Case Histories on 2D/3D Underground Stratification Using Sparse Machine Learning

Takayuki Shuku, *Associate Professor, Dept. of Environmental Management Engineering, Okayama University, Okayama, Japan; email: shuku@cc.okayama-u.ac.jp*
Jianye Ching, *Distinguished Professor, Dept. of Civil Engineering, National Taiwan University, Taipei, Taiwan; email: jyching@gmail.com*

ABSTRACT: *This paper showcases novel underground stratification based on sparse machine learning (SML) methods, including sparse Bayesian learning (SBL) and least absolute shrinkage selection operator (Lasso). The SML methods proposed by the authors were applied to two- and three-dimensional underground stratification of actual sites, Odagawa Riverbank (Okayama, Japan) and New Lock (Terneuzen, the Netherlands), to demonstrate their performances. Cone penetration test (CPT) data were available in both sites, and they were converted to the soil behavior type (SBT) index for the underground stratification analysis. The trends of I_c/SBT profiles and their distribution colormaps estimated by the two methods were compared to discuss the methodological characteristics. For the Odagawa Riverbank case, the detection ratio of SBT obtained by the two methods was also compared to investigate the estimation accuracy in terms of stratification ability.*

KEYWORDS: underground stratification, sparse machine learning, cone penetration test, soil stratification, soil behavior type index.

SITE LOCATION: Geo-Database

INTRODUCTION

In many countries across the world, the subsurface space is gradually becoming an integral part of urban planning. This requires adequate subsurface models for the design of geotechnical structures such as foundations of buildings, bridges, tunnels, etc. In particular, the need for three-dimensional (3D) subsurface modeling is rapidly increasing nowadays, because it provides more spatial insights, more precise and objective representations of real-world phenomenon, and better interpretation of spatial relations. A necessary step in subsurface modeling is underground stratification.

Underground stratification based on data can be formulated as a machine learning task, and several machine learning (ML) methods can be useful for underground stratification. One of the ML methods known as sparse machine learning (SML) (e.g., MacKay 1992; Tibshirani 1996; Tipping 2001) has recently received much attention for its ability of managing several data processing and mining tasks. According to the general principle of sparsity, a phenomenon should be represented with as few variables as possible. This approach, which essentially favors simple models over more complex ones, is central to many research fields, and it can also be promising in geotechnical engineering.

Ching and Phoon (2017) proposed a method for characterizing one-dimensional (1D) spatial variation of soil property in the depth direction based on Sparse Bayesian Learning (SBL, Mackey 1992; Tipping 2001). This method characterizes three types of uncertainties: (1) the functional form (shape) of the trend function; (2) the parameters of the trend function (e.g., intercept and gradient); and (3) the random field parameters describing spatial variation about the trend function, namely standard deviation (σ) and scale of fluctuation (δ), within a consistent Bayesian framework. Recently, they extended this SBL method to 3D settings (Ching et al. 2020) and non-lattice data set (Ching et al. 2021). Shuku (2019) and Shuku et al. (2020) focused on another SML method, the least absolute shrinkage selection operator (Lasso, Tibshirani 1996), and developed a method for consistently estimating trends and detecting layer boundaries in depth-dependent soil data. An extension of the

Submitted: 30 October 2020; Published: 25 October 2021
Reference: Shuku T., and Ching J. (2021). Case Histories on 2D/3D Underground Stratification Using Sparse Machine Learning. International Journal of Geoengineering Case Histories, Volume 6, Issue 4, pp. 35-47, doi: 10.4417/IJGCH-06-04-03

Lasso method to practical 3D settings was achieved by Shuku and Phoon (2021). SML-based methods have recently received much attention in geotechnical engineering: Wang and Zhao (2017) proposed Bayesian compressive sampling for estimating soil property profile and its uncertainty using limited number of data, and Hu et al. (2020) and Zhao and Wang (2020) applied Bayesian supervised learning to interpolation and stratification of multi-layer soil property profile.

Many past studies on underground stratification mainly focused on 1D settings, and there is limited research on multi-dimensional underground stratification despite its significant demand (Hu and Wang 2020; Wang et al. 2020). The purpose of this paper is to showcase real case histories on 2D/3D underground stratification using the latest SML methods proposed by the authors and to compare between the two methods based on the "law of parsimony." The main purpose of this paper is not to present the details of the theories and derivations. Please refer to the articles published by the authors for more theoretical details.

This paper is structured as follows. Section 2 briefly introduces the two SML methods. Section 3 outlines two real sites for underground stratification, Odagawa Riverbank site (Okayama, Japan) for a 2D example and New Lock site (Terneuzen, the Netherlands) for a 3D example. Section 4 shows the results of underground stratification by the two methods and briefly discusses the differences between the two methods. The summary is presented in Section 5.

SPARSE MACHINE LEARNING

This section outlines the two SML methods, SBL (MacKay 1992; Tipping 2001) and Lasso (Tibshirani 1996; Shuku 2019; Shuku et al. 2020; Shuku and Phoon 2021). For simplicity, we focus on a 1D setting herein. The details for 2D/3D settings can be found in Ching and Phoon (2017) and Ching et al. (2020, 2021) for SBL and in Shuku (2019), Shuku et al. (2020), and Shuku and Phoon (2020) for Lasso.

Suppose that we are given the 1D dataset of $\{z_i, y_i\}_{i=1}^N$ where z_i is depth, y_i is soil property, and N is the number of data points. Let us denote $\mathbf{y} = [y_1\ y_2\ \dots\ y_N]^T \in \mathbf{R}^{N\times 1}$, and $\mathbf{z} = [z_1\ z_2\ \dots\ z_N]^T \in \mathbf{R}^{N\times 1}$. The trend function, $\mathbf{t}(\mathbf{z},\mathbf{w}) \in \mathbf{R}^{N\times 1}$, is modeled as the linear combination of a collection of basis functions (BFs):

$$\mathbf{t}(\mathbf{z},\mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{z}) \tag{1}$$

where w_j is an unknown coefficient; $\mathbf{w} = [w_1\ \dots\ w_M]^T \in \mathbf{R}^{M\times 1}$; ϕ_j is the j^{th} BF; and $\phi_j(\mathbf{z}) = [\phi_j(z_1)\ \dots\ \phi_j(z_N)]^T \in \mathbf{R}^{N\times 1}$. There are many possible choices for the BFs, such as polynomials, sigmoidal, wavelet, and Legendre polynomials (Bishop 2006). The data y is modeled as the summation between the trend and noise:

$$\mathbf{y} = \mathbf{t}(\mathbf{z},\mathbf{w}) + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\varepsilon} \in \mathbf{R}^{N\times 1}$ is a zero mean Gaussian random noise with variance σ^2. The likelihood function of the model is given by:

$$p(\mathbf{y} \mid \mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\right\|_2^2\right\} \tag{3}$$

where $\boldsymbol{\Phi}$ is an $N \times M$ matrix, called the design matrix:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(z_1) & \phi_2(z_1) & \cdots & \phi_M(z_1) \\ \phi_1(z_2) & \phi_2(z_2) & & \phi_M(z_2) \\ \vdots & \vdots & & \vdots \\ \phi_1(z_N) & \phi_2(z_N) & \cdots & \phi_M(z_N) \end{bmatrix} \tag{4}$$

A straightforward approach to estimate $\mathbf{w}$ and σ^2 is by maximizing Eq. (3). This, however, usually leads to an excessively complex model that over-fits the data. The Bayesian approach has been widely used to avoid this over-fitting problem. Based on Bayes' rule, the posterior probability density function (PDF) of $\mathbf{w}$ and σ, $p(\mathbf{w}, \sigma \mid \mathbf{y})$, is given by:

$$p(\mathbf{w},\sigma\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{w},\sigma)\,p(\mathbf{w},\sigma)}{p(\mathbf{y})} \tag{5}$$

where $p(\mathbf{w}, \sigma)$ is the prior PDF of $\mathbf{w}$ and σ, and $p(\mathbf{y})$ is the PDF of $\mathbf{y}$. Both SBL and Lasso can be derived from Bayes' rule, Eq. (5), and the fundamental concepts of the methods are given in the following sub-sections.

