Quasi-site-specific soil property prediction Based on Hierarchical Bayesian Model-Affinity Propagation

Shiying Zheng[1], Menglu Huang*[2] and Yuxiang Ren[3]

[1]*Graduate school of environmental and life science, Okayama University, 1-1-1 Tsushima-naka, Kita-ku, Okayama 700-8530, Japan. Email: px153mwq@s.okayama-u.ac.jp*
[2]*Graduate school of environmental and life science, Okayama University, 1-1-1 Tsushima-naka, Kita-ku, Okayama 700-8530, Japan. Email: menglu_huang@s.okayama-u.ac.jp*
[3]*Graduate school of environmental and life science, Okayama University, 1-1-1 Tsushima-naka, Kita-ku, Okayama 700-8530, Japan. Email: py3g9cvm@s.okayama-u.ac.jp*

Abstract: In the realm of data-centric geotechnics, the characterization of specific site properties poses notable challenges. This study presents a novel approach by introducing a Hierarchical Bayesian Model (HBM) in conjunction with Jensen-Shannon Divergence (JSD) and Affinity Propagation (AP) to enhance the efficiency of site clustering and prediction.
Keywords: Data-driven site characterization, Hierarchical Bayesian model (HBM), Affinity propagation, Geotechnical site clustering, Soil property prediction.

1. Methodology

1.1 Hierarchical Bayesian model (HBM)

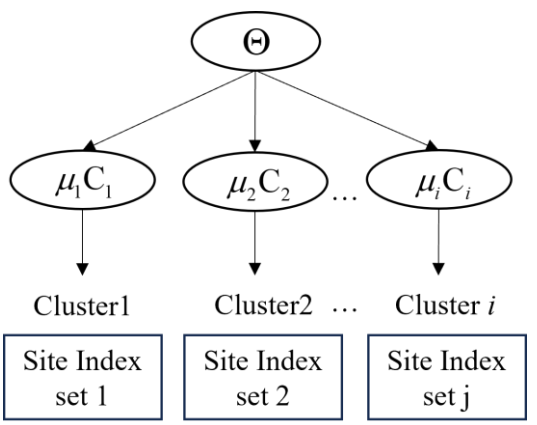

Figure 1. Hierarchical Bayesian model structure.

The data-driven site characterization plays a pivotal role in Data-centric Geotechnics. However, soil property data of a specific site is typically multivariate, uncertain and unique, sparse, and incomplete (MUSIC) (Phoon et al. 2019). On the other hand, it is widely recognized that indirect generic data (database) are plentiful and easily accessible. Therefore, 'borrowing' data from similar sites within the database is an effective means for characterizing a specific site. In order to deal with this issue, Ching et al (2021) proposed an efficient HBM-Gibbs approach with closed-form conditional probabilities, as shown the Fig. 1. And the HBM-Gibbs approach contains learning stage and inference stage.

1.2 Jensen-Shannon divergence (JSD)

The probability density functions (PDFs) for each site can be obtained by the HBM-Gibbs method. The similarity of each site is measured based on the PDFs and Jensen-Shannon Divergence (JSD). The expression for the similarity matrix *s* between sites is as follows.

$$s(i,j) = s(j,i) = 1 - D_{JS}\left(f_{i|X_d} \middle\| f_{j|X_d}\right) \quad (1)$$

Where D_{JS} represents the JSD between two probability distributions, the values range from 0 to 1. As the dissimilarity between the two distributions decreases, the JSD approaches 0.

1.3 Affinity Propagation (AP)

Affinity Propagation (AP) is a graph-based cluster algorithm which proposed by Brendan J (2007). It does not require the number of clusters to be specified in advance, assumes that each data is treated as a potential exemplar (means center of cluster). The exemplar of each cluster and the data point attributed to this exemplar can be determined by calculating the affinity between the sites based on similarity matrix $s(i, k)$.

