\mathbf{X}^o)$.

Ching and Phoon (2020) further incorporated the spatial correlation of soil parameters (1D) into the model proposed by Ching and Phoon (2019a). This is to say the model proposed by Ching and Phoon (2020) considers not only the cross-correlation but also spatial correlation, contributing to simulating 1D profiles of soil parameters. More details of this model can be found in Section 3.1. More recently, Ching, Phoon, et al. (2022) extended to 3D multivariate probabilistic characterization based on the work done by Ching and Phoon (2019a, 2020).

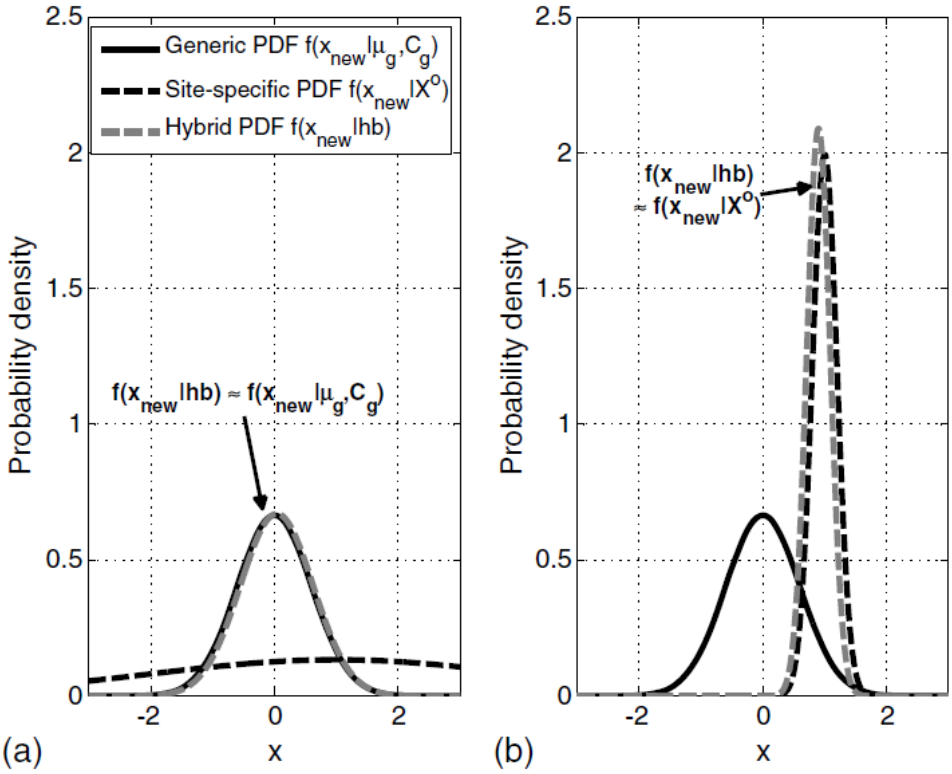


Fig. 2.3. Hybridization effect: (a) sparse site-specific data; and (b) abundant site-specific data (Ching & Phoon, 2019a).

2.2.3. Non-stationary probabilistic transformation models

Wang and Zhao (2017) and Wang and Zhao (2016) offered a different scheme to characterize a target soil parameter. The prior knowledge, which is important in the aforementioned PTMs (reflected by the engineering judgments or generic databases), is not necessarily needed in this scheme. Moreover, it can be seen that in PTMs, the distribution of a soil parameter follows a normal distribution with stationary mean and standard deviation at most times, whereas in such a scheme the distribution of a soil parameter is no longer stationary and Gaussian but non-stationary non-Gaussian. It is based on the Bayesian compressive sampling (BCS) method, and the theory behind is compressive sampling which asserts that a signal can be reconstructed from a few measurements of the signal itself, utilizing the fact that many natural signals are compressible (i.e., they can be represented by a weighted summation of only several pre-specified basis functions). The BCS model is able to provide the profile of the target soil parameter directly from the sparse measurements and quantify the statistical uncertainty. Wang et al. (2018) extended BCS further by combing it with a random field generator, Karhunen–Loève (KL) expansion, to BCS-KL which can generate random field samples. It successfully bypassed estimation on random field parameters such as mean, standard deviation, and correlation functions. Zhao and Wang (2018) utilized BCS-KL to generate cross-correlated random field samples, which considered not only the spatial auto-correlation but also the cross-correlation between two correlated parameters (i.e., two random fields). Hu and Wang (2020) extended the BCS-KL model to 2D and utilized it for soil stratification.

2.2.4. Conclusion

Through the literature review of PTMs, it can be found that in most cases, site-specific sparse geotechnical data are acceptable to characterize soil parameters. In addition, the uncertainty of the simulated results can be explicitly quantified. However, every coin has two sides. The computational efficiency of PTMs is generally low because exact sampling is always needed. This will become a problem when a lot more variables are simulated (e.g., extension of probabilistic methods to higher dimensional site characterization).

2.3. Machine learning methods

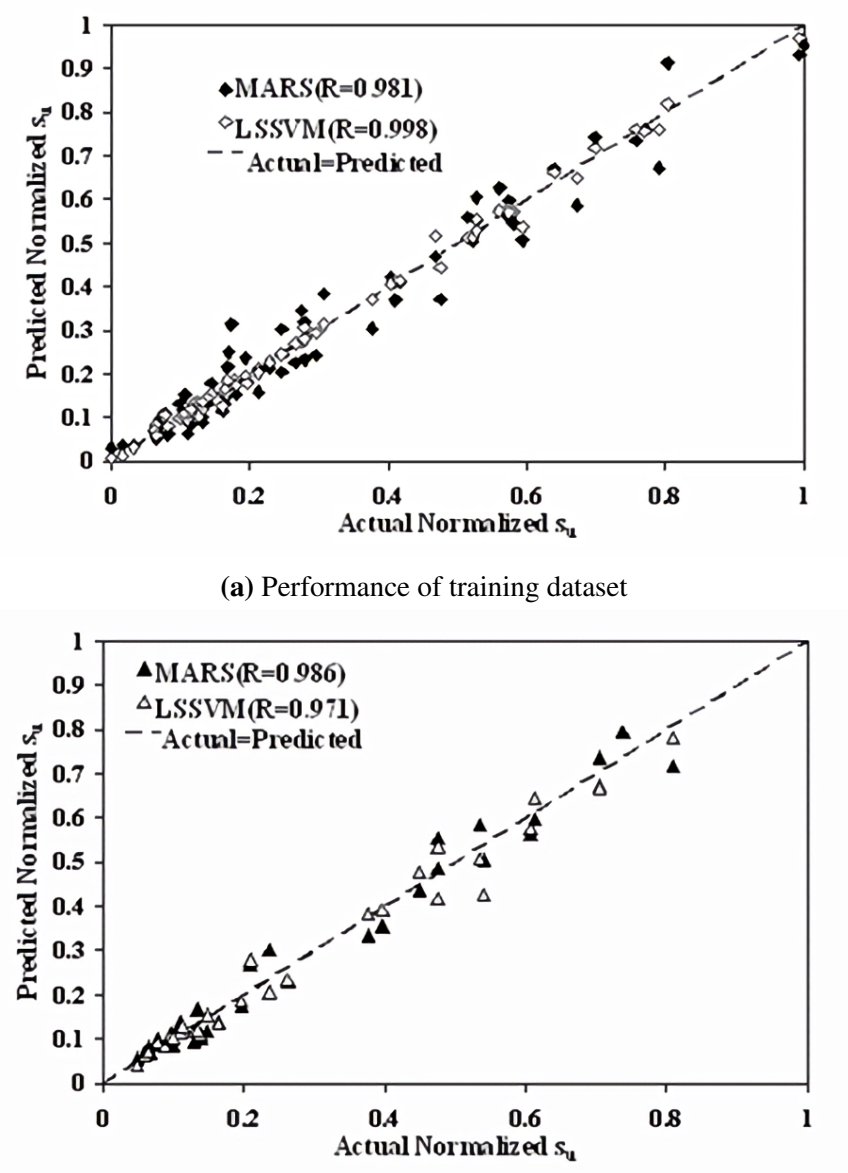
Besides PTMs, ML is another improved method to better predict and characterize soil parameters. It has been increasing popular because it has demonstrated superior predictive ability compared to traditional methods (Shahin, 2013). Simply speaking, a higher accuracy can be expected from ML compared to transformation models. This section summarizes the studies on applying basic and more advanced ML models to soil shear strength predictions.

2.3.1. Basic ML methods

Basic ML technics were utilized in geotechnical engineering initially, such as artificial neural network (ANN), support vector machine (SVM), classification and regression tree (CART), and multivariate adaptive regression spline (MARS) (Das et al., 2011; Kanungo et al., 2014; Ly & Pham, 2020; Samui & Kurup, 2012). At this stage, the procedures were relatively simple and direct: selecting learning sets including possible soil parameters as input and shear strength as output and splitting them into training datasets, validation datasets, and testing datasets; selecting a single machine learning model; learning and evaluating the accuracy. To give some examples, Samui and Kurup (2012) used corrected cone resistance, vertical total stress, hydrostatic pore pressure, pore pressure at the cone tip, pore pressure just above the cone base as the input and S_u as the output, which were measured from massive clay samples in several countries around the world through laboratory tests and in-situ tests. MARS and least square SVM were respectively utilized. The coefficient of correlation (R) was used to evaluate the accuracy of those two machine learning models. The results obtained in this study are shown in Fig. 2.4. Similarly, Ly and Pham (2020) used moisture content, clay content, void ratio, plastic limit, liquid limit, and specific gravity as input and shear strength as output, which were measured from 538 soil samples collected from Long Phu 1 power plant project, Vietnam through laboratory tests. Then a single machine learning model, SVM, was utilized. Finally, various statistical indicators, including R, root mean squared error (RMSE), and mean absolute error (MAE), were used to evaluate the performance of the model. The results obtained in this study are shown in Table 2.2.

2.3.2. More advanced ML methods

With the development of computer science and artificial intelligence, more advanced machine learning schemes are applied to the prediction of soil shear strength in order to improve accuracy. The first scheme is to hybridize the model with optimization algorithms (Pham et al.,



(a) Performance of training dataset

(b) Performance of testing dataset

Fig. 2.4. Performance of MARS and LSSVM for (a) training dataset; and (b) testing dataset (Samui & Kurup, 2012).

Table 2.2. Summary of the prediction performance of SVM over 200 random simulations for the training and testing datasets (Ly & Pham, 2020).

Part	Values	RMSE	MAE	R	Error Std
Train dataset	Average	0.0988	0.058	0.8824	0.0989
-	Min	0.0707	0.0494	0.5088	0.0708
-	Max	0.305	0.0764	0.9399	0.3051
Test dataset	Average	0.082	0.0555	0.9164	0.0818
-	Min	0.0616	0.0451	0.722	0.0615
-	Max	0.1788	0.079	0.9537	0.178

2020; Pham et al., 2018). In machine learning models, there are hyper-parameters which should be set before the learning process. Those hyper-parameters have a significant influence on the performance of the predictive model. Therefore, combining optimization algorithms, which can tune and optimize the hyper-parameters, is essential. For example, Pham et al. (2018) respectively utilized particle swarm optimization - adaptive network based Fuzzy inference system (PANFIS), genetic algorithm - adaptive network based Fuzzy inference system (GANFIS), support vector regression (SVR), and ANN to predict soil shear strength with plastic index, liquid limit, moisture content, and clay content as input, which were measured from 188 plastic clay soil samples collected in Nhat Tan and Cua Dai bridges, Vietnam. The former two models were combined with two meta-heuristic optimization algorithms: particle swarm optimization (PSO) and genetic algorithm (GA). The performances were evaluated by RMSE and R. As a result, PANFIS showed the highest prediction accuracy (RMSE = 0.038 and R = 0.601), then GANFIS (RMSE = 0.04 and R = 0.569), followed by two baseline models without optimization algorithms, SVR (RMSE = 0.044 and R = 0.549) and ANN (RMSE = 0.047 and R = 0.49). The benefits from hybridizing the optimization algorithms can be demonstrated.

The second scheme is to ensemble multiple machine learning algorithms (Mbarak et al., 2020; W. Zhang et al., 2021). Given the fact that the datasets used to train ML models usually contain soil parameters measured from soil samples collected from different sites, with different types and characteristics, ensemble methods for ML in geotechnical engineering are necessary since the ensemble learning can process different hypotheses to form a better hypothesis, contributing to a better prediction performance (Nascimento et al., 2014). Ensemble methods can be broadly categorized into two methods according to their structures (W. Zhang et al., 2021): bagging and boosting. The bagging ensemble method combines different algorithms independently. This is to say those learning algorithms do not interact with each other, and their results are combined to make the final result. The boosting ensemble method builds up a learning algorithm based on the predecessor algorithm to reduce the error from the predecessor algorithm. This means that different learning algorithms in the boosting ensemble framework interact with each other. W. Zhang et al. (2021) utilized two ensemble learning methods: random forest (RF) and extreme gradient boosting (XGBoost), respectively corresponding to the bagging and boosting category of ensemble methods to predict S_u. Three baseline machine learning models (SVM, MARS, and multilayer perceptron) were adopted to compare with the former two. Preconsolidation stress, vertical effective stress, liquid limit, plastic limit, and natural water content were used as input. The learning data were from a TC304 database. RMSE, coefficient of determination (R^2), bias factor (b), and mean absolute percentage error (MAPE) were employed to evaluate the performance of those models. The results in Table 2.3 indicated XGBoost and RF methods outperformed the other three baseline models.

Table 2.3. Comparisons among model predictive modeling results (W. Zhang et al., 2021).

Evaluation index	RMSE (kPa)		R^2		MAPE (%)		Bias	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
XGBoost	2.38	4.4	0.92	0.73	10.85	19.23	0.99	1.01
RF	2.51	4.6	0.91	0.7	10.93	19.63	0.99	1
SVR	3.6	4.82	0.82	0.67	13.69	20.56	1.02	1.04
MLPR	4.19	4.91	0.75	0.66	19.41	21.38	1	1.02
MARS	4.42	4.89	0.73	0.66	20.23	22.43	1.01	1.04

2.3.3. Conclusion

The ML models reviewed in this section exhibit precise predictive outcomes but they require relatively large datasets for training purposes. It can be seen that usually hundreds of data pairs (input-output) are needed. While meeting this requirement may be possible in some real projects only in terms of volume, geotechnical data possess another "ugly" property that they are incomplete, which will make it pretty challenging to obtain a satisfactory dataset from a project. For example, a ML model is constructed to predict soil parameter A based on (B, C, D). In principle (A, B, C, D) data at the same depth and reasonably close locations are needed. However, it is very rare that such complete multivariate data points are available during a common site investigation program (Ching & Phoon, 2019a). This problem may impede the practical utilization of ML in geotechnical engineering. In addition, even if an ML model is trained well based on a generic dataset containing relatively abundant geotechnical data (e.g., those offered by TC304 ISSMGE), when applied to a specific site, the accuracy will undoubtedly decrease (Yu, 2022).

2.4. Fusion of geotechnical and geophysical data

Some progress has been made recently to take advantage of the fusion of geotechnical data and geophysical data in site characterization, as these two data sources are complementary and highly likely to correlate with each other. Complementary refers to that geotechnical data are accurate but sparse while geophysical data are inaccurate (back analysis is needed) but abundant. Even though it is still at an early stage (Phoon et al., 2022), such a scheme seems promising to extend site characterization to multidimensional space and is able to provide more realistic and reliable ground models (Xie et al., 2022). This section presents different fusion methods based on ML and PTMs.

Zuada Coelho and Karaoulis (2022) utilized the fusion of multi-source data (geotechnical data, geophysical data as well as geological data) in order to make 3D subsoil schematizations based on machine learning methods (neural network and RF). The input is the position (coordinates), geophysical data (electrical resistivity), and geological data (geological formation entity), and the output is the soil behavior type index (IC), which is estimated by CPTs and can be used to classify soils (Robertson, 2010). In this case, RF outperforms neural network and it is applied to perform the regional subsoil schematization of a part of Zeeland, Netherlands, which can be seen in Fig. 2.5.

Xie et al. (2022) combined the shear wave velocity (V_s, geophysical data) obtained from multichannel analysis of surface waves (MASW) and cone resistance (q_c) obtained from CPTs to develop a 2D synthetic field for q_c. Compared to the method proposed by Huang et al. (2018), which also intended to combine V_s and q_c, the method by Xie et al. (2022) requires less computational effort and the empirical transformation models between those two data are not needed. There are two aspects of the development of the 2D field assisted by the geophysical data in this study: 1) the estimation of the horizontal scale of fluctuation (SOF); 2) the estimation of q_c at unobserved locations. For the first aspect, the horizontal SOF, which is difficult to estimate from sparse CPT data, is estimated from V_s 2D measurements. This horizontal SOF is directly used as the horizontal SOF of q_c as they are highly correlated. For the second aspect, the intrinsic collocated co-kriging (ICCK) method is adopted to estimate q_c in the 2D field by

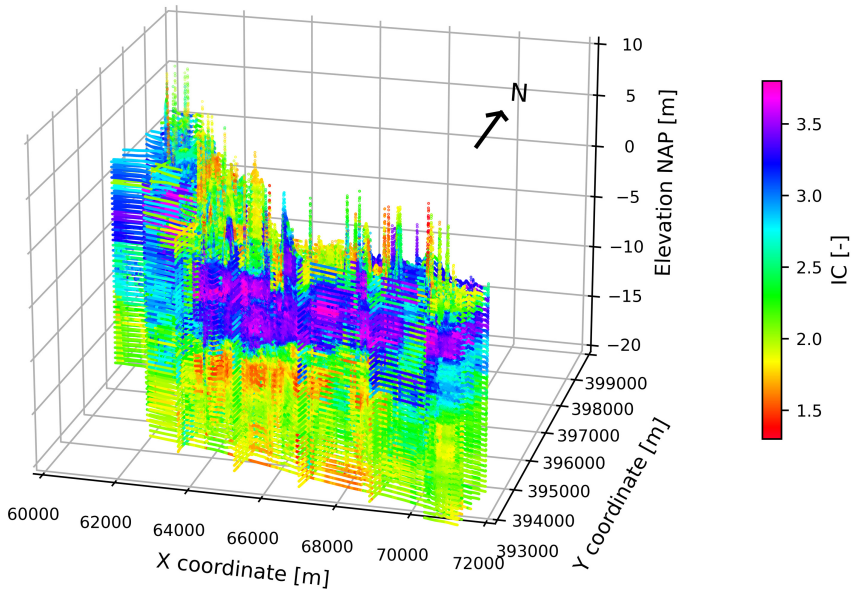


Fig. 2.5. Predicted IC values for Zeeland following the data fusion method based on RF (Zuada Coelho & Karaoulis, 2022).

linearly combining observed V_s and q_c, which avoids the tedious modelling of cross-covariance relations between these two sources of data compared to traditional cokriging. The proposed method is applied to a case study in the Christchurch area, New Zealand, where there are 4 observed q_c profiles and a 2D measurement of V_s, and cross validation is adopted to assess the accuracy of the estimated result. Cross validation in this case refers to removing one observed q_c profile to estimate and calculating the mean square error (MSE) between this observed profile and the estimated profile. By comparing the cross validation result of ICCK (MSE = 4.17) with that of ordinary kriging (OK) (MSE = 5.82), which only uses q_c itself to estimate the 2D field, the improvement of combining the geophysical data is demonstrated. The comparison can be viewed in Fig. 2.6. More details of ICCK can be found in Section 3.2.

Xu et al. (2022) proposed a multi-source BCS (MS-BCS) method to combine the geotechnical data with Vs obtained from MASW to characterize the site in a 2D field. The method does not require an empirical equation between these two types of data nor the prior information on the geotechnical data. In the original 2D BCS method, which does not consider geophysical data, the determination of mean vector and covariance matrix for the weights and selection of nontrivial basis functions can be achieved by estimating a hyperparameter α_t based on the Bayesian framework (weights and basis functions are two significant components in BCS). However, if the site-specific measurements are too sparse, it is hard to achieve this process. So geophysical data are incorporated in this study and specifically contribute to two aspects: 1) offering the nontrivial basis functions (the final nontrivial functions are the union of nontrivial functions from geotechnical data and geophysical data); 2) helping to estimate the hyperparameter α_t. The effect of combining geophysical and geotechnical data is demonstrated by comparing the simulation result from MS-BCS and BCS in one numerical example. This example provides a synthetic S_u and V_s 2D domain. Sparse S_u data points are selected from the 2D domain as measurements. The comparison is shown in Fig. 2.7. It can be seen that MS-BCS,

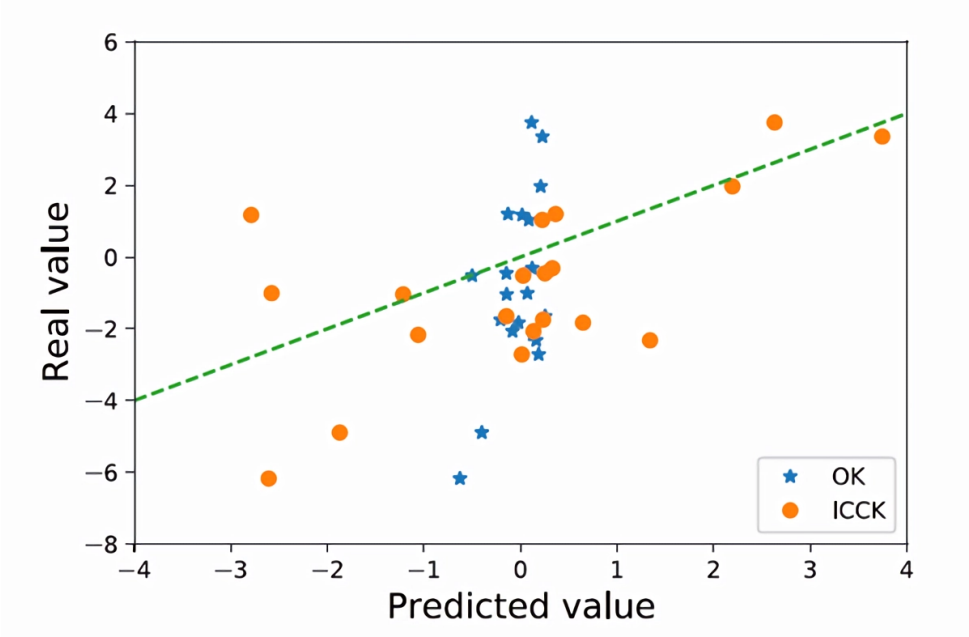


Fig. 2.6. Cross validation result comparison between the ICCK and OK model (Xie et al., 2022).

which combines geophysical data, performs better to reproduce the original S_u 2D domain.

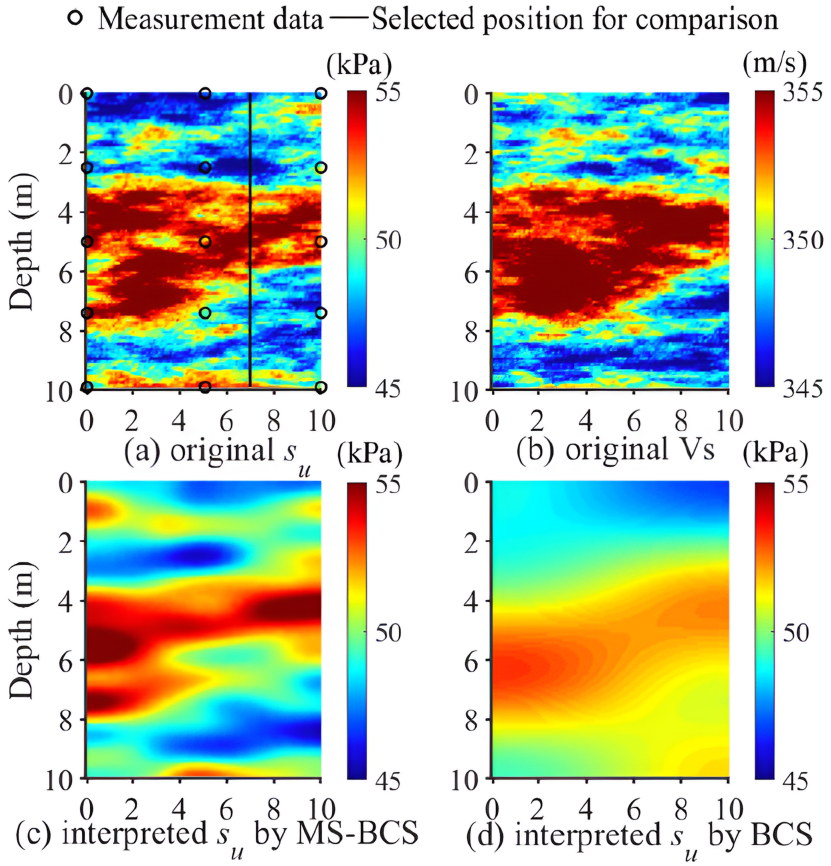


Fig. 2.7. Comparison between the original and interpreted S_u from MS-BCS and BCS (Xu et al., 2022).

3. Methodology

This chapter starts with elucidating MUSIC-X (Ching & Phoon, 2020), the multivariate PTM adopted in this study to preliminarily characterize S_u (i.e., simulating 1D S_u profiles). The theoretical basis of this model is introduced, followed by practical procedures to construct it and finally, modifications on the model to make it fit with this study are shown. Subsequently, the intrinsic collocated cokriging (ICCK) model (Babak & Deutsch, 2009b) is presented, by which the simulated S_u profiles (primary data) and auxiliary 2D measurements of V_{int} (secondary data) are combined linearly with different kriging weights to estimate a 2D cross section of S_u. The weights are determined by the spatial (cross) correlation of these two parameters, which is specifically elaborated.

3.1. MUSIC-X

3.1.1. Theoretical basis

Most PTMs, including MUSIC-X, are based on a Bayesian framework. The Bayesian theorem describes a generalized concept to update probabilities of prior events when given new evidence. When applying a Bayesian framework to the soil parameter characterization in geotechnical engineering, it is a powerful tool to improve the prior probability distribution of the characterization parameters for a soil parameter, based on the new measurements from the geotechnical investigation at this specific site, to the posterior probability distribution of the characterization parameters, which can more accurately characterize the soil parameter at this site. To better understand, the mathematical formula is explicitly expressed as:

$$P(C \mid Data) = \frac{P(Data \mid C)P(C)}{P(Data)} \tag{3.1}$$

$P(C \mid Data)$ is the posterior/conditional probability distribution function (PDF) of the characterization parameters (including the mean μ and standard deviation σ) of a soil parameter (X) given the observed data. The characterization parameters can be used to characterize the distribution of X (i.e., to calculate $P(X \mid C, Data)$, or simply denoted as $P(X \mid C)$). $P(Data \mid C)$ is the likelihood function to reflect how the characterization parameters fit with the observed data. $P(C)$ is the prior PDF of the characterization parameters. $P(Data)$ is the prior PDF of producing the observed data. The likelihood function can be determined once the distribution type of Data are determined (e.g., normal distribution). $P(Data)$ is hard to calculate on its own. Instead it is regarded as a normalizing constant which makes the total probability (i.e., the integral of the area under $P(C \mid Data)$) equal to one. The prior distribution information is from engineers' judgement, available site characterization materials from nearby projects, and generic geotechnical databases which contain the information about a soil parameter from site investigations

all over the world. Usually, the derived posterior distribution $P(X \mid C)$ is too complicated to be analytically solved especially when it upgrades to multivariate. Therefore, some statistical methods such as Markov chain Monte Carlo (MCMC) are applied to sampling from the posterior PDF. These samples are distributed following the PDF and can be applied to reflecting the uncertainty and probabilistic characterization.

The MUSIC-X model follows the basic Bayesian theorem but utilizes it in a novel way. Firstly it is a multivariate model. As mentioned earlier, an ideal PTM can make multivariate inputs and outputs, which is basically achieved in this model. So in this case, X initially becomes $\mathbf{x} = (X_1, X_2, \ldots, X_n$, n is the number of different geotechnical parameters used in the model), Data contains observed data for n geotechnical parameters, and the characterization parameters become the mean vector (μ_s) and covariance matrix ($\mathbf{C}_s$). The subscript s means the characterization parameters are derived in a site-specific way. Moreover, due to the fact that the size of a typical database in geotechnical engineering is small, a multivariate PDF (i.e., $P(\mathbf{x} \mid C)$) cannot be constructed based on such limited information unless it is a multivariate normal PDF. Therefore the core assumption of the MUSIC-X model is that $P(\mathbf{x} \mid C)$ is multivariate normal.

Secondly, the vertical spatial correlation is integrated in the model. $\mathbf{x}$ further becomes $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, m is the number of depths involved in the model) and Data changes accordingly. $\mathbf{X}$ and Data now can be imaged as an excel sheet with n columns (multivariate geotechnical parameters) and m rows (depths). The Data sheet definitely involves NaN cells which means that a certain parameter is not observed at this depth because the geotechnical measurements are always sparse and incomplete. The way to deal with this incompleteness can be found in Section 3.1.2. Moreover, as for how to integrate the spatial correlation, the characterization parameter $\mathbf{C}_s$ becomes $\mathbf{C}_s \otimes \mathbf{R}$, where $\mathbf{R}$ is the autocorrelation matrix which can inform the spatial correlation; $\otimes$ is the Kronecker product. $\mathbf{R}$ can be calculated by autocorrelation functions (ACFs) and Whittle-Matern ACF is specifically adopted in this case. Based on such an integration, 1D profiles of a target soil parameter can be directly simulated by taking the corresponding column in the $\mathbf{X}$ samples drawn from $P(\mathbf{X} \mid C)$ and this simulation utilizes not only the cross correlation (from $\mathbf{C}_s$), but also the spatial correlation (from $\mathbf{R}$), contributing to a high accuracy simulation. From X to $\mathbf{x}$ to $\mathbf{X}$, the development of the MUSIC-X model can be learned.

Thirdly, it takes the non-informative priors (i.e., $P(C)$ is non-informative). The non-informative setting in the Bayesian model always requires a lager computational power but it avoids prior preference and bias and it is more friendly to non-professionals. Typically people can adopt uniform distributions with a specific range to make the prior distribution non-informative (flat). However, the uniform prior distribution is not conjugate to the multivariate normal model, leading to inefficiency in the later exact sampling. The conjugacy means the prior distribution and posterior distribution (i.e., $P(C)$ and $P(C \mid Data)$) should have a same type. Considering this problem, the MUSIC-X model takes conjugate prior distributions of μ_s and $\mathbf{C}_s$ to the multivariate normal model and tune their parameters to make them non-informative. How to find the conjugate prior distributions, how to demonstrate the conjugacy and how to tune to be non-informative are illustrated by Ching and Phoon (2019a).