Sparse Bayesian Learning (SBL)

The idea of SBL was originally proposed by MacKay (1992). In SBL, a zero-mean Gaussian prior distribution is assigned for $\mathbf{w}$:

$$p(\mathbf{w}\,|\,\mathbf{s}) = \prod_{i=1}^{M} N(w_i\,|\,0, s_i) \tag{6}$$

where s_i represents the standard deviation of w_i, and $\mathbf{s}$ denotes $(s_1, s_2, \ldots, s_M)^{\mathrm{T}}$. The optimal values of $\mathbf{s}$ and σ are determined by maximizing the following marginal likelihood function (Tipping 2001):

$$p(\mathbf{y}\,|\,\mathbf{s},\sigma^2) = \int p(\mathbf{y}\,|\,\mathbf{w},\sigma^2)\,p(\mathbf{w}\,|\,\mathbf{s})\mathrm{d}\mathbf{w} = (2\pi)^{-N/2}\,|\,\sigma^2\mathbf{I} + \mathbf{\Phi\Omega\Phi}^{\mathrm{T}}\,|^{-1/2}\,\exp\left\{-\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\sigma^2\mathbf{I} + \mathbf{\Phi\Omega\Phi}^{\mathrm{T}})^{-1}\mathbf{y}\right\} \tag{7}$$

where $\mathbf{I}$ is identity matrix, and $\mathbf{\Omega} = \mathrm{diag}\{s_1^2, s_2^2, \ldots, s_M^2\}$. In the resulting optimal $(s_1, s_2, \ldots, s_M)$, many s_i's go to zero (Tipping 2001). The BFs associated with these zero s_i's play no role in the predictions and so are pruned out, resulting in a sparse model.

Ching and Phoon (2017) re-formulated the original SBL proposed by Tipping (2001) to incorporate correlated noise as well as to quantify the statistical uncertainties in the trend function and scale of fluctuation. Ching et al (2020) extended the SBL method to 3D, and this SBL can also simulate conditional random fields of $\mathbf{y}$. Moreover, Ching et al (2021) further revised the SBL method such that it can handle incomplete sounding data.

Least Absolute Shrinkage Selection Operator (Lasso)

Lasso was originally proposed by Tibshirani (1996) and has been widely used in statistical science and image/vision analysis. In Lasso, the following variables $\{w_1, w_2, \ldots, w_n\}^{\mathrm{T}}$ are used to discretize the trend function into a piecewise function:

$$\mathbf{t} = \mathbf{w} = \{w_1, w_2, \ldots, w_n\}^{\mathrm{T}} \tag{8}$$

Namely, $y_i = t_i + \varepsilon_i = w_i + \varepsilon_i$. In addition, Lasso assumes the following prior PDF for $\mathbf{w}$:

$$p(\mathbf{w}\,|\,\kappa) \propto \exp\left(-\kappa\|\mathbf{w}\|_1\right) \tag{9}$$

where κ is the diversity parameter in the Laplace prior PDF. In Eq. (5), the normalization term $p(\mathbf{y})$ is often left out, and if the residual $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{t}$ is independent of $\mathbf{w}$, the posterior PDF of $(\mathbf{w}, \kappa)$ can be written as:

$$p(\mathbf{w},\kappa\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\mathbf{w},\kappa)\,p(\mathbf{w},\kappa) \tag{10}$$

By substituting Eqs. (3) and (8) into (9), we finally get the maximum *a posteriori* (MAP) estimate of $\mathbf{w}$ as:

$$\boldsymbol{w}_{\mathrm{MAP}} = \underset{\boldsymbol{w}}{argmax}\left\{-\tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{w}\|_2^2 - \lambda\|\boldsymbol{w}\|_1\right\} \tag{11}$$

where $\lambda = \kappa\sigma^2$; $\|\cdot\|_1$ is an ℓ_1 norm, which stands for sum of the absolute values of $\mathbf{w}$. Unlike a Gaussian prior, the Laplace prior has "corner (non-differentiable) points" and it encourages parameter vector $\mathbf{w}$ to be sparse; i.e., most of the elements of vector $\mathbf{w}$ become zero.

Shuku et al. (2020) developed a method for estimating trends and detecting layer boundaries in depth-dependent soil data based on Lasso. Shuku and Phoon (2020) developed a Lasso method for 3D geotechnical subsurface modeling, called geotechnical Lasso (GLasso), and an efficient algorithm to solve 3D problems.

In the following, the method developed by Ching and Phoon (2017) and Ching et al (2020, 2021) is simply referred to as SBL, whereas the method developed by Shuku et al. (2020) and Shuku and Phoon (2021) is simply referred to as Lasso.

CASE HISTORIES

SBL and Lasso, as outlined in the previous section, were applied to two real case histories with cone penetration test (CPT) data. This section briefly outlines these case histories.

Odagawa Riverbank, Okayama, Japan

Odagawa Riverbank is located approximately 27 km west of Okayama City, Okayama, Japan, on the north side of the Odagawa River. CPTs were performed on the berm, and the layout of CPTs is shown in Figure 1. The layout consists of 15 CPTs at a spacing of 5 meters. The cone tip resistance (q_t), sleeve friction (f_s), and pore pressure (u) were recorded with 5 cm depth resolution for each sounding. Some q_t and f_s values were negative, and this negative-value issue was resolved by the procedure indicated in the APPENDIX. The depth profile of the soil behavior type index (I_c) and soil behavior type (SBT) are shown in Figures 2 and 3. The I_c and SBT are based on the soil classification system proposed by Robertson (1990, 2016) and Robertson and Wride (1998), as summarized in Table 1.

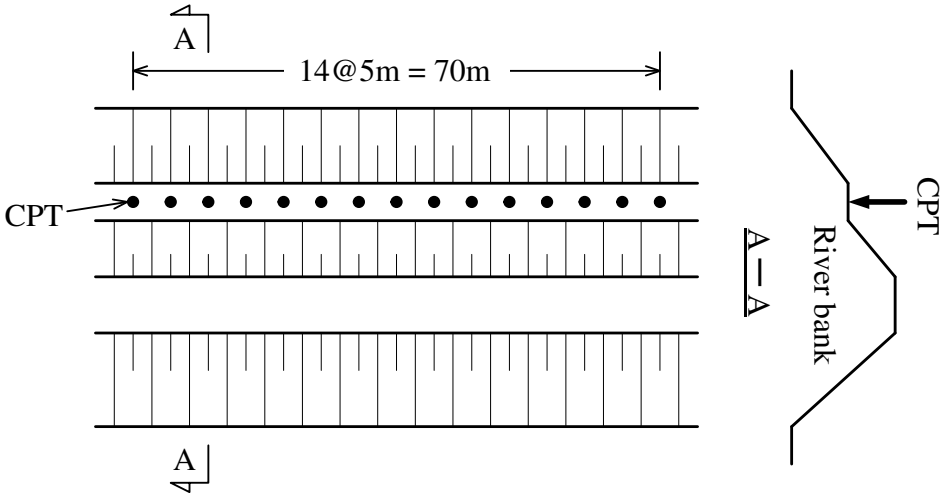


Figure 1. Layout of CPTs at Odagawa Riverbank site.

