Site Characterization Employing Self-Organizing Map Technique with SHANSEP Parameters

Kyeongmo Koo
Graduate Student, Department of Civil and Environmental Engineering, Kookmin University, Seongbuk-gu, Seoul, 02707, South Korea; E-mail: kmkoo807@kookmin.ac.kr
Seongwook Han
Undergraduate course, Department of Civil and Environmental Engineering, Kookmin University, Seongbuk-gu, Seoul, 02707, South Korea; E-mail: hswk0618@naver.com

ABSTRACT
This study aims to improve the data-driven site characterization using the Global database. Self-Organizing Map (SOM), a clustering algorithm that can cover geotechnical spatial variation by utilizing knowledgeable and constraint data together, is utilized and a total of four features were selected for model training: SHANSEP parameters S and m, modified $\log(Q_t)$, and PI. The result of comparing the inter-node similarity based on the Euclidean distances between the weight vectors of the nodes in the optimized 12 by 12 SOM model shows the top three Sites most similar to Site #17 were Sites #13, #14, and #10 in that order. According to the statistical analyses for the closest 30 neighboring samples of the additional samples from Site #17 with the missing features, the missing properties were found to follow log-normal distribution then the useful statistics could be determined to estimate them.
Keywords: Data-driven site characterization, Clustering, Self-organizing map

1. INTRODUCTION
This paper aims (Report to the TC304/TC309 Student Contest on Clustering Applied to A Global Database in FOMLIG 2023 on 4 Dec. 2023) to introduce and provide solutions for approaching a global database using our sparse data. The solutions have to include (1) the characterization of the unknown site based on the given global five clay properties and (2) a probabilistic methodology for this purpose. A brief overview of the given data is introduced in Table 1.

Table 1. Statistics of the given global DB (N=133)

	Description	Mean [kPa]	STD. [kPa]	CoV [%]
S_u	Undrained shear strength	29.9	16.5	55.2
σ_p'	Pre-consolidation stress	153.3	109.1	71.2
q_t	Corrected tip resistance	681.3	465.3	68.3
PI	Plasticity index	37.5	13.9	37.0
σ_v'	Effective vertical stress	106.3	90.6	85.3

In the geotechnical field, the uncertainty and variability of soil properties present the most significant issues in solving engineering problems. This situation highlights the necessity for Data-Driven Site Characterization (DDSC) in the context of our limited resources, where the stability of the data itself becomes the most critical point. However, the data we acquire is known to possess the characteristics termed MUSIC-3X (Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted with 3-dimensional spatial variation), as identified by Phoon & Ching (2021). To characterize such an 'ugly' global database, numerous computing techniques and ML algorithms have been employed. In the context of geological engineering, variations of a soil property with depths in a soil layer often share similar statistics and may be taken as one group (Wang et al., 2014). Therefore, clustering techniques can perform classification considering spatial variation without being confined to the geoinformation of the data and are conceptualized as semi-supervised, given their ability to utilize not only knowledgeable data but also constrained data.

In this study, we utilized the Self-Organizing Map (SOM), a typical clustering algorithm, to characterize unknown locations. This paper is organized as follows: Chapter 2 includes the pre-processing and feature selection for SOM training, Chapter 3 details the methodology for the solution, Chapter 4 describes the solution process for questions 1 and 2, and finally, Chapter 5 summarizes.

2. FEATURE SELECTION
The five given properties each contain unique characteristics. For example, while PI represents a physical property that indicates the site characteristics, strength parameters such as S_u, σ_p', q_t, and σ_v' may not be suitable for site classification as they express site

conditions. Therefore, for this problem redefined as geotechnical site classification, we pre-processed the four features considering the geotechnical context.