Finally, the Gibbs sampler (GS), which is a special instance of MCMC is used to sample from the PDFs. GS can decompose random parameters into groups and sequentially draw samples for each group from its full conditional PDF (i.e., posterior distribution). Such a PDF is conditioned on the remaining groups. Taking the characterization parameters (μ_s and $\mathbf{C}_s$) for example, these two parameters are not sampled from $P(C \mid Data)$ integrally. Instead, it draws a μ_s sample from $P(\mu_s \mid Data, \mathbf{C}_s)$ first. Then it draws a $\mathbf{C}_s$ sample from $P(\mathbf{C}_s \mid Data, \mu_s)$. The μ_s

in $P(\mathbf{C}_s \mid Data, \mu_s)$ is the μ_s sample drawn in the first step. The conditional PDF of μ_s will be updated in the same manner. This process will continue for several times and the samples are asymptotically distributed as the desired distribution. Compared to MCMC simulation, GS has a higher efficiency. The detailed GS scheme can be seen in Section 3.1.2.

3.1.2. Practical construction

According to the theory mentioned above, Fig. 3.1 presents the procedures to construct the MUSIC-X model in practice. Each step will be elaborated afterwards and the step with * is associated with an additional flow chart.

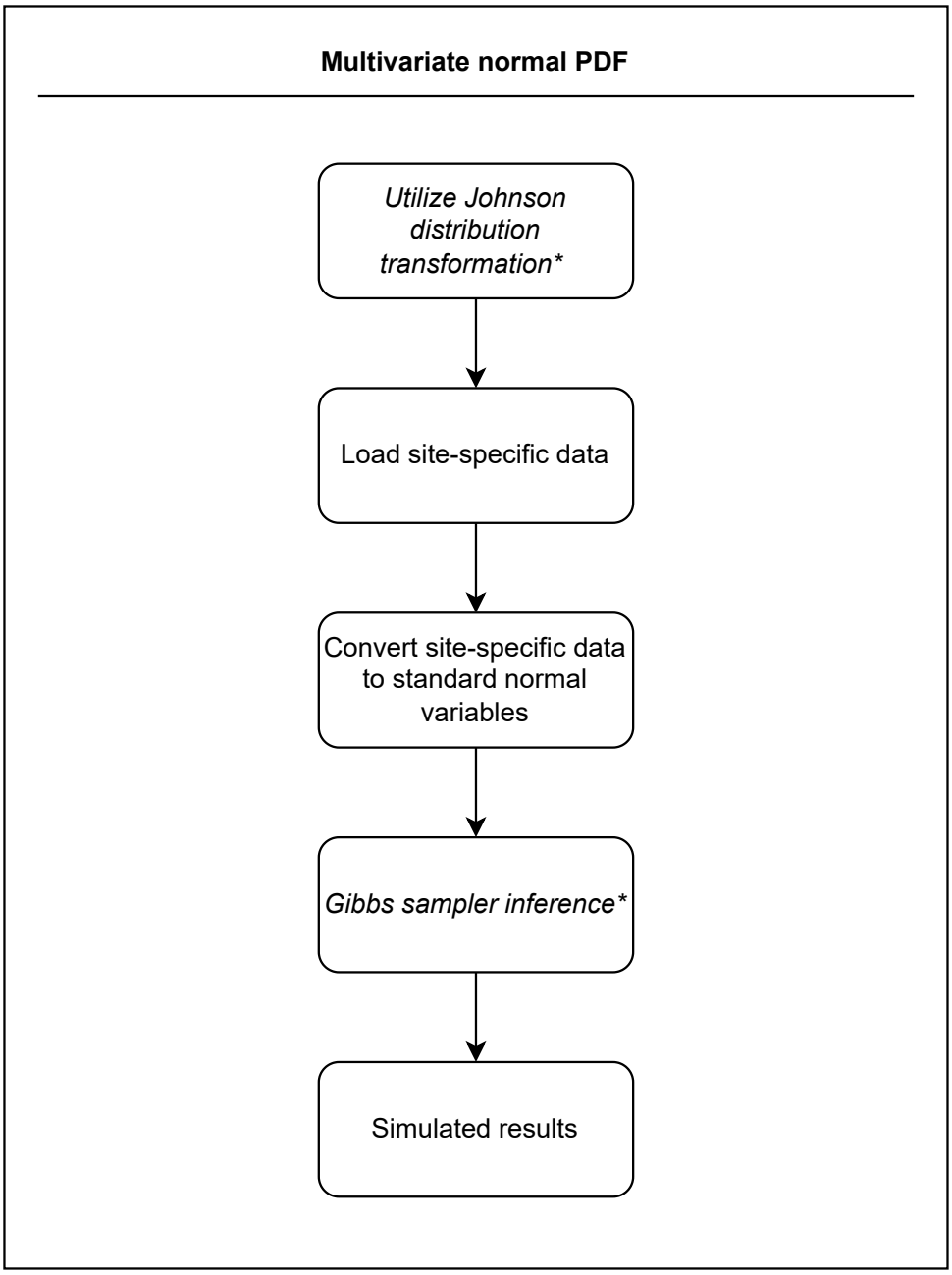


Fig. 3.1. Flow chart of constructing the MUSIC-X model.

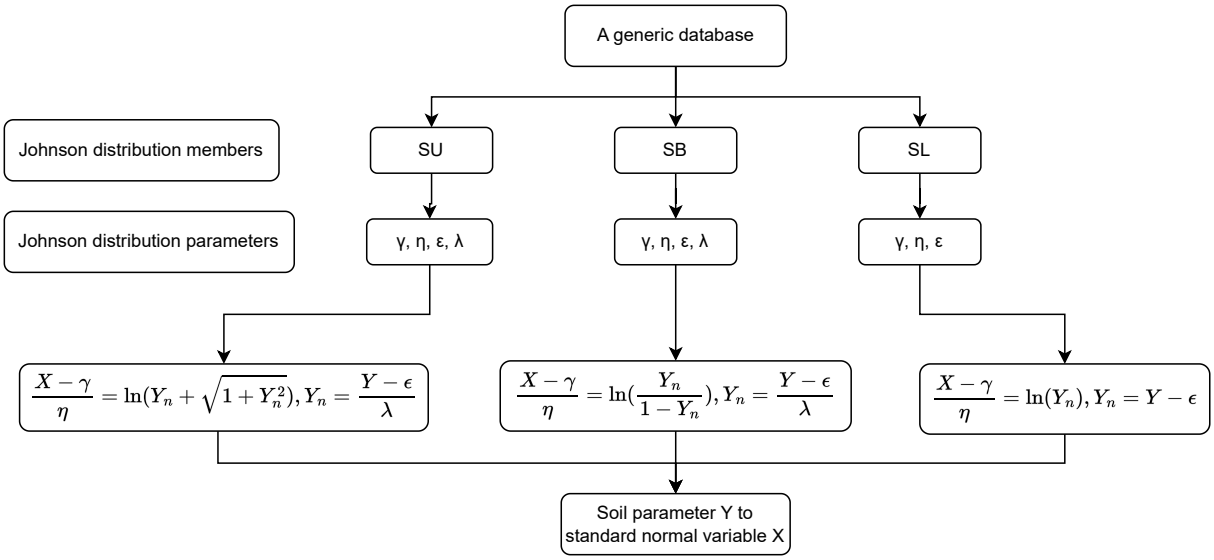


Fig. 3.2. Flow chart of JD transformation.

First of all, the essence of the MUSIC-X model is a multivariate normal PDF. In order to construct the PDF, Data (i.e., site-specific X_1, X_2, ..., X_n) should be multivariate normal and the marginal distribution of a single parameter (e.g., the distribution of X_1) must follow a standard normal distribution. However, a soil parameter is typically non-normal, which requires to be converted into a standard normal random variable at first. The Johnson distribution (JD) is adopted. JD is a family of probability distribution that can be used to transform a non-normal variable into a standard normal variable. So the first step in the MUSIC-X model is to utilize JD transformation to convert non-normal soil parameters into marginally standard normal variables. How to implement JD transformation is shown in Fig. 3.2. Firstly, a generic database is needed to find which member a certain soil parameter belongs to. There are three members (or types) in JD, respectively SU (unbounded system), SB (bounded system) and SL (lognormal system). Basically some percentiles of the values of this parameter are calculated based on the generic database and can be used to estimate which member it should be categorized to. After determining the type of this soil parameter, the JD parameters (for SU and SB, γ, η, ε and λ; for SL, only the former three), can be determined also based on the generic database. For different JD members, the mathematical formulas to calculate JD parameters are different. The way to find which JD member a soil parameter belongs to and the formulas to calculate JD parameters for different JD members are explicitly shown in Ching and Phoon (2014a). Once the JD parameters are determined, the mathematical function of JD can be applied to convert an non-normal soil parameter Y_i to a standard normal variable X_i. These transformation functions are all in closed form so they are easy to use. In the MUSIC-X model, CLAY/10/7490 database is used to derive the JD transformation for 11 geotechnical parameters[1] (i.e., the MUSIC-X

[1]In Ching and Phoon (2020), the MUSIC-X model involves 10 geotechnical parameters. It is updated by Prof. Ching <jyching@gmail.com> and Prof. Phoon <kkphoon@sutd.edu.sg>.

model is 11 dimensional):

$$
\begin{aligned}
Y_1 &= \ln(\mathrm{LL}) \\
Y_2 &= \ln(\mathrm{PI}) \\
Y_3 &= \mathrm{LI} \\
Y_4 &= \ln\left(\sigma'_v/P_a\right) \\
Y_5 &= \ln\left(\sigma'_p/P_a\right) \\
Y_6 &= \ln\left(S_u/\sigma'_v\right) \\
Y_7 &= B_q \\
Y_8 &= \ln\left[\left(q_t-\sigma_v\right)/\sigma'_v\right] = \ln\left(q_{t1}\right) \\
Y_9 &= \ln(\mathrm{C_c}) \\
Y_{10} &= \ln(\mathrm{C_s}) \\
Y_{11} &= \ln\left(N_{60}/(\sigma'_v/P_a)\right)
\end{aligned}
\tag{3.2}
$$

where LL = liquid limit; PI = plasticity index; LI = liquidity index; σ'_{v} = vertical effective stress; σ'_{p} = preconsolidation stress; $\mathrm{S_u}$ = mobilized undrained shear strength; $\mathrm{B_q}$ = pore pressure ratio; $\mathrm{q_t}$ = (corrected) cone tip resistance; $\mathrm{C_c}$ = compression index; $\mathrm{C_s}$ = swelling index; $\mathrm{N_{60}}$ = SPT N values; $\mathrm{P_a}$ = atmospheric pressure. Note except LI and $\mathrm{B_q}$, which can be potentially negative, the rest of them are logarithm transformed. Their corresponding JD transformation functions are expressed as:

$$
\begin{aligned}
\frac{X_1+2.647}{3.684} &= \sinh^{-1}\left(\frac{Y_1-3.002}{1.259}\right) \\
\frac{X_2-0.117}{2.128} &= \sinh^{-1}\left(\frac{Y_2-3.399}{1.245}\right) \\
\frac{X_3+0.817}{1.539} &= \sinh^{-1}\left(\frac{Y_3-0.293}{0.819}\right) \\
\frac{X_4-0.396}{2.530} &= \sinh^{-1}\left(\frac{Y_4-0.380}{2.294}\right) \\
\frac{X_5+1.054}{3.197} &= \sinh^{-1}\left(\frac{Y_5+0.531}{3.011}\right) \\
\frac{X_6+0.833}{1.732} &= \sinh^{-1}\left(\frac{Y_6+1.730}{1.046}\right) \\
\frac{X_7-62.681}{4.381} &= \ln\left(\frac{Y_7+0.791}{2063399.889-Y_7}\right) \\
\frac{X_8+1.143}{1.591} &= \sinh^{-1}\left(\frac{Y_8-1.215}{0.729}\right) \\
\frac{X_9-0.082}{4.041} &= \sinh^{-1}\left(\frac{Y_9+0.756}{3.720}\right) \\
\frac{X_{10}-0.170}{2.107} &= \sinh^{-1}\left(\frac{Y_{10}+2.710}{1.525}\right) \\
\frac{X_{11}+0.270}{1.776} &= \ln\left(\frac{Y_{11}+2.311}{6.082-Y_{11}}\right)
\end{aligned}
\tag{3.3}
$$

where $\sinh^{-1}(x) = \ln\left[x+\left(1+x^2\right)^{0.5}\right]$ is the inverse hyperbolic sine function.

Returning to Fig. 3.1, after figuring out how to transform soil parameters $\mathrm{Y_i}$ to standard

normal variables X_i, site-specific data are loaded and transformed based on JD functions. Subsequently μ_s and $\mathbf{C}_s$ should be derived from the transformed data to construct the multivariate normal PDF. Due to the sparsity of the site-specific data, direct inference of μ_s and $\mathbf{C}_s$ contains significant amount of statistical uncertainty. So the Bayesian inference based on GS is adopted to derive μ_s and $\mathbf{C}_s$. In this case, the random parameters are (μ_s, $\mathbf{C}_s$, $\mathbf{a}$), which will be decomposed and sampled sequentially by GS. The former two are the characterization parameters while $\mathbf{a}$ is a hyperparameter to make the conjugate prior distribution of $\mathbf{C}_s$ non-informative. It will also be continuously updated in the sampling process as $\mathbf{C}_s$ keeps updating. Since it is associated with $\mathbf{C}_s$, it can be regarded a characterization parameter as well. After completing the sampling process once, the multivariate normal PDF for $\mathbf{X}$ can be sampled subsequently. So it attaches to the last step of GS inference and actually the random parameters are (μ_s, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}$). Moreover, $\mathbf{X}$ actually consists of $\mathbf{X}^u$ (unobserved data) and $\mathbf{X}^o$ (observed data, i.e., Data in Section 3.1.1). $\mathbf{X}^u$ is extracted out of $\mathbf{X}$ because $\mathbf{X}^o$ is not required to simulate. Finally the random parameters in the GS inference are (μ_s, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$). The exemplary table shown in Table 2.1 can give an intuitional impression of $\mathbf{X}$, $\mathbf{X}^o$ and $\mathbf{X}^u$. The whole table refers to $\mathbf{X}$. $\mathbf{X}^u$ is NaN cells while $\mathbf{X}^o$ is cells with values. As for the way to derive the conditional PDFs for μ_s, $\mathbf{C}_s$ and $\mathbf{a}$, they have the same distribution formats as their prior distribution formats due to the conjugacy. For example, the conjugate prior distribution for μ_s (i.e., $P(\mu_s)$) is multivariate normal thus its conditional distribution $P(\mu_s \mid \mathbf{C}_s, \mathbf{a}, \mathbf{X})$ is also multivariate normal. For the conditional PDF of $\mathbf{X}^u$, it is multivariate normal as mentioned so it can be also derived. Moreover, it is worth noting that the auto-correlation matrix $\mathbf{R}$ (shown in Eq. (3.4)) is also integrated in the conditional PDFs to reflect the vertical spatial correlation.

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1m} \\ & 1 & \rho_{23} & \cdots & \rho_{2m} \\ & & 1 & & \\ & & & \ddots & \vdots \\ SYM. & & & & 1 \end{bmatrix} \tag{3.4}$$

where ρ_{ij} (i, j = 1, 2, ..., m, m is the number of depths involved in the model) is the spatial correlation coefficient between data at the ith and jth depth separated by a lag distance h_{ij}. The spatial correlation can be determined by a prescribed auto-correlation model Whittle–Matérn (W-M), expressed as:

$$\rho(h) = \frac{2}{\Gamma(\nu)} \cdot \left(\frac{\sqrt{\pi} \cdot \Gamma(\nu + 0.5) \cdot |h|}{\Gamma(\nu) \cdot \delta} \right)^{\nu} \times K_{\nu} \left(\frac{2\sqrt{\pi} \cdot \Gamma(\nu + 0.5) \cdot |h|}{\Gamma(\nu) \cdot \delta} \right) \tag{3.5}$$

where ν is the smoothness parameter, $\Gamma(\cdot)$ is the gamma function, $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind with order ν, δ is the vertical scale of fluctuation. The parameters in Eq. (3.5) can be identified by the site-specific CPT data and the identified autocorrelation structure is assumed to be the same for all soil parameters in the MUSIC-X model. Actually $\mathbf{R}$ is only integrated to $\mathbf{C}_s$ ($\mathbf{C}_s \otimes \mathbf{R}$), but since $\mathbf{C}_s$ is the condition for all the rest random parameters, it is involved in the derivation of conditional PDFs for all random parameters. The specific derivation of these conditional PDFs can be found in the appendix of Ching and Phoon (2020) and the specific conditional PDFs can be seen in Appendix A in this thesis. Finally the scheme to run GS is shown in Fig. 3.3. The first step in GS inference is to initialize these random variables at arbitrary values. After initialization, there are no NaN values in $\mathbf{X}$. This is how the MUSIC-X model solves the incompleteness problem. For example μ_s vector is set to be a zero vector. Then samples of (μ_s, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$) are drawn from their corresponding conditional PDFs

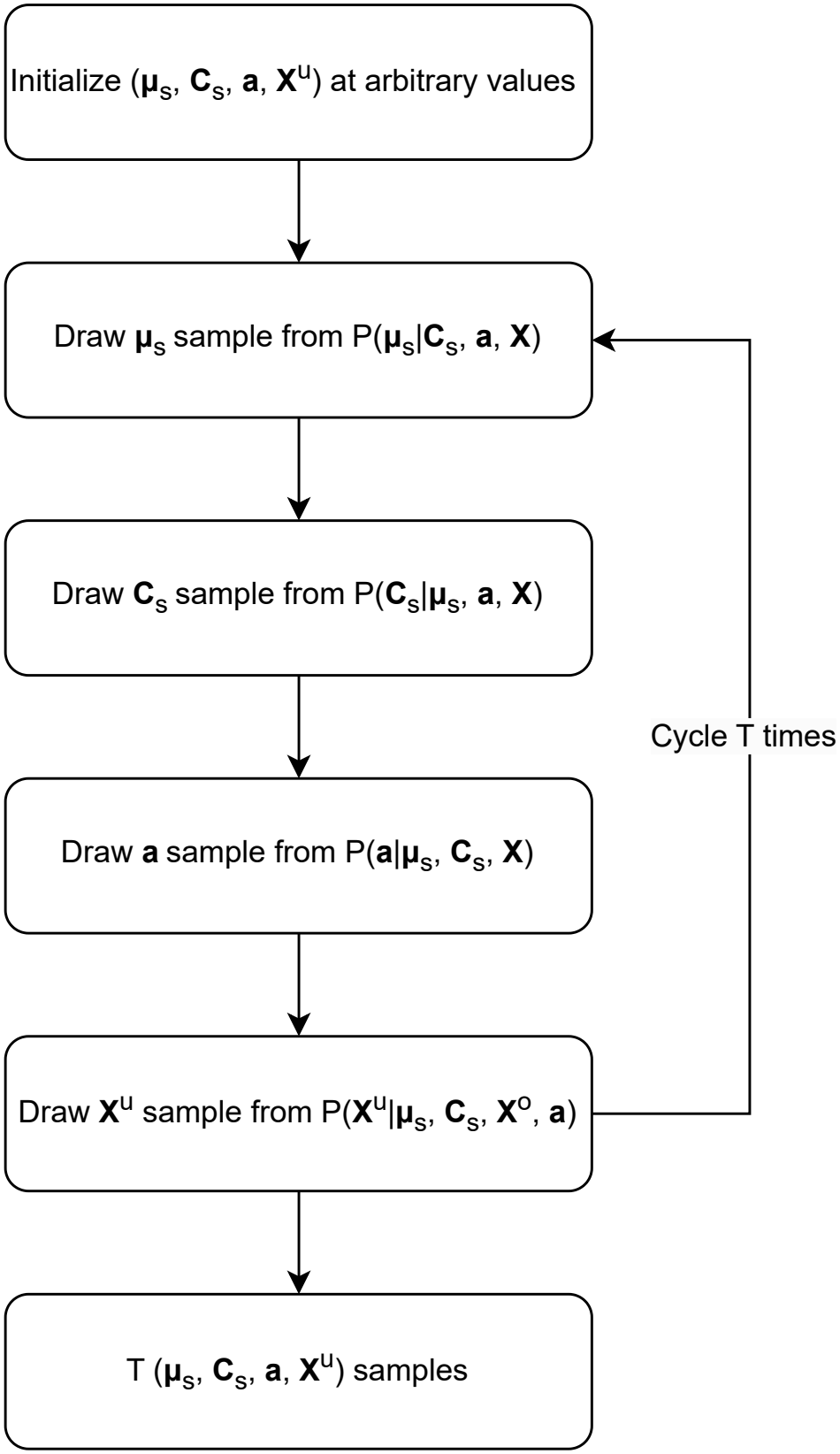


Fig. 3.3. Flow chart of GS inference.

sequentially. The sampling process cycles for T times and finally T (μ_s, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$) samples are obtained. Just notice these T samples cannot be applied to probabilistic analysis directly because at the initial period of the modelling, sampling is not stable. The period is called burn-in period ($t_{burn\text{-}in}$) and it should be removed from the samples. This is to say, finally ($T - t_{burn\text{-}in}$) samples are used.

3.1.3. Modification

The MUSIC-X model is modified in some aspects to be more applicable to this study. Firstly, since it is proposed to integrate V_{int} into the MUSIC-X model to test whether the model can be improved, $\mathbf{X}$ is extended from 11 dimensional to 12 dimensional. Basically the GS inference part does not have to be modified but the JD transformation for V_{int} has to be derived. The process to construct the JD transformation functions for V_{int} is the same as Fig. 3.2, only that one extra generic database for V_{int} should be added. In this study, V_{int} is exhaustively sampled and are abundant to make such a database, which will be shown later in Section 4.1.1. Therefore this modification can be achieved. The second modification is to change the ACF. The ACF originally used is W-M model, which is shown in Eq. (3.5). Two parameters should be derived to construct the model, namely SOF (δ) and smoothness (ν). SOF is key to infer the soil properties considering the spatial variability (Xie et al., 2022). It describes the distance over which the parameters of a soil or rock are similar or correlated (Cami et al., 2020). It is the main parameter to reflect the spatial correlation in this study. The smoothness controls the type of ACF. For example, when $\nu = 0.5$, the ACF is the single exponential function. Ching and Phoon (2017) combined the W-M model with sparse Bayesian learning, which makes it possible to find the best structure of the ACF for the site-specific data. It is more accurate than what people usually do: predefining an ACF. However, in this study, this point is not the main concern, so the issue is simplified to compare a few simpler candidate ACFs and use the best-fitted one for the site-specific data. Specifically, the ACFs under consideration are: the single exponential (SExp), second order Markov (SMK) and squared exponential (QExp) ACF, which express the spatial correlation between two data points separated by a lag distance h as (Ching & Phoon, 2019b):

$$\rho(h) = \exp(-2 \times |h|/\delta) \quad \text{SExp} \tag{3.6}$$

$$\rho(h) = (1 + 4|h|/\delta) \times \exp(-4|h|/\delta) \quad \text{SMK} \tag{3.7}$$

$$\rho(h) = \exp\left(-\pi \times h^2/\delta^2\right) \quad \text{QExp} \tag{3.8}$$

The W-M model can be converted to these three ACFs by respectively by setting ν in Eq. (3.5) equal to 0.5, 1.5 and infinity. It can be seen that among these ACFs, only the SOF is needed, which can be estimated by curve fitting between theoretical ACFs and the experimental estimator for ACFs. Specifically, the experimental estimator can be calculated by:

$$\rho(h) = \frac{1}{\mathrm{n(h)}} \sum_{i}^{\mathrm{n(h)}} U(z_i)\, U(z_i + h) \tag{3.9}$$

where U is standard normal data to estimate the spatial correlation (e.g., detrended and normalized q_t from CPTs); $U(z_i)$ and $U(z_i + h)$ refer to the data pairs with a lag distance h; n(h) is

the total number of data pairs at a certain lag distance. The theoretical ACFs are fitted with the experimental curve and the best-fitted SOF is chosen as the "true" SOF. Meanwhile, the ACF best fits with the experimental curve is chosen to use.

3.2. ICCK

Given that ICCK is an advancement in fundamental kriging methods, directly delving into it may pose challenges in understanding. Thus this section starts from some basic kriging methods and traces their progression towards ICCK, followed by details on ICCK.

3.2.1. Development of kriging

Kriging is a spatial interpolation method that can give a prediction of unknown values of a parameter at unobserved locations based on the weighted linear combinations of the observed values of such a parameter (van Beers & Kleijnen, 2003), which is widely accepted in geo-statistics. Initially, simple kriging (SK) and ordinary kriging (OK) are developed and their corresponding formulas can be view in Eq. (3.10) and Eq. (3.11):

$$[Z*(u_0)-m]=\sum_{\alpha=1}^{n}\lambda_{Z,\alpha}\,[Z(u_\alpha)-m] \tag{3.10}$$

$$Z*(u_0)=\sum_{\alpha=1}^{n}\lambda_{Z,\alpha}Z(u_\alpha) \tag{3.11}$$

where $Z*(u_0)$ is the estimation of the parameter of interest Z at a target location u_0, $Z(u_\alpha)$ is a known value of the parameter at a location surrounding the target location u_α, α = 1, ..., n, n is the number of observed data surrounding the target location, $\lambda_{Z,\alpha}$ is the kriging weight for the αth observed data, m is the global mean of the parameter over the considered domain. Both of them assume a constant global covariance (i.e., the covariance between two data points in space remains constant). The difference is that SK assumes the global mean is a constant and known (as m is explicitly shown in Eq. (3.10)) while OK assumes the global mean is unknown. Therefore, SK assumes a second-order stationarity and OK assumes a quasi-stationarity. Since satisfying the second-order stationarity is difficult in a realistic domain, OK is more commonly used. The kriging weights in OK can be calculated by the following equation:

$$\sum_{\alpha=1}^{n}\lambda_{Z,\alpha}C_z\left(u_\alpha-u_\beta\right)+\lambda_L=C_z\left(u_\beta-u_0\right)\quad\beta=1,...,n \tag{3.12}$$

or

$$\begin{bmatrix} C_{Z,11} & \cdots & C_{Z,1n} & 1\\ \vdots & \ddots & \vdots & \vdots\\ C_{Z,n1} & \cdots & C_{Z,\mathrm{nn}} & 1\\ 1 & \cdots & 1 & 0 \end{bmatrix}\cdot\begin{bmatrix} \lambda_{Z,1}\\ \vdots\\ \lambda_{Z,n}\\ \lambda_L \end{bmatrix}=\begin{bmatrix} C_{Z,10}\\ \vdots\\ C_{Z,n0}\\ 1 \end{bmatrix} \tag{3.13}$$

where $C_z(u_\alpha-u_\beta)$ and $C_z(u_\beta-u_0)$ are the covariances between data at u_α, u_β and u_0, separated by the lag distance $h_{\alpha\beta}$ and $h_{\beta0}$ (i.e., $(u_\alpha-u_\beta)$ and $(u_\beta-u_0)$ represent the lag distance), u_α and

u_β are locations surrounding the target location u_0, α, β = 1, ..., n, λ_L is the Lagrange multiplier. The variance at the estimated location of OK can be calculated by the following equation:

$$\sigma^2_{\mathrm{OK}}(u_0) = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} C_z(u_\alpha - u_0) \tag{3.14}$$

where σ^2 is the variance of the data.

Later, considering in a real project, the database usually contains measurement data of other soil parameters, which may provide additional information to benefit the estimation of the parameter of interest, the kriging method has been extended to cokriging to deal with multivariate geological problems. The formula of cokriging is expressed as:

$$Z*(u_0) = \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} Z(u_\alpha) + \sum_{\alpha=1}^{n_y} \lambda_{Y,\alpha} Y(u_\alpha) \tag{3.15}$$

where Z is still the parameter of interest (primary parameter) and Y is the auxiliary parameter to estimate Z (secondary parameter). Basically it is similar to Eq. (3.11), just that one additional term for secondary parameter Y is added. Cokriging is useful but it could be tedious to run, which mainly results from the covariance matrix. The covariance matrix shown in Eq. (3.13) only needs the covariance between the primary data, while cokriging with two parameters needs three types of covariance: the covariance for the primary and secondary data, as well as the cross covariance between the primary and secondary data. This leads to burdensome work especially when inferring the cross covariance. Considering this problem, collocated cokriging has been developed. Collocated cokriging utilizes the Markov model (Almeida & Journel, 1994), which will be shown later in Section 3.2.2, to simplify the estimation of covariance. Typically simple collocated cokriging is used, expressed as:

$$Z*(u_0) = \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} Z(u_\alpha) + \lambda_{Y,0} Y(u_0) \tag{3.16}$$

where $Y(u_0)$ is a known value of the secondary parameter at the target location and $\lambda_{Y,0}$ is its kriging weight. This model is popular because it is simple (Babak & Deutsch, 2009a), however, it may lead to a problem called variance inflation. To be specific, under a Markov model, simple collocated cokriging holds an assumption that weighting the datum collocated with location being estimated is sufficient (Babak & Deutsch, 2009a). In other words, secondary data surrounding the target location are deemed to bring no additional information and only the secondary data collocated in the target location is applied, which can be clearly seen in Eq. (3.16). This feature results in that the kriging variance may be slightly too high. Moreover, as each additional simulation point is added along the sequential path (i.e., with more observed primary data added during modelling), the variance in the estimation increases, which refers to variance inflation. As a result, it leads to a biased estimation. To solve this problem, intrinsic collocated cokriging (ICCK) is proposed (Babak & Deutsch, 2009a, 2009b). ICCK uses an intrinsic model of coregionalization to reduce variance inflation. Instead of only considering the secondary data at target location for cokriging, ICCK employs full simple cokriging based on the intrinsic model. This means that secondary data at all primary data locations are used (i.e., extending the concept of "collocated"), and the intrinsic coregionalization model is used to calculate local distributions (Babak & Deutsch, 2009a). By doing so, ICCK ensures that the correlation between primary and secondary data is accurately reproduced, and variance inflation is eliminate. In this study, ICCK is adopted to estimate the primary parameter S_u with the assistance of the secondary parameter V_{int}.