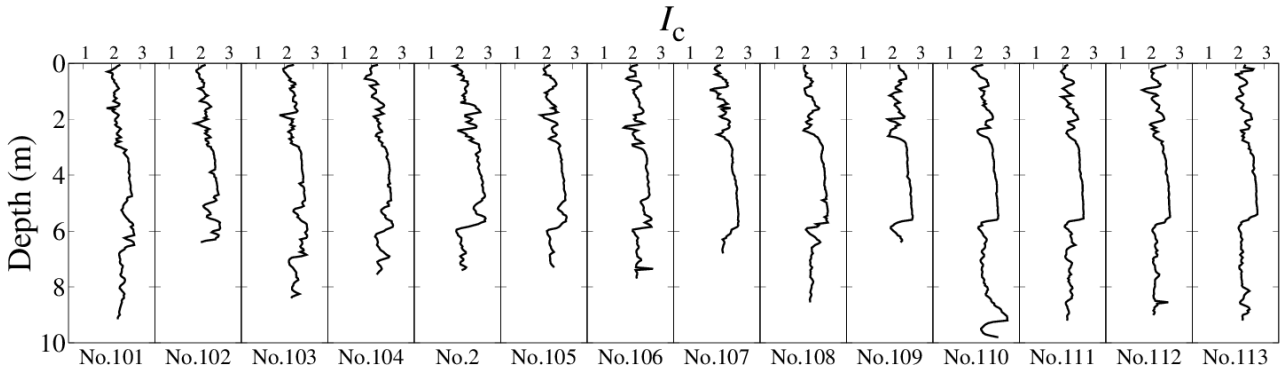


Figure 2. Depth profile of I_c at Odagawa Riverbank site.

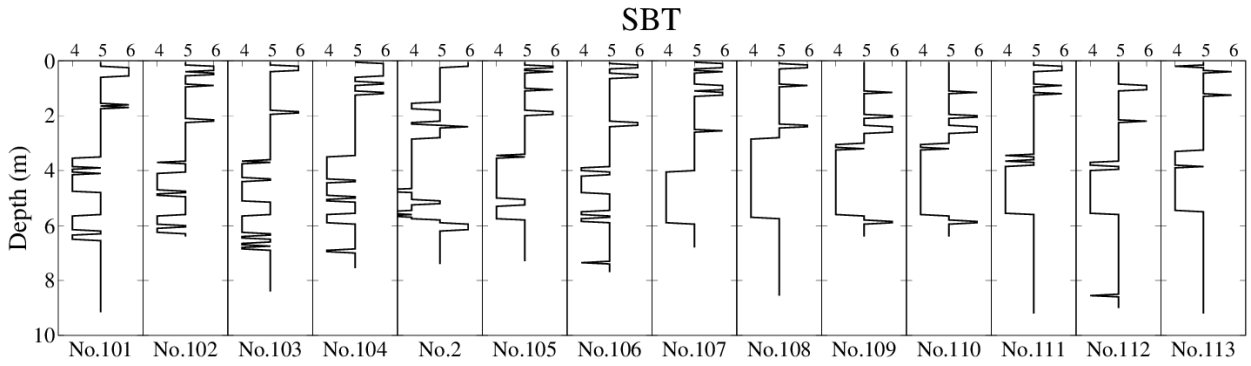


Figure 3. Depth profile of SBT at Odagawa Riverbank site.

Table 1. Soil behavior type index by Robertson (1990, 2016) and Robertson and Wride (1998).

Soil behavior type index, I_c	Zone	Soil behavior type (SBT)
$I_c < 1.31$	7	Gravelly sand to dense sand
$1.31 < I_c < 2.05$	6	Sands: clean sand to silty sand
$2.05 < I_c < 2.60$	5	Sand mixtures: silty sand to sandy silt
$2.60 < I_c < 2.95$	4	Silt mixtures: clayey silt to silty clay
$2.95 < I_c < 3.60$	3	Clays: silty clay to clay
$I_c > 3.60$	2	Organic soils: peats

New Lock, Terneuzen, the Netherlands

The New Lock site is located approximately 100 km southwest of Rotterdam, the Netherlands. The 427-meter-long New Terneuzen Lock is being constructed on the existing Terneuzen locks' complex and is designed to provide better access to the ports of Ghent and Terneuzen, as well as to promote a faster flow of shipping between the Netherlands, Belgium, and France. The layout of the CPTs is shown in Figure 4. The depth of the CPTs ranges from 3 to 70 m, and the horizontal spacing of the CPTs ranges from 0.36 to 2,835 m. In total, 98 CPTs were performed at this site. We focused on a small area containing 51 CPTs shown in Figure 4 for the underground stratification. The I_c profiles of A–A and B–B cross sections are shown in Figure 5.

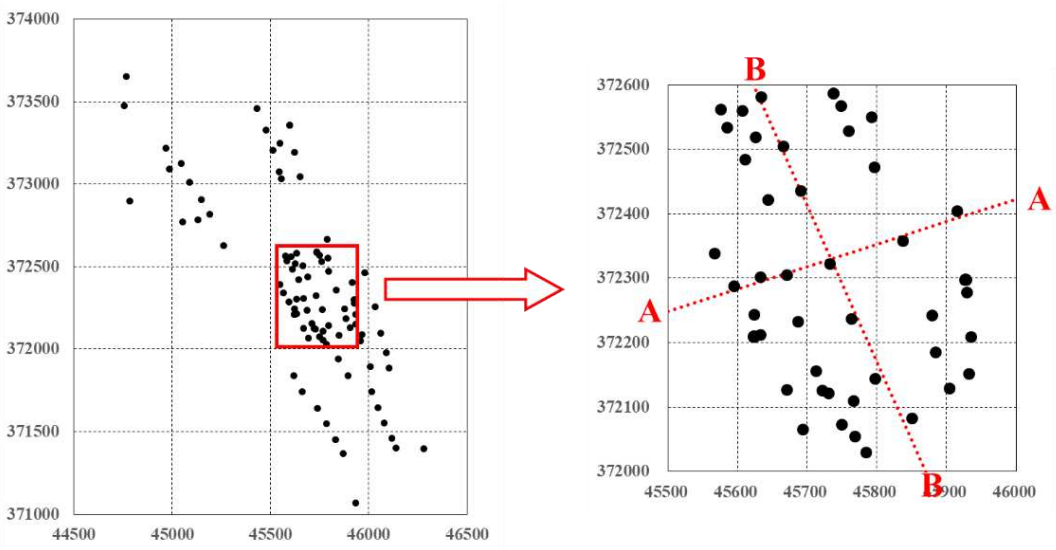
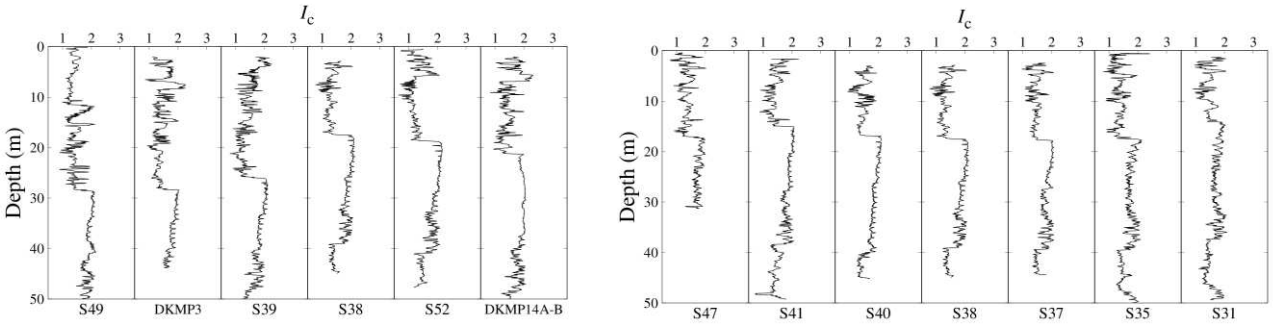


