Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore. ©2025 The Author(s), CC-BY 4.0 licence.
doi: 10.3850/GRF-25280825_isgsr-154-P122-cd

DEFINING NEW STATISTICAL FEATURES FOR GEOTECHNICAL PROPERTIES: EXPLORING HIGHER-ORDER DEPENDENCIES IN MIXED DOMAIN SPACES WITH THE MINIMUM INFORMATION DEPENDENCE MODEL

Taiga Saito

Department of Civil and Environmental Engineering, Tohoku University, 6-6-06, Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi, 980-8579, Japan. E-mail: taiga.saito.r3@dc.tohoku.ac.jp

Yu Otake

Department of Civil and Environmental Engineering, Tohoku University, 6-6-06, Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi, 980-8579, Japan. E-mail: yu.otake.b6@tohoku.ac.jp

Stephen Wu

Research Organization of Information and Systems, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, 190-8562, Japan. E-mail: stewu@ism.ac.jp

Keisuke Yano

Research Organization of Information and Systems, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, 190-8562, Japan. E-mail: yano@ism.ac.jp

Understanding the complex interdependencies among soil parameters is crucial for designing and maintaining civil engineering structures. Traditional modeling often assumes linearity and normality, potentially overlooking the nonlinear and higher-order relationships inherent in geotechnical data. This study evaluates the effectiveness of Minimum Information Dependence Modeling (MIDM), a novel joint probability distribution model, in analyzing dependencies within heterogeneous geotechnical datasets comprising both quantitative and qualitative variables. Applying MIDM to the real soil database, we computed conditional scores using a pseudolikelihood estimator based on Besag's approach to estimate strata labels. The results demonstrate that MIDM successfully captures complex non-linear relationships, providing a more nuanced understanding of subsurface conditions and enhancing anomaly detection and model validation in geotechnical risk assessments. Despite the high computational costs associated with the sampling-based approach, MIDM shows promise for transforming traditional practices toward more data-driven and quantitative methodologies. Future research will focus on optimizing computational efficiency and integrating MIDM with other data-driven techniques to construct high-density three-dimensional ground models, aiming to establish MIDM as a reliable tool in geotechnical investigations.

Keywords: Minimum information dependence model, Mixed-domain, Big indirect database, Data-centric geotechnics, Patterns extraction, 3D subsurface geological model, Stratigraphy.

1. Introduction

The performance and safety of geotechnical structures, such as foundations, retaining walls, and tunnels, are critically influenced by the properties of the surrounding ground. Natural ground conditions exhibit inherent heterogeneity due to diverse topographical and geological processes, resulting in significant spatial variability in soil properties even within confined geographical areas. This variability poses substantial challenges for the effective design, risk assessment, and maintenance management of construction projects. Accurate characterization of soil heterogeneity is thus essential for mitigating risks and optimizing cost estimations in geotechnical engineering.

Historically, probabilistic methods have been employed to account for soil variability and uncertainty in geotechnical analyses. Ching et al. (2021) advanced mathematical modeling techniquesfor geotechnical parameter estimation, such as hierarchical Bayesian models, emphasizing the importance of large databases in enhancing model reliability. Building on this, some research focuses on leveraging extensive datasets through similarity assessments to further refine parameter estimation. While these methods are theoretically robust and rely on Gibbs sampling under normality assumptions, they remain largely confined to linear frameworks—limiting their ability to capture non-linear or higher-order dependencies in complex soil systems. This shortcoming is often exacerbated by the scarcity of extensive quantitative data.

Based on the above background, this study employs Minimum Information Dependence Modeling (MIDM), a mathematically novel joint probability distribution model proposed by Sei and Yano (2023), to evaluate dependencies among heterogeneous data, including both quantitative and qualitative variables. MIDM

distinguishes itself by quantifying complex, non-linear relationships that traditional models struggle to capture, particularly in mixed-domain datasets. Its high interpretability and flexibility make it a promising tool for uncovering nuanced dependency structures in geotechnical data.

We aim to evaluate the effectiveness of MIDM in geotechnical engineering by examining dependencies within heterogeneous datasets derived from ground investigations. Specifically, we apply MIDM to soft clay soil data from Tokyo-CLAY/14/67760—a comprehensive soil database characterized by significant variability and complex stratification, as detailed by Saito et al. (2024)—and compute conditional scores using a pseudolikelihood estimator based on Besag's approach. These conditional scores serve as indicators of how well data points align with the model's dependency structure, thereby enabling effective anomaly detection and model validation. By leveraging conditional scores, our approach enhances the robustness and reliability of geotechnical risk assessments, facilitating more informed decision-making in engineering practices. Furthermore, from a geotechnical engineering perspective, we highlight the potential for identifying non-linear causal relationships that conventional models cannot represent, and discuss future directions in data-driven geotechnical engineering research.