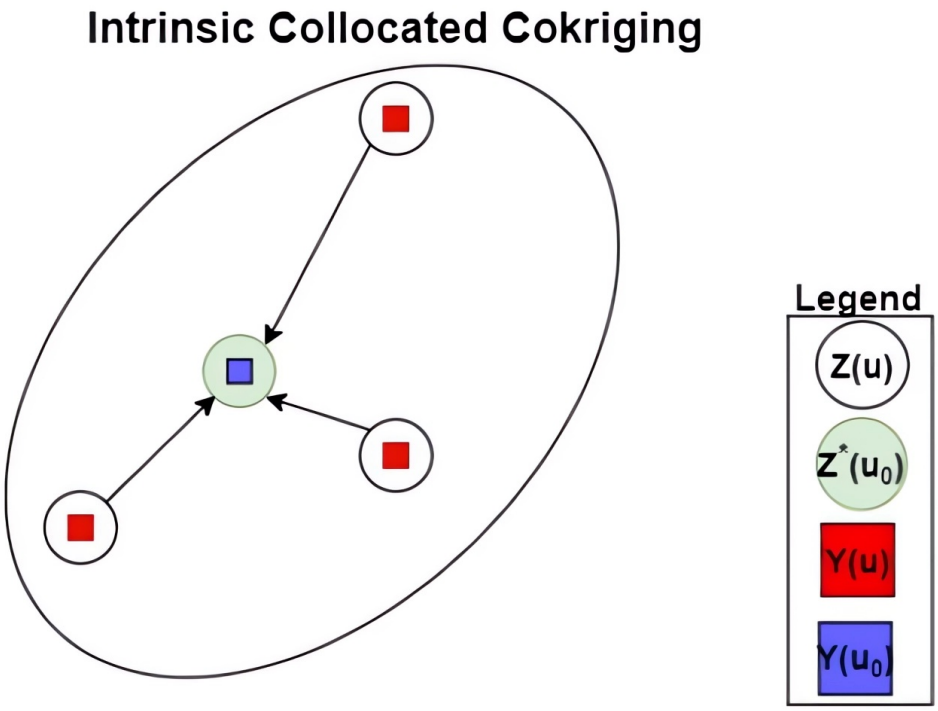


Fig. 3.4. Schematic diagram of ICCK (Samson & Deutsch, 2020).

3.2.2. ICCK

The formula of ICCK (Samson & Deutsch, 2020) is expressed as:

$$Z*(u_0) = \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} Z(u_\alpha) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha} Y(u_\alpha) + \lambda_{Y,0} Y(u_0) \tag{3.17}$$

$Z*(u_0)$ is the estimation of the primary parameter at a target location, which is unknown; $Z(u_\alpha)$ is a known value of the primary parameter at a location surrounding the target location; $Y(u_\alpha)$ is a known value of the secondary parameter at a location surrounding the target location; $Y(u_0)$ is a known value of the secondary parameter at the target location. $\lambda_{Z,\alpha}$ is the kriging weight for the αth primary data; $\lambda_{Y,\alpha}$ is the kriging weight for the αth secondary data, $\alpha = 1, \ldots,$ n, n is the number of observed primary/secondary data surrounding the target location; $\lambda_{Y,0}$ is the kriging weight for the secondary data at the target location. A schematic diagram is shown in Fig. 3.4 to better understand the formula. The green circle is the ICCK estimation for the primary parameter, which is based on the open circles (surrounding observed primary data), red squares (surrounding observed secondary data) and a blue square (secondary data observed at the target location). It can be seen that these primary data and secondary data are paired to use, which is called collocated. The arrow refers to the lag distance. Basically, if the kriging weights are determined, the ICCK estimation can be calculated. As for how to obtain the kriging weights,

they can be calculated based on the following system of equations:

$$\begin{cases} \sum_{\alpha=1}^{n} \lambda_{z,\alpha}\rho_z\left(u_\alpha - u_\beta\right) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha}\rho_{yz}\left(u_\alpha - u_\beta\right) + \lambda_{Y,0}\rho_{yz}\left(u_\beta - u_0\right) = \rho_z\left(u_\beta - u_0\right) \\ \beta = 1, ..., n \\ \sum_{\alpha=1}^{n} \lambda_{z,\alpha}\rho_{yz}\left(u_\alpha - u_\beta\right) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha}\rho_{y}\left(u_\alpha - u_\beta\right) + \lambda_{Y,0}\rho_{y}\left(u_\beta - u_0\right) = \rho_{yz}\left(u_\beta - u_0\right) \\ \beta = 1, ..., n \\ \sum_{\alpha=1}^{n} \lambda_{z,\alpha}\rho_{yz}\left(u_\alpha - u_0\right) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha}\rho_{y}\left(u_\alpha - u_0\right) + \lambda_{Y,0} = \rho_{yz}(0) \end{cases} \tag{3.18}$$

where $\rho_z(u_\alpha - u_\beta)$ and $\rho_z(u_\beta - u_0)$, $\alpha, \beta = 1, \ldots, n$, are the spatial correlation between primary data; $\rho_y(u_\alpha - u_\beta)$, $\rho_y(u_\alpha - u_0)$, $\rho_y(u_\beta - u_0)$ are the spatial correlation between secondary data; $\rho_{yz}(u_\alpha - u_\beta)$, $\rho_{yz}(u_\beta - u_0)$ are the spatial cross correlation between primary data and secondary data; $\rho_{yz}(0)$ is the cross correlation coefficient between the primary and secondary parameter (as lag distance is 0). Among these ρ terms, $(u_\alpha - u_\beta)$ refers to the lag distance $h_{\alpha\beta}$ between data surrounding the target location, while $(u_\alpha - u_0)$ and $(u_\beta - u_0)$ refer to the lag distance $h_{\alpha 0}$ and $h_{\beta 0}$ between the surrounding data and the data at the target location. The spatial correlation essentially corresponds to the covariance, merely with an additional normalization step involving the variance of the data (see Eq. (3.21)). It can be seen that the key to solve this system is to accurately calculate the spatial (cross) correlation terms, which will be elaborated in the following sections.

Estimation of spatial correlation

The spatial correlation ρ_z and ρ_y can be measured by three commonly used methods: variograms, covariograms and correlograms (a.k.a., ACFs used in Section 3.1). The former two do not directly inform the spatial correlation coefficient (e.g., covariograms measure the covariance) but they three are interchangeable and basically they contain equivalent information. In general, variograms measure the dissimilarity while covariograms and correlograms measure the similarity between the data pair separated by a lag distance in space. In this study, variograms are applied to estimating the spatial correlation for the ICCK model. A variogram can be converted into a correlogram, which directly informs the spatial correlation needed in Eq. (3.18), through the process shown below. Firstly, taking primary data Z for example, the relationship between the variogram and covariogram is shown in Eq. (3.19):

$$\gamma_z(h) = C_z(0) - C_z(h) \tag{3.19}$$

where $\gamma_z(h)$ is the variogram, $C_z(h)$ is the covariogram, $C_z(0)$ is equal to the variance. All the primary data and secondary data should be detrended and normalized to standard normal variables (mean = 0, standard deviation = 1) before ICCK interpolation so Eq. (3.19) can be written as:

$$C_z(h) = 1 - \gamma_z(h) \tag{3.20}$$

On the other hand, the relationship between the covariogram and correlogram is shown in Eq. (3.21).

$$\rho_z(h) = \frac{C_z(h)}{\mathrm{var}(z)} \tag{3.21}$$

Also due to the standardization of data, Eq. (3.21) is can be expressed as:

$$\rho_z(h) = C_z(h) \tag{3.22}$$

So based on Eq. (3.20) and Eq. (3.22), the relationship between the variogram and correlogram is expressed as:

$$\rho_z(h) = 1 - \gamma_z(h) \tag{3.23}$$

It can be used to convert a variogram to a correlogram to estimate the spatial correlation.

The way to use variograms is similar to correlograms that three commonly used variograms are adopted to find the best-fitted one for the site-specific data. Note the theoretical functions are different for the variograms. Specifically in this study, the spherical, exponential and Gaussian variogram are adopted, which are expressed as:

$$\gamma(\mathrm{h}) = \begin{cases} b + C_0 \cdot \left(1.5 \cdot \frac{h}{a} - 0.5 \cdot \frac{h^3}{a}\right) & \text{, if } 0 \leq h \leq \mathrm{SOF} \\ b + C_0 & \text{, if } h > \mathrm{SOF} \end{cases} \quad \text{spherical} \tag{3.24}$$

$$\gamma(\mathrm{h}) = b + C_0 \cdot \left(1 - e^{-\frac{h}{a}}\right) \quad \text{exponential} \tag{3.25}$$

$$\gamma(\mathrm{h}) = b + C_0 \cdot \left(1 - e^{-\frac{h^2}{a^2}}\right) \quad \text{Gaussian} \tag{3.26}$$

where b is the nugget, C_0 is the sill and a is the range parameter. It should be noted that a is related to SOF but not necessarily equal to SOF. For the spherical variogram, SOF = a, for the exponential variogram, SOF = 3a and for the Gaussian variogram, SOF = $\sqrt{3}$a. The explanation will be given in Section 3.2.2. The variogram parameters (i.e., b, C_0 and a) can be similarly estimated by curving fitting between theoretical and experimental variograms and SOF can be estimated based on a. The experimental estimator for the variograms[2] is expressed as:

$$\gamma(h) = \frac{1}{2 \times \mathrm{n(h)}} \sum_{i=1}^{\mathrm{n(h)}} \left(U\left(z_i\right) - U\left(z_i + h\right)\right)^2 \tag{3.27}$$

where these notations have the same meaning as they have in Eq. (3.9).

Estimation of spatial cross correlation

As for how to calculate the spatial cross correlation ρ_{yz}, similar to the spatial correlation, theoretically people can also use the experimental variogram to fit with the theoretical variogram model to measure the spatial cross correlation. However, there is an additional condition that any linear combination of the variables is itself a regionalized variable, and its variance must be positive or zero (Xie et al., 2022). Simply speaking, in a regionalized setting, primary data and secondary data are combined linearly with different weights to get the estimation (i.e., Eq. (3.17)). Those estimated variables inherit the original properties from their corresponding regions and their variance must be non-negative. Usually a linear model of coregionalization (LMC) is adopted to ensure the resulting kriging matrix, which contains the spatial correlation terms on the left side of the system of equations, is positive definite and thus to fulfill the condition. However, in practice, fitting the variograms required for an LMC is tedious, especially given the fact that primary data are under sampled and secondary data are exhaustively sampled. Considering this problem, Markov model I is adopted to simplify the estimation of cross

[2] Actually it is for semivariograms, indicated by 1/2. But the term variogram is more commonly used.

correlation (Almeida & Journel, 1994). The Markov model I expresses the cross correlogram as:

$$\rho_{y_z}(h) = \rho_{y_z}(0)\rho_z(h) \tag{3.28}$$

where $\rho_{y_z}(0)$ is the Pearson correlation coefficient between the primary parameter and the secondary parameter. Furthermore, in this model, it assumes that the correlogram of the secondary parameter is equal to that of the primary parameter, shown as:

$$\rho_y(h) = \rho_z(h) \tag{3.29}$$

Fenton and Griffiths (2003) and Fenton et al. (2005) agreed to use the identical soil autocorrelation structure for different soil parameters by arguing that the spatial correlation of a soil is governed largely by the spatial variability in, for example, its source materials, weathering patterns, stress, and formation history, so that one would expect that all the soil properties would vary similarly within an interval. So this assumption is reasonable. With the help of Markov model I, the spatial (cross) correlation is much easier to estimate and thus the system of equations can be easily constructed to solve for the kriging weights.

Anisotropy modelling

In a realistic geotechnical setting, the spatial correlation is different in the horizontal and vertical direction. In this study, geometric anisotropy is adopted to model the anisotropy in the spatial correlation, following the method in Xie et al. (2022). An example is given here to elaborate how to implement geometric anisotropy specifically. Firstly assuming the exponential variogram is adopted, which is already shown in Eq. (3.25). b, C_0 and a can be regarded as constants. It should be noted that a is the range parameter, which is NOT SOF. SOF means a distance beyond which there is almost no correlation. In other words, out of this distance the variance (reflected by the variogram) reaches the plateau. It can be found that in Eq. (3.25), when $h = 3a$, the equation inside the bracket becomes $1 - e^{-3} = 0.95 \approx 1$, which means when $h > 3a$, the variogram will reach a plateau. So SOF is equal to 3a instead of a. For the spherical and Gaussian variogram, the relationship between a and SOF is derived in the same manner. By substituting SOF = 3a back to the original equation, it can be written as:

$$\gamma(h) = b + C_0 \left[1 - \exp(-3\frac{h}{SOF})\right] \tag{3.30}$$

The dimensionless term $\frac{\mathrm{h}}{\mathrm{SOF}}$ is denoted as h′, which is the distance scalar in one dimension:

$$h' = \frac{h}{SOF} \quad (1D) \tag{3.31}$$

So finally Eq. (3.30) can be written as:

$$\gamma(h) = b + C_0 \left[1 - \exp(-3h')\right] \tag{3.32}$$

Now considering the anisotropy in two dimension, the lag distance $\mathbf{h} = [\Delta x, \Delta z]$ is actually a vector. It is proposed to standardize the anisotropic distance **h** to the dimensionless scalar distance h′ by SOFs in the horizontal and vertical direction, which is expressed as:

$$h' = \sqrt{\left(\frac{\Delta x}{SOF_h}\right)^2 + \left(\frac{\Delta z}{SOF_v}\right)^2} \quad (2D) \tag{3.33}$$

This distance scalar contains the anisotropy and by substituting this h' back to Eq. (3.32), the calculated variogram is also anisotropic. This is the geometric anisotropy method. In this study, the vertical SOF of S_u is estimated from CPT data while the horizontal SOF of S_u is proposed to estimate from V_{int} data under the assumption that correlated soil parameters share similarities in the spatial correlation.

4. Case study

The scheme proposed in this study will be demonstrated by a case study in this chapter. It starts with the introduction of the database to use, including basic information, specific data and how to process the data. Then the processed data are further applied to estimating important information required for the MUSIC-X and ICCK model, including the cross correlation between S_u and V_{int} and spatial correlation. Finally the implementation of the MUSIC-X and ICCK model are elucidated by summarizing the input data and presenting the validation and comparison strategy regarding the simulation results.

4.1. Basic information of the database

The database proposed in this study is from the site investigation (including geotechnical and geophysical surveys) of Hollandse Kust (west) Wind Farm Zone. This wind farm zone is located approximately 51 km off the west coast of Netherlands, covering an area of around 175 km^2. Netherlands enterprise agency (RVO) offers free access to the site investigation data, which can be found at https://offshorewind.rvo.nl/cms/view/f4f39d87-68f8-4925-97c4-c49f6a07a001/soll-hollandse-kust-west. The site investigation area highlighted by the red line can be viewed in the Fig. 4.1.

In this area, in total there are 118 CPT locations and 46 BH locations, from where geotechnical parameters can be measured. Meanwhile, there are several ultra-high resolution multichannel seismic (UHR-MCS) reflection survey lines covering this area, from where the geophysical parameters can be measured. For geotechnical data, this site investigation enjoys an advantage that there is always a CPT conducted adjacent to a BH. In order words, CPT-BH clusters are always available, which exactly meets the needs of the MUSIC-X model to combine multisource geotechnical data. For ICCK modelling, the horizontal range of the estimated 2D domain should not be too large in order to avoid extremely large estimation uncertainties. Therefore, the CPT-BH clusters, which are used to simulate 1D S_u profiles in MUSIC-X modelling, should be relatively close to each other (i.e., the simulated profiles should be close to each other) because the simulated S_u profiles are the input of the ICCK model. Considering this point, after looking over the layout of the geotechnical investigations, four clusters at position 97 (#97), position 98 (#98), position 104 (#104) and position 107 (#107) are considered. The relative positions of these clusters can be viewed in Fig. 4.2. Furthermore, since there are fewer lab tests conducted in the samples from BH 107, preliminarily geotechnical data at #97, #98 and #104 are proposed to use. As for geophysical data, the layout of the survey lines in the three-cluster area is shown in Fig. 4.3. Survey line 2X596, 2X595, 2X594 are just through #97, #98 and #104 respectively, which can provide geophysical data at these three positions. So finally the geotechnical and geophysical data at #97, #98 and #104 are all available and will be used in this study.

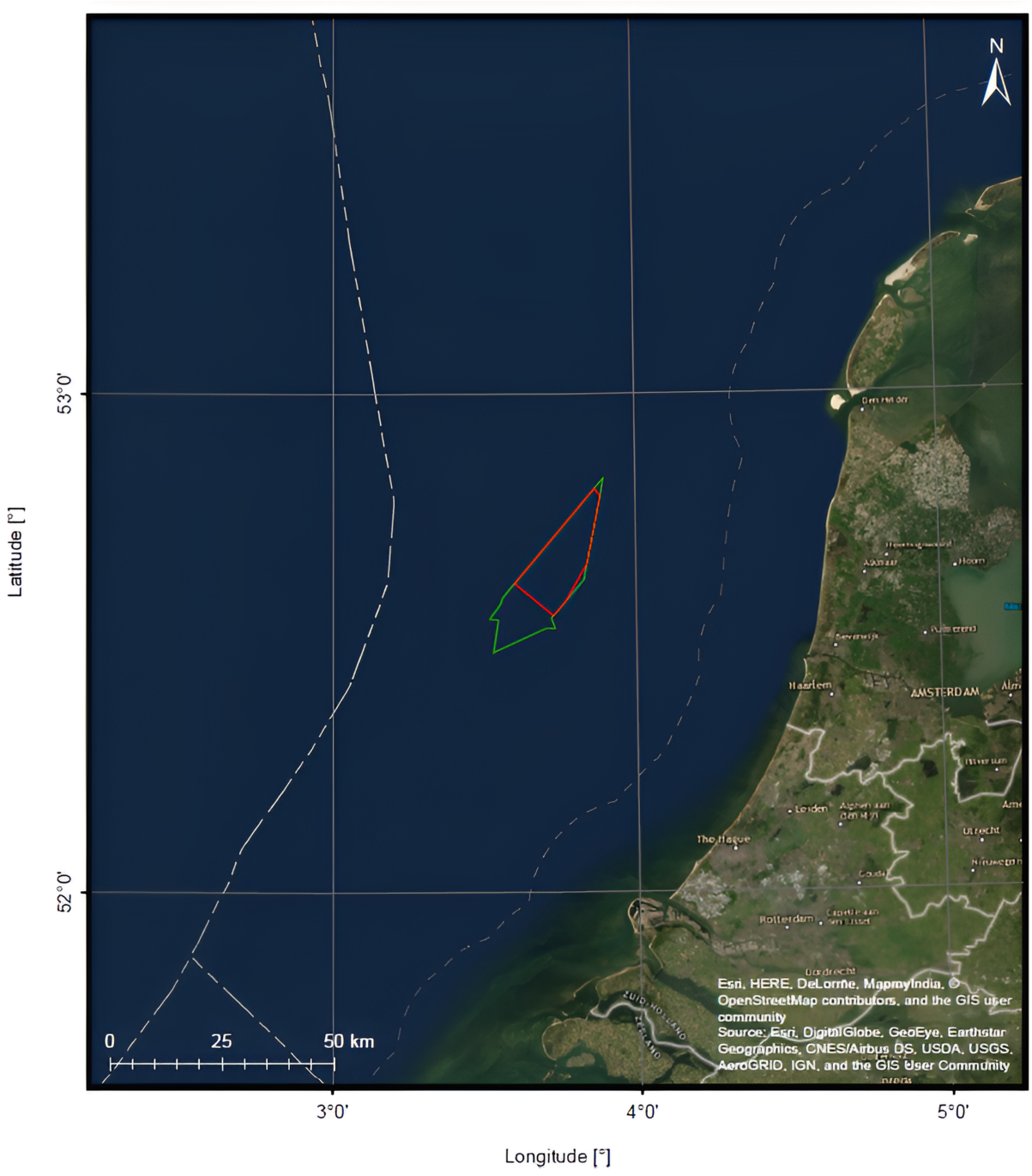


Fig. 4.1. Satellite map of Dutch coast wind farm zone. The investigation area of the Hollandse Kust (west) Wind Farm Zone is highlighted by the red line.

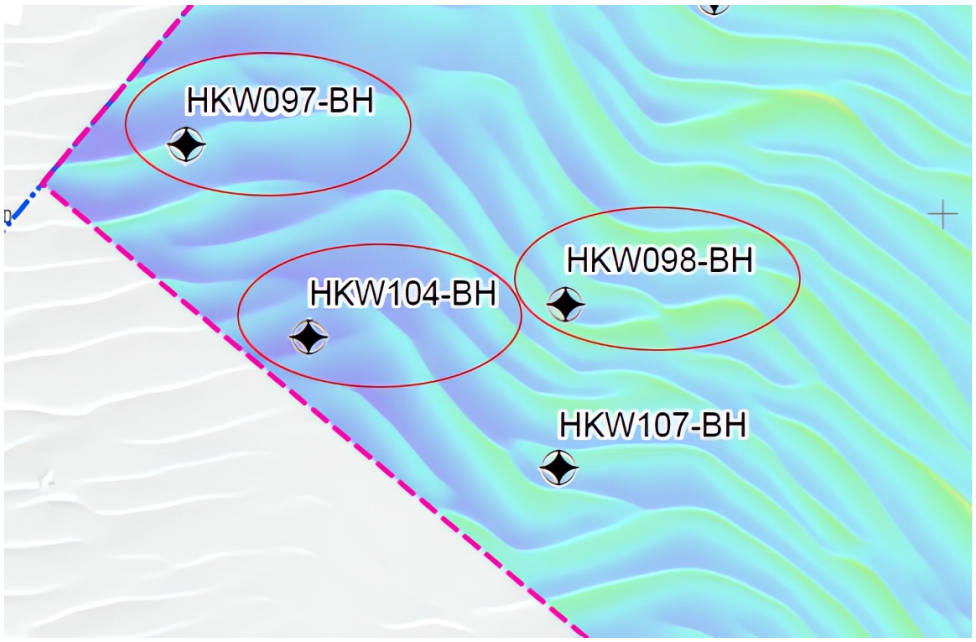


Fig. 4.2. Relative positions of considered CPT-BH clusters. Note at each BH position, there is a CPT conducted, which is not shown explicitly.

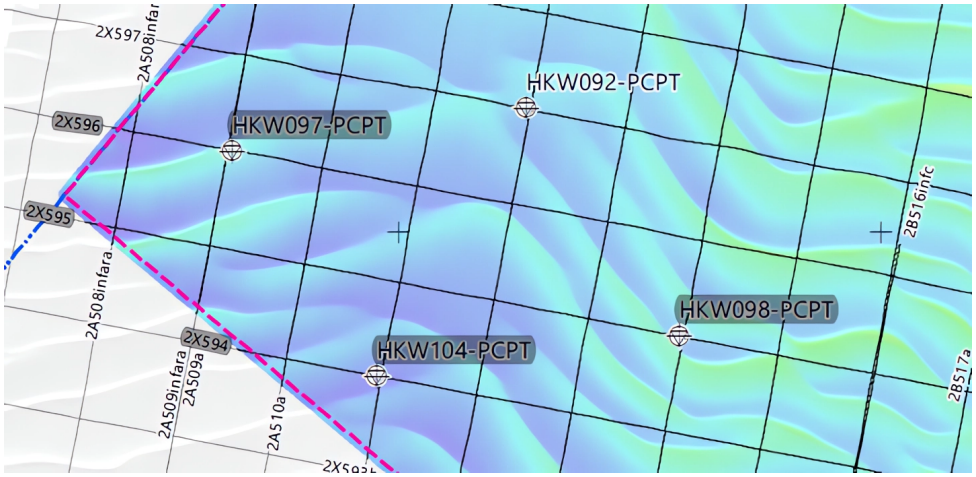


Fig. 4.3. UHR-MCS reflection survey gird in the considered area.

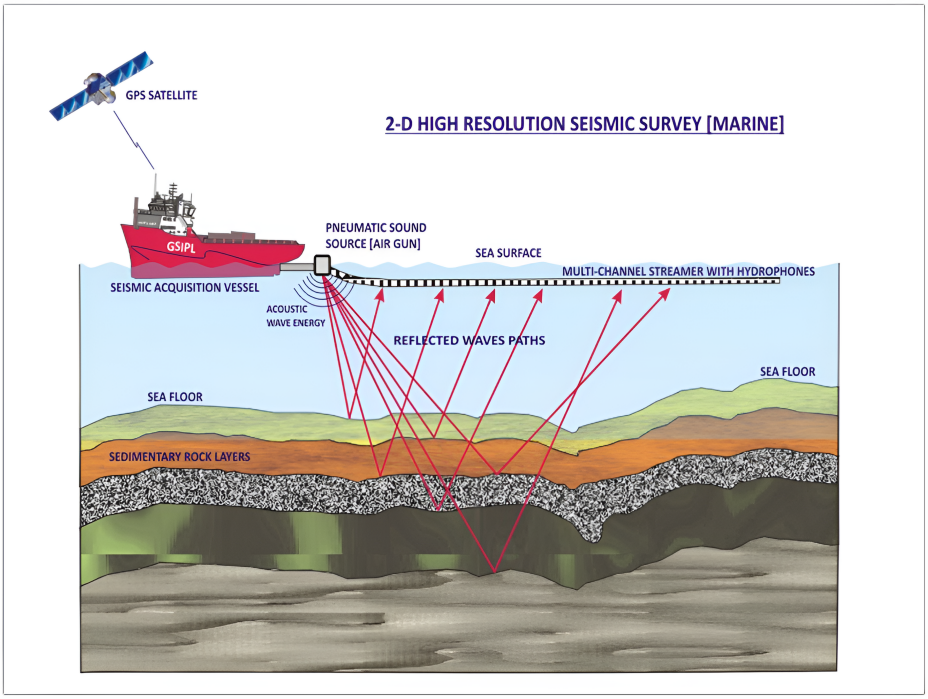


Fig. 4.4. Schematic diagram of UHR-MCS survey ("Schematic diagram of UHR-MCS.", n.d.).

In the following data processing part, geophysical data will be processed prior to geotechnical data because geophysical data usually have a lower resolution. In this study, since integration of these two data sources is met in both MUSIC-X and ICCK part, upscaling of geotechnical data is needed. Therefore geophysical data are initially processed to inform to which extent geotechnical data should be upscaled.

4.1.1. Geophysical data

In this study, the geophysical parameter to use is V_{int} and it is needed in both MUSIC-X modelling and ICCK modelling. It should be noted that V_{int} is not directly measured by UHR-MCS reflection survey lines but interpreted from raw seismic data measured by UHR-MCS reflection survey lines. A schematic diagram is shown in Fig. 4.4 to enlighten how seismic data are measured. As can be seen, a vessel with sound sources and a floating multichannel streamer travels along a prescribed track. During the travel, seismic waves are generated from the sources, passing through soil or rock layers, which are the reflectors, and finally bouncing back to the receivers (geophones) in the streamer. During this process, seismic data along the survey line are collected.

Afterwards, seismic attributes such as V_{int} considered in this study are interpreted based on these directly measured raw data including the amplitude and seismic reflection data. A flowchart shown in Fig. 4.5 illustrates this process. Firstly by analyzing the amplitude of seismic reflections, people can identify the presence and positions of horizons and by analyzing seismic reflection data, the root mean square velocity (V_{rms}) can be interpreted based on the common depth point (CDP) method. V_{rms} refers to a speed at which seismic waves travel though subsurface layers of different V_{int} along a specific ray path. It can be understood as

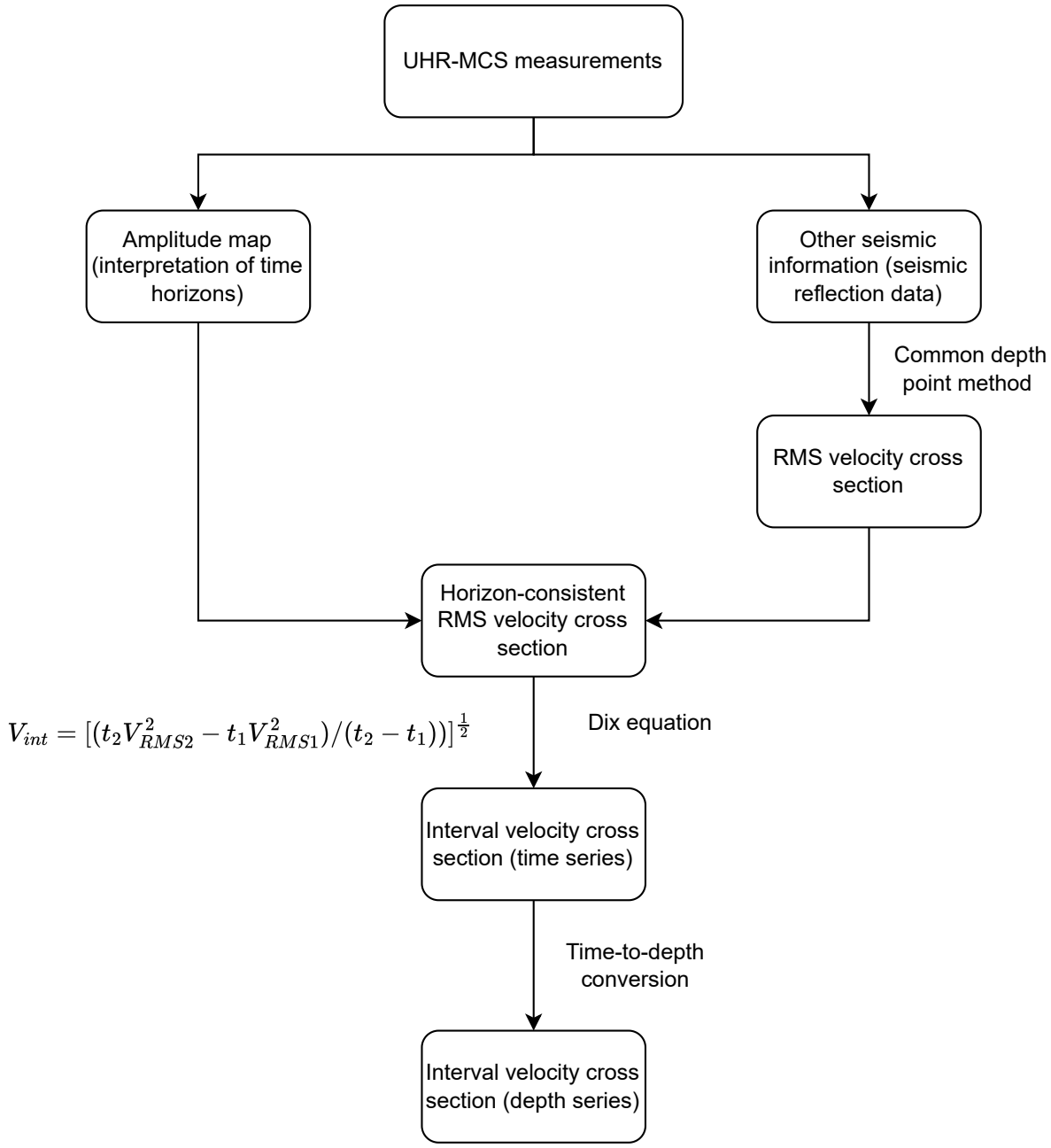


Fig. 4.5. Flow chart of processing UHR-MCS measurements to derive V_{int}.

the weighted average of V_{int} (the weight is the depth). The CDP method is used to stack data to enhance the signal-to-noise ratio of the seismic data and improve the accuracy. Specifically some sets of seismic reflection data, which are from different source-receiver pairs, are sorted into a common depth point (or a gather). The data in the gather is then stacked to find the stacking velocity. The stacking velocities are directly equal to the RMS velocities if soil/rock layers are horizontal or gently dipping, while some correction methods are needed to derive RMS velocities from the stacking velocities if there is a large variation laterally because under the circumstances they may differ substantially. After V_{rms} is interpreted, its profiles are picked at specified analysis locations along the survey line and intersected with the time horizon to derive a horizon-consistent V_{rms} cross section. Subsequently the Dix equation is employed to convert the V_{rms} cross section to V_{int} cross section. To be specific, based on the seismic wave travel time to the first layer (t_1) and to the second layer (t_2), as well as V_{rms} to the first layer (V_{rms1}) and to the second layer (V_{rms2}), V_{int} within the layer can be calculated. Such a cross section is based on time series. In other words the y axis of this cross section is the reflection time instead of depth. Therefore, the final step is to conduct a time-to-depth conversion to obtain the V_{int} cross section. The depth is easy to calculate by multiplying V_{int} and a half of the reflection time (two way travel). This is the general process to interpret V_{int}. It will be further processed to satisfy the MUSIC-X and ICCK model. It should be noted that during this process, inversion uncertainties exist, which is still a challenging problem. This source of uncertainty is not considered in this study.