Figure 4. Layout of CPTs at the New Lock site.



(a) A – A *(b) B – B*

Figure 5. Depth profiles of I_c at A-A and B-B sections.

RESULTS OF UNDERGROUND STRATIFICATION USING SBL AND LASSO

In this section, the analysis results of SBL and Lasso for the case histories are demonstrated. 2D underground stratification was performed for the riverbank site, and 3D analysis was performed for the New Lock site. Brief descriptions on the numerical setup of SBL and Lasso are as follows: In SBL, the estimation results are obtained from the following three steps: 1) selection of BFs, 2) Drawing w, σ, and scale of fluctuation samples, and 3) conditional 3D random field simulation. In this paper, shifted Legendre polynomials were used as BFs, and BFs were selected by maximizing the marginal likelihood function. In step 2, the samples were drawn from marginal likelihood using an improved version of transitional Markov Chain Monte Carlo (iTMCMC) (Ching and Cheng 2007; Betz et al. 2016; Ching and Wang 2017). In Lasso, the estimates were obtained by maximizing/minimizing objective function (Eq. (11)), and we used the alternating direction method of multiplier (ADMM) (Boyd et al. 2010) to optimize the function. The regularization parameters λ for all the simulations were selected using the L-curve method (Hansen 1992) in this paper. As previously stated, this paper mainly demonstrates the analysis results of SBL and Lasso, and the details of their analysis setups are not presented.

To evaluate the performance of the two methods, the root-mean-square error (RMSE) of I_c and detection ratio (DR) of SBT are provided, which are defined by the following equations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - x_i)^2}\ , \tag{12}$$

$$\mathrm{DR} = \frac{1}{m}\sum_{i=1}^{m} I_i\ ,\ I_i = \begin{cases} 1 & \mathrm{SBT}(y_i) = \mathrm{SBT}(x_i) \\ 0 & \mathrm{SBT}(y_i) \neq \mathrm{SBT}(x_i) \end{cases}. \tag{13}$$

where y_i is the measured I_c data at the i^{th} depth, x_i is the estimated I_c value at the i^{th} depth, and SBT(y_i) and SBT(x_i) are the SBTs correspond to y_i and x_i. In order to compute the RMSE, a point estimate for I_c is needed. Lasso can readily produce a point estimate for I_c, which is the maximum a posteriori (MAP) estimate. However, SBL does not directly produce a point estimate. Instead, SBL produces samples of I_c. In principle, the significance of the sample mode of the I_c samples is similar to that of MAP. However, it is technically more challenging to compute the sample mode for the I_c samples of SBL than the sample mean or sample median. The sample mode is not a robust point estimate, compared to the sample mean and sample median. In this study, as opposed to adopting the sample mode of the SBL I_c samples, the sample median is adopted as the point estimate for SBL since: 1) the sample median is easy to compute, 2) the sample median is less noisy than the sample mode, and 3) it is found that the sample median of the SBL I_c samples is close to their sample mode for the two case histories.

Odagawa Riverbank Site

In this case history, we performed leave-one-out cross-validation (LOOCV) to evaluate the performance of the methods, which is widely used in the machine learning community to evaluate the performance of machine learning methods. In LOOCV, the I_c data of one CPT are used for validation and the I_c data for the remaining CPTs are used for training. Figure 6 shows the results of LOOCV, which compares the I_c results estimated by the two SML methods with the measured I_c profiles. The black lines indicate the measured I_c data, the blue continuous lines and dashed lines indicate median and 95% upper/lower

bounds of the conditional I_c random fields simulated by SBL, and the red continuous lines indicate the MAP estimates by Lasso. Both methods reasonably capture the measured I_c trends.

Table 2 summarizes the RMSEs for LOOCV on the I_c profiles. The MAP by Lasso and median by SBL were used in computing these RMSEs. Since the I_c profiles estimated by the two methods are similar, the RMSEs for the two methods are also analogous. The SBL method provides the posterior PDF of the estimates, which is more informative than the point estimate by Lasso. Lasso, nevertheless, can effectively detect sharp changes (such as layer boundaries) in the I_c trend due to the nature of the ℓ_1 norm in Eq. (10). Moreover, the I_c trend identified by Lasso is "cleaner" (fewer oscillating details) than by SBL. Unlike continuous BF-based methods, Lasso does not require BFs. This is a notable advantage of Lasso.

Figure 7 compares the SBT profiles calculated based on the I_c results in Figure 6. In the figure, the black lines indicate the measured SBT, whereas the red and blue lines are the SBT profiles by SBL and Lasso, respectively. The SBT profile for Lasso is calculated by the I_c – SBT mapping (Table 1) based on the I_c result for Lasso in Figure 6. The determination of the SBT profile for SBL is more complicated. First, the conditional I_c random fields simulated by SBL are converted to SBT random fields. The most probable SBT at a location is determined as the mode of the SBT samples at that location. The SBT profile for SBL is simply the profile of the most probable SBT. Due to the sparsity property of SBL and Lasso, the stratification results estimated by SBL and Lasso (red and blue lines) tend to be simpler than those directly estimated by the measured SBT (black lines).

Table 3 summarizes the detection ratios (DRs) of SBL and Lasso. The average DRs for SBL and Lasso are 0.83 and 0.81, respectively. Figures 8 and 9 show the colormaps of the I_c and SBT distributions estimated by SBL and Lasso. In Figure 8, Lasso tends to produce an I_c distribution that is cleaner (fewer oscillating details) than that produced by SBL.

