Integration of Multivariate Normal Distribution Model in Geotechnical Analysis of Clay Properties

Kosei Kawata[1]

[1]*Technical Research Institute, Toyo Construction Co., Ltd. Email: kawata-kosei@toyo-const.co.jp*

Abstract: This study demonstrates an integrated approach using a multivariate normal distribution model in conjunction with Johnson Transformation and Mahalanobis distance to analyze clay properties from a global database. This method provides a robust framework for comparative site analysis and predictive modeling of incomplete records, offering enhanced accuracy and reliability in geotechnical assessments.

Keywords: clay properties, multivariate normal distribution, Johnson distribution, Mahalanobis distance.

1. Introduction

This study focuses on leveraging a global database of clay properties and employs statistical methods for comparative analysis and predictive modeling in site evaluation. It is centered around the application of multivariate normal distribution principles to comprehensively understand the interactions among complex geotechnical properties. The primary objective of constructing a multivariate normal distribution model in this context is to accurately represent the correlations that exist between various clay properties.

An integral part of this analysis is the use of the Mahalanobis distance formula. This formula is instrumental in comparing probability distributions and is widely recognized for its effectiveness in statistical analysis. It is particularly adept at quantifying the dissimilarity between two probability distributions, each characterized by unique means and variances. (Similarity measure (Dec. 1, 2023, 13:00, (UTC+0900)). In Wikipedia) This feature is very useful in identifying similarities within a data set and is an important aspect of geotechnical data analysis.

Recent advancements in the field have seen numerous researchers developing databases and constructing multivariate normal distribution models to better understand soil behavior ((Ching and Phoon 2014) for detailed examples). A significant contribution to this area of study is the Johnson System of Distributions, introduced by N.L. Johnson in 1949. This system encompasses a variety of distribution types that are particularly effective in transforming non-normal data distributions into a normal form. In this paper, we specifically apply the Johnson System to model clay properties, thereby enhancing the predictive accuracy and reliability in geotechnical engineering applications.

2. Methods

2.1 Clay properties

In this study, five clay properties (x1: undrained shear strength (s_u), x2: preconsolidation stress (σ'_p), x3: corrected cone tip resistance (q_c), x4: plastic index (PI), x5: vertical effective stress (σ'_v)). A total of 133 data points are available for modeling in global database. Also, there is a target site that contains 10 records (Target record 1, Target record 2, …, Target record 10). Target record 8, Target record 9, Target record 10 record only includes qt, PI and σ'_v.

2.2 Jonson distribution fit for x.

The above five parameters are normalized by the Johnson transform in Slifker and Shapiro (1980). The distribution types and distribution parameters for x1, x2, x3, x4, and x5 obtained from the above procedure are summarized in Table 1. The fitted Johnson distributions are plotted in Figure 1, along with empirical histograms constructed from the x data. The Johnson distribution will be employed in the analysis that follows.

Table 1. Distribution type and distribution parameters for x1, x2, x3, x4, and x5.

Random variable	Clay properties	Distribution type	Distribution parameters				
			ζ	η	γ	λ	ε
x1	s_u	SU	0.5005	1.736	-3.401	6.609	2.385
x2	σ'_p	SU	0.5240	0.712	-0.836	17.907	84.105
x3	q_c	SU	0.5240	0.948	-1.821	86.170	238.949
x4	PI	SU	0.5240	1.564	-0.406	18.125	31.039
x5	σ'_v	SU	0.5300	1.103	-2.409	15.245	7.596

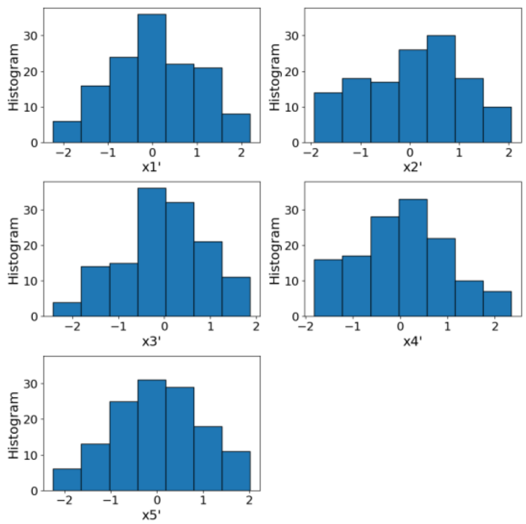

Figure 1. Example of a figure

2.3 Multivariate normal distribution model

The standard normal distribution model is used as the model, expressed by the following equation.

$$N(\boldsymbol{z}|\boldsymbol{C}) = \frac{1}{\sqrt{(2\pi)^M|\boldsymbol{C}|}} exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{C}^{-1}\boldsymbol{z}\right) \quad (1)$$

Where z is the parameter vector after Johnson transform of x and C is the covariance matrix. The final covariance matrix is shown below.

$$C = \begin{pmatrix} 1 & 0.83 & 0.75 & 0.30 & 0.63 \\ 0.83 & 1 & 0.82 & 0.06 & 0.71 \\ 0.75 & 0.82 & 1 & 0.06 & 0.91 \\ 0.30 & 0.06 & 0.06 & 1 & -0.07 \\ 0.63 & 0.71 & 0.91 & -0.07 & 1 \end{pmatrix} \quad (2)$$

2.4 Task 1: Site similarity analysis

We employed a multivariate normal distribution model to assess the similarities between the target site and database sites. This process entailed the calculation of covariance matrices and their corresponding inverses. Subsequently, we utilized the Mahalanobis distance to facilitate a detailed comparative analysis between sites. (Bishop, C.M. (2006).)