MUSIC-X

In the MUSIC-X part, V_{int} 1D profiles adjacent to #97, #98 and #104 are needed to integrate with geotechnical data profiles. Taking #97 for example, the V_{int} cross section along line 2X596, which exactly goes through #97 can be viewed in Fig. 4.6. The x axis is the x coordinates (in easting-northing coordinate system) of CDPs along the survey line and the y axis is the depth. Different colors stand for different values of V_{int} with a unit of m/s. Based on x-coordinates, the V_{int} 1D profile at #97 is extracted and shown in Fig. 4.7 and some processing is done on it subsequently. Firstly this profile will be truncated to a segment covering the depth in consistent with the depth of geotechnical profiles used at #97. It should be noted that the upper part of the profile has a constant value around 1500m/s, indicating that is the water layer. When picking up a depth segment to analyze, this layer should be removed. Secondly, since the geotechnical data in MUSIC-X model are all dimensionless, for consistency, V_{int} is nondimensionalized as well. It is divided by 1500 m/s, the V_{int} of water. Furthermore, the resolution of this V_{int} profile (i.e., vertical resolution) is interpolated to 0.1m. The original resolution is 0.111m, which makes it hard to integrate with geotechnical data. Thus, a Python script is written to derive the 1D interpolation function from the original depths and V_{int} by `interp1d`. Then this function is applied to the new depths with a resolution of 0.1m and obtain the Vint corresponding to the new depths. The interpolation does not impact much because V_{int} varies slightly when the depth interval is not very large. For example at depth z = 44.4m, V_{int} = 1685.95m/s, at z = 44.511m, V_{int} = 1686.32m/s. V_{int} interpolated at z = 44.5m is 1686.283m/s. Even though there is difference between the interpolated and "true" value at z = 44.5m, the difference is pretty small.

In addition, as mentioned earlier, a generic database of V_{int} is needed to derive its Johnson distribution parameters in the MUSIC-X model. Just Fig. 4.6 contains more than 30 million data points, which are definitely enough to derive the Johnson distribution. Of course not that

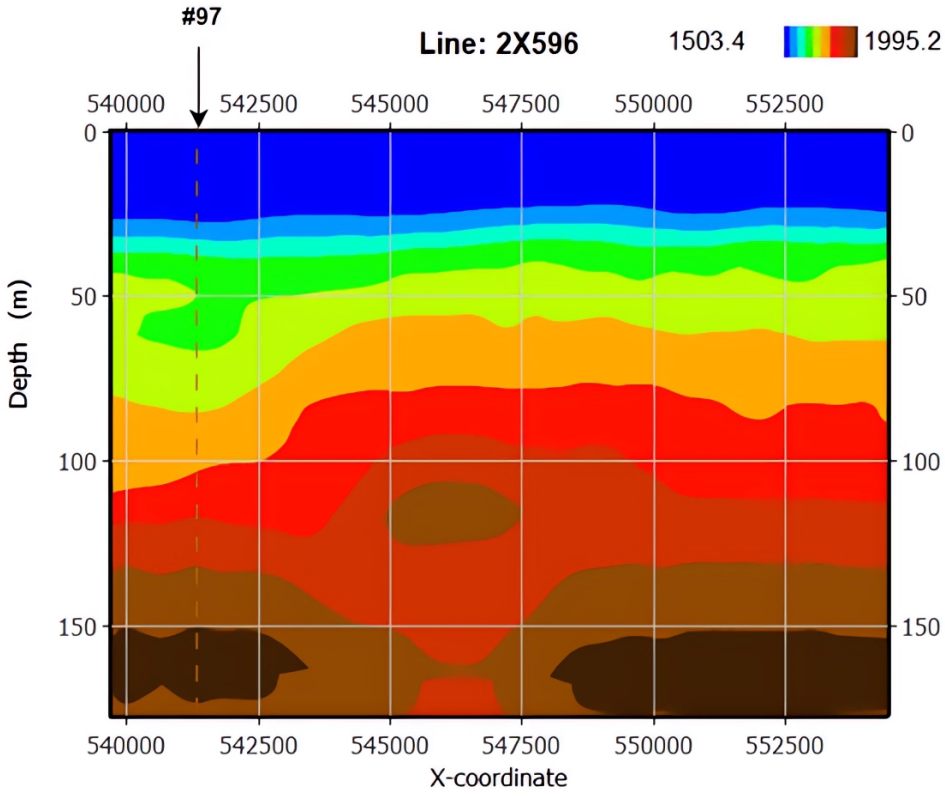


Fig. 4.6. V_{int} cross section along UHR-MCS survey line 2X596. The original latitude/longitude in the UTM Zone 31N (EPSG 25831) projection is transformed into the y/x coordinate in the easting-northing coordinate system.

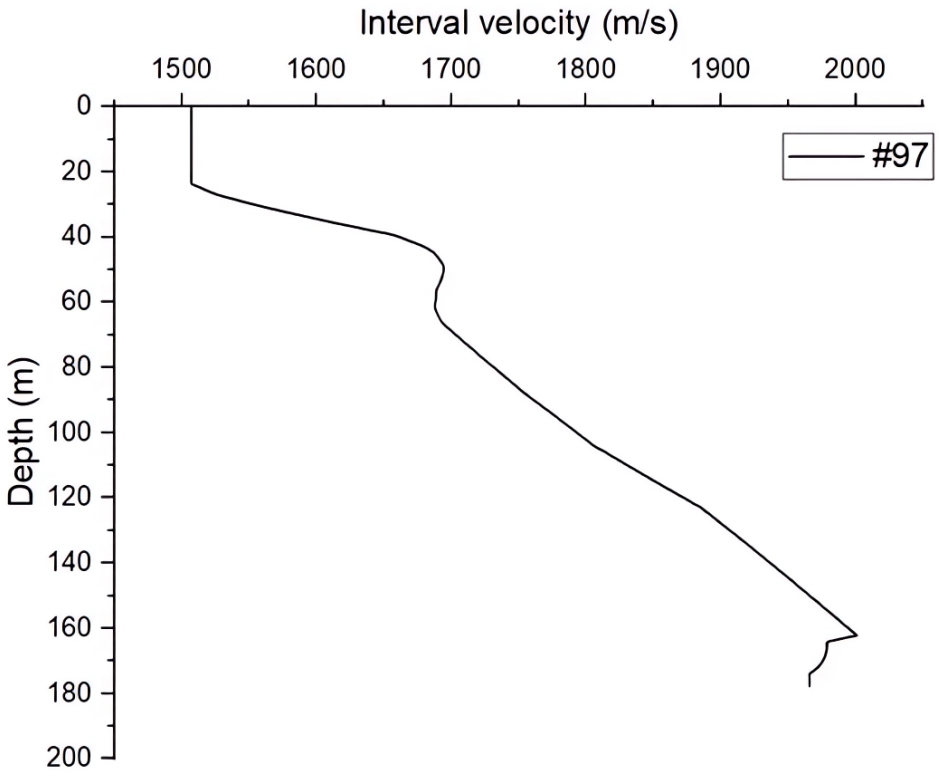


Fig. 4.7. V_{int} profile at #97.

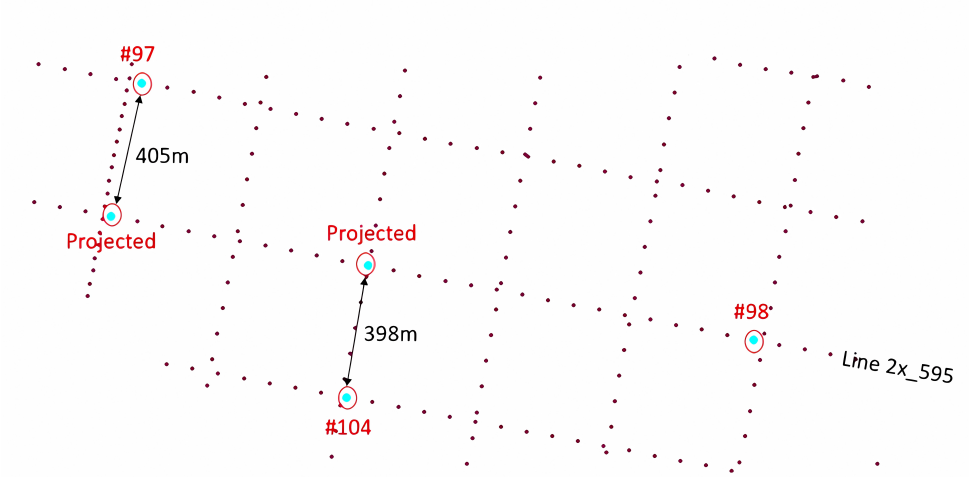


Fig. 4.8. Distances between original #97/#104 and their projected positions in survey line 2X596.

many data points are used, only thousands of V_{int} points from cross sections along survey line 2X596, 2X595, 2X594 are chosen to construct the database.

ICCK

In the ICCK part, a cross section of V_{int} is needed to estimate the cross section of S_u. Based on the layout shown in Fig. 4.3, it can be seen that there is no V_{int} cross section exactly through these three positions (actually they three are not in a line), so it is proposed to align #97 and #104 to survey line 2X595 and use a segment of 2X595 which covers aligned #97, #98 and aligned #104. As for how to realize the alignment, usually direct projection is used. However in this case, the distance between the original position and projected position is very large, around 400 meters, which can be seen in Fig. 4.8. So it is not very accurate to directly make such a projection. Instead, since the V_{int} profiles are known at #97 and #104, these two profiles are compared with the V_{int} profiles near their corresponding projected positions and the position with the most similar V_{int} profile to the original profile is chosen as the alignment location. As for how to compare to find the most similar V_{int} profile, all interval velocities are normalized first by min-max normalization:

$$V'_{int} = \frac{V_{int} - \min(V_{int})}{\max(V_{int}) - \min(V_{int})} \tag{4.1}$$

Then the mean square error (MSE) between the original V_{int} profile and V_{int} profiles at possible alignment positions are calculated. The position with the minimal MSE has a V_{int} profile which is the most similar to the V_{int} profile at the original position and this position is the aligned position.

Furthermore, considering in the original coordinate system (easting-northing, shown in Fig. 4.9), it is hard to find the projected location and calculate the distance, the coordinate system is rotated clockwise (i.e., X′-Y′ in Fig. 4.9) to make these survey lines exactly along the x axis. The coordinates are transformed by the following equation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4.2}$$

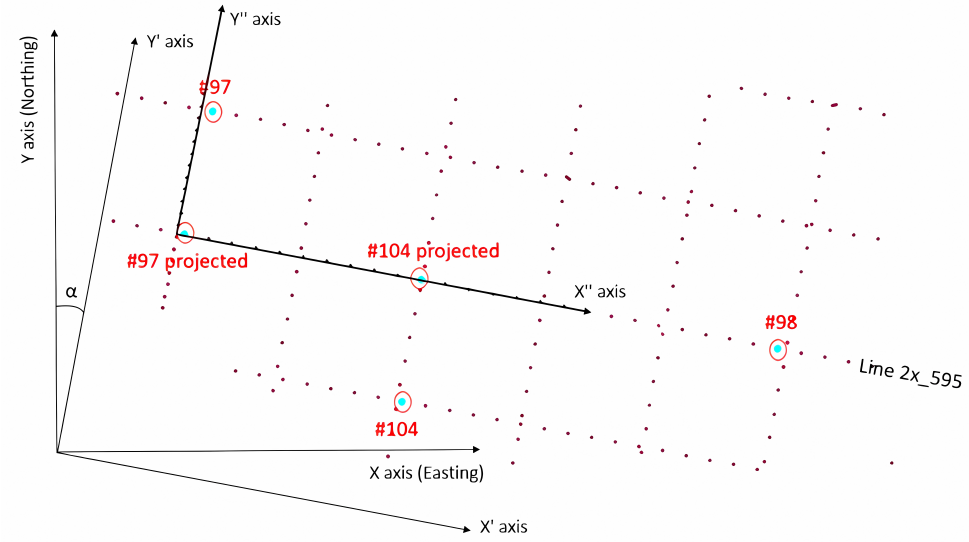


Fig. 4.9. Coordinate system transformation: X-Y is original; X″-Y″ is eventually used.

where x and y are the original coordinates while x′ and y′ are the transformed coordinates. α is the rotation angle, which is 11 degrees in this case. Then this rotated coordinate system moves to the left side of the line 2X595 (i.e., X″-Y″ in Fig. 4.9), so in the position alignment process, only x″-coordinates are focused because all y″-coordinates are zero along this line.

After rotating the coordinate system, #97 is taken as an example to elaborate how to find the aligned position. The directly projected position is at x″ = 5m. Then the most similar V_{int} profile is searched within 5m on the both side (i.e., x″ $\in (0, 10)$). The resolution of the V_{int} in the horizontal direction is 1m so there are ten possible aligned locations and their corresponding MSEs can be viewed in Fig. 4.10. It can be found that the projected location does not have a minimum MSE instead at x″ = 1m, the MSE is the minimum. That is the aligned location for #97. Similarly for #104, the aligned location is at x″ = 775m, shown in Fig, 4.11. For #98, it does not need to be moved and it is at x″ = 1969m.

So based on the above processing, the V_{int} cross section along 2X595 starting from x″ = 0m to x″ = 1969m is extracted for the ICCK model. Its vertical resolution is 0.1m while the horizontal resolution is 1m. One more process, to truncate the cross section to a segment covering the depth in consistent with the depth of simulated S_u profiles, will be done after determining the depth of simulated S_u profiles.

4.1.2. Geotechnical data

MUSIC-X

In the MUSIC-X part, taking #97 for example, the available geotechnical data measured by BH and CPT, are organized and can be viewed in Table 4.1. These 11 soil parameters are exactly in accordance with the parameters used in the MUSIC-X model. With regards to this table, there are five points to mention. The first point is the explanation for why such a segment is chosen. Firstly by browsing the geotechnical data at #97, it is found that most lab test data are concentrated from 20m to 40m depth. Secondly, according to the investigation report, the soils in the investigation area have been categorized as 9 units: A, B1, B2, C1, C2, D, E, F, G. The

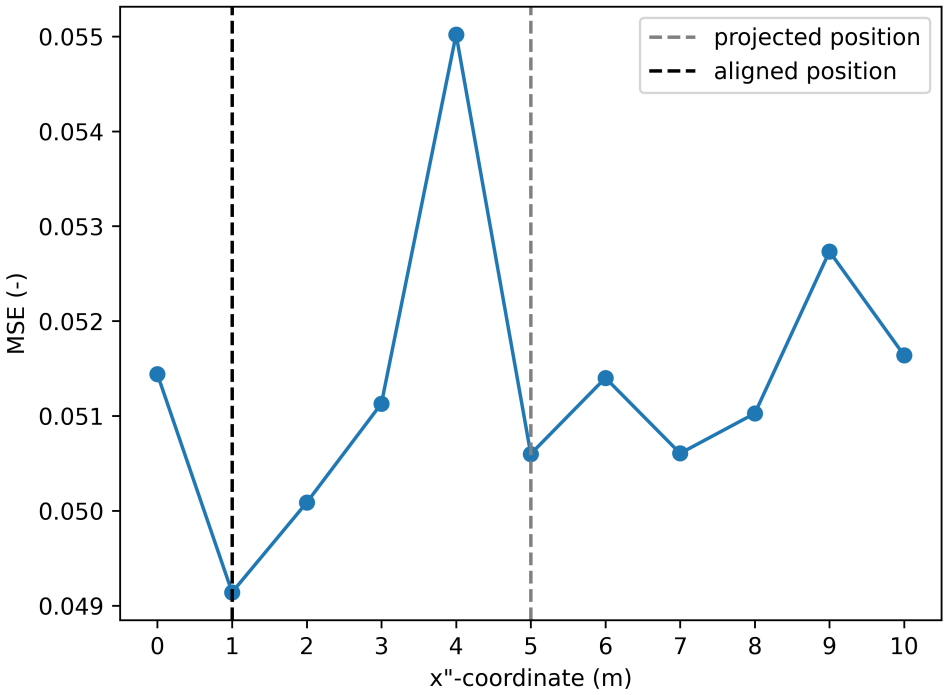


Fig. 4.10. MSEs of potential alignment positions for #97.

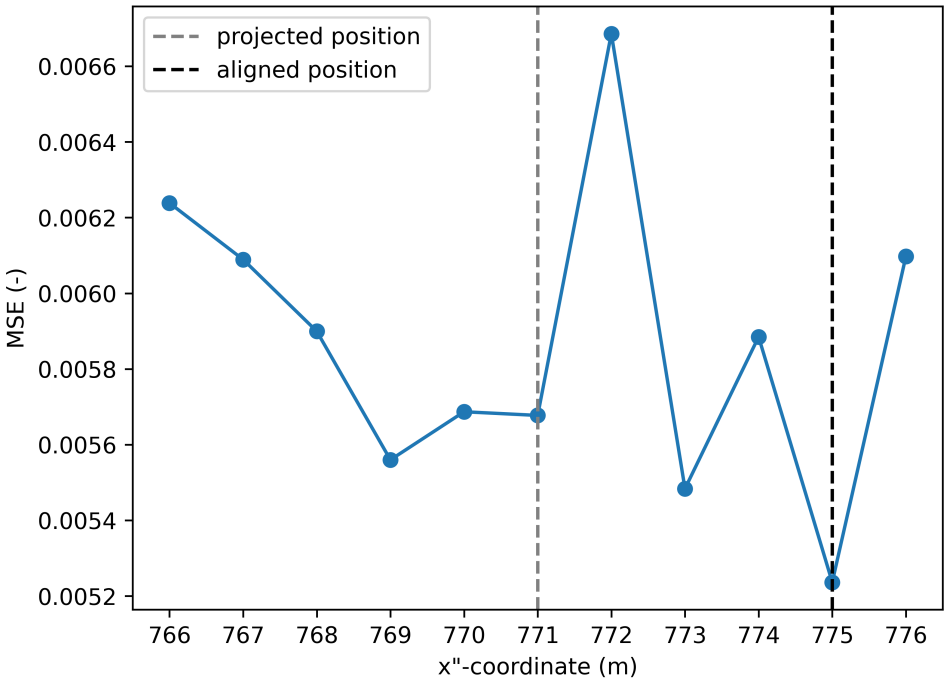


Fig. 4.11. MSEs of potential alignment positions for #104.

Table 4.1. Available geotechnical data at #97

Depth(m)	S_u(PP)(kN/m^2)	S_u(mob)(kN/m^2)	LL	PI	LI	σ'_v/P_a	σ'_p/P_a	S_u/σ'_v	B_q	q_{t1}	C_c	C_s	$N_{60}/(\sigma'_v/P_a)$
			Y_1=ln(LL)	Y_2=ln(PI)	Y_3=LI	Y_4=ln(σ'_v/P_a)	Y_5=ln(σ'_p/P_a)	Y_6=ln(S_u/σ'_v)	Y_7=B_q	Y_8=ln(q_{t1})	Y_9=ln(C_c)	Y_{10}=ln(C_s)	Y_{11}=ln($N_{60}/(\sigma'_v/P_a)$)
21.6	117.000	139.622	-	-	-	2.104	-	0.655	0.097	19.188	-	-	-
23.2	100.000	141.172	-	-	-	2.259	-	0.617	-0.124	9.509	-	-	-
23.6	-	-	59.000	32.000	-0.419	2.298	-	-	-0.031	12.436	-	-	-
24.2	-	-	33.000	8.000	4.100	2.346	-	-	-0.017	41.649	-	-	-
24.8	242.000	185.361	-	-	-	2.385	-	0.767	0.214	13.787	-	-	-
25.0	-	-	36.000	21.000	0.100	2.400	-	-	0.015	12.467	-	-	-
25.4	192.000	177.541	-	-	-	2.443	-	0.717	0.122	9.822	-	-	-
28.8	104.000	165.364	-	-	-	2.761	-	0.591	0.415	8.106	-	-	-
29.4	171.000	191.262	-	-	-	2.817	-	0.670	0.336	11.007	-	-	-

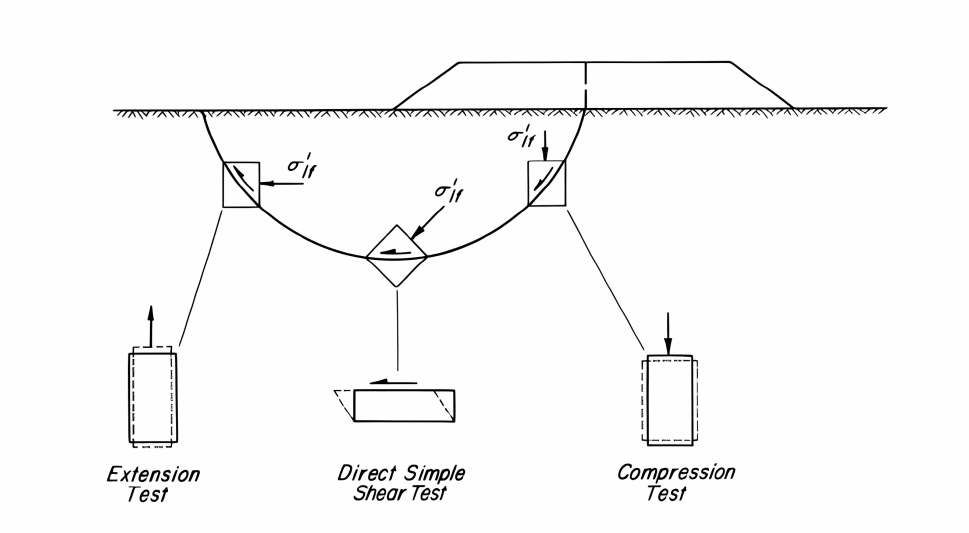


Fig. 4.12. Relevance of laboratory shear tests to modes of shear on a surface of sliding in the field (Terzaghi et al., 1996).

simulation is better to conduct in the same soil unit. At #97, from 11.5m to 31.5m depth, all soils belong to unit F. So comprehensively considering these two factors, the segment to use at #97 is from 21.6m to 29.4m depth.

The second is that, in order to integrate with V_{int} profiles, which have a resolution of 0.1m in the vertical direction, only geotechnical parameters observed at depths with one decimal place are used and complied in the table. It should be admitted that it leads to the loss of information.

The third is that, as can be seen in the table, there are many missing cells, but this is not a problem because MUSIC-X model can accommodate incomplete input.

The fourth is that actually from 21.6m to 29.4m depth, the σ'_v/P_a, B_q and q_{t1} profiles are continuously measured with a resolution of 0.1m (the original resolution is 0.02m but upscaled to 0.1m). For simplicity, in the table only a part of them at depths where other parameters (e.g. LL, PI and LI) are observed, are shown. The table to input in the model is actually augmented by rows and after being augmented, it has (29.4-21.6)/0.1 +1 = 79 rows. For σ'_v/P_a, B_q and q_{t1}, they are known in the augmented rows while for the other soil parameter, they are unknown at the augmented rows.

The last point, which is the most important is that S_u compiled in the table is mobilized S_u (S_u(mob)), which is required for the MUSIC-X model but S_u available in this study is S_u measured by pocket penetrometer tests (S_u(PP)). There is a transformation between S_u(PP) and S_u(mob). S_u(mob) means the shear strength mobilized in a full-scale undrained failure in the field instead of S_u directly measured by any kind of lab tests. Specifically in the MUSIC-X model, S_u measured from different tests are all converted to S_u(mob) for an embankment failure. The S_u(mob) in the embankment failure is illustrated by Fig. 4.12. According to the failure mechanism of an embankment, a triaxial compression, triaxial extension and direct simple shear test are recommended to simulate the shear of an embankment, therefore the mobilized S_u for an embankment failure can be expressed as the following equation, which is basically the average

of these three tests:

$$S_u(\text{mob}) = \frac{S_u\,(CK_0UC) + S_u(\text{DSS}) + S_u\,(CK_0UE)}{3} \cdot \mu_t \tag{4.3}$$

where $S_u(CK_0UC)$ is S_u measured from K_0 consolidated undrained compression test, $S_u(DSS)$ is S_u measured from direct simple shear test and $S_u(CK_0UE)$ is S_u measured from K_0 consolidated undrained extension test; μ_t is the stain rate. Since these three lab tests are nonroutine in a realistic project, several studies have been done to develop transformation equations to convert S_u measured by a certain lab test to mobilized S_u, which can be seen in Table 4.2 (Ching & Phoon, 2014b).

Table 4.2. Developed transformations models to estimate S_u(mob) (Ching & Phoon, 2014b).

Available S_u info	Transformation model
FV	$S_u(\text{mob}) \approx S_u(\text{field}) \approx [S_u(\text{FV})]\,\mu$
UC	$S_u(\text{mob}) \approx S_u(\text{UC})$
UU	$S_u(\text{mob})/\sigma'_v \approx S_u(\text{UC})/\sigma'_v \approx -0.073 + 1.018 S_u(\text{UU})/\sigma'_v$
CIUC	$S_u(\text{mob})/\sigma'_v \approx S_u(\text{UC})/\sigma'_v \approx -0.278 + 1.172 S_u(\text{CIUC})/\sigma'_v$
CK_0UC, DSS, CK_0UE	$S_u(\text{mob}) \approx \{[S_u\,(CK_0UC) + S_u(\text{DSS}) + [S_u\,(CK_0UE)]\,/3\}\,\mu_t$
CK_0UC, CK_0UE	$S_u(\text{mob}) \approx \{[S_u\,(CK_0UC) + S_u\,(CK_0UE)]\,/2\}\,\mu_t^*$
DSS	$S_u(\text{mob}) \approx [S_u(\text{DSS})]\,\mu_t^*$
CK_0UC	$S_u(\text{mob}) \approx [S_u(\text{DSS})]\,\mu_t \approx [S_u\,(CK_0UC)]\,[0.67\mu_t]$
CK_0UE	$S_u(\text{mob}) \approx [S_u(\text{DSS})]\,\mu_t \approx [S_u\,(CK_0UE)]\,\left[1.53^{\dagger}\,(\mu_t)\right]$

Note: FV is field vane test; UC is unconfined compression test; UU is unconsolidated undrained compression test; CIUC is isotropically consolidated undrained compression test.

However, so far there is no transformation model developed for PP tests. Such a model is specially constructed in this study. Firstly PP tests are found to be similar to unconfined compression (UC) tests. Their working principle is the same: compressing a sample from the top unconfinedly. Additionally there are some studies to correlate S_u measured by PP tests and UC tests. On the other hand, some studies have been done to investigate in the correlation between S_u(UC) and S_u(mob), such as the deterministic equation shown in Table 4.2. So it is proposed to transform S_u(PP) to S_u(UC) first then transform S_u(UC) to S_u(mob). Here comes a problem that in these two stages, unavoidably there are transformation uncertainties. It is better if they can be quantified. Eventually, it is proposed to integrate the S_u(PP)-S_u(UC) and S_u(UC)-S_u(mob) transformation model in a probabilistic way. The specific procedures are as follows.