2. Minimum Information Dependence Model (MIDM)

2.1. *Theoretical Background*

In MIDM, the probability density function of a d-dimensional random variable $\mathbf{x} = \{x_1, \dots, x_d\}$is expressed as follows:

$$p(\mathbf{x}; \underline{\theta}, \underline{\nu}) = \exp\left(\underline{\theta}^{\top} h(\mathbf{x}) - \sum_{i=1}^{d} a_i(x_i; \underline{\theta}, \underline{\nu}) - \psi(\underline{\theta}, \underline{\nu})\right) \prod_{i=1}^{d} r_i(x_i; \nu_i). \#(1)$$

Here, $r_1(\underline{x}_1; \nu_1), \dots, r_d(\underline{x}_d; \nu_d)$ are the marginal densities od each variable, and $\underline{\nu} = (\nu_1, \dots, \nu_d)$ represents their parameters.The exponential term exp(·) adjusts the joint probability density from the case where the variables are independent tomatch the actual joint probability density. The first term $\underline{\theta}^{\top} h(\boldsymbol{x})$ inside the exponential represents the dependency among randomvariables, whereas the second and third terms, called the adjustment functions $a_i(x_i; \underline{\theta}, \underline{\nu})$ and the potential function $\psi(\underline{\theta}, \underline{\nu})$, are functions determined simultaneously when the first term isspecified.

Focusing on the term $\underline{\theta}^{\top} h(\mathbf{x})$, where$h: x \to \mathbb{R}^K$ and $\underline{\theta} \in \mathbb{R}^K$ the dependence parameter vector. Although the form of $\underline{\theta}^{\top} h(\boldsymbol{x})$ can be any function, we adopt the following polynomial function in this study:

$$\underline{\theta}^{\top} h(\mathbf{x}) = \sum_{g=2}^{d} \sum_{1 \le i_1 < i_2 < \cdots < i_g \le d} \underline{\theta}_{i_1 i_2 \dots i_g} x_{i_1} x_{i_2} \dots x_{i_g} . \#(2)$$

MIDM can capture complex nonlinear dependencies that go beyond simple pairwise correlations by including terms in $h(\mathbf{x})$ that represent relationships among three or more variables. The parameters $\underline{\theta}$ signify the influence of each term constituting $h(\mathbf{x})$. We can understand the degree of influence of the individual interaction terms by examining the magnitude of these parameters.

2.2. *Estimation for Dependence Parameters*

Eq.(1) can be calculated analytically by performing maximum likelihood estimation for $\underline{\theta}$ and $\underline{\nu}$. However, these calculations require solving simultaneous integral equations, which entail significant analytical costs. To address this, the full likelihood can be decomposed into the product of a conditional likelihood and a marginal likelihood that is independent of the parameter $\underline{\nu}$. By applying the MCMC exchange algorithm using the conditional likelihood, $\underline{\theta}$ can be calculated quasi-analytically without the need for explicit identification of ½ For more details, please refer to Sei and Yano. (2024).

2.3. *Calculation of Conditional Scores*

In this section, we propose to compute the conditional scores of a pseudolikelihood estimator based on Besag's approach. The score (gradient) of the pseudolikelihood function for a pair of multivariate samples is calculated using the following equation:

$$g_{\theta} = \frac{1}{n^2} \sum_{s=1}^{n} \sum_{t \neq s, t=1}^{n} \sum_{i=1}^{d} \frac{h\left(x^i(s[t])\right) + h\left(x^i(t[s])\right) - h\left(x(s)\right) - h(x(t))}{1 + \exp\left\{\underline{\theta}^{\top}\left(h\left(x^i(s[t])\right) + h\left(x^i(t[s])\right) - h\left(x(s)\right) - h(x(t))\right)\right\}}, \#(3)$$

where n denotes the number of observations of d-dimensional data and$x^i(t[s])$ denotes the vector $x(s)$ with its i-th elementreplaced by the i-th element of $x(t)$, and $x^i(t[s])$ denotes thevector$x(t)$ with its i-th element replaced by the i-th elementof $x(s)$.This score reflects how changes in the data affect the estimated parameters and provides insight into the dependency structure captured by the model.Specifically, the conditional score takes a small value when a combination of variables conforms to the model's dependency structure, indicating that the data point is "plausible" for the model. Conversely, if the score takes a large value, it indicates a data point that deviates from

the dependent relationship between the variables, suggesting an outlier or a sample that does not fit the model. Thus, the magnitude of the conditional score is an indicator of the extent to which the data points deviate from the model's expected dependency trend. Thisproperty of the score allows us to use it for anomaly detection and model validation, and to quantitatively assess whether the model adequately captures the dependency structure. In addition, by identifying data points with high conditional scores, we can detect anomalous data that deviate from the dependent trend and provide important information for model refinement.