Two kinds of message, responsibility $r(i, k)$ and availability $a(i, k)$ determine the affinity of the inter-sites, which former indicates that site k serves as the accumulated evidence for site i exemplar, and the latter indicates that site i selects site k as the accumulated evidence for exemplar. The final clustering result is obtained by updating responsibility and availability through a certain number of iterations. The specific formulas are shown below.

$$r(i,k) = s(i,k) - \max_{k',k' \neq k} \{a(i,k') + s(i,k')\} \quad (2)$$

$$a(i,k) = \min\{0, r(k,k) + \sum_{i',i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \quad (3)$$

$$a(k,k) = \sum_{i',i' \notin \{i,k\}} \max\{0, r(i',k)\} \quad (4)$$

The more critical parameters in the AP algorithm are the similarity matrix, "preference" value and "damping" value. Similarity matrix $s(i, k)$ includes the similarity of intra-site and inter-site, which is determined by HBM model. "Preference" refers to $s(k, k)$, as the basis for selection of exemplar. The larger the preference value, the more likely the point is to be exemplar, which directly affects the size of the cluster. To prevent oscillations during the iteration process, the damping factor is introduced to limit the value of each update.

Silhouette Index (SI) is used to evaluate the performance of cluster. SI is a method for assessing the validity of a clustering solution. It measures how closely related an object is to its own cluster compared to other clusters, producing a score that ranges from -1 to 1. A high SI value indicates strong intra-cluster similarity and good separation between different clusters. This metric helps in identifying the optimal number of clusters by comparing the silhouette scores for different clustering solutions.

Based on the three aforementioned methods, the research methodology is summarized as follows.

Step 1: Transform the original target site data Y_t and the database Y_d (including clay properties $\ln(\mathrm{PI})$, $\ln\left(\sigma'_V/P_a\right)$, $\ln\left(s_u/\sigma'_V\right)$, $\ln\left(\sigma'_P/P_a\right)$, $\ln\left(q_c/\sigma'_V\right)$) into X_t (including complete data X_t^c and incomplete data X_t^s) and X_d using the Johnson transformation proposed by Ching and Phoon (2014).

Step 2: Utilize HBM-Gibbs and JDS to calculate the similarity pairwise matrix s for sites within X_d.

Step 3: Perform cluster analysis of X_d based on the s and AP method.

Step 4: Determine the cluster label for X_t based on the X_t^c.

Step 5: Perform HBM-Gibbs inference to predict missing data using the entire X_t and transform it back into the original data.

2. Research Outcomes

2.1 Cluster results for the database

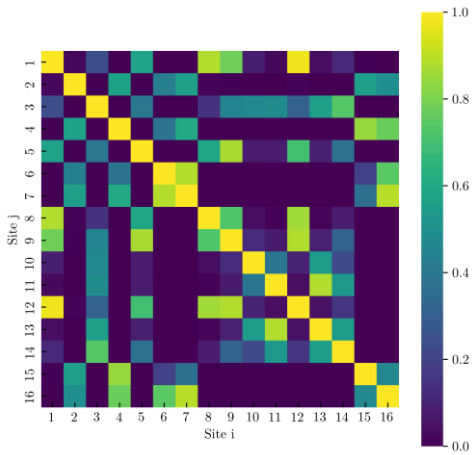

Figure 2. Similarity matrix of sites.

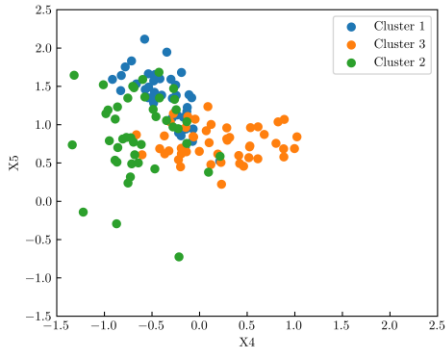

Figure 3. Cluster of site index obtained by AP.

Fig. 2. illustrates the similarity matrix of 16 different sites in the database. Based on the matrix and the AP method, we have successfully partitioned the database into three clusters: Cluster 1 (1, 5, 8, 9, 12); Cluster 2 (3, 10, 11, 13, 14); Cluster 3 (2, 4, 6, 7, 15, 16), where the values in parentheses represent the set of site indices included in each cluster. Fig. 3 shows the clustering results, demonstrating the effectiveness of the method. And Fig. 4 shows the values of $\boldsymbol{\mu}_i$ and $\boldsymbol{C}_i$ for the 16 sites sampled from $p(\boldsymbol{\mu}_s, \boldsymbol{C}_s|\theta)$.