In S_u(PP)-S_u(UC) part, Budak et al. (2022) have found the correlation between the S_u(PP) and S_u(UC) based on 293 remodeled clay samples from 10 different sites in Turkey and the results are shown in Fig. 4.13. The mean of (S_u(PP) / S_u(UC)) is equal to 1.54 and the standard deviation of (S_u(PP) / S_u(UC)) is equal to 0.36. Based on some mathematical calculations, the mean and standard deviation of $\ln(S_u(\text{PP})/S_u(\text{UC}))$ are found to be 0.405 and 0.231. Assuming S_u(PP) and S_u(UC) are lognormal variables, which is typical for soil parameters, S_u(PP) / S_u(UC) is also assumed to be lognormal. Then $\ln(S_u(\text{PP})/S_u(\text{UC}))$ can be expressed as:

$$\ln\left[\frac{S_u(PP)}{S_u(UC)}\right] = \mu + \sigma Z = 0.405 + 0.231 Z_1 \tag{4.4}$$

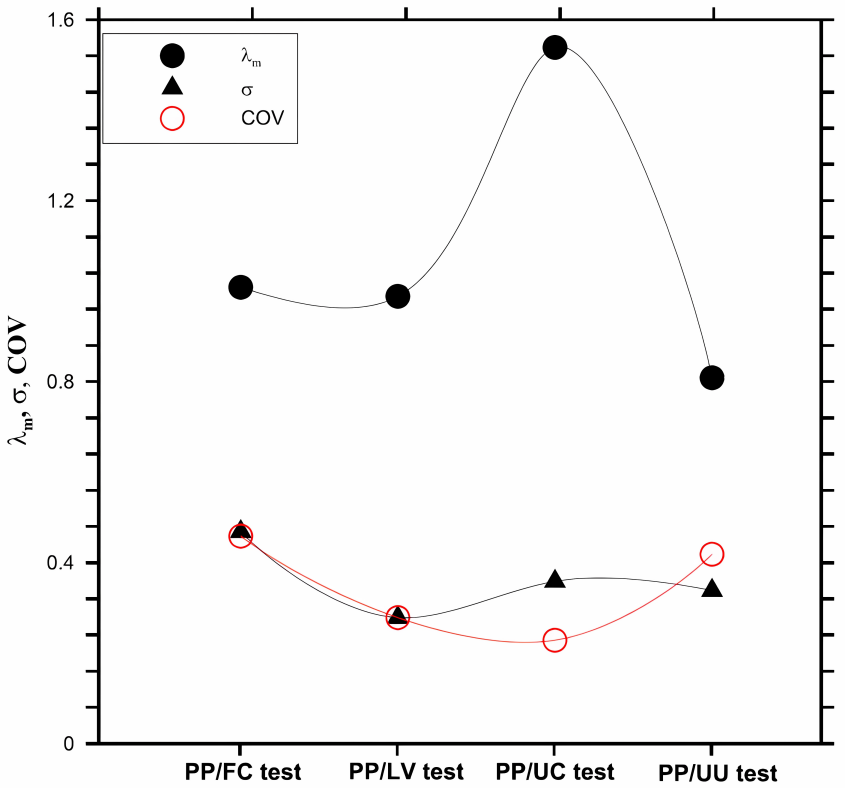


Fig. 4.13. Correlation behaviors between S_u(PP) and S_u measured by other laboratory tests including UC (Budak et al., 2022). λ_m is the mean, σ is the standard deviation and COV is the coefficient of variance.

where Z_1 is a standard normal variable. So the probabilistic transformation between S_u(PP) and S_u(UC) is expressed as:

$$\ln[S_u(UC)] = \ln[S_u(PP)] - 0.405 - 0.231Z_1 \tag{4.5}$$

In the S_u(UC)-S_u(mob) part, although there is a deterministic transformation equation, it should not be used as the transformation uncertainty cannot be quantified. Instead, a probabilistic transformation equation to convert S_u(UC) to S_u(mob) proposed by Ching and Phoon (2015) is employed. The development of this transformation model is based on a generic database CLAY/7/6310. It is expressed as:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-1.047 + 0.263\ln\left[\frac{S_u(UC)}{\sigma_v'}\right] + 0.531\ln(OCR) + 0.081\ln(\frac{PI}{20}) + \ln(\mu_t)}_{\text{mean}} + \underbrace{0.21Z_0}_{\text{std}} \tag{4.6}$$

where μ_t is the strain rate factor, equal to $1 + 0.1 \times \log_{10}$r, r is the strain rate in field, which is recommended to be 60 for UC tests (Ching et al., 2013). Z_0 is a standard normal variable.

By substituting S_u (UC) in Eq. (4.6) with the expression of S_u(UC) in Eq. (4.5), a proba-

bilistic transformation between S_u(PP) and S_u(mob) can be developed and expressed as:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-1.154+0.263\ln\left[\frac{S_u(PP)}{\sigma_v'}\right]+0.531\ln(OCR)+0.081\ln(\frac{PI}{20})+\ln(\mu_t)}_{\text{mean}} + 0.21Z_0 - 0.061Z_1 \tag{4.7}$$

In Eq. (4.7), both $0.21Z_0$ and $-0.061Z_1$ are normal variables, so these two terms can be combined to a variable with a standard deviation of $std(Z') = \sqrt{0.21^2+(-0.061^2)} = 0.219$. Thus Eq. (4.7) can be simplified to:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-1.154+0.263\ln\left[\frac{S_u(PP)}{\sigma_v'}\right]+0.531\ln(OCR)+0.081\ln(\frac{PI}{20})+\ln(\mu_t)}_{\text{mean}} + \underbrace{0.219}_{\text{std}}Z' \tag{4.8}$$

where Z′ is a standard normal variable. With Eq. (4.8), the mean of $S_u(\text{mob})/\sigma_v'$ can be calculated by the following equation:

$$\text{mean of } S_u(\text{mob})/\sigma_v' = \exp\left(\mu_U' + \sigma_U'^2/2\right) \tag{4.9}$$

where μ_U' and σ_U' are the mean and std of $\ln(S_u(\text{mob})/\sigma_v')$. Finally the mean of S_u(mob) can be calculated. The mean value of S_u(mob) is used as a representative value. If necessary, it is also possible to calculate S_u(mob) at lower 5% quantile to follow Eurocode 7. Different from the deterministic transformation, the probability of transformed S_u(mob) is known.

Additionally, it can be found that OCR and PI are needed in Eq. (4.8). In case there are no OCR and/or PI values measured together with S_u(PP), these two terms can be replaced with probabilistic equations. Specifically, ln(OCR) and ln(PI/20) can be regarded as two normal variables, which can be expressed as:

$$\ln(OCR) = \frac{\ln(OCR_L)+\ln(OCR_U)}{2} + \frac{\ln(OCR_U)-\ln(OCR_L)}{2\times 1.96}Z_{OCR} \tag{4.10}$$

$$\ln(PI/20) = \frac{\ln(PI_L/20)+\ln(PI_U/20)}{2} + \frac{\ln(PI_U/20)-\ln(PI_L/20)}{2\times 1.96}Z_{PI} \tag{4.11}$$

Subscripts U and L respectively stand for upper bound and lower bound, Z_{ocr} and Z_{PI} are two standard normal variables. In this site investigation, OCR is found to be between 2.6 and 5.7 and PI is found to be between 7 and 34. By substituting the bound values into Eq. (4.10) and Eq. (4.11) then substituting ln(OCR) and ln(PI/20) in Eq. (4.8) with Eq. (4.10) and Eq. (4.11), the probabilistic transformation between S_u(PP) and S_u(mob) in the case that PI and OCR are both unknown can be expressed as:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-0.459+0.263\ln\left[\frac{S_u(PP)}{\sigma_v'}\right]+\ln(\mu_t)}_{\text{mean}} + \underbrace{0.245}_{\text{std}}Z' \tag{4.12}$$

where Z′ is a standard normal variable. In similar fashion, the probabilistic transformation between S_u(PP) and S_u(mob) in the case that PI is unknown can be expressed as:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-1.175+0.263\ln\left[\frac{S_u(PP)}{\sigma_v'}\right]+0.531\ln(OCR)+\ln(\mu_t)}_{\text{mean}} + \underbrace{0.221}_{\text{std}}Z' \tag{4.13}$$

The probabilistic transformation between S_u(PP) and S_u(mob) in the case that OCR is unknown can be expressed as:

$$\ln\left[\frac{S_u(mob)}{\sigma_v'}\right] = \underbrace{-0.438 + 0.263\ln\left[\frac{S_u(PP)}{\sigma_v'}\right] + 0.081\ln(\frac{PI}{20}) + \ln(\mu_t)}_{\text{mean}} + \underbrace{0.243}_{\text{std}}Z' \tag{4.14}$$

In conclusion, the S_u(mob) shown in Table 4.1 is transformed from S_u(PP) using Eq. (4.8), Eq. (4.12), Eq. (4.13) and Eq. (4.14).

ICCK

In the ICCK part, for geotechnical parameters, only S_u is needed but S_u data are not the site-specific data such as S_u shown in Table 4.1. Instead simulated S_u profiles at #97, #98 and #104, the output of the MUSIC-X model, are used. They will be shown after MUSIC-X implementation. These simulated profiles are not in a common cross section, so they are aligned to a common cross section where V_{int} is measured because the ICCK model uses data in a collocated way. The alignment positions have been shown in Section 4.1.1.

4.2. Estimation of cross correlation between S_u and V_{int}

Having processed the geotechnical data and geophysical data, the cross correlation between S_u and V_{int} should be estimated from S_u-V_{int} data pairs at #97, #98 and #104. This information helps to preliminarily judge whether it is meaningful to integrate V_{int} into the MUSIC-X model to enhance S_u simulation performance and whether it is reasonable to use the spatial correlation of V_{int} as that of S_u. In addition, the correlation coefficient is necessary in the ICCK model. The linear regression fitted from the data pairs is shown in Fig. 4.14. It shows a moderate positive correlation and the correlation coefficient is 0.63.

4.3. Estimation of spatial correlation

The spatial correlation is necessary to be estimated from the processed geotechnical and geophysical data as well, for both the MUSIC-X and ICCK model. In the MUSIC-X model, correlograms (a.k.a., ACFs) are adopted to estimate the spatial correlation while in the ICCK model, variograms are adopted. As mentioned earlier, in the considered correlograms, SOF is only needed while in the considered variograms, b, C_0 and SOF (a) are needed. In this study, the estimation of the spatial correlation is basically the estimation of these parameters and search for the best-fitted correlogram/variogram in the candidates. The vertical spatial correlation can be estimated from q_t data, for the correlograms considered in the MUSIC-X model and for the variograms considered in the ICCK model. The horizontal spatial correlation can be estimated from V_{int} data as S_u has been found to be correlated with V_{int} in Section 4.2, only for the variograms in the ICCK model. The following subsections show the results of parameters estimated by curve fitting of correlograms or variograms and the best-fitted correlogram or variogram in these two models.

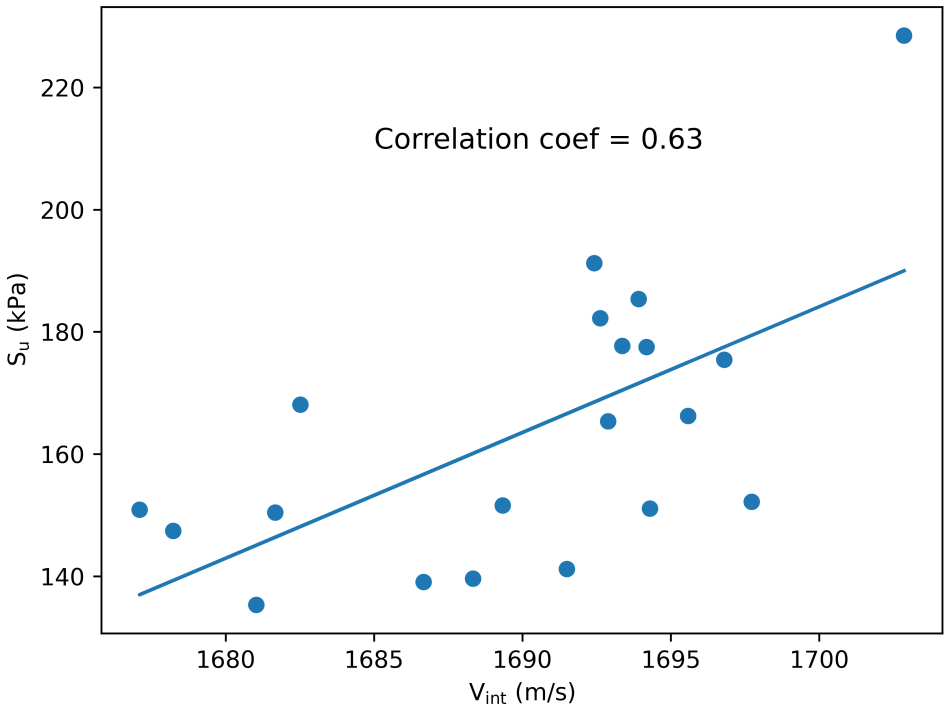


Fig. 4.14. Cross correlation between S_u and V_{int} data.

4.3.1. Vertical spatial correlation

MUSIC-X

For the correlograms in the MUSIC-X model, q_t profiles at #97, #98 and #104 are combined to estimate the vertical SOF (SOF_v). Specifically, these three profiles are combined to derive a common trend based on the 1D linear regression. Then this trend is applied to detrending these three q_t profiles. Subsequently the profiles are normalized by dividing the common standard deviation. Finally the normalized profiles are used to estimate the vertical SOF based on correlogram curve fitting. It should be noted that detrending by a deterministic trend function (e.g., linear or quadratic regression) involves additional uncertainties and it is better to conduct probabilistic modelling of the spatial trend. There will be a trial to probabilistically model the trend in Section 6.2, but basically it is beyond the scope of this study and all the spatial trends used in this study are still deterministic.

The curve fitting results for the SExp, QExp and SMK correlogram are shown in Fig. 4.15. It is hard to compare the fitting performance based on visual judgement so the coefficient of determination (R^2) is adopted to quantify it. In this case, QExp fits as well as SMK, both having a R^2 equal to 0.91, followed by SExp with a R^2 equal to 0.86. Finally the SMK is chosen as the correlogram used in the MUSIC-X model. The best-fitted SOF_v for SMK is 1.83m.

ICCK

For the variograms in the ICCK model, q_t profiles are also combined to estimate the parameters (b, C_0, SOF_v) as what is done for correlograms in the MUSIC-X model. The only difference is that the trend is derived from the 2D linear regression instead of 1D linear regression. In the

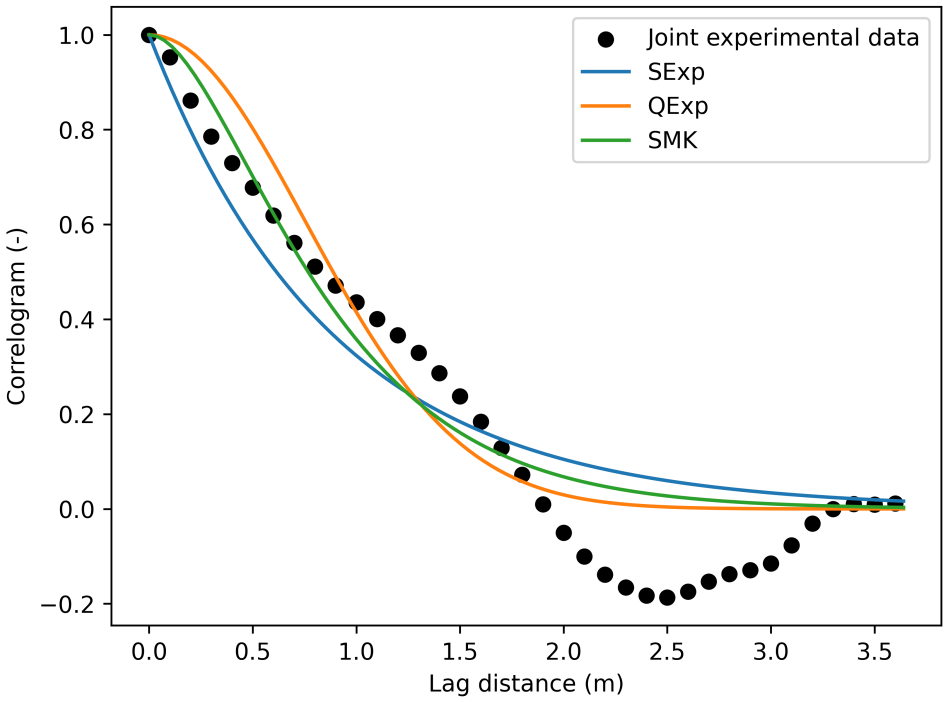


Fig. 4.15. Curve fitting for the correlograms in the vertical direction.

ICCK modelling stage, since these three profiles are all aligned to survey line 2X595 (location alignment can be found in Section 4.1.1), they can be regarded as a 2D domain and in this case it is found 2D linear regression works better than 1D. The GSTool (Müller et al., 2021) is applied to variogram curve fitting.

The curve fitting results for the Gaussian, exponential and spherical variogram are shown in Fig. 4.16. The best fitted variogram is the spherical variogram (R^2 = 0.995), followed by Gaussian (R^2 = 0.986) and exponential (R^2 = 0.982). The best fitted SOF_v of the spherical variogram is 2.63m with b $\approx$ 0 and C_0 = 1.01.

4.3.2. Horizontal spatial correlation

The V_{int} horizontal profiles along survey line 2X595 are also combined (i.e., the cross V_{int} cross section) to estimate the (b, C_0, horizontal SOF (SOF_h)) for the variograms in the ICCK model. Notice it is not a full cross section but the segment mentioned in Section 4.1.1. This cross section consists of several horizontal profiles along the depth and each profile is 1969m long with a resolution of 1m. V_{int} data are also detrended based on the 2D linear regression, then normalized by dividing the standard deviation and finally applied to the curve fitting. The curve fitting results are shown in Fig. 4.17.

The best-fitted variogram is Gaussain with R^2 equal to 0.998 followed by the spherical variogram (R^2 = 0.986) and exponential variogram (R^2 = 0.963). The best-fitted SOF_h of the Gaussian variogram is 407m with b = 0.026 and C_0 = 1.27.

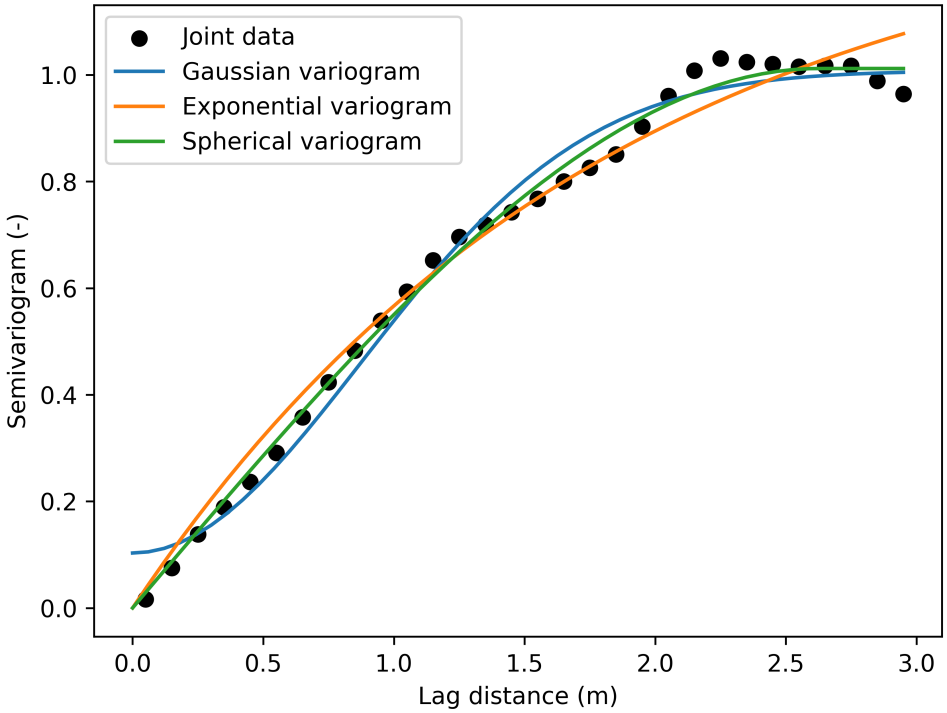


Fig. 4.16. Curve fitting for the variograms in the vertical direction.

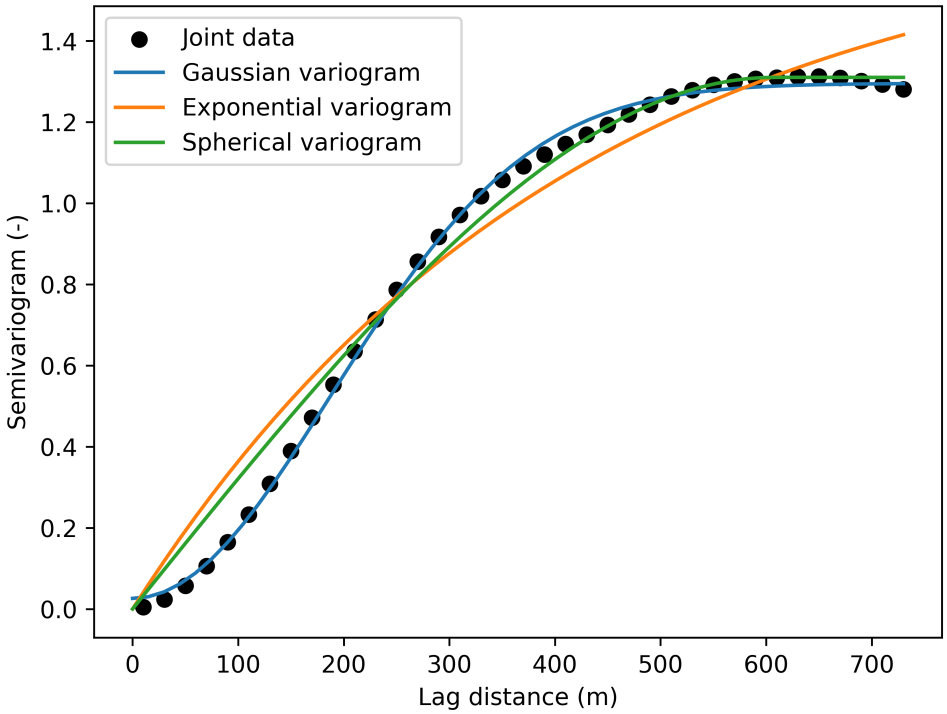


Fig. 4.17. Curve fitting for the variograms in the horizontal direction.

4.4. MUSIC-X implementation

4.4.1. Input setting

Based on the processed data and derived information, the input for the V_{int} integrated MUSIC-X model is summarized here. The geotechnical data at #97, #98 and #104 are integrated respectively with V_{int} data at these three positions. Taking #97 for example, the final input is shown in Table 4.3. The additional database of V_{int} is added to derive the JD transformation for V_{int}. The correlogram used in the model is the SMK correlogram with a SOF_v equal to 1.83m. This is commonly applied to the simulation at #97, #98 and #104. Finally the model will run for 25000 times and the first 5000 times are regarded as the burn-in period. In other words, 20000 S_u profile samples will be obtained for probabilistic analysis.

In the case that the original MUSIC-X is implemented to check whether the simulation performance of S_u is enhanced, the input differs as only geotechnical parameters are used (just like using Table 4.1 instead of Table 4.3) and the generic database of V_{int} is not added.

4.4.2. Validation and comparison strategy

After the implementation of the MUSIC-X model, the results at #97, #98 and #104 will be validated first. Specifically, based on the 20000 simulated S_u profiles, the 95% confidence interval (CI) of the simulated S_u can be found. The range of 95% CI can reflect the simulation uncertainty. In addition, the coefficient of variance (COV, equal to std/mean) of simulated S_u at every depth will be calculated and the average of COV (COV_{ave}) at all depths along the profile is used as an index to quantify the uncertainty of the whole profile. Furthermore, the simulated median S_u profiles at these three positions are chosen to be the input for ICCK implementation. The median profile at each position will be cross validated by each time removing a measured S_u data point to simulate and finally calculating the root mean square error (RMSE) between the simulated results and measured results. The RMSE is calculated for S_u/σ'_v instead of S_u because the dimensionless term is what the model directly simulates.

Additionally, in order to test whether the simulation performance of S_u will be enhanced after integrating V_{int} data, the simulation results from the V_{int} integrated and not integrated scenario at these three positions are compared. The aspects to compare are in accordance with the validation scheme. For example, at #97, the range of 95% CI of simulated S_u profiles, the COV_{ave} and the RMSE for both scenarios will be compared. The scenario with a narrower 95% CI, a lower COV_{ave} and a lower RMSE is considered to have a lower uncertainty and better simulation performance.

4.5. ICCK implementation

4.5.1. Input setting

Based on the processed data and derived information, the input for the ICCK model is summarized here. As for the "observed" primary data, they are simulated median S_u 1D profiles at #97, #98 and #104 from the MUSIC-X modelling, which are available after the implementation

Table 4.3. Available geotechnical and geophysical data at #97

Depth(m)	LL	PI	LI	σ'_v/P_a	σ'_p/P_a	S_u/σ'_v	B_q	q_{t1}	C_c	C_s	$N_{60}/(\sigma'_v/P_a)$	$V_{int}/V_{int,water}$
	Y_1=ln(LL)	Y_2=ln(PI)	Y_3=LI	Y_4=ln(σ'_v/P_a)	Y_5=ln(σ'_p/P_a)	Y_6=ln(S_u/σ'_v)	Y_7=B_q	Y_8=ln(q_{t1})	Y_9=ln(C_c)	Y_{10}=ln(C_s)	Y_{11}=ln($N_{60}/(\sigma'_v/P_a)$)	Y_{12}=ln($V_{int}/V_{int,water}$)
21.6	-	-	-	2.104	-	0.655	0.097	19.188	-	-	-	1.12556
23.2	-	-	-	2.259	-	0.617	-0.124	9.509	-	-	-	1.12766
23.6	59.000	32.000	-0.419	2.298	-	-	-0.031	12.436	-	-	-	1.12813
24.2	33.000	8.000	4.100	2.346	-	-	-0.017	41.649	-	-	-	1.12873
24.8	-	-	-	2.385	-	0.767	0.214	13.787	-	-	-	1.12927
25	36.000	21.000	0.100	2.400	-	-	0.015	12.467	-	-	-	1.12944
25.4	-	-	-	2.443	-	0.717	0.122	9.822	-	-	-	1.12945
28.8	-	-	-	2.761	-	0.591	0.415	8.106	-	-	-	1.12859
29.4	-	-	-	2.817	-	0.670	0.336	11.007	-	-	-	1.12828

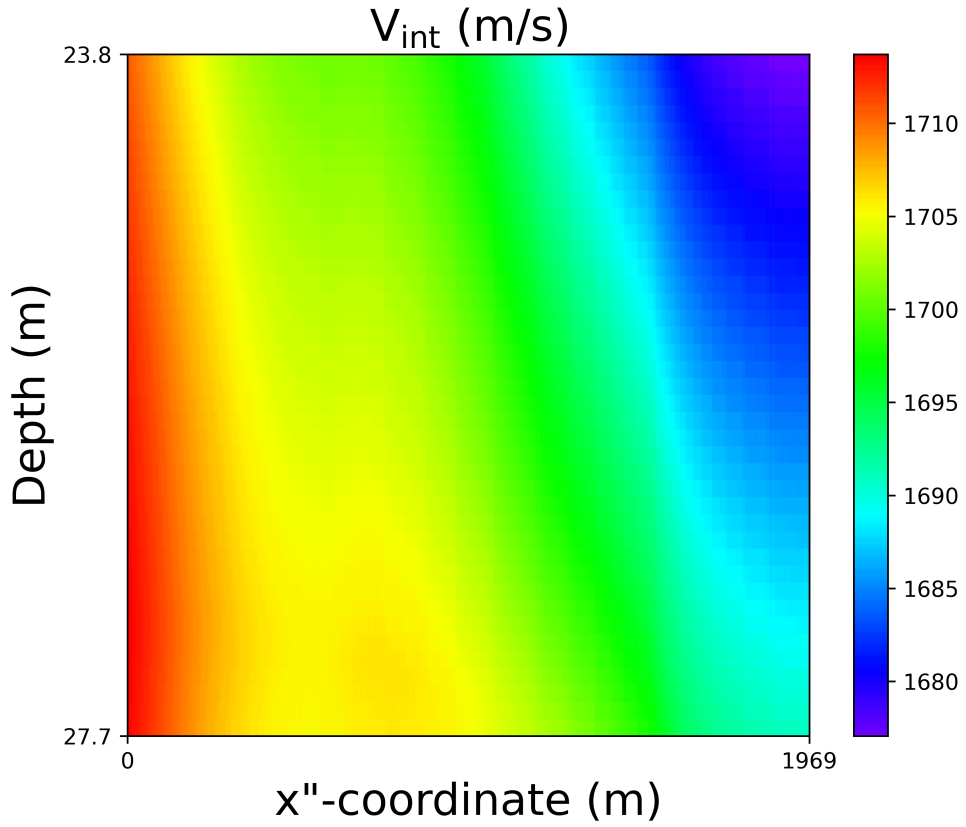


Fig. 4.18. 2D measurement of V_{int} used in the ICCK model.

of the MUSIC-X model. As for which model (the V_{int} integrated MUSIC-X or the original MUSIC-X model) the median profiles are from, it depends on the validation results of these two profiles. The median profile with a lower RMSE will be adopted. In addition, the depth segments of simulated S_u profiles at these three positions are different based on the segment selection scheme shown in Section 4.1.2. They are respectively from 21.6m to 29.4m depth, 23.8m to 29.1m depth and 20.3m to 27.7m depth. These three simulated profiles are truncated to a common depth segment, from 23.8m to 27.7m depth when being applied to the ICCK model as the primary data. As for the observed secondary data, they are a segment of V_{int} cross section along survey line 2X595. The full cross section is truncated to a segment covering aligned #97, #98 and aligned #104 in the horizontal direction and covering 23.8m to 27.7m depth in the vertical direction, which can be viewed in Fig. 4.18. It should be noted that both S_u and V_{int} have to be detrended and normalized to standard normal variables before being applied to ICCK implementation. For the cross correlation between S_u and V_{int}, the correlation coefficient is found to be 0.63. Finally for the spatial correlation, there is a small change since the best-fitted variograms are not the same in the vertical direction and horizontal direction, which can be seen in the Section 4.3. Considering the horizontal spatial correlation is more important in this case because the horizontal distance is very large and the major direction of the variance is horizontal, it is proposed to use the Gaussian variogram, which estimates the horizontal spatial correlation better, to estimate the spatial correlation in the ICCK model. Therefore, the best-fitted horizontal variogram and its associated parameters remain unchanged (SOF_h = 407m, b = 0.026, C_0 = 1.27) while the variogram used to reflect the vertical spatial correlation has been changed from the best-fitted spherical variogram to Gaussian variogram. The parameters fitted for the Gaussian variogram in the vertical direction are: SOF_v = 1.89m, b = 0.103 and C_0 =

0.905.