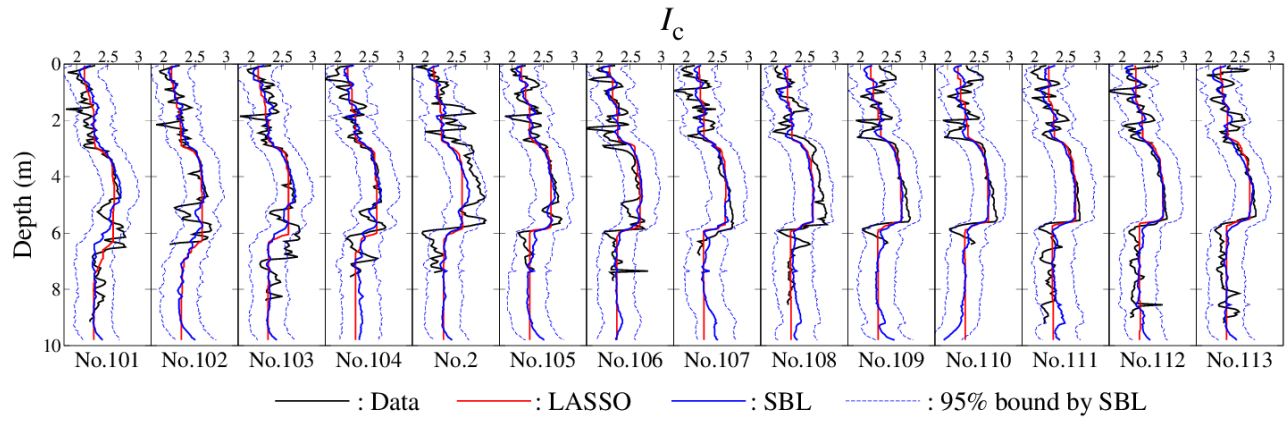


Figure 6. Estimated I_c profiles by SBL and Lasso.

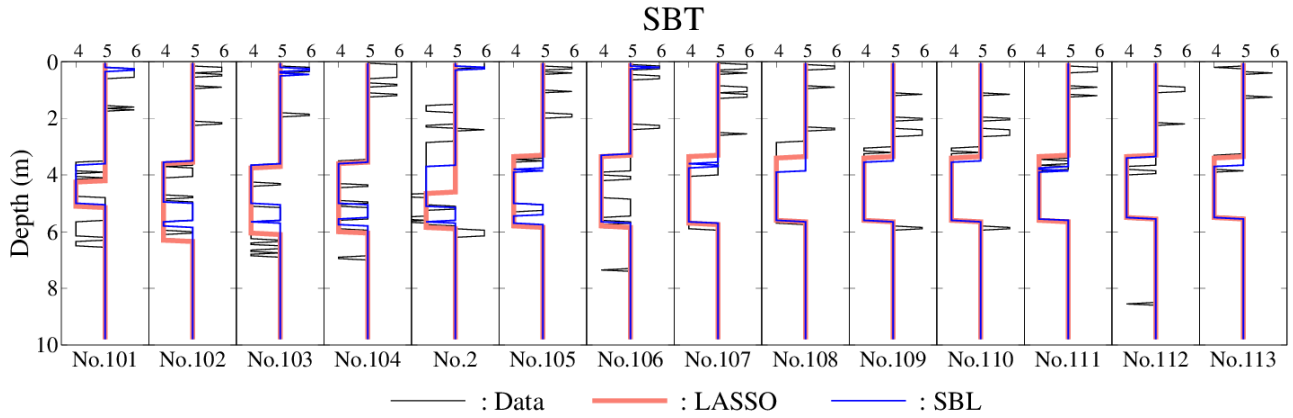


Figure 7. Estimated SBT profiles by SBL and Lasso.

Table 2. RMSEs for the I_c profiles (Odagawa Riverbank site).

No.	101	102	103	104	2	105	106	107	108	109	110	111	112	113	3
SBL	0.020	0.015	0.023	0.013	0.042	0.011	0.021	0.014	0.018	0.027	0.009	0.015	0.013	0.014	0.011
Lasso	0.013	0.018	0.023	0.017	0.050	0.011	0.023	0.016	0.017	0.024	0.010	0.011	0.014	0.012	0.017

Table 3. Detection rates for the SBT profiles (Odagawa Riverbank site).

No.	101	102	103	104	2	105	106	107	108	109	110	111	112	113	3
SBL	0.83	0.77	0.83	0.81	0.68	0.88	0.73	0.81	0.83	0.85	0.86	0.94	0.91	0.93	0.84
Lasso	0.76	0.71	0.83	0.78	0.57	0.89	0.74	0.77	0.89	0.87	0.88	0.92	0.91	0.96	0.81

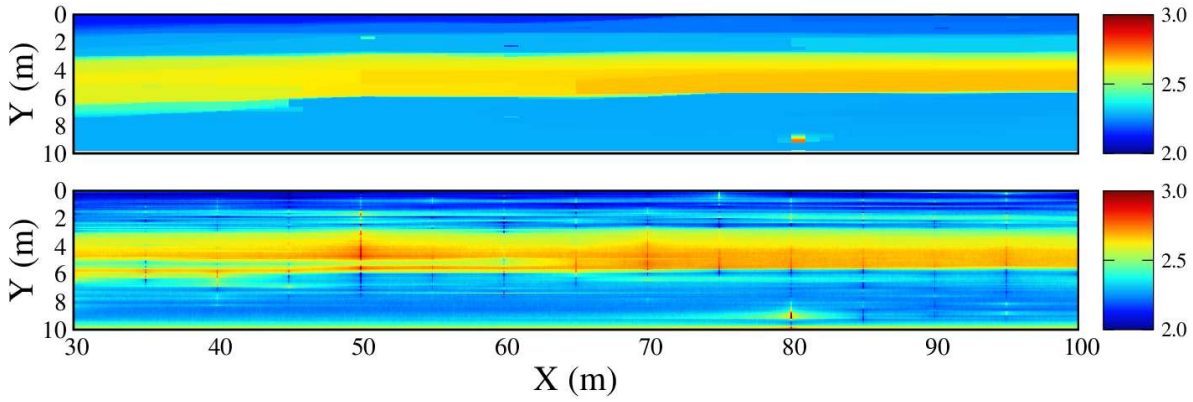


Figure 8. Color map of estimated I_c by Lasso (top) and SBL (bottom).