Firstly, to replace and condense S_u, σ_p', and σ_v' with suitable features, we adopted the Stress History and Normalized Soil Engineering Properties (SHANSEP) model as expressed in Eq. (1) (Ladd & Foote, 1974).

$$S_u/\sigma_v' = S \cdot OCR^m \qquad \text{Eq. (1)}$$

We fitted the SHANSEP model for each site, determining S and m values from the results to use as new features. Consequently, samples within the same site share identical S and m values, making them site-specific parameters representing the properties of each site. Secondly, regarding q_t, we converted it into a $\log(Q_t^*)$ as Eq. (2), which plays the same role as the Normalized cone tip resistance $\log(Q_t)$ used by Robertson (2009) in proposing the SBT-index-based classification system for CPT data.

$$\log(Q_t{}^*) = \log\left(\frac{q_t - \sigma_v'}{\sigma_v'}\right) \qquad \text{Eq. (2)}$$

Finally, by directly using Plasticity Index PI, we prepared a total of four training features: S, m, $\log(Q_t^*)$, and PI.

3. METHODOLOGY

This study compares the similarity between known and unknown samples based on the Self-Organizing Map (Kohonen, 1982), a typical clustering algorithm. SOM is an algorithm that clusters samples with similar feature values in the input space into the same or neighbor nodes in a 2-dimensional output space. As learning progresses, the nodes on the SOM update their weight vectors, which correspond to the concept of position vectors. Input samples are assigned to nodes with the most similar weight vectors, and the distance between these determines the distance (similarity). To achieve this, we aim to develop a model where all samples have their position vectors, optimizing a 12 by 12 SOM (similar in size to the total number of samples) based on clustering performance metrics with hyper-parameter tuning.

4. RESULTS & ANALYSES

4.1 Sites Listing Based on Similarity

Fig. 1 displays the optimized SOM model trained and optimized using the data from 16 sites and 7 target samples (Site #17-1, 2, 3, 4, 5, 6, and 7) together.

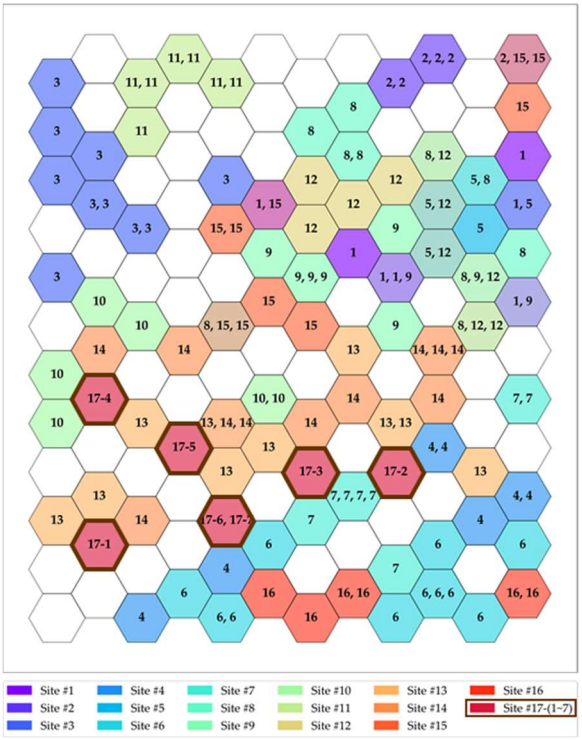

Fig. 1. Optimized 12 by 12 SOM model.

After replacing the weight vectors of each node with the representative values of the samples clustered within them, we calculated and compared the distances between position vectors based on each of the 7 target samples, and the results are summarized in Table 2.

Table 2. Calculated mean distances to the Sites in global DB from the Site #17 in the trained SOM model

#1	#2	#3	#4	#5	#6	#7	#8
0.515	0.611	0.451	0.405	0.539	0.450	0.422	0.505
#9	**#10**	**#11**	**#12**	**#13**	**#14**	**#15**	**#16**
0.479	0.345	0.963	0.508	0.237	0.244	0.470	0.592

4.2 Stochastic Estimation of Missing Values from Neighboring Samples

By the same processes, the SOM model was optimized by utilizing 3 samples from the Site#17 (Site #17-8, 9, and 10) with missing values together. The three samples containing the missing values are characterized based on 30 neighboring samples selected in order of the distances from all the other samples to determine the trustworthy statistical analysis (Hogg et al., 2012). Although missing values exist, the same site has the same S and m values, so all input features can be defined. When learning is completed, it is possible to make statistical analyses by changing the representation of the model for the missing original features. Fig. 2 shows the Cumulative Density Function (CDF) curves obtained by inverse

distance weighting based on the distances between samples and each of the 30 neighboring samples and the fitting results for various statistical distribution models.