4.5.2. Validation and comparison strategy

After developing a cross section of S_u based on ICCK modelling, the estimation result will be validated in two aspects. The first is that the ICCK estimation variance will be calculated based on the following equation (Xie et al., 2022):

$$\sigma^2{}_{ICCK}(u_0) = 1 - \left(\sum_{\alpha=1}^{n} \lambda_{z,\alpha} \rho_z(u_\alpha - u_0) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha} \rho_{yz}(u_\alpha - u_0) + \lambda_{Y,0} \rho_{yz}(0) \right) \tag{4.15}$$

By calculating the variance at all estimated positions, a variance map will be constructed, which helps to analyze the uncertainty of the cross section. The second is that the estimation result will be cross validated by removing the "observed" S_u profile at #98 to estimate and calculating the RMSE between the "observed" profile and the estimated profile. The RMSE can be also used to judge the uncertainty of the estimation result.

In addition, two additional schemes are tested and compared with the scheme proposed in this study to investigate in it further. The first scheme is to eliminate V_{int} and only use the simulated S_u profiles from the MUSIC-X model to estimate the cross section of S_u. The ordinary kriging (OK) method will be applied because it returns to a univariate estimation. In this way, the effect of utilizing the geophysical parameter V_{int} to characterize S_u can be analyzed. The second scheme is to directly use measured S_u data points and V_{int} data to estimate the cross section of S_u. This is to say, the result from the MUSIC-X model is not utilized. V_{int} data are still involved so ICCK will be applied. In this way, the effectiveness of combining the MUSIC-X with ICCK can be analyzed. The additional schemes will be validated in the two aspects mentioned above as well and the comparison between these three schemes will be done with regards to the validation results.

5. Results and discussion

5.1. MUSIC-X modelling

Fig. 5.1 reveals the MUSIC-X simulation results for both the V_{int} integrated and not integrated scenario at #97, #98 and #104. In each subplot, red dashed lines show the 95% CI for the scenario to integrate V_{int} while black dashed lines show the 95% CI for the scenario not to integrate V_{int}. Additionally, the red solid line and black solid line respectively show the median profile for the integrated and not integrated scenario. The measured S_u data points are highlighted by blue dots. They can be regarded as conditions in a conditional 1D random field. So it can be seen that at these measured positions, all profiles converge to a single point. The simulation at each position for a single scenario takes around 15 minutes.

It is found that at #104, the 95% CI is narrower for the integrated scenario, which indicates that the integration of V_{int} reduces the uncertainty of simulated S_u profiles. However, at #97 and #98, the 95% CI becomes a little bit wider after integrating V_{int} into the model. It is deemed to be caused by the intrinsic property of V_{int}. V_{int} is not exactly the compressional wave velocity or the shear wave velocity which are properties of the soils/rocks. Instead it is the average value within a certain depth interval, which more generally reflects the spatial trend. Therefore, although there is a moderate correlation shown between V_{int} and S_u, due to the averaging effect, V_{int} may not accurately match the underlying distribution of S_u and it cannot guide the model to generate an accurate distribution of S_u as a CPT profile or shear wave velocity profile does. Especially, sudden changes appear in the q_t profile at #98, which is shown in Fig. 5.2. As q_t is highly correlated with S_u, in general S_u is highly possible to have such sudden changes as well. Due to the averaging effect in V_{int}, it cannot accurately capture such a change, leading to a worse simulation result. In a word, what these results indicate is that the additional cross correlation between S_u and V_{int} is not effective to improve the simulation performance of S_u in a multivariate PTM (i.e., MUSIC-X).

On the other hand, although the cross correlation between S_u and V_{int} may not benefit, the abundance of V_{int} data along the depth may reflect the general spatial trend. Thus potentially, when simulating S_u through correlating V_{int} to S_u (i.e., integrating V_{int} to the MUSIC-X model), V_{int} can serve as a supplement to reflect the vertical spatial correlation. In other words, the spatial correlation can be more accurately incorporated in the model to simulate S_u, finally contributing to more accurate simulations. So it is motivated to investigate in whether the spatial correlation information inside V_{int} can benefit simulating S_u. In order to conduct the test, the spatial correlation estimated from combined q_t profiles in the MUSIC-X model will be modified to be less accurate, making the effect of the spatial correlation information from V_{int} more discernible. Specifically, the spatial correlation for #97, #98 and #104 will be re-estimated based on their individual q_t profiles, which should be less accurate as there are less data points. Taking #97 for example, the correlogram curve fitting result solely based on the q_t profile at #97 is shown in Fig. 5.3. The best-fitted correlogram is SMK with R^2 equal to 0.54,

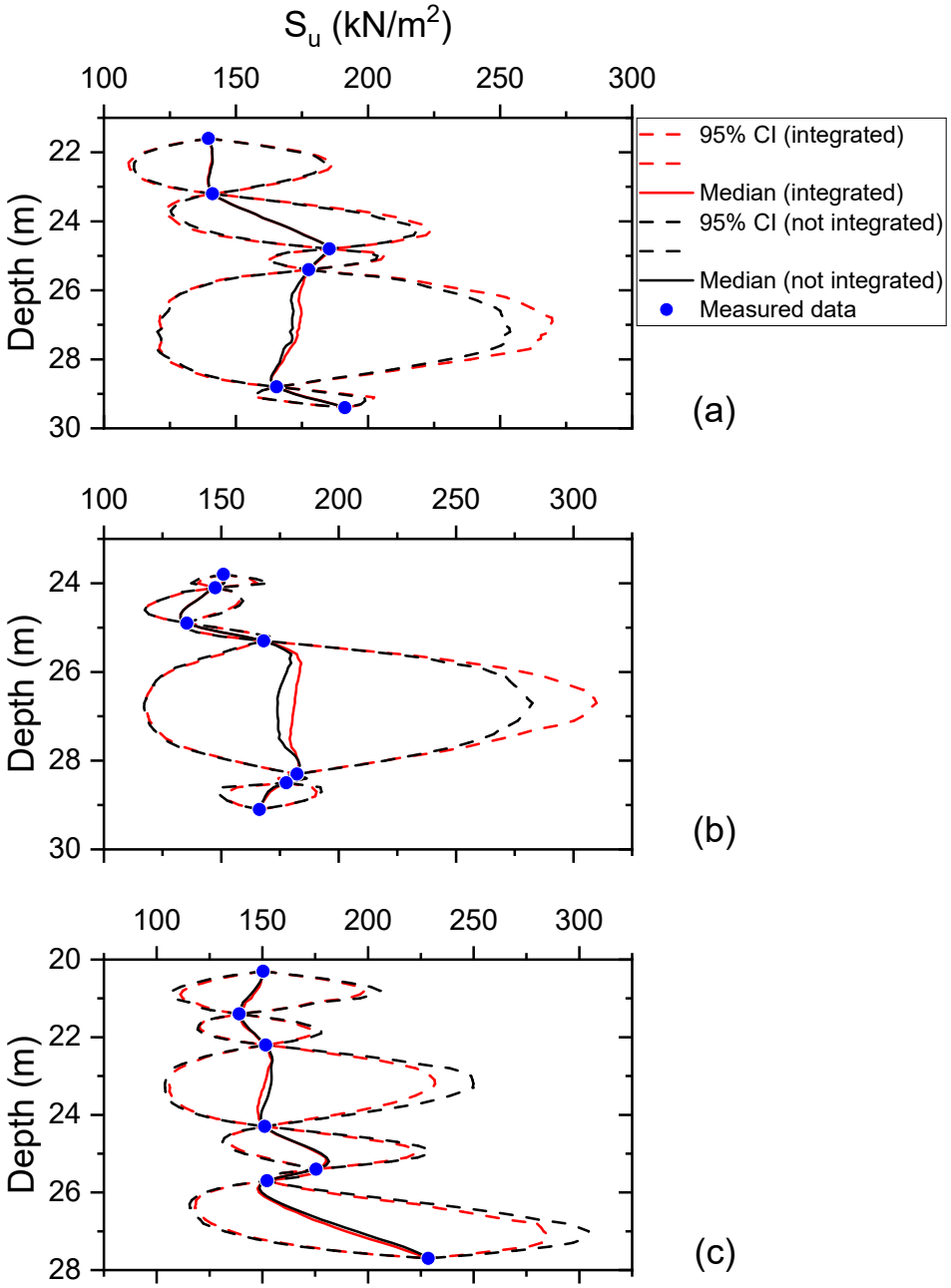


Fig. 5.1. MUSIC-X simulation results of S_u profiles for both the V_{int} integrated and not integrated scenario at (a) #97; (b) #98; and (c) #104.

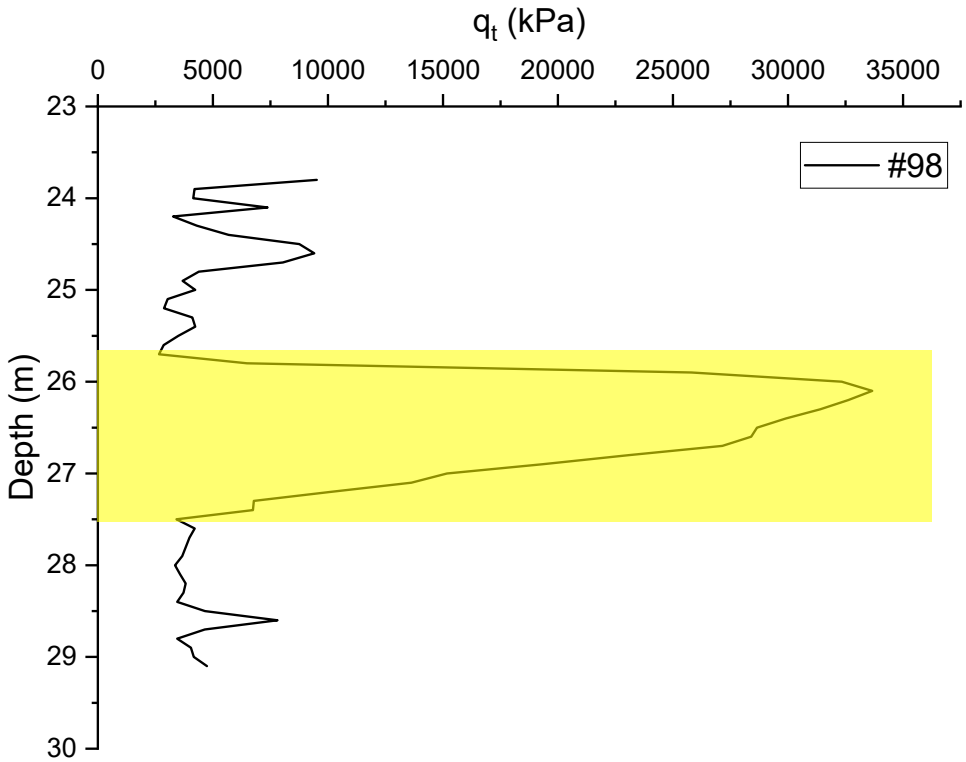


Fig. 5.2. Measured q_t profile at #98. The yellowish rectangle indicates an embedded sand layer.

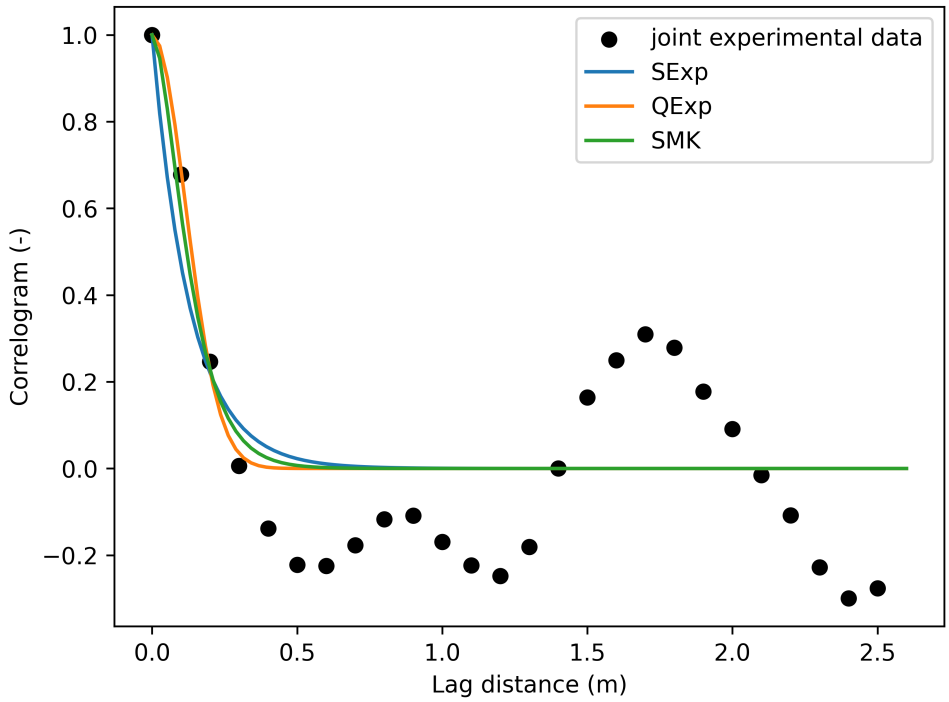


Fig. 5.3. Curve fitting for vertical correlograms at #97 based on its individual q_t profile.

which is relatively low compared to the R^2 resulted from combined q_t profiles in Section 4.3.1 and indicates inaccurate spatial correlation. For #98 and #104, their spatial correlation will be re-estimated in similar fashion. Finally the S_u profiles at these positions will be re-simulated by MUSIC-X based on the inaccurate spatial correlation. In this way, it is able to investigate in whether the integration of V_{int} can supply accurate spatial correlation to S_u and reduce the uncertainty of simulated S_u. The results are shown in Fig. 5.4. The format of Fig. 5.4 is the same as Fig. 5.1. It can be seen that after integrating V_{int}, at #97, the 95% CI for the integrated scenario becomes narrower than that of the not integrated scenario and at #104, the 95% CI for the integrated scenario is more noticeably narrower. #98 is excluded from analysis as the thin embedded soil layer is hard to capture by V_{int} and unavoidably it is more uncertain. These results basically demonstrate the conjecture that the integration of V_{int} can reduce the uncertainty of simulated S_u through supplying the spatial correlation information.

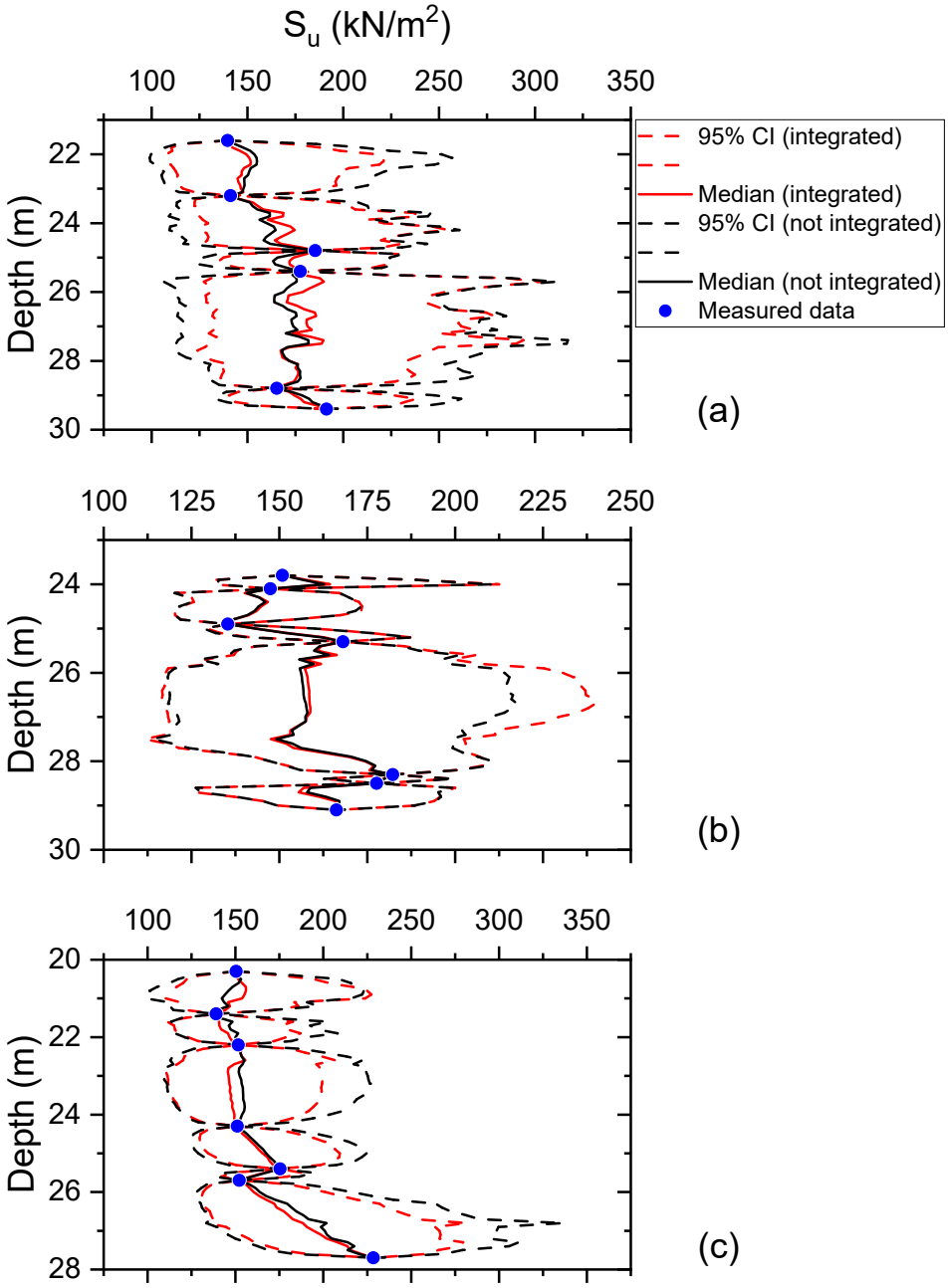
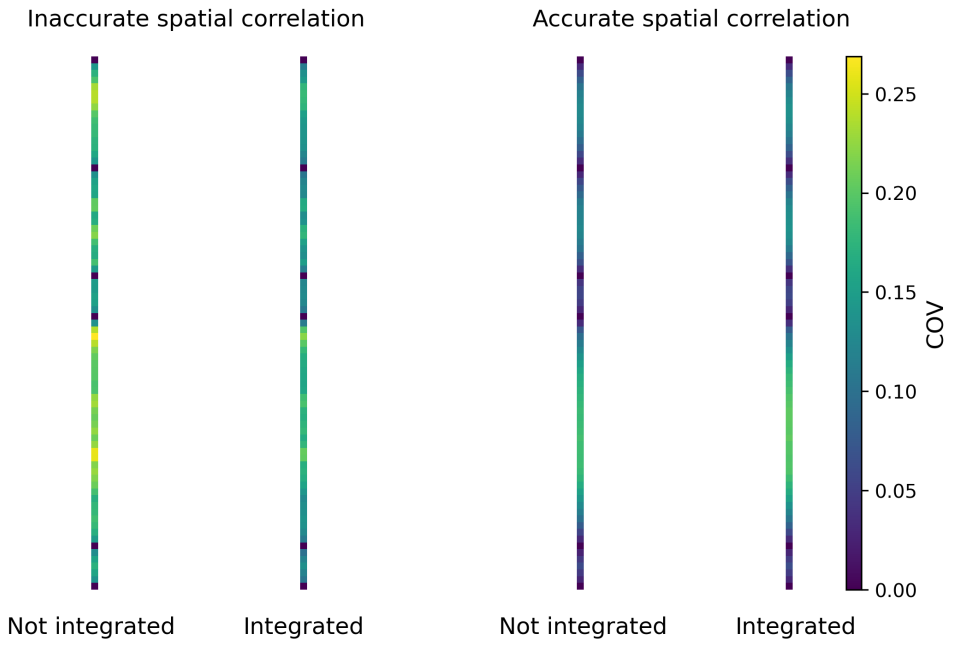
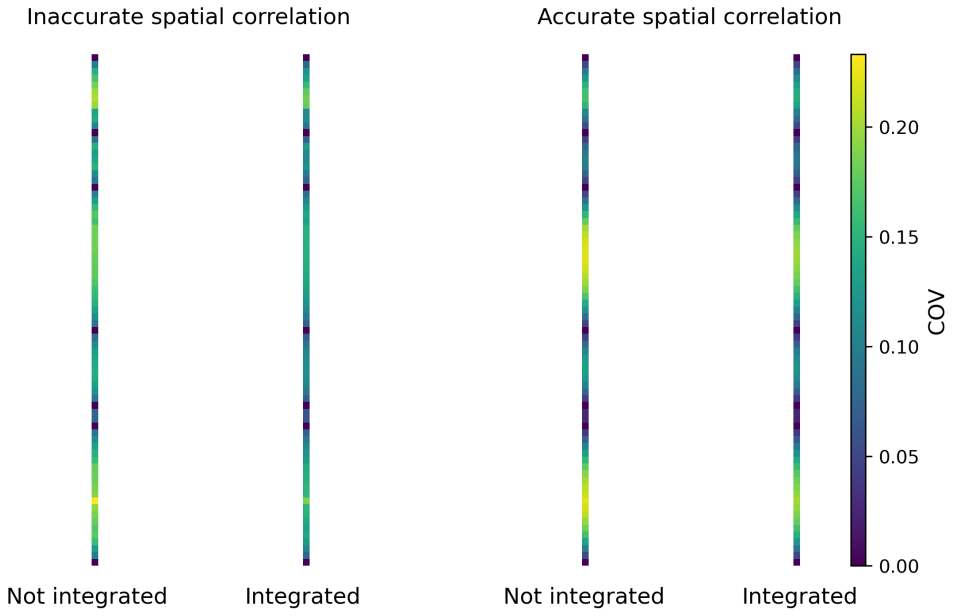


Fig. 5.4. MUSIC-X simulation results of S_u profiles for both the V_{int} integrated and not integrated scenario using inaccurate spatial correlation at (a) #97; (b) #98; and (c) #104.

In order to quantify the change in the uncertainty of simulated S_u at #97 and #104, the COVs along the depth are calculated and shown in Fig. 5.5. The COV_{ave} over the depth are calculated and shown in Table 5.1. Note in the figure and table, the accurate spatial correlation estimated from combined q_t profiles is relative to the inaccurate spatial correlation estimated from the individual q_t profiles. The high R^2 can only tell that there is a good fitting result between the experimental and theoretical correlograms, not necessarily revealing that the spatial correlation is accurately estimated in a strict sense. In Table 5.1, it can be explicitly seen that when the spatial correlation is accurately estimated, after integrating V_{int} to the MUSIC-X model the



(a) COV along the depth at #97.



(b) COV along the depth at #104.

Fig. 5.5. COV of simulated S_u from MUSIC-X along the depth for V_{int} integrated/not integrated scenario and for accurate/inaccurate spatial correlation scenario at (a) #97; and (b) #104.

Table 5.1. Average of COV over the depth for V_{int} integrated/not integrated scenario and for accurate/inaccurate spatial correlation scenario at #97 and #104. The reduction of COV_{ave} after integrating V_{int} is shown in the last column.

	Position	COV_{ave} (not integrated)	COV_{ave} (integrated)	Reduction ratio (%)
Accurate spatial correlation	#97	0.100	0.108	-7.751
	#104	0.119	0.107	10.206
Inaccurate spatial correlation	#97	0.174	0.138	20.934
	#104	0.134	0.109	18.593

simulated S_u may be more uncertain. When the spatial correlation is inaccurately estimated, the integration of V_{int} reduces the uncertainty of simulated S_u by around 20%.

Moreover, the simulated S_u profiles in this stage are further needed in later the ICCK stage as input (primary data). Preliminarily the median profiles of simulated S_u (based on accurate spatial correlation; the inaccurate spatial correlation modelling is just for investigation in the effect of V_{int}) at #97, #98 and #104 are proposed to use. The cross validation of the median S_u profiles for both scenarios can be visualized in Fig. 5.6. In each subplot, the red squares stand for the measured S_u data to test the simulation performance. Open circles show the cross validation results for the integrated scenario. They are companied with error bars to show the 95% CI of the integrated scenario. It can be seen that the simulated values on 95% CI are either too optimistic and pessimistic so median data are better to use. Black circles show the cross validation results for the not integrated scenario. The cross validation results of the integrated and not integrated scenario are explicitly compared in Table 5.2 by RMSE. It can be seen that

Table 5.2. RMSE of the median S_u profile during cross validation for both the V_{int} integrated and not integrated scenario.

	#97	#98	#104
Integrated	0.07	0.06	0.13
Not integrated	0.07	0.06	0.14

these two scenarios perform similarly and actually Fig. 5.1 shows the median profiles for both scenarios at these positions do not differ significantly. Finally the median S_u profiles from the V_{int} integrated MUSIC-X model are chosen as the input for the ICCK model. Due to the RMSEs of the median profiles are relatively small, in spite of existent uncertainties, these profiles are regarded as the "observed" profiles in the ICCK modelling.

5.2. ICCK modelling

Prior to presenting the ICCK estimation result, the input primary data (i.e., the simulated median S_u profiles from the V_{int} integrated MUSIC-X model) and secondary data (i.e., V_{int} cross section) is firstly shown in Fig. 5.7. They will be detrended and normalized into standard normal variables when inputting.

Fig. 5.8 shows the estimation result of the S_u cross section from the ICCK model. The x axis shows the x''-coordinate in the X''-Y'' coordinate system and the y axis is the depth. Different colors standard for different values of S_u. The positions where S_u profiles are known are marked by the dashed lines. The estimation takes around 10 minutes.

Subsequently the validation results with regards to this cross section are shown. Firstly the variance map can be viewed in Fig. 5.9. It can be seen that the variance is zero at positions where there are known S_u profiles. In addition, there is an increase in variance when the distance from the S_u known positions increases, so the largest variance is in the middle part between two S_u known positions. The largest variance is around 0.47 in this ICCK modelling. Secondly the cross validation result is shown in Fig. 5.10. The x axis is the predicted (estimated) values and the y axis is the true ("observed") values at #98. The dashed line indicates where predicted

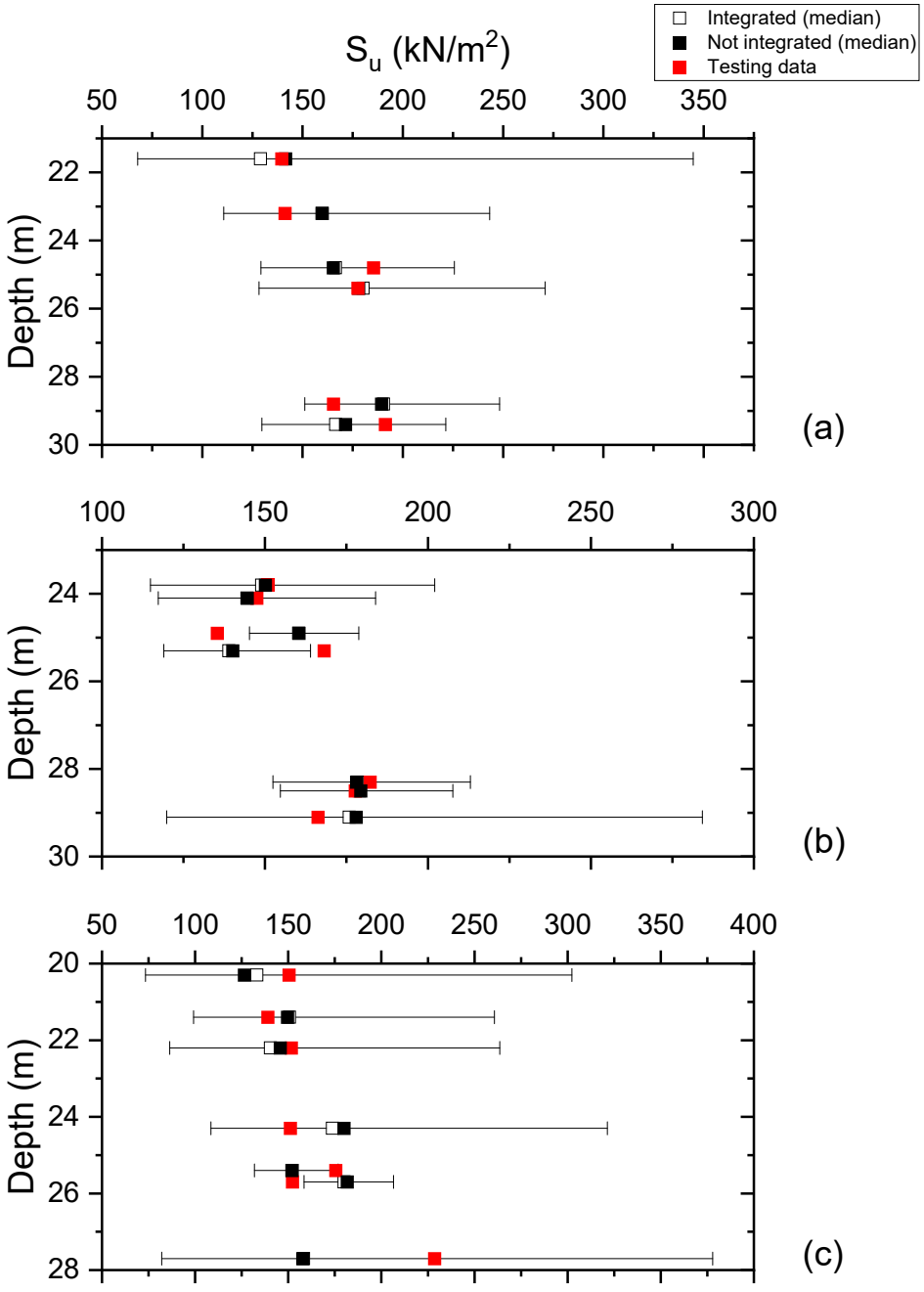
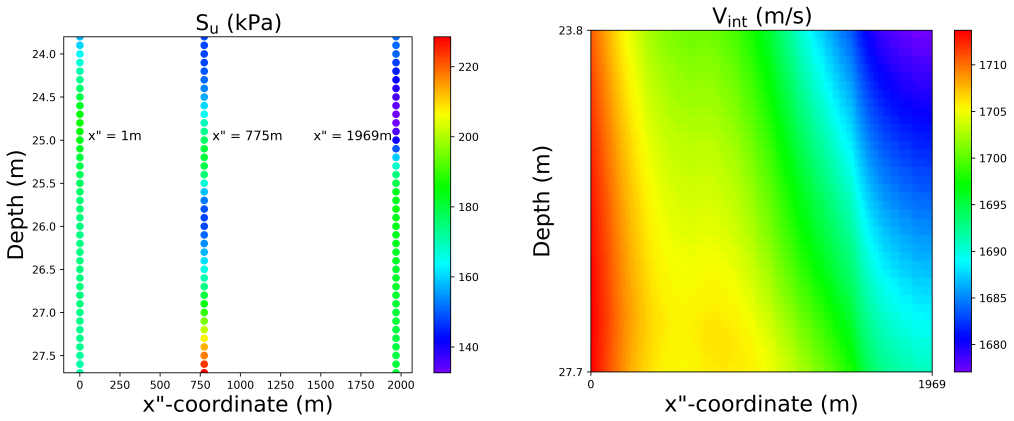


Fig. 5.6. Cross validation result of the median S_u profile for both the V_{int} integrated and not integrated scenario at (a) #97; (b) #98; and (c) #104. For the integrated scenario, error bars are given to show the 95% CI.