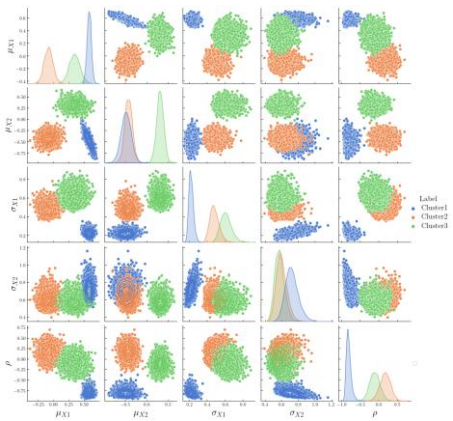

Figure 4. Pair-wise plots of the values of $\boldsymbol{\mu}$ and $\boldsymbol{C}$ (5000 samples)

2.2 Determination of a cluster label for the target site (Q1)

HBM model was trained using complete data from the target database, resulting in PDF of the target site, the Jensen-Shannon Divergence (JSD) was employed to calculate the similarity between X_t^c and clusters 1, 2, and 3, yielding similarity values of 1.02×10^{-7}, 0.62, and 2.85×10^{-5}, respectively. Consequently, the label assigned to the target site is cluster 2，indicating its similarity to **sites [3,10,11,13,14]**, while the similarity with other instances in clusters 2 and 3 is nearly zero.

2.3 Missing data prediction (Q2)

Once the target sites have obtained cluster labels, there are two strategies for training the prior HBM. Strategy 1 involves utilizing only the same cluster data from the database (Sharma et al. 2019), while Strategy 2 involves using all the data from the database. The model is then updated based on all the X_t (posterior HBM) Fig. 5 shows the prior and posterior PDFs for the standard deviation of X_5 based on the two strategies. Finally, the posterior model is employed to predict missing data. Tables 1 and 2 respectively present the median of the predicted values and the upper and lower bounds of the 95% confidence interval for the two strategies.

Table 1. Predictions based on cluster 2

index	Median of S_u	95% CI of S_u	Median of σ'_p	95% CI of σ'_p
1	20.4958	[13.0507, 32.0219]	91.9170	[53.8737, 167.0050]
2	12.2571	[7.9095, 18.7105]	58.2757	[36.7071, 97.7599]
3	10.4493	[6.7050, 16.8320]	52.8680	[33.2049, 88.9062]

Table 2. Predictions based on the database.

index	Median of S_u	95% CI of S_u	Median of σ'_p	95% CI of σ'_p
1	16.8615	[10.3520, 26.8551]	113.1170	[61.6551, 227.9830]
2	12.6940	[8.7248, 18.3717]	57.7199	[37.8517, 92.7097]
3	12.8317	[8.0982, 21.2991]	45.8361	[28.4739, 80.0998]

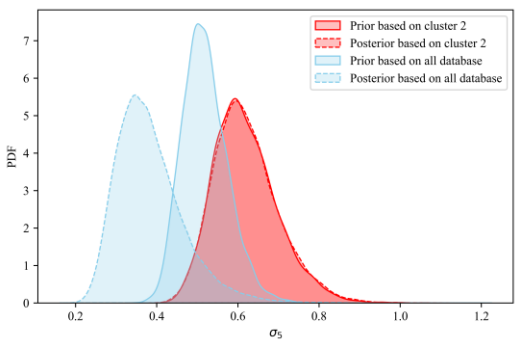

Figure 5. Prior and Posterior PDF of σ_5

3. Conclusions

The proposed methodology, integrating Hierarchical Bayesian model, Jensen-Shannon

divergence, and Affinity Propagation, provides a robust framework for data-driven geotechnical site characterization. The efficient clustering of sites, and accurate prediction of missing data demonstrate the effectiveness of methodology. The study contributes to advancing data-centric geotechnics, offering a reliable approach for site characterization and informed decision-making in geotechnical engineering.

Acknowledgement

The authors would like to thank Professor Nishimura and Shibata, as well as the organizers of the contest.

References

Frey, B.J., and Delbert, D. 2007. Clustering by passing messages between data points. Science, 315: 972-976.

Ching, J., and Phoon, K. K. (2014). Correlations among some clay parameters-the multivariate distribution. Canadian Geotechnical Journal, 51(6), 686-704.

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

Phoon, K. K., J, Ching, and Y, Wang. Managing risk in geotechnical engineering–from data to digitalization[C]//Proc., 7th Int. Symp. on Geotechnical Safety and Risk (ISGSR 2019). 2019. 13-34.

Sharma, A., Ching, J., and Phoon, K. K. 2023. A spectral algorithm for quasi-regional geotechnical site clustering. Computers and Geotechnics, 161, 105624.