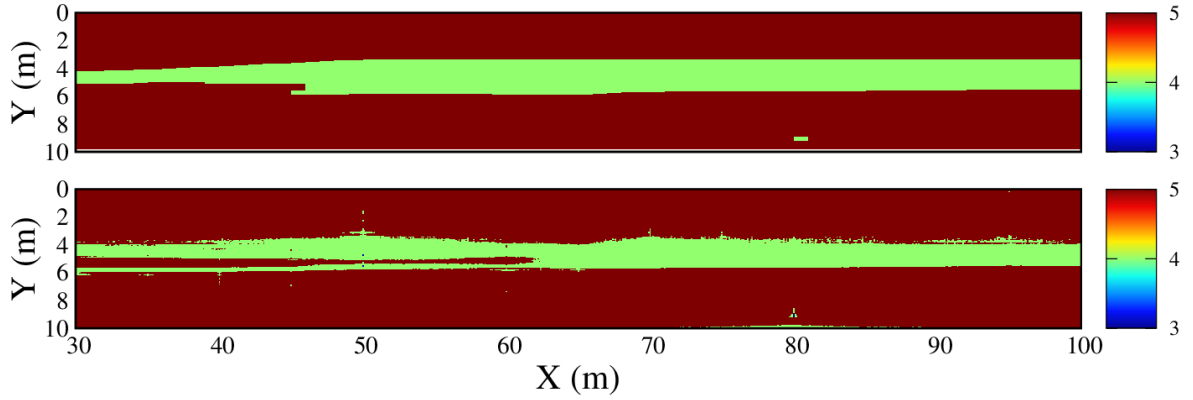


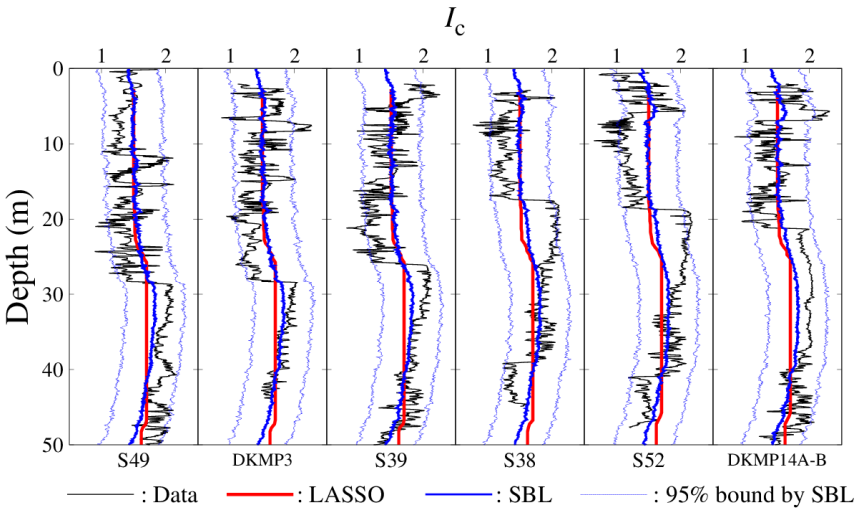
Figure 9. Color map of estimated SBT by Lasso (top) and SBL (bottom).

New Lock Site

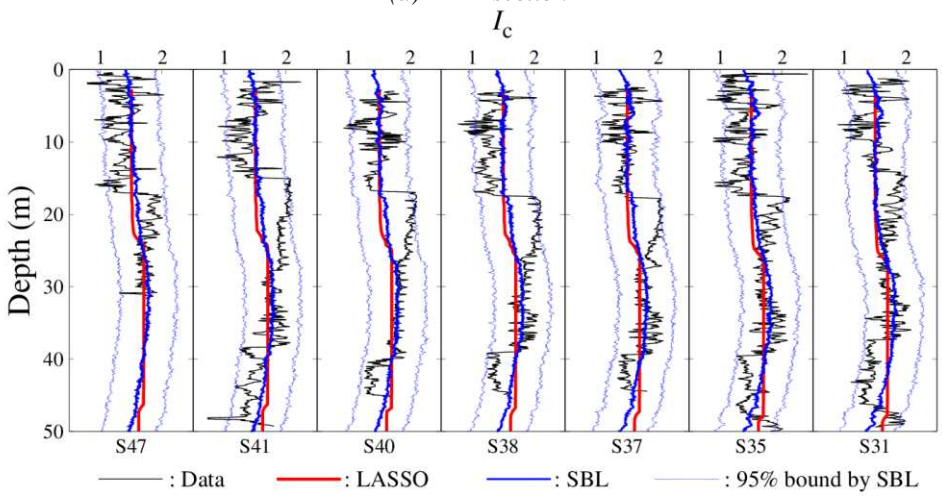
In this example, LOOCV was not performed because of the high computational cost, and we simply conducted 3D underground stratification using all 51 CPTs. Therefore, the results presented in this section do not reflect the predictive ability of the estimated model. Developing the method that can achieve training and validation for realistic 3D problems within a reasonable timeframe is a critical future objective.

For SBL, 10 CPTs with short depth ranges were not analyzed (only 41 CPTs were analyzed) to speed up the SBL algorithm significantly. These 10 CPTs all have long CPTs in their close proximity, so the removal of these short CPTs should only have minimal impact. For Lasso, all 51 CPTs were analyzed. Figure 10 shows the estimated I_c profiles on the A – A and B – B sections. In the figure, the black lines indicate the measured I_c data, the blue continuous and dashed lines indicate the median and 95% bounds of the conditional I_c random fields simulated by SBL, and the red lines indicate the trends estimated by Lasso. Some CPTs (such as S49 and DKMP3) that are close to the A – A and B – B sections are projected to the nearest respective locations, and the measured I_c profiles of these projected CPTs are also shown in Figure 10 for comparison. The I_c median and 95% bounds obtained by SBL tend to follow the general trends for the measured I_c profiles of nearby CPTs. In contrast, the trends estimated by Lasso tend to be simple (close to constant). In particular, Lasso detects a change point (a possible layer boundary) around the depth of 25 m.

Figure 11 compares SBT profiles calculated based on the I_c results in Figure 10. Although the measured SBT identifies many thin layers, SBL and Lasso provide simple stratification results: only a single layer. Tables 4 and 5 summarize the DRs for the A – A and B – B sections, respectively. The average value of DRs for SBL and Lasso are identical, which is 0.79. Note that LOOCV is not performed for the New Lock case, so the DRs in Tables 4 and 5 may not reflect the actual prediction capacity.

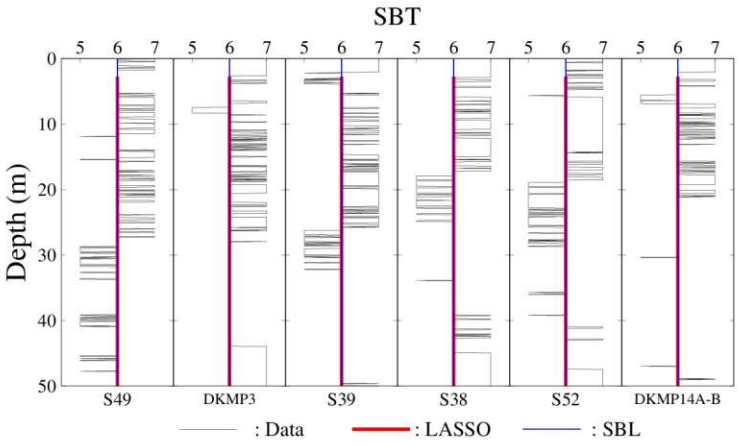


(a) A – A section

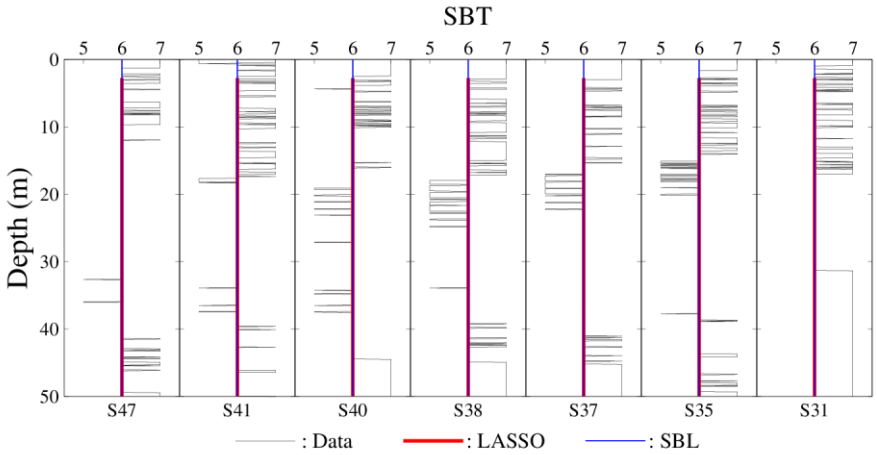


(b) B – B section

Figure 10. Estimated I_c profiles by SBL and Lasso.



(a) A – A section



(b) B – B section

Figure 11. Estimated SBT profiles by SBL and Lasso.