3. Results for Real Example

3.1. *Problem setting*

We performed an analysis using Tokyo-CLAY/14/67760 data. Specifically, we examined a validation site within Tokyo-CLAY/14/67760 spanning a 300/m × 300/m area that was intensively surveyed due to the presence of major structures, such as the airport terminal described by Otake et al. (2024). The area contains 50 boreholes within a 300m × 300m section, with a total of 973 data sets available.The subject site consists of three clay layers with different properties. The layers are labeled “Ysu”, “Ycu”, and “Ycl” from the uppermost layer downward (see Otake et al. (2024) for detailed descriptions of these labels). The site contains a total of eight variables: four geotechnical survey variables (s_u, PL, LL, w), three-dimensional coordinate data (X, Y, Z), and the strata label $Lb \in \{$“Ysu”, “Ycu”, “Ycl”$\}$. Finally, the seven variables (s_u, PL, LL, w, X, Y, Z) were standardized, and the categorical variable Lb was encoded by assigning 0 when Lb = “Ysu”, 1 when Lb = “Ycu” and 2 when Lb = “Ycl”.Figure 1 shows the three-dimensional subsurface geological model of the validation site used for validation. The soil classification was made based on comprehensive engineering judgment by the facility managers, as described in Otake et al. (2024). Subsequently, boundary surfaces were created to connect the change points of the soil layers. Based on these results, spatial interpolation was performed using simple Gaussian process regression. From this dataset $X \in \mathbb{R}^{924\times8}$, excluding the two boreholes with IDs 636 ($X_{s1} \in \mathbb{R}^{20\times8}$) and 684 ($X_{s2} \in \mathbb{R}^{29\times8}$), which are used as validation boreholes in this study, the dependent parameters $\underline{\theta} \in \mathbb{R}^{247}$ are estimated using equations (1) and (2).

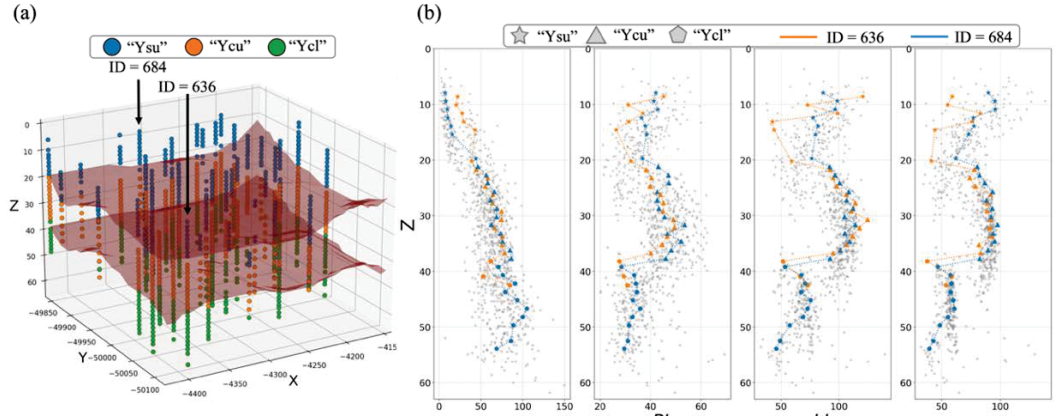

Fig. 1.(a) Visualization of all borehole data in the validation area, showing the locations of the boreholes and their respective layer labels. (b) Depth distribution of soil parameters in the validation area. The gray points represent the depth distribution of all boreholes shown in (a), while the orange and blue points indicate the data used for validation boreholes.

3.2. *Estimation of the strata label with conditional scores*

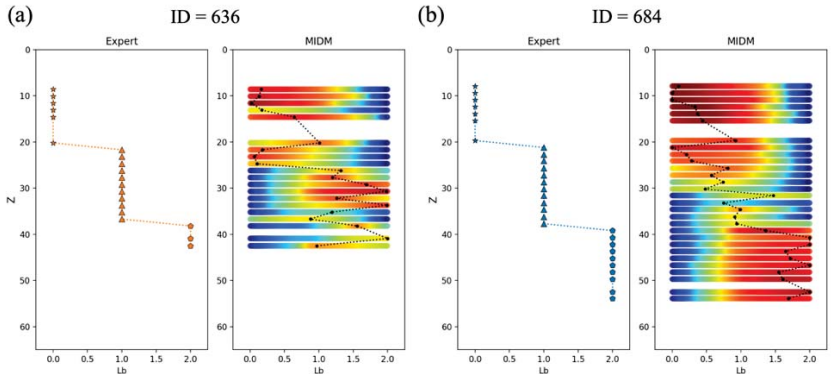

Fig.2. Soil layer classification results based on conditional scores. The first and third figures from the left show deterministic judgments by the facility managers, while the second and fourth figures show heat maps of the estimated results by MIDM. In the heat maps, red indicates lower scores, and blue represents higher scores. (a) Estimation results for Borehole ID=636, (b) Estimation results for Borehole ID=684.