(a) Primary data **(b)** Secondary data

Fig. 5.7. Input for ICCK with (a) simulated S_u profiles from MUSIC-X modelling as primary data and (b) measured V_{int} cross section as secondary data.

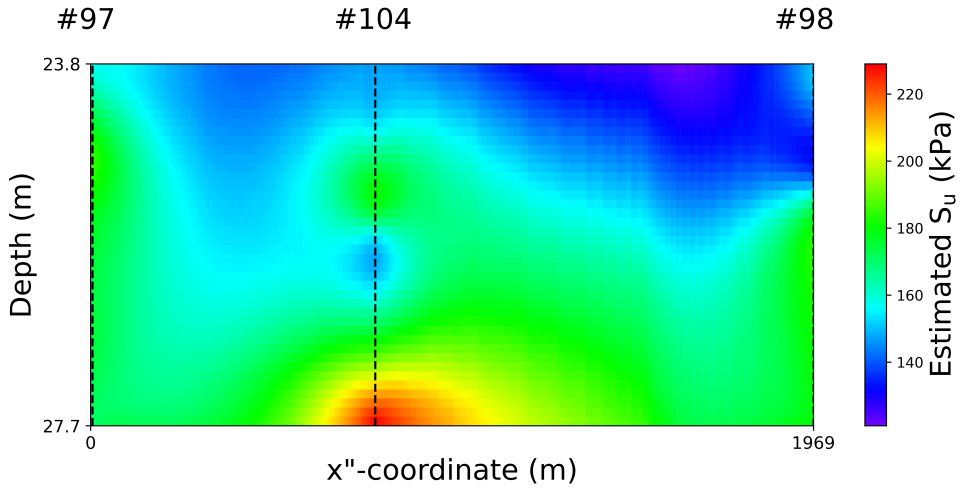


Fig. 5.8. Result of the S_u cross section estimated by ICCK.

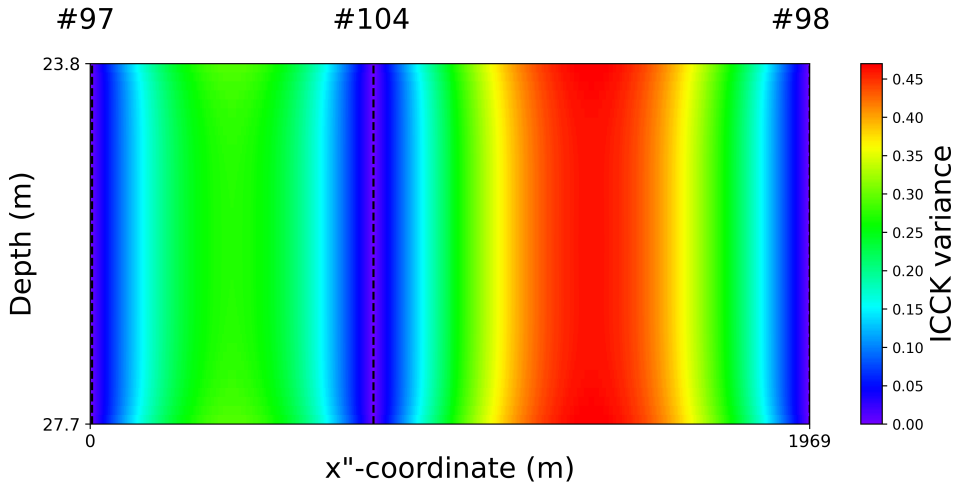


Fig. 5.9. Estimation variance of the S_u cross section from ICCK modelling.

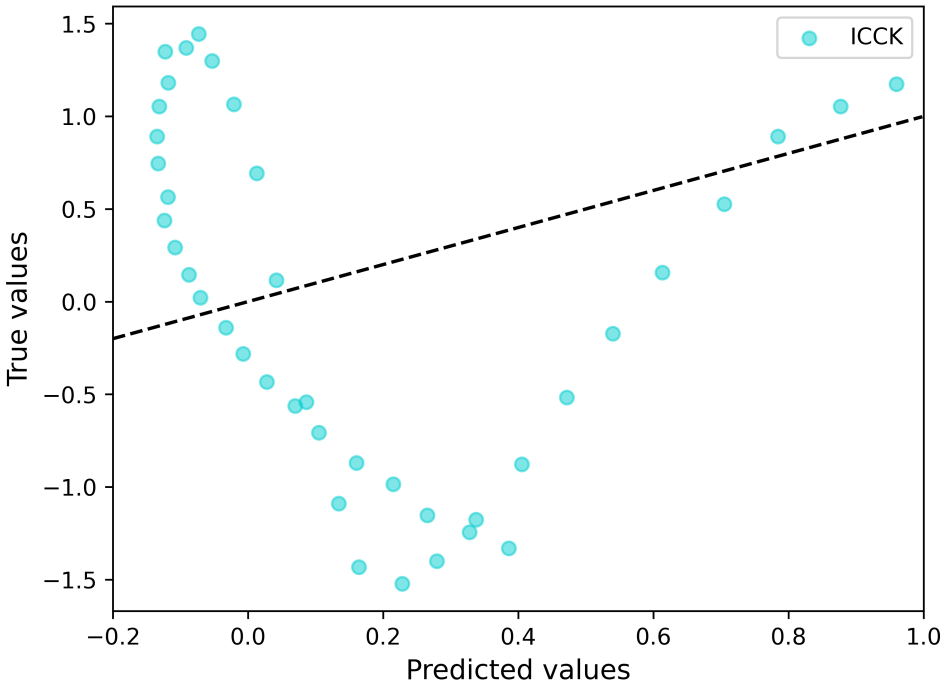


Fig. 5.10. Cross validation result for the ICCK estimation. Values refer to normalized S_u.

values are exactly the same as the true values. The RMSE between the ICCK estimation and the "observed" S_u profiles at #98 is 1.04.

Finally the comparison between the original scheme (i.e., combing MUSIC-X with ICCK) and two additional schemes in terms of the validation results are presented. For the estimation variance, since the variance mainly depends on the horizontal direction (as Fig. 5.9), the variance map can be simplified to a plot with average variance over the depth as y axis and the horizontal position as the x axis. The estimation variance comparison between these three schemes in such a figure can be viewed in Fig. 5.11. The blue line stands for the estimation variance in the scheme proposed in this study. It is evident that this scheme has the least variance along the horizontal direction, followed by the ICCK model without MUSIC-X modelling and OK model (no V_{int} data).

For the cross validation, the comparison is shown in Fig. 5.12. The blue dots, red squares and the green stars respectively stand for the cross validation results from the ICCK model with MUSIC-X modelling, the ICCK model without MUSIC-X modelling and OK model. The results in this figure are divided into two parts to discuss. The first part is the comparison between the ICCK model with MUSIC-X modelling and the OK model. As both of them use the result from MUSIC-X modelling, when discussing them, the reminder of MUSIC-X modelling is omitted. The RMSE for the ICCK and OK model is respectively 1.04 and 0.94. It should be noted that this does not indicate that it is better to use OK rather than ICCK and not to integrate geophysical data because firstly the horizontal spatial correlation used in the OK model is still from V_{int} data and secondly as shown in Fig. 5.11 the estimation variance of the OK model is pretty high. What this result indicates is actually that the cross correlation between V_{int} and S_u is not beneficial for the estimation of S_u. Since the spatial correlation is the same in these two models, the main difference lies in that the ICCK model utilizes the cross correlation between V_{int} and S_u while the OK does not use it. Further investigation has been conducted to demonstrate it. The correlation coefficient between S_u and V_{int}, which was 0.63 (as Fig. 4.14),

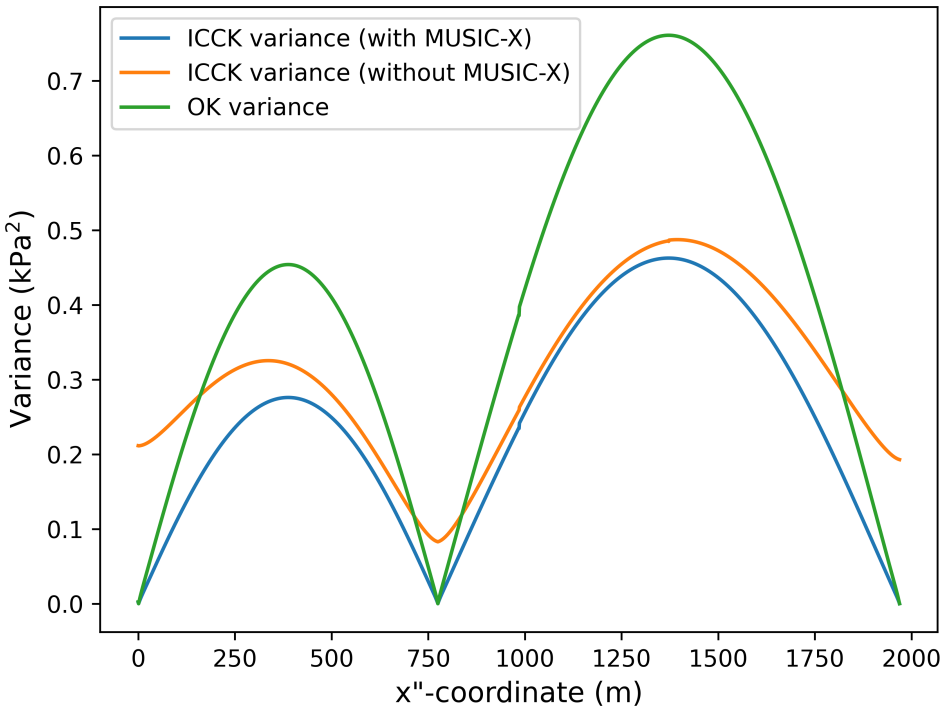


Fig. 5.11. Comparison between the three schemes in terms of estimation variance. The variance at each horizontal position is the average variance over the depth of this position.

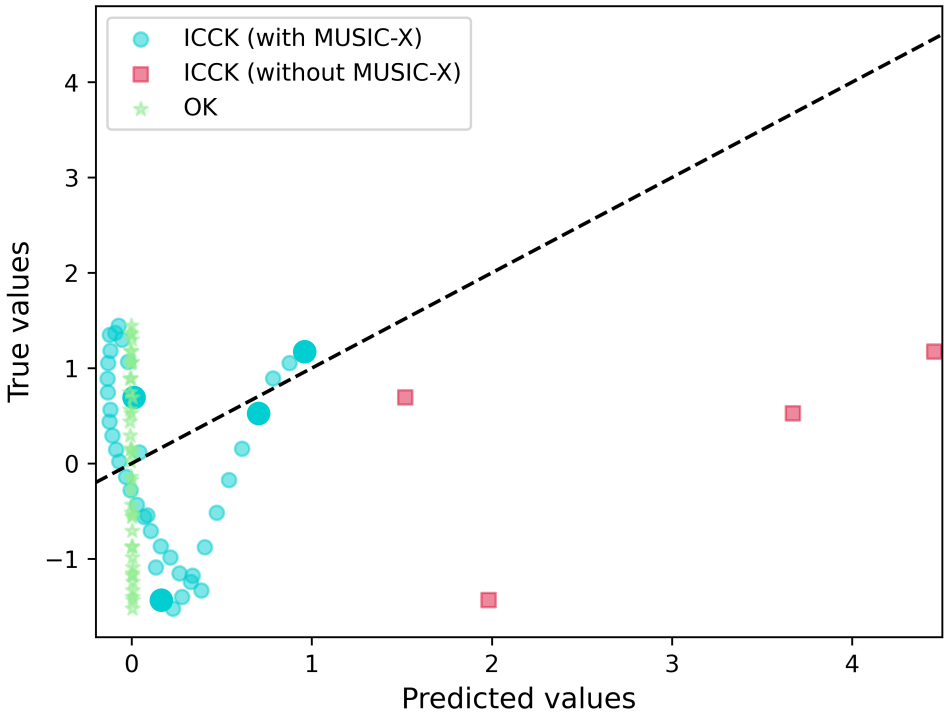


Fig. 5.12. Comparison between the three schemes in terms of cross validation result. Values refer to normalized S_u.

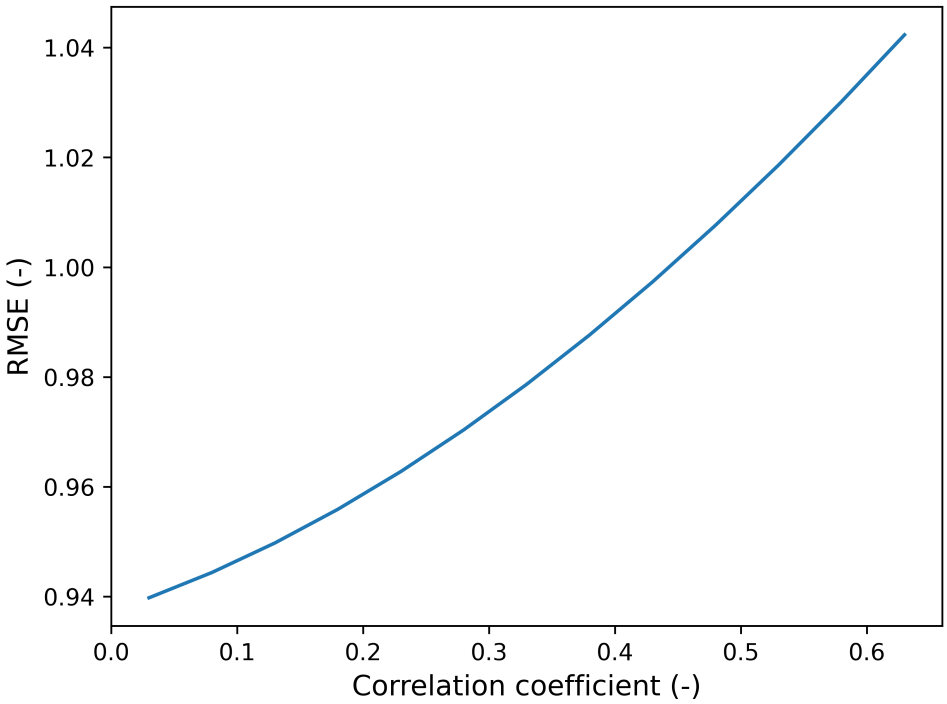


Fig. 5.13. Effect of reducing the correlation coefficient between S_u and V_{int} on the cross validation result of the original scheme.

is set to be smaller values. In this way, the estimation of S_u in the ICCK model depends less on V_{int}. The resultant RMSEs of the ICCK model with smaller correlation coefficients can be viewed in Fig. 5.13. It can be seen that as the ICCK estimation of S_u depends less on V_{int}, RMSE decreases and finally approaches to 0.94, when there is almost no correlation between S_u and V_{int} (as OK). This result accords with the founding in the MUSIC-X part that the cross correlation between S_u and V_{int} is not beneficial to reducing the uncertainty of simulated S_u but may enlarge it. The second part is the comparison between the ICCK model with MUSIC-X modelling and the ICCK model without MUSIC-X modelling. For the ICCK model without MUSIC-X modelling, there are only four observed data points at #98, so the cross validation can only be conducted based on these four data points. The cross validation result of the ICCK model with MUSIC-X modelling corresponding to these four data points are highlighted by larger size blue dots in Fig. 5.12. The RMSE for the ICCK model without MUSIC-X modelling is 2.87 which is relatively high. The RMSE for the ICCK model with MUSIC-X modelling with regards to these four points is only 0.88. This indicates the contribution of combining MUSIC-X with ICCK.

6. Conclusions

6.1. Conclusions

In this thesis a multivariate PTM, MUSIC-X, is combined with a kriging method, ICCK, to probabilistically characterize S_u in 2D space based on the fusion of geotechnical and geophysical data. It utilizes the cross correlation between S_u and other relevant geotechnical parameters and the geophysical parameter V_{int}. Meanwhile, it utilizes the spatial correlation in both vertical and horizontal direction, which are respectively estimated from q_t and V_{int} data. The performance of this scheme is demonstrated by a real case study, which shows the applicability of applying such a scheme to an offshore site characterization and how it contributes to a realistic and accurate characterization of S_u. The effectiveness of this scheme proves the soundness to combine MUSIC-X and ICCK for 2D S_u characterization and introduces a new perspective, focusing on the fusion of geotechnical and geophysical data, to characterize soil parameters in a higher dimensional spatial context. In practice, the developed S_u cross section and its quantified uncertainty can help people to better characterize S_u at the specific site thus design more efficiently and reliably. One additional contribution of this study is that a probabilistic model to transform S_u measured by PP tests to mobilized S_u has been developed.

After the investigation in the proposed scheme, the answers to the research questions in this study have been concluded:

- How can a 2D cross section of S_u be developed by a multivariate PTM based scheme?

Firstly a well-established multivariate PTM, MUSIC-X, is utilized. V_{int} is integrated to the original MUSIC-X model additionally to reduce the uncertainty of simulation results. Based on the cross correlation between S_u and other 10 geotechnical parameters and V_{int}, as well as the vertical spatial correlation estimated from q_t profiles, the MUSIC-X model can simulate 1D profiles of S_u. Among the simulation results, since the median profile of S_u has been found to involve low uncertainty by cross validation, it is regarded as an "observed" profile at a certain position. Subsequently a few nearby simulated "observed" profiles of S_u from MUSIC-X modelling are aligned to a common cross section where there are abundant 2D measurements of V_{int} data. Finally these 1D profiles of S_u and the 2D cross section of V_{int} are input into the ICCK model, respectively as the primary data and secondary data. The ICCK model can estimate the primary data at unknown positions in 2D space with the assistance of secondary data through the cross correlation between these two parameters and the spatial correlation in both vertical and horizontal direction. The vertical spatial correlation is estimated from q_t profiles as well while the horizontal spatial correlation is estimated from the V_{int} cross section. As a result, the 2D cross section of S_u can be developed. In brief, the MUSIC-X model is utilized together with the ICCK model to simulate a 2D cross section of S_u.

To complete the answer further, the performance of the scheme is concluded as well. The combination of the MUSIC-X and ICCK model is a sensible choice and indeed contributes given

that the estimated S_u cross section of this scheme has a significantly higher accuracy compared to that of the scheme without the MUSIC-X model. In addition, this scheme requires relatively low computational cost. If the horizontal spatial correlation is somehow directly integrated to the MUSIC-X model to simulate 2D S_u cross sections, the computational cost will increase dramatically. For example, when being applied to the same case study, it has to simulate around 2000 soundings in the horizontal direction and 40 soundings in the vertical direction for 12 soil parameters. Then there are in total $2000 \times 40 \times 12 = 960000$ variables of interest. The GS will unavoidably become infeasible.

- How can the horizontal spatial correlation be inferred for the scheme effectively?

The horizontal SOF, which is the key to reflecting the horizontal spatial correlation, can be estimated from the geophysical data V_{int}. The V_{int} data in the horizontal direction are abundant enough to conduct curve fitting between experimental and theoretical variograms and find the horizontal SOF. The estimated horizontal SOF has a good fitting performance so the horizontal spatial correlation can be generally accurately reflected. Additionally, V_{int} and S_u have been found to be correlated with each other at the specific site in case study so it is reasonable to assume S_u share the same horizontal SOF with V_{int} in this study. In brief, the horizontal spatial correlation estimated from V_{int} is accurate and effectively adopted by the ICCK model to estimate the cross section of S_u.

- What is the effect of utilizing V_{int} to characterize S_u?

The effect of V_{int} can be divided into two aspects to conclude. On the one hand, the cross correlation between V_{int} and S_u is not beneficial for the characterization of S_u no matter in 1D or 2D space based on the results from MUSIC-X and ICCK modelling. Although there is a moderate positive correlation between S_u and V_{int} found in the database, due to the averaging effect of V_{int}, it cannot accurately reflect the distribution of S_u especially at places where there is a sudden change in material property. Therefore, the utilization of cross correlation between S_u and V_{int} may result in the increase of uncertainty of simulated S_u.

On the other hand, the spatial correlation information inside V_{int} is valuable for S_u characterization. V_{int} can reflect the general spatial trend in the vertical direction and become a supplement to integrate the spatial correlation into the simulation of S_u when the spatial correlation is not accurately estimated. In other words, it is recommended to cross correlate S_u with V_{int} to reduce the uncertainty from the vertical spatial correlation. In addition, estimating the horizontal spatial correlation for S_u from V_{int} can be a good alternative when geotechnical testing data in the horizontal direction are limited.

6.2. Future work

The scheme proposed in this thesis succeeds in the development of 2D characterization of S_u. However there are still some limitations and more investigations can be conducted to improve the scheme in the future:

- Probabilistic trend

When estimating the spatial correlation in this study, a deterministic trend derived from linear regression is always applied to detrending. This is a commonly used method, however, recent studies (Ching & Phoon, 2017; Ching, Yoshida, et al., 2022) have found that the spatial

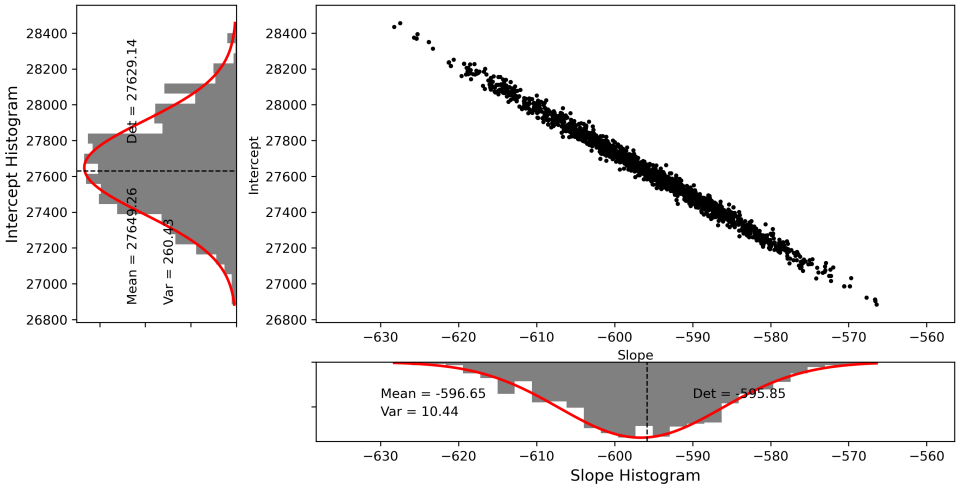


Fig. 6.1. Cross plot between the intercept and slope of the 1D linear trend. Their random samples are drawn from the posterior distribution in the Bayesian model. The marginal distribution of the slope and intercept are given in subplots.

trend also varies and detrending in a deterministic way actually involves additional uncertainties. Using probabilistic models for the trend functions is recommended for future research. A preliminary trial is shown here to give an insight into the probabilistic trend. Taking the 1D linear regression for example, it is determined by two parameters, namely the slope and intercept. A simple Bayesian model is constructed in Python based on `PyMC3` package, which is able to conduct Bayesian modelling and random sampling, to simulate these two parameters and finally develop the probabilistic trends. The process is straightforward. Firstly the prior distribution of the slope and intercept is set: normal distribution, mean equal to the deterministic slope or intercept, standard deviation equal to a large value (i.e., 10000). Secondly the likelihood function is constructed as the type of distribution, mean, standard deviation and observed data are known. The observed data refer to the three CPT profiles at #97, #98 and #104 in this study. Finally the posterior distribution of the slope and intercept is established and samples can be drawn. The distribution of random slopes and intercepts can be found in Fig. 6.1. Based on the random slopes and intercepts, the probabilistic trends can be developed, which are shown in Fig. 6.2. The "true" trend is inside the range of the trends but not necessarily to be the deterministic one.

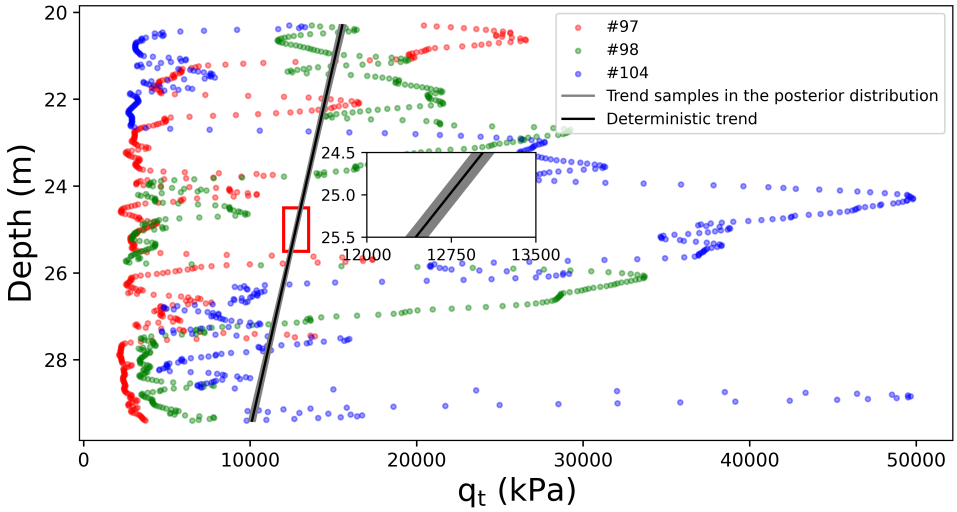


Fig. 6.2. Posterior predictive trends for the q_t profile based on q_t profiles at #97, #98 and #104.

However, there are many limitations for such a simple model. The first is that the predefined 1D linear trend may not be applicable to this site, introducing bias that will ultimately propagate to the design outcomes (Ching & Phoon, 2017). The second is that the prior mean and standard deviation should be set in a more reasonable way (e.g., non-informative priors). The third is that large computational power is required for this model. There are only two parameters to model but it takes more than 4 hours to run. Optimizations are needed to improve the computational efficiency. Finally, there is not a method proposed in the model to identify which probabilistic trend to use. For slopes and intercepts, it is irrational to take values at 5% quantile because they are not like S_u, that people can identify it should be big or small. So in this study still the deterministic trends are utilized. Anyway, this is just a simple probabilistic model, more advanced models such as Gaussian regression process and sparse Bayesian learning should be studied for the probabilistic trend in the future.

- Uncertainty integration

During the process to develop the S_u cross section, some uncertainties are eliminated manually. When transforming S_u(PP) to S_u(mob) for the MUSIC-X model, the mean values of S_u(mob) are taken as "true" values due to low uncertainty. In addition, the median of simulated S_u profiles from the MUSIC-X model are taken as "observed" profiles also due to low uncertainty. Obviously the uncertainty of the final S_u cross section is underestimated. Therefore it is recommended to integrate all possible uncertainties together in the scheme and the uncertainty of the finally developed cross section can be accurately reflected.

One proposal is shown here to enlighten how to integrate the uncertainty. Taking the uncertainty from the simulated median S_u profiles from MUSIC-X modelling for example, this uncertainty directly makes the observed primary data $Z(u_\alpha)$ used in ICCK estimation (see Eq. (3.17)), not exactly accurate. It can be imaged as the measurement error. This error can be integrated to the ICCK model by adding an error term ε_z to $Z(u_\alpha)$. ε_z can be set to be a normal variable with a zero mean and a variance equal to the value obtained from the MUSIC-X simulations. In this way, the uncertainty from the simulated S_u profiles can be integrated and ICCK estimation variance at target locations will be relatively larger. It should be noted that, the kriging matrix to derive the kriging weights will be changed accordingly. The variance of

measurement will be added to the original variogram of the primary data and thus the spatial correlation terms in the kriging matrix will be reduced.

- Extension to 3D

The characterization of S_u is more desirable if it is 3D and actually the scheme proposed in this study can be extended to 3D space. The spatial correlation between an observed data (no matter primary and secondary) and the ICCK estimation of the primary parameter, which basically reflects how this data contributes to the estimation, depends on the lag distance. The distance is a scalar, regardless of the space dimensions and the distance between two points in 3D space is still easy to calculate. So if there is an ideal database containing 3D measurements of V_{int} and relevant lab tests and CPTs are conducted in the same 3D domain, the effectiveness of this scheme can be demonstrated in 3D space.

- More suitable geophysical data

It can be seen that a considerable amount of uncertainties in the developed S_u cross section comes from the geophysical parameter V_{int}. In the future study, other geophysical parameters, which have been found to be highly correlated with S_u and beneficial for S_u characterization such as shear wave velocity, can be applied to S_u characterization in the same manner. It is expected that the superiority in accuracy of the scheme can be more evidently shown through comparing with the two alternative schemes.