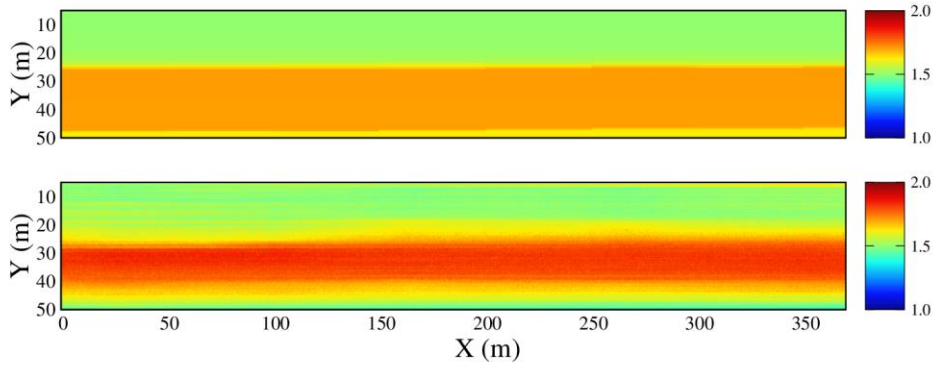
Table 4. Detection rates for the SBT profiles (A – A, New Lock site).

No.	S49	DKMP3	S39	S38	S52	DKMP14 A-B
SBL	0.77	0.78	0.71	0.70	0.63	0.86
Lasso	0.79	0.83	0.71	0.68	0.58	0.90

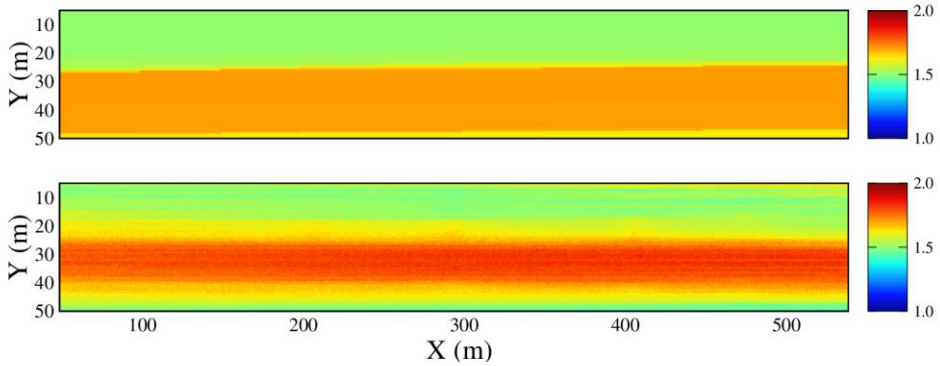
Table 5. Detection rates for the SBT profiles (B – B, New Lock site).

No.	S47	S41	S40	S38	S37	S35	S31
SBL	0.76	0.86	0.87	0.70	0.91	0.86	0.91
Lasso	0.93	0.73	0.95	0.68	0.90	0.78	0.81

Figures 12 and 13 show the colormaps of the I_c and SBT distributions on the A – A and B – B sections estimated by SBL and Lasso. For SBL, only the colormap for median I_c is shown (95% bounds are not shown). In Figure 12, Lasso tends to produce an I_c distribution that is cleaner (with sharp changes) than that produced by SBL. In the SBT colormap (Figure 13), both methods produce simple stratification and identify only a single layer (SBT 6: clean sand to silty sand).

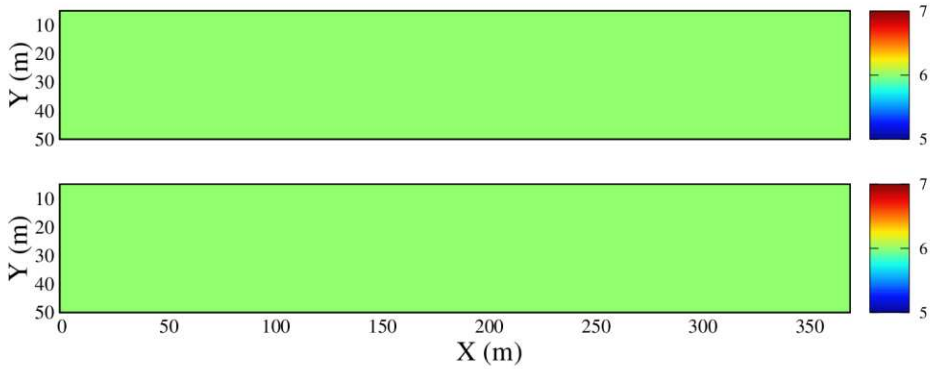


(a) A–A section

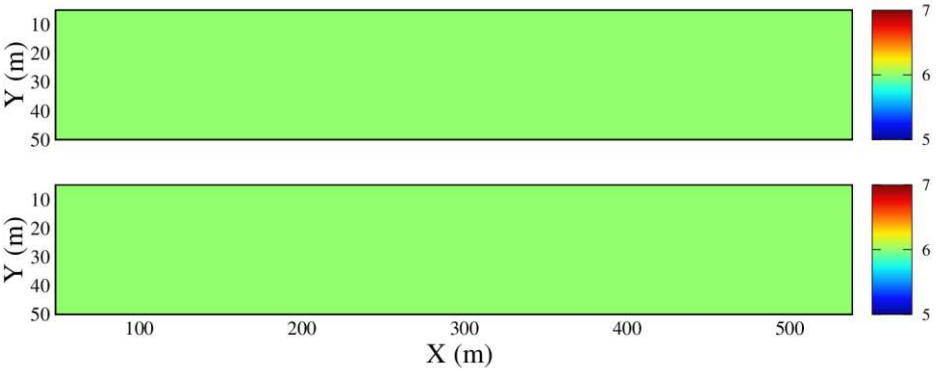


(b) B–B section

Figure 12. Colormap of estimated I_c by Lasso (top) and SBL (bottom).



(a) A–A section



(b) B–B section

Figure 13. Colormap of estimated SBT by Lasso (top) and SBL (bottom).

CONCLUSION

This paper showcased novel underground stratification based on SBL and Lasso for two case histories: the Odagawa Riverbank (Okayama, Japan) and New Lock (Terneuzen, the Netherlands). Cone penetration test (CPT) data were available in both sites, and the Robertson system and soil behavior type (SBT) index were used for the underground stratification. The I_c/SBT profiles and their spatial distribution were estimated by SBL and Lasso. Although both methods provided similar results, Lasso tends to produce simpler stratification results than SBL. However, SBL can produce the posterior PDF of estimates, which can be useful for reliability-based design. Lasso, on the other hand, is capable of detecting layer boundaries without the need to choose basis functions. This is a notable advantage for underground stratification. Some real case histories have been used in the recent papers on SML-based methods (Hu and Wang 2020; Wang et al. 2020), and analyzing their case histories using SBL and Lasso for comparison is another interesting topic for future study.

ACKNOWLEDGMENTS

The authors would like to thank the members of the TC304 Committee on Engineering Practice of Risk Assessment & Management of the International Society of Soil Mechanics and Geotechnical Engineering for developing the database 304dB (http://140.112.12.21/issmge/tc304.htm?=6) used in this study and making it available for scientific inquiry. This research was partly supported by JSPS KAKENHI Grant Number JP18K05880.