The conditional scores given in Eq. (3) for the two validation boreholes are used to estimate the strata labels. We generate dummy variables for *Lb* in increments of 0.01 from 0 to 2 and calculate their scores under the condition that all other variables remain observed. Because "Ysu," "Ycu," and "Ycl" are encoded as 0, 1, and 2, respectively, examining how the dummy variables' scores vary allows us to determine which layer they are most similar to. Although the score is originally a vector comprising 247 combinations of the considered pairs, for simplicity we sum these vector components into a scalar score. Figure 2 presents a heatmap of the two validation boreholes after calculating and normalizing these scores. From the heatmaps, we observe depth-dependent changes in both boreholes that correspond to strata transitions. Expert judgments, illustrated in the depth profiles of Figure 1(b), rely on qualitative assessments of trends and abrupt changes in soil properties. In contrast, MIDM integrates this information seamlessly, facilitating strata label estimation by identifying deviations from the validation site's dataset trends. Notably, the model allows for continuous and probabilistic strata labeling, rather than relying solely on discrete classifications, thus enabling a more nuanced assessment of layer boundaries.

Overall, the model captures variations in borehole data, reflecting the inherent complexity of subsurface conditions. In contrast to expert-based classifications that often assume regular layer boundaries, our model indicates that strata can shift gradually with depth, detecting subtle transitions and revealing transitional zones rather than sharply defined boundaries. By providing a continuous score distribution, MIDM accommodates unique features within each borehole, suggesting its capacity for adapting to localized variations rather than imposing a uniform classification scheme.

Alignments or discrepancies between expert and MIDM classifications suggest how reliable MIDM could be for automated soil classification. If MIDM's predictions generally align with expert-identified layers, it strengthens the model's viability in practical geotechnical investigations. Its probabilistic outputs (via heatmaps) offer nuanced insights for engineering design decisions—particularly in projects sensitive to soil heterogeneity, such as foundation or slope stability work. By comparing these heatmaps with expert classifications, engineers can pinpoint areas where expert judgment might be subjective or inconsistent, and leverage MIDM's standardized approach to improve efficiency and consistency across multiple sites.

In engineering practice, relying on such models could reduce the time and cost associated with expert-dependent soil classification, especially when applied to large-scale projects with multiple borehole data.The ability to visualize soil classification confidence across depths provides valuable insights into soil layer heterogeneity, which could be beneficial for geotechnical analysis and decision-making in construction and engineering projects. Further validation studies could assess MIDM's accuracy and robustness across diverse soil conditions and geographical regions.

4. Conclusion

This study demonstrates that Minimum Information Dependence Modeling (MIDM) effectively evaluates dependencies in heterogeneous geotechnical datasets. By applying MIDM to the real soil data, we successfully estimated strata labels using conditional scores, capturing complex non-linear relationships often overlooked by traditional models. MIDM offers a more nuanced understanding of subsurface conditions, enhancing anomaly detection and model validation in geotechnical risk assessments.

However, MIDM's sampling-based approach results in high computational costs, which must be addressed for large-scale applications. Additionally, our analysis relied on a dataset with relatively small dimensionality; incorporating a larger and more diverse set of geotechnical variables is essential to fully realize the model's potential. Future research will focus on optimizing MIDM's computational efficiency through advanced algorithms or approximations. We also plan to broaden the model's applicability by integrating it with other data-driven techniques (e.g., Lyu et al., 2024) to construct high-density three-dimensional ground models. Validating MIDM across various soil conditions and regions will be crucial to establishing its reliability as a potential replacement for qualitative judgments in geotechnical investigations.This approach holds promise for transforming traditional practices toward more data-driven and quantitative methodologies.

References

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

Lyu, B., Wang, Y., and Shi, C. (2024). Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge.*Computers and Geotechnics 170*, 106336.

Otake, Y., Saito, T., Wu, S., Yoshida, I., and Takano, D. (2024). Exploring challenges via analysis of multivariate geotechnical properties: insights from large-scale local sampling of Japanese marine clay. InK. Phoon and C. Tang (Eds.),*Databases for Data-Centric Geotechnics*, Chapter 11. Taylor & Francis.

Saito, T., Otake, Y., Wu, S., Takano, D., Sugiyama, Y., and Yoshida, I. (2025). What defines a “site” in geotechnical engineering?: A comparative study between local and global big indirect databases.*Computers and Geotechnics 177(A)*, 106826.

Sei, T., and Yano, K. (2024). Minimum information dependence modelling. *Bernoulli 30(4)*, 2623-2643.