A. Full conditional PDFs in MUSIC-X

This appendix first shows the specific conditional (posterior) PDFs corresponding to the four random parameters (μ_s, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$) in GS sampling in the MUSIC-X model. To make these parameters clear again, $\mathbf{X}$ is an m $\times$ n matrix, where m is the number of depths and n is the multivariate soil parameters considered in the MUSIC-X model. Within this matrix, some entries are observed (denoted as $\mathbf{X}^o$) while some are unobserved (denoted as $\mathbf{X}^u$), which is needed to simulate. The union of $\mathbf{X}^o$ and $\mathbf{X}^u$ is the complete matrix $\mathbf{X}$. The characterization parameters of $\mathbf{X}$ is the mean vector μ_s and covariance matrix $\mathbf{C}_s$. $\mathbf{a}$ = (a_1, a_2, ..., a_n) is the hyperparameter to make the prior of $\mathbf{C}_s$ non-informative. Then the complete multivariate PDF, by which these four parameters can be sampled integrally, is shown. This complete PDF is not directly used in the MUSIC-X model because it is too complicated to sample. The derivation of these PDFs is contributed by Ching and Phoon (2020).

A.1. Conditional PDF $\mathrm{P}(\boldsymbol{\mu}_\mathbf{s} \mid \mathbf{X}, \mathbf{C}_\mathbf{s}, \mathbf{a})$

$$\begin{aligned}\mathrm{P}(\boldsymbol{\mu}_s \mid \mathbf{X}, \mathbf{C}_s, \mathbf{a}) = N\{\boldsymbol{\mu}_s;\ &\left[\mathbf{C}_0^{-1} + \left(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}\right)\mathbf{C}_s^{-1}\right]^{-1}\left[\left(\mathbf{1}^T\mathbf{R}^{-1}\right)\otimes\mathbf{C}_s^{-1}\right]\mathbf{X},\\ &\left[\mathbf{C}_0^{-1} + \left(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}\right)\mathbf{C}_s^{-1}\right]^{-1}\Big\}\end{aligned} \tag{A.1}$$

where N means normal distribution; $\mathbf{C}_0$ is the prior covariance matrix for $\boldsymbol{\mu}_s$; $\mathbf{1}$ denotes an (m $\times$ 1) vector containing ones; $\mathbf{R}$ is the autocorrelation matrix. $\mathbf{C}_0$ is taken to be a diagonal matrix with very large elements (e.g., 10^4) to make it non-informative.

A.2. Conditional PDF $\mathrm{P}(\mathbf{C}_\mathbf{s} \mid \mathbf{X}, \mu_\mathbf{s}, \mathbf{a})$

$$\mathrm{P}(\mathbf{C}_\mathrm{s} \mid \mathbf{X}, \boldsymbol{\mu}_\mathrm{s}, \mathbf{a}) = \mathrm{IW}\left\{\mathbf{C}_\mathrm{s}; \boldsymbol{\Sigma} + \left[\mathrm{mat}(\mathbf{X}) - \boldsymbol{\mu}_\mathrm{s}\mathbf{1}^\mathrm{T}\right]\mathbf{R}^{-1}\left[\mathrm{mat}(\mathbf{X}) - \boldsymbol{\mu}_\mathrm{s}\mathbf{1}^\mathrm{T}\right]^\mathrm{T}, m+n+1\right\} \tag{A.2}$$

where IW means inverse-Wishart distribution and $\boldsymbol{\Sigma}$ is the scale matrix. $\boldsymbol{\Sigma} = 2(\nu_{\mathrm{IW}} - n + 1) \times diag(1/a_1, 1/a_2, ..., 1/a_n)$, where ν_{IW} is the degree of freedom and diag($\cdot$) denotes a diagonal matrix. ν_{IW} is taken to be $n+1$ to make it non-informative.

A.3. Conditional PDF $\mathrm{P}(\mathrm{a_i} \mid \mathbf{X}, \boldsymbol{\mu}_\mathbf{s}, \mathbf{C}_\mathbf{s}, \mathbf{a}_{\backslash\mathbf{i}})$

$$\mathrm{P}(a_i \mid \mathbf{X}, \boldsymbol{\mu}_\mathbf{s}, \mathbf{C}_\mathbf{s}, \mathbf{a}_{\backslash\mathbf{i}}) = \mathrm{IG}(a_i; \frac{n+2}{2}, 10^{-4} + 2\mathbf{C}_{s,ii}^{-1}) \tag{A.3}$$

where IG means the inverse-gamma distribution; $\mathbf{a}_{\backslash\mathbf{i}}$ denotes (a_1, ..., a_{i-1}, a_{i+1}, ..., a_n); $\mathbf{C}_{s,ii}^{-1}$ denotes the (i, i) entry in the $\mathbf{C}_s^{-1}$ matrix.

A.4. Conditional PDF $\mathrm{P}(\mathbf{X}^{\mathrm{u}} \mid \mathbf{X}^{\mathrm{o}}, \mu_{\mathrm{s}}, \mathbf{C}_{\mathrm{s}}, \mathbf{a})$

$$
\begin{aligned}
&P(\mathbf{X}^u \mid \mathbf{X}^o, \mu_s, \mathbf{C}_s, \mathbf{a}) = N(\mathbf{X}^u; E, COV) \\
&E(\mathbf{X}^u \mid \mathbf{X}^o, \mu_s, \mathbf{C}_s, \mathbf{a}) = (\mathbf{1} \otimes \mu_s)^u + (\mathbf{R} \otimes \mathbf{C}_s)^{uo} \times \left[(\mathbf{R} \otimes \mathbf{C}_s)^o\right]^{-1} \left[\mathbf{X}^o - (\mathbf{1} \otimes \mu_s)^o\right] \\
&\mathrm{COV}(\mathbf{X}^u \mid \mathbf{X}^o, \mu_s, \mathbf{C}_s, \mathbf{a}) = (\mathbf{R} \otimes \mathbf{C}_s)^u - (\mathbf{R} \otimes \mathbf{C}_s)^{uo} \times \left[(\mathbf{R} \otimes \mathbf{C}_s)^o\right]^{-1} \left[(\mathbf{R} \otimes \mathbf{C})^{uo}\right]^T
\end{aligned}
\tag{A.4}
$$

where $(\mathbf{1} \otimes \mu_s)^o$ and $(\mathbf{R} \otimes \mathbf{C}_s)^o$ are the mean vector and covariance matrix for $\mathbf{X}^o$; $(\mathbf{1} \otimes \mu_s)^u$ and $(\mathbf{R} \otimes \mathbf{C}_s)^u$ are the mean vector and covariance matrix for $\mathbf{X}^u$; and $(\mathbf{R} \otimes \mathbf{C}_s)^{uo}$ is the covariance matrix between $\mathbf{X}^o$ and $\mathbf{X}^u$. The vectors $(\mathbf{1} \otimes \mu_s)^o$ and $(\mathbf{1} \otimes \mu_s)^u$ can be found by partitioning $(\mathbf{1} \otimes \mu_s)$. The matrices $(\mathbf{R} \otimes \mathbf{C}_s)^o$, $(\mathbf{R} \otimes \mathbf{C}_s)^u$, and $(\mathbf{R} \otimes \mathbf{C}_s)^{uo}$ can also be found by partitioning $(\mathbf{R} \otimes \mathbf{C}_s)$.

A.5. Complete multivariate PDF

$$
\begin{aligned}
\mathrm{P}(\mathbf{X}, \boldsymbol{\mu}_{\mathrm{s}}, \mathbf{C}_{\mathrm{s}}, \mathbf{a}) &= \mathrm{P}(\mathbf{X} \mid \boldsymbol{\mu}_{\mathrm{s}}, \mathbf{C}_{\mathrm{s}}) \cdot \mathrm{P}(\boldsymbol{\mu}_{\mathrm{s}}) \cdot \mathrm{P}(\mathbf{C}_{\mathrm{s}} \mid \mathbf{a}) \cdot \mathrm{P}(\mathbf{a}) \\
&= \mathrm{P}(\mathbf{X} \mid \boldsymbol{\mu}_{\mathrm{s}}, \mathbf{C}_{\mathrm{s}}) \cdot \mathrm{N}(\boldsymbol{\mu}_{\mathrm{s}}; \boldsymbol{\mu}_0, \mathbf{C}_0) \cdot \mathrm{IW}(\mathbf{C}_{\mathrm{s}}; \boldsymbol{\Sigma}, \nu_{\mathrm{IW}}) \cdot \left[\prod_{i=1}^{n} \mathrm{IG}(a_i; \alpha, \beta)\right] \\
&= |\mathbf{R} \otimes \mathbf{C}_{\mathrm{s}}|^{-\frac{1}{2}} \cdot (2\pi)^{-\frac{mn}{2}} \cdot \mathrm{e}^{-\frac{1}{2}(\mathbf{X} - \mathbf{1} \otimes \boldsymbol{\mu}_{\mathrm{s}})^{\mathrm{T}} (\mathbf{R} \otimes \mathbf{C}_{\mathrm{s}})^{-1} (\mathbf{X} - \mathbf{1} \otimes \boldsymbol{\mu}_{\mathrm{s}})} \\
&\times |\mathbf{C}_0|^{-\frac{1}{2}} \cdot (2\pi)^{-\frac{n}{2}} \cdot \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}_{\mathrm{s}} - \boldsymbol{\mu}_0)^{\mathrm{T}} \mathbf{C}_0^{-1} (\boldsymbol{\mu}_{\mathrm{s}} - \boldsymbol{\mu}_0)} \\
&\times \frac{|\boldsymbol{\Sigma}|^{\nu_{\mathrm{IW}}/2}}{2^{n \times \nu_{\mathrm{IW}}/2} \cdot \Gamma_n(\nu_{\mathrm{IW}}/2)} \cdot |\mathbf{C}_{\mathrm{s}}|^{-\frac{\nu_{\mathrm{IW}} + n + 1}{2}} \cdot \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma} \times \mathbf{C}_{\mathrm{s}}^{-1}\right)} \times \left[\prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot a_i^{-\alpha - 1} \cdot \mathrm{e}^{-\frac{\beta}{a_i}}\right] \\
&\propto |\mathbf{R} \otimes \mathbf{C}_{\mathrm{s}}|^{-\frac{1}{2}} \cdot |\mathbf{C}_{\mathrm{s}}|^{-\frac{2n+2}{2}} \cdot |\boldsymbol{\Sigma}|^{(n+1)/2} \\
&\cdot \left(\prod_{i=1}^{n} a_i^{-\alpha - 1}\right) \mathrm{e}^{-\frac{1}{2}(\mathbf{X} - \mathbf{1} \otimes \boldsymbol{\mu}_{\mathrm{s}})^{\mathrm{T}} (\mathbf{R} \otimes \mathbf{C}_{\mathrm{s}})^{-1} (\mathbf{X} - \mathbf{1} \otimes \boldsymbol{\mu}_{\mathrm{s}}) - \frac{1}{2} \boldsymbol{\mu}_{\mathrm{s}}^{\mathrm{T}} \mathbf{C}_0^{-1} \boldsymbol{\mu}_{\mathrm{s}} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma} \times \mathbf{C}_{\mathrm{s}}^{-1}\right) - \sum_{i=1}^{n} \frac{\beta}{a_i}}
\end{aligned}
\tag{A.5}
$$

where μ_0 is the prior mean vector for $\boldsymbol{\mu}_s$, set to be a zero vector; $\Gamma_n(\cdot)$ is the multivariate gamma function with dimension = n; tr$(\cdot)$ is the matrix trace; α and β are respectively the shape parameter and scale parameter for the IG distribution of $\mathbf{a}$, set to be 0.5 and 10^{-4}.

References

Almeida, A. S., & Journel, A. G. (1994). Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology*, *26*(5), 565–588. https://doi.org/10.1007/BF02089242

Babak, O., & Deutsch, C. V. (2009a). An intrinsic model of coregionalization that solves variance inflation in collocated cokriging. *Computers Geosciences*, *35*(3), 603–614. https://doi.org/https://doi.org/10.1016/j.cageo.2008.02.025

Babak, O., & Deutsch, C. V. (2009b). Improved spatial modeling by merging multiple secondary data for intrinsic collocated cokriging. *Journal of Petroleum Science and Engineering*, *69*(1), 93–99. https://doi.org/10.1016/j.petrol.2009.08.001

Berre, T. (1982). Triaxial Testing at the Norwegian Geotechnical Institute. *Geotechnical Testing Journal*, *5*(1/2), 3–17. https://doi.org/10.1520/GTJ10794J

Budak, T. O., Gurbuz, A., & Eksioglu, B. (2022). Practical transitions among undrained shear strengths of remolded samples from pocket penetrometer tests and other laboratory tests. *CATENA*, *213*, 106148. https://doi.org/10.1016/j.catena.2022.106148

Cami, B., Javankhoshdel, S., Phoon, K.-K., & Ching, J. (2020). Scale of fluctuation for spatially varying soils: Estimation methods and values. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *6*(4), 03120002. https://doi.org/10.1061/AJRUA6.0001083

Cao, Z., & Wang, Y. (2014). Bayesian Model Comparison and Characterization of Undrained Shear Strength. *Journal of Geotechnical and Geoenvironmental Engineering*, *140*(6), 04014018. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001108

Ching, J., & Phoon, K.-K. (2012). Establishment of generic transformations for geotechnical design parameters. *Structural Safety*, *35*, 52–62. https://doi.org/10.1016/j.strusafe.2011.12.003

Ching, J., & Phoon, K.-K. (2014a). Correlations among some clay parameters — the multivariate distribution. *Canadian Geotechnical Journal*, *51*(6), 686–704. https://doi.org/10.1139/cgj-2013-0353

Ching, J., & Phoon, K.-K. (2014b). Transformations and correlations among some clay parameters — the global database. *Canadian Geotechnical Journal*, *51*(6), 663–685. https://doi.org/10.1139/cgj-2013-0262

Ching, J., & Phoon, K.-K. (2015). Reducing the Transformation Uncertainty for the Mobilized Undrained Shear Strength of Clays. *Journal of Geotechnical and Geoenvironmental Engineering*, *141*(2), 04014103. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001236

Ching, J., & Phoon, K.-K. (2017). Characterizing Uncertain Site-Specific Trend Function by Sparse Bayesian Learning. *Journal of Engineering Mechanics*, *143*(7), 04017028. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001240

Ching, J., & Phoon, K.-K. (2019a). Constructing Site-Specific Multivariate Probability Distribution Model Using Bayesian Machine Learning. *Journal of Engineering Mechanics*, *145*(1), 04018126. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001537

Ching, J., & Phoon, K.-K. (2019b). Impact of Autocorrelation Function Model on the Probability of Failure. *Journal of Engineering Mechanics*, *145*(1), 04018123. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001549

Ching, J., & Phoon, K.-K. (2020). Constructing a Site-Specific Multivariate Probability Distribution Using Sparse, Incomplete, and Spatially Variable (MUSIC-X) Data. *Journal of Engineering Mechanics*, *146*(7), 04020061. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001779

Ching, J., Phoon, K.-K., & Lee, W.-T. (2013). Second-Moment Characterization of Undrained Shear Strengths from Different Test Procedures, 308–320. https://doi.org/10.1061/9780784412763.025

Ching, J., Phoon, K.-K., Yang, Z., & Stuedlein, A. W. (2022). Quasi-site-specific multivariate probability distribution model for sparse, incomplete, and three-dimensional spatially varying soil data. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, *16*(1), 53–76. https://doi.org/10.1080/17499518.2021.1971256

Ching, J., Wu, T.-J., Stuedlein, A. W., & Bong, T. (2018). Estimating horizontal scale of fluctuation with limited CPT soundings. *Geoscience Frontiers*, *9*(6), 1597–1608. https://doi.org/10.1016/j.gsf.2017.11.008

Ching, J., Yoshida, I., & Phoon, K.-K. (2022). Comparison of trend models for geotechnical spatial variability: Sparse Bayesian Learning vs. Gaussian Process Regression. *Gondwana Research*. https://doi.org/10.1016/j.gr.2022.07.011

Crawford, M. M., Bryson, L. S., Woolery, E. W., & Wang, Z. (2018). Using 2-D electrical resistivity imaging for joint geophysical and geotechnical characterization of shallow landslides. *Journal of Applied Geophysics*, *157*, 37–46. https://doi.org/10.1016/j.jappgeo.2018.06.009

Das, S., Samui, P., Khan, S., & Sivakugan, N. (2011). Machine learning techniques applied to prediction of residual strength of clay. *Open Geosciences*, *3*(4), 449–461. https://doi.org/10.2478/s13533-011-0043-1

D'Ignazio, M., Phoon, K.-K., Tan, S. A., & Länsivaara, T. T. (2016). Correlations for undrained shear strength of finnish soft clays. *Canadian Geotechnical Journal*, *53*(10), 1628–1645. https://doi.org/10.1139/cgj-2016-0037

Duan, W., Cai, G., Liu, S., & Puppala, A. J. (2019). Correlations between Shear Wave Velocity and Geotechnical Parameters for Jiangsu Clays of China. *Pure and Applied Geophysics*, *176*(2), 669–684. https://doi.org/10.1007/s00024-018-2011-x

Feng, X., & Jimenez, R. (2015). Estimation of deformation modulus of rock masses based on Bayesian model selection and Bayesian updating approach. *Engineering Geology*, *199*, 19–27. https://doi.org/10.1016/j.enggeo.2015.10.002

Fenton, G. A., Griffiths, D. V., & Williams, M. B. (2005). Reliability of traditional retaining wall design. *Géotechnique*, *55*(1), 55–62. https://doi.org/10.1680/geot.2005.55.1.55

Fenton, G. A., & Griffiths, D. V. (2003). Bearing-capacity prediction of spatially random c soils. *Canadian Geotechnical Journal*, *40*(1), 54–65. https://doi.org/10.1139/t02-086

Hansbo, S. (1957). *New approach to the determination of the shear strength of clay by the fall-cone test*. Statens geotekniska institut.

Hegazy, Y. A., & Mayne, P. W. (2006). A global statistical correlation between shear wave velocity and cone penetration data. In *Site and geomaterial characterization* (pp. 243–248).

Hicks, M. A., & Samy, K. (2002). Influence of heterogeneity on undrained clay slope stability. *Quarterly Journal of Engineering Geology and Hydrogeology*, *35*(1), 41–49.

Hong, S., Lee, M., Kim, J., & Lee, W. (2010). Evaluation of undrained shear strength of busan clay using cpt. *Proceedings of the 2nd International Symposium on Cone Penetration Testing, CPT*, *10*.

Hu, Y., & Wang, Y. (2020). Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation. *Computers and Geotechnics*, *124*, 103634. https://doi.org/10.1016/j.compgeo.2020.103634

Huang, J., Zheng, D., Li, D.-Q., Kelly, R., & Sloan, S. W. (2018). Probabilistic characterization of two-dimensional soil profile by integrating cone penetration test (CPT) with multi-channel analysis of surface wave (MASW) data. *Canadian Geotechnical Journal*, *55*(8), 1168–1181. https://doi.org/10.1139/cgj-2017-0429

Hussien, M. N., & Karray, M. (2016). Shear wave velocity as a geotechnical parameter: An overview. *Canadian Geotechnical Journal*, *53*(2), 252–272. https://doi.org/10.1139/cgj-2014-0524

JAMIOLKOWSKI, M., Ladd, C., Germaine, J., & LANCELLOTTA, R. (1985). New developments in field and laboratory testing of soils. proceedings of the eleventh international conference on soil mechanics and foundation engineering, san francisco, 12-16 august 1985. *Publication of: Balkema (AA)*.

Kanungo, D. P., Sharma, S., & Pain, A. (2014). Artificial Neural Network (ANN) and Regression Tree (CART) applications for the indirect estimation of unsaturated soil shear strength parameters. *Frontiers of Earth Science*, *8*(3), 439–456. https://doi.org/10.1007/s11707-014-0416-0

Karlsrud, K., & Hernandez-Martinez, F. G. (2013). Strength and deformation properties of Norwegian clays from laboratory tests on high-quality block samples. *Canadian Geotechnical Journal*, *50*(12), 1273–1293. https://doi.org/10.1139/cgj-2013-0298

Kulhawy, F. H., & Mayne, P. W. (1990). *Manual on estimating soil properties for foundation design* (tech. rep.). Electric Power Research Inst., Palo Alto, CA (USA); Cornell Univ., Ithaca, NY (USA). Geotechnical Engineering Group.

Lacasse, S., & Berre, T. (1988). Triaxial testing methods for soils. *Advanced triaxial testing of soil and rock, ASTM STP*, *977*, 264–289.

Ladd, C. C., & Foott, R. (1974). New Design Procedure for Stability of Soft Clays. *Journal of the Geotechnical Engineering Division*, *100*(7), 763–786. https://doi.org/10.1061/AJGEB6.0000066

Li, D.-Q., Qi, X.-H., Cao, Z.-J., Tang, X.-S., Zhou, W., Phoon, K.-K., & Zhou, C.-B. (2015). Reliability analysis of strip footing considering spatially variable undrained shear strength that linearly increases with depth. *Soils and Foundations*, *55*(4), 866–880. https://doi.org/10.1016/j.sandf.2015.06.017

Ly, H.-B., & Pham, B. T. (2020). Prediction of Shear Strength of Soil Using Direct Shear Test and Support Vector Machine Model. *The Open Construction & Building Technology Journal*, *14*(1). https://doi.org/10.2174/1874836802014010268

Mayne, P. W. (1985). A Review of Undrained Strength in Direct Simple Shear. *Soils and Foundations*, *25*(3), 64–72. https://doi.org/10.3208/sandf1972.25.3_64

Mbarak, W. K., Cinicioglu, E. N., & Cinicioglu, O. (2020). SPT based determination of undrained shear strength: Regression models and machine learning. *Frontiers of Structural and Civil Engineering*, *14*(1), 185–198. https://doi.org/10.1007/s11709-019-0591-x

Mesri, G. (1975). Discussion of "New Design Procedure for Stability of Soft Clays". *Journal of the Geotechnical Engineering Division*, *101*(4), 409–412. https://doi.org/10.1061/AJGEB6.0005026

Müller, S., Schüler, L., Zech, A., & Heße, F. (2021). Gstools v1. 3: A toolbox for geostatistical modelling in python. *Geoscientific Model Development Discussions*, *2021*, 1–33.

Nakase, A., & Kamei, T. (1983). Undrained Shear Strength Anisotropy of Normally Consolidated Cohesive Soils. *Soils and Foundations*, *23*(1), 91–101. https://doi.org/10.3208/sandf1972.23.91

Nascimento, D. S. C., Coelho, A. L. V., & Canuto, A. M. P. (2014). Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study. *Neurocomputing*, *138*, 347–357. https://doi.org/10.1016/j.neucom.2014.01.027

Ng, I.-T., Yuen, K.-V., & Lau, C.-H. (2015). Predictive model for uniaxial compressive strength for Grade III granitic rocks from Macao. *Engineering Geology*, *199*, 28–37. https://doi.org/10.1016/j.enggeo.2015.10.008

Nguyen, T. S., & Likitlersuang, S. (2021). Influence of the Spatial Variability of Soil Shear Strength on Deep Excavation: A Case Study of a Bangkok Underground MRT Station. *International Journal of Geomechanics*, *21*(2), 04020248. https://doi.org/10.1061/(ASCE)GM.1943-5622.0001914

Omar, T., & Sadrekarimi, A. (2014). Effects of multiple corrections on triaxial compression testing of sands. *Journal of GeoEngineering*, *9*(2), 75–83.

Otoko, G. R., Manuel, I., Igwagu, M., & Edoh, C. (2016). Empirical cone factor for estimation of undrained shear strength. *Electronic Journal of Geotechnical Engineering*, *21*(18), 6069–6076.

Ou, C., & Liao, J. (1987). Geotechnical engineering research report. *Rep. No. GT96008. Taipei, Taiwan: National Taiwan Univ. of Science and Technology*.

Peri, E., Ibsen, L. B., & Nielsen, B. V. N. (2019). Influence of sample slenderness and boundary conditions in triaxial test–a review. *7th International Symposium on Deformation Characteristics of Geomaterials, IS-Glasgow 2019*, 02009.

Pham, B. T., Qi, C., Ho, L. S., Nguyen-Thoi, T., Al-Ansari, N., Nguyen, M. D., Nguyen, H. D., Ly, H.-B., Le, H. V., & Prakash, I. (2020). A Novel Hybrid Soft Computing Model Using Random Forest and Particle Swarm Optimization for Estimation of Undrained Shear Strength of Soil. *Sustainability*, *12*(6), 2218. https://doi.org/10.3390/su12062218

Pham, B. T., Son, L. H., Hoang, T.-A., Nguyen, D.-M., & Tien Bui, D. (2018). Prediction of shear strength of soft soil using machine learning methods. *CATENA*, *166*, 181–191. https://doi.org/10.1016/j.catena.2018.04.004

Phoon, K.-K. (2018). Probabilistic site characterization. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *4*(4), 02018002.

Phoon, K.-K., Ching, J., & Shuku, T. (2022). Challenges in data-driven site characterization. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, *16*(1), 114–126. https://doi.org/10.1080/17499518.2021.1896005

Phoon, K.-K., & Kulhawy, F. H. (1999). Evaluation of geotechnical property variability. *Canadian Geotechnical Journal*, *36*(4), 625–639.

Phoon, K.-K., & Kulhawy, F. H. (2001). Characterization of geotechnical variability and Evaluation of geotechnical property variability: Reply. *Canadian Geotechnical Journal*, *38*(1), 214–215. https://doi.org/10.1139/t00-090

Rémai, Z. (2013). Correlation of undrained shear strength and CPT resistance. *Periodica Polytechnica Civil Engineering*, *57*(1), 39–44. https://doi.org/10.3311/PPci.2140

Rezaei, S., Shooshpasha, I., & Rezaei, H. (2018). Empirical Correlation between Geotechnical and Geophysical Parameters in a Landslide Zone (Case Study: Nargeschal Landslide). *Earth Sciences Research Journal*, *22*(3), 195–204. https://doi.org/10.15446/esrj.v22n3.69491

Robertson, P. K. (2010). Soil behaviour type from the cpt: An update. *2nd International symposium on cone penetration testing*, *2*(56), 8.

Samson, M., & Deutsch, C. (2020). Collocated cokriging. *Learning*.

Samui, P., & Kurup, P. (2012). Multivariate Adaptive Regression Spline and Least Square Support Vector Machine for Prediction of Undrained Shear Strength of Clay. *International Journal of Applied Metaheuristic Computing (IJAMC)*, *3*(2), 33–42. https://doi.org/10.4018/jamc.2012040103

Schematic diagram of UHR-MCS. (n.d.). http://geostar-surveys.com/map%20-methodology%20-%20High%20Resolution%20Seismic%20surveys.html

Shahin, M. A. (2013). Artificial intelligence in geotechnical engineering: Applications, modeling aspects, and future directions. *Metaheuristics in water, geotechnical and transport engineering*, *169204*.

Terzaghi, K., Peck, R. B., & Mesri, G. (1996). *Soil mechanics in engineering practice*. John wiley & sons.

Thakur, V. K. S., Fauskerud, O. A., Gjelsvik, V., Christensen, S. O., Oset, F., Nordal, S., Viklund, M., & Strand, S.-A. (2016). A procedure for the assessment of the undrained shear strength profile of soft clays. *Proceedings of the 17th Nordic Geotechnical Meeting*.

Trafford, A., & Long, M. (2020). Relationship between Shear-Wave Velocity and Undrained Shear Strength of Peat. *Journal of Geotechnical and Geoenvironmental Engineering*, *146*(7), 04020057. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002298

van Beers, W. C. M., & Kleijnen, J. P. C. (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society*, *54*(3), 255–262. https://doi.org/10.1057/palgrave.jors.2601492

Wang, Y., & Zhao, T. (2017). Statistical interpretation of soil property profiles from sparse data using bayesian compressive sampling. *Géotechnique*, *67*(6), 523–536. https://doi.org/10.1680/jgeot.16.P.143

Wang, Y., & Akeju, O. V. (2016). Quantifying the cross-correlation between effective cohesion and friction angle of soil from limited site-specific data. *Soils and Foundations*, *56*(6), 1055–1070. https://doi.org/10.1016/j.sandf.2016.11.009

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

Wang, Y., & Zhao, T. (2016). Interpretation of soil property profile from limited measurement data: A compressive sampling perspective. *Canadian Geotechnical Journal*, *53*(9), 1547–1559. https://doi.org/10.1139/cgj-2015-0545

Wang, Y., Zhao, T., & Phoon, K.-K. (2018). Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. *Canadian Geotechnical Journal*, *55*(6), 862–880. https://doi.org/10.1139/cgj-2017-0254

Xie, J., Huang, J., Lu, J., Burton, G. J., Zeng, C., & Wang, Y. (2022). Development of two-dimensional ground models by combining geotechnical and geophysical data. *Engineering Geology*, *300*, 106579. https://doi.org/10.1016/j.enggeo.2022.106579

Xu, J., Wang, Y., & Zhang, L. (2022). Fusion of geotechnical and geophysical data for 2D subsurface site characterization using multi-source Bayesian compressive sampling. *Canadian Geotechnical Journal*, *59*(10), 1756–1773. https://doi.org/10.1139/cgj-2021-0323

Yan, W. M., Yuen, K.-V., & Yoon, G. L. (2009). Bayesian Probabilistic Approach for the Correlations of Compression Index for Marine Clays. *Journal of Geotechnical and Geoenvi-*

ronmental Engineering, *135*(12), 1932–1940. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000157

Yu, B. (2022). *Machine learning for prediction of undrained shear strength from cone penetration test data* (Master's thesis). Delft University of Technology. http://resolver.tudelft.nl/uuid:f5d58ecc-2b15-46c1-8ecb-1f56e1edf0f2

Zein, A. K. M. (2017). Estimation of undrained shear strength of fine grained soils from cone penetration resistance. *International Journal of Geo-Engineering*, *8*(1), 9. https://doi.org/10.1186/s40703-017-0046-y

Zhang, L., Li, D.-Q., Tang, X.-S., Cao, Z.-J., & Phoon, K.-K. (2018). Bayesian model comparison and characterization of bivariate distribution for shear strength parameters of soil. *Computers and Geotechnics*, *95*, 110–118. https://doi.org/10.1016/j.compgeo.2017.10.003

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

Zhao, T., & Wang, Y. (2018). Simulation of cross-correlated random field samples from sparse measurements using Bayesian compressive sensing. *Mechanical Systems and Signal Processing*, *112*, 384–400. https://doi.org/10.1016/j.ymssp.2018.04.042

Zuada Coelho, B., & Karaoulis, M. (2022). Data fusion of geotechnical and geophysical data for three-dimensional subsoil schematisations. *Advanced Engineering Informatics*, *53*, 101671. https://doi.org/10.1016/j.aei.2022.101671