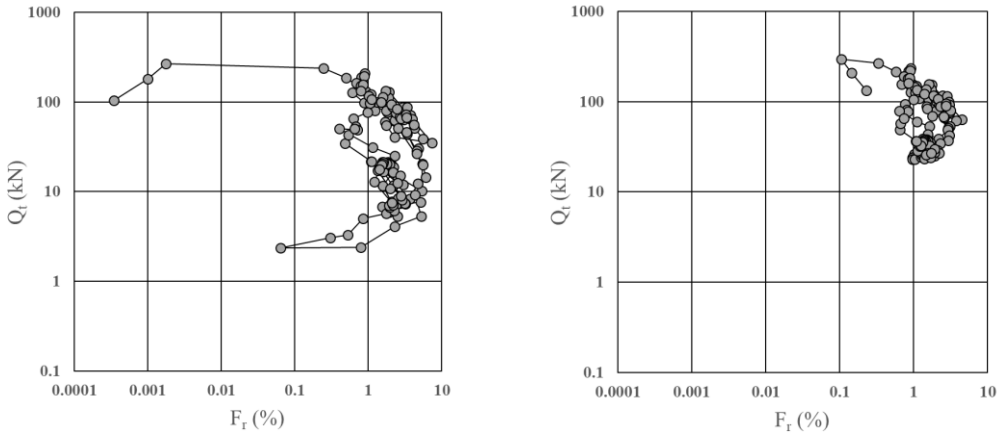
APPENDIX

Negative values of q_t and f_s can be obtained in practice, and the CPT data measured at the Odagawa Riverbank site have this negative-value issue. This issue usually happens in very soft normally consolidated soils and highly compressible peat (e.g., Sandven 2010). Boylan et al. (2008) investigated this issue and reported that the temperature gradient can result in significant (positive or negative) shifts in the cone measurements. This APPENDIX describes our approach to handle the negative values of q_t and f_s.

The soil behavior type index (I_c) proposed by Robertson (1990) depends on the normalized cone resistance (Q_t) and friction ratio (F_r). The bounds of $1 \le Q_t \le 1{,}000$ and $0.1 \le F_r \le 10$ are considered practical. In fact, these are the bounds for the Q_t–F_r chart developed by Robertson (1990). Figure 14a shows the Q_t–F_r plot for the data points for a CPT that contains negative q_t and f_s values. Some Q_t–F_r data points are located outside of the practical bounds. It is assumed that these data points are contaminated by erroneous shifts in q_t and f_s. By adding a minimal amount of counter-shifts (denoted by Δq_t and Δf_s) to the original q_t and f_s values, it is possible to contain all Q_t–F_r data points inside the bounds. The procedure is as follows.

1. Calculate Q_t and F_r based on the original data with negative q_t and f_s values.
2. For each CPT, identify the minimum values of Δq_t and Δf_s such that all Q_t–F_r data fall into the bounds, $1 \le Q_t \le 1000$ and $0.1 \le F_r \le 10$.
3. Repeat the prior two steps for all CPTs and find the Δq_t and Δf_s values that work for all CPTs.

We ultimately got Δq_t = +1700 kN and Δf_s = +18 kN, and Figure 14b shows the Q_t–F_r plot based on the shifted q_t and f_s.



(a) Original data *(b) Shifted data*

Figure 14. Q_t–F_r diagram.

REFERENCES

Betz, W., Papaioannou, I., and Straub, D. (2016). "Transitional Markov chain Monte Carlo: Observations and improvements." ASCE Journal of Engineering Mechanics, 142(5).

Bishop, C.M. (2006). *Pattern recognition and machine learning*, Springer, 738.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). "Distributed optimization and statistical learning via alternating direction method of multipliers." Found Trends Machine Learning, 3(1).

Boylan, N., Mathijssen, F., Long, M., and Molenkamp, F. (2008). "Accuracy of Piezocone Testing in Organic Soils." *Proc. 11th Baltic Conf. Geotechnics in Maritime Engineering*, Gdansk, Poland, 367-375.

Ching, J., and Chen, Y.C. (2007). "Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection and model averaging." *ASCE Journal of Engineering Mechanics*, 133(7), 816-832.

Ching, J., and Wang, J. S. (2017). "Discussion: Transitional Markov Chain Monte Carlo: Observations and improvements." *ASCE Journal of Engineering Mechanics*, 143(9).

Ching, J. and Phoon, K.K. (2017). "Characterizing uncertain site-specific trend function by sparse Bayesian learning." *ASCE Journal of Engineering Mechanics*, 143(7).

Ching, J., Huang, W.H., and Phoon, K.K. (2020). "Three-dimensional probabilistic site characterization by sparse Bayesian learning." *ASCE Journal of Engineering Mechanics*, 146(12).

Ching, J., Yang, Z., and Phoon, K.K. (2021). "Dealing with non-lattice data in three-dimensional probabilistic site characterization." *ASCE Journal of Engineering Mechanics*, 147(5)

Hansen, P.C. (1992). "Analysis of discrete ill-posed problems by means of L-curve." SIAM Review, 34, 516-580.

Hu, Y. and 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.

Hu, Y., Wang, Y., Zhao, T., and Phoon, K. K. (2020). "Bayesian supervised learning of site-specific geotechnical spatial variability from sparse measurements." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 6(2).

MacKay, D. J. C. (1992). "Bayesian interpolation." *Neural Comput*., 4(3), 415-447.

Robertson, P.K. (1990). "Soil classification using the cone penetration test." *Canadian Geotechnical Journal*, 27(1), 151-158.

Robertson, P. K. (2016). "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system - an update." *Canadian Geotechnical Journal*, 53(12): 1910-1927.

Robertson, P. K., and Wride, C. E. (1998). "Evaluating cyclic liquefaction potential using the cone penetration test." *Canadian Geotechnical Journal*, 35, 442-459.

Sandven, R. (2010). "Influence of test equipment and procedures on obtained accuracy in CPTU." *2nd International Symposium on Cone Penetration Testing*, Huntington Beach, CA, USA, 1-26.

Shuku, T. (2019). "Sparse modeling in geotechnical engineering." *Proc. of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR)*.

Shuku, T., Phoon, K.K., Yoshida, I. (2020). "Trend estimation and layer boundary detection in depth-dependent soil data using sparse Bayesian lasso.", *Computers and Geotechnics*, 128, 103845.

Shuku, T. and Phoon, K.K. (2020). "Three-dimensional subsurface modeling using geotechnical lasso." *Computers and Geotechnics*, 133, 1034068.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *J. Royal Statist. Soc. B*, 58(1), 267-288.

Tipping, M. E. (2001). "Sparse Bayesian learning and the relevance vector machine." *J. Mach. Learn. Res.*, 1, 211-244.

Wang, Y., and Zhao, T. (2017). "Statistical interpretation of soil property profiles from sparse data using Bayesian Compressive Sampling." *Geotechnique*, 67(6), 523-536.

Wang, Y., Hu, Y., and Zhao, T. (2020). "CPT-based subsurface soil classification and zonation in a 2D vertical cross-section using Bayesian compressive sampling." *Canadian Geotechnical Journal*, 57(7), 947-958.

Zhao, T., and Wang, Y. (2020). "Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods." *Engineering Geology*, 265.

Zhao, T., Xu, L., and Wang, Y. (2020). "Fast non-parametric simulation of 2D multi-layer cone penetration test (CPT) data without pre-stratification using Markov Chain Monte Carlo simulation." *Engineering Geology*, 273.

The open access Mission of the International Journal of Geoengineering Case Histories is made possible by the support of the following organizations:

Access the content of the ISSMGE International Journal of Geoengineering Case Histories at:
https://www.geocasehistoriesjournal.org