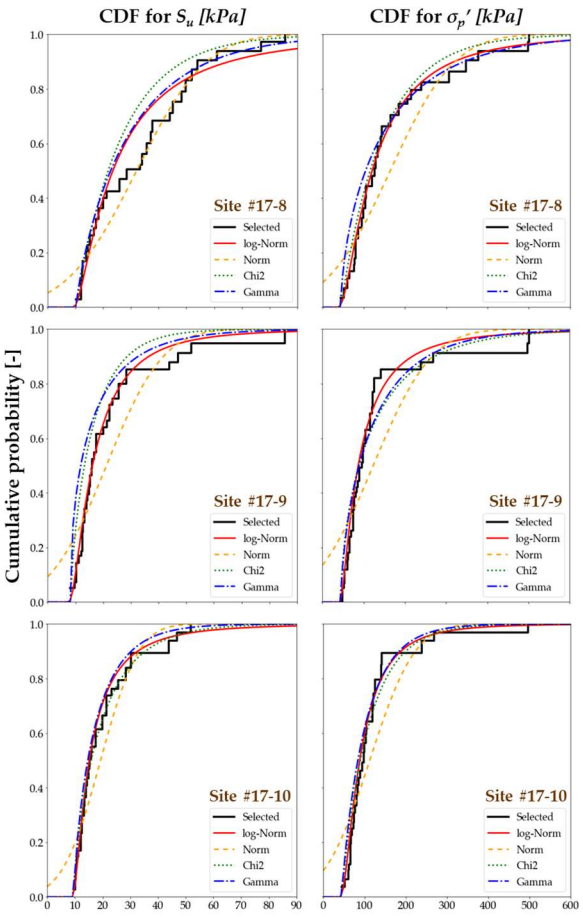

Fig. 2. Empirical and Fitted CDF for probable missing properties.

As a result, the best Kolmogorov–Smirnov test statistics were shown under the assumption that the two missing properties follow log-normal distribution in all cases, and the 95% Confidence Interval (CI) and median on the log-normal CDF according to this were summarized in Table 3.

Table 3. Estimated statistics for the missing properties.

Site	S_u [kPa]		σ_v' [kPa]	
	95% CI	Median	95% CI	Median
#17-8	10.8~124.9	23.5	48.4~573.3	120.1
#17-9	8.7~62.2	16.1	46.5~403.4	86.4
#17-10	9.9~57.3	15.2	47.6~306.0	88.2

5. SUMMARY & CONCLUSIONS

This study used a Self-Organizing Map method, a representative clustering algorithm, to characterize the unknown Site #17 based on the soil property database of the 16 global Sites. For SOM model training, a total of four features were selected: SHANSEP parameters S and m determined through the best-fit technique for each site, $\log(Q_t^*)$ from the conventional Robertson's CPT soil classification chart, and PI.

Based on the clustering performance evaluation index, the 12 by 12 SOM model was optimized and the results were analyzed. As a result, the top three Sites most similar to the Site #17 were the Sites #13, #14, and #10, in that order. In addition, as a result of performing statistical analysis on 30 neighboring samples of the three additional samples from the Site #17 in the optimized SOM model, the missing properties of the target samples were assumed to follow the log-normal distribution and useful statistics such as 95% CI and median could be obtained.

REFERENCES

Wang, Y., Huang, K., & Cao, Z. (2014). Bayesian identification of soil strata in London clay. Géotechnique, 64(3), 239-246.

Ching, J., Wu, S., & Phoon, K. K. (2021). Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model. Journal of Engineering Mechanics, 147(10), 04021069.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological cybernetics, 43(1), 59-69.