The first step in our analysis was to compute the mean vector for each site's dataset within the database. This mean vector, denoted as $\bar{x}$, represents the average value across all measured variables for a given site. Following this, we calculated the Mahalanobis distance for each site relative to the target site. The Mahalanobis distance, $\mathrm{D(x)}$, is defined as:

$$D(x) = \sqrt{(x-\bar{x})^T S^{-1}(x-\bar{x})} \quad (3)$$

where x represents the data vector for the target site, $\bar{x}$ is the mean vector for a database site, and S^{-1} is the inverse of the covariance matrix.

2.5 Task 2: Predictive modeling for incomplete records

The multivariate normal distribution model can be interpreted as a probability-based transformation equation that allows for the estimation of unknown parameters given a given parameter, provided that the parameter is included in the model. For the incomplete records (Target record 8, Target record 9, Target record 10), we expanded the multivariate normal distribution model by recalculating covariance matrices for the parameters of identified sites. We utilized the mean and variance of the conditional distributions derived from known values to estimate missing properties.

3. Results

3.1 Task 1: Site similarity analysis

The results of Task 1of 7 records (Target record 1, Target record 2, …, Target record 7) are summarized in the Table 2 below.

Table 2. Distribution type and distribution parameters for x1, x2, x3, x4, and x5.

Target record	Most similar site index	Mahalanobis distance
Target record 1	13	1.772
Target record 2	3	0.658
Target record 3	13	1.129
Target record 4	10	0.299
Target record 5	13	0.416
Target record 6	3	0.377
Target record 7	13	0.401

3.2 Task 2: Predictive modeling for incomplete records

The prior and posterior distributions for x1' and x2' in Target record 8 are shown in Figure 2,3. Prior in the figure represents the constructed multivariate normal distribution model, and posterior represents the probability distribution model given x3': corrected cone tip resistance (qc), x4': plastic index (PI), x5': vertical effective stress (σ'v). The same was done for Target record 9 and Target record 10, and the respective mean_x1 (kN/m^2) and 95%CL_x1 (kN/m^2) and mean_x2 (kN/m^2) and 95%CL_x2 (kN/m^2) are obtained in Table 3.

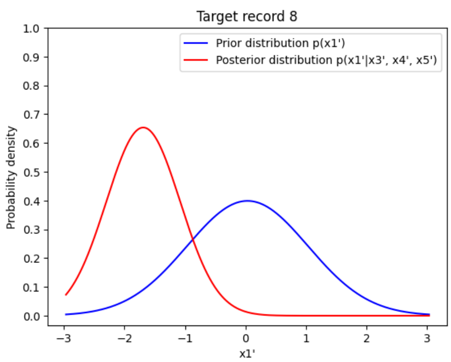

Figure 2. Conditional probability density distribution based on x3' = 0.030, x4' = 1.077, and x5' = 0.368 for normalized undrained shear strength(x1')

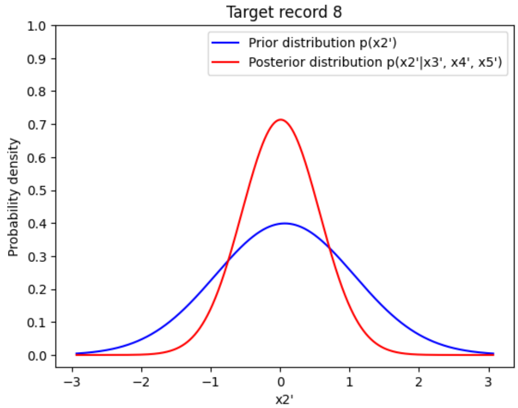

Figure 3. Conditional probability density distribution based on x3' = 0.030, x4' = 1.077, and x5' = 0.368 for normalized preconsolidation stress (x2')

Table 3. Distribution type and distribution parameters for x1, x2, x3, x4, and x5.

Target record	*mean_x1 (kN/m^2)*	*95%CL_x1 (kN/m^2)*	*mean_x2 (kN/m^2)*	*95%CL_x2 (kN/m^2)*
Target record 8	*29.60*	*[15.57, 57.73]*	*110.75*	*[77.68, 220.50]*
Target record 9	*14.69*	*[7.11, 27.93]*	*81.57*	*[35.40, 116.82]*
Target record 10	*10.01*	*[4.38, 19.56,]*	*67.93*	*[0, 97.76]*

4. Conclusions

This study introduces a combination of Johnson Transformation, Mahalanobis distance, and multivariate normal distribution models in geotechnical engineering. Our approach not only enhances the accuracy of site similarity analysis but also provides a reliable method for predicting missing properties in geological assessments.

References

Similarity measure: https://en.wikipedia.org/wiki/Similarity_measure

Ching J, Phoon KK (2014) Correlations among some clay parameters — the multivariate distribution. Can Geotech J 51:686–704.

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 811 36, 149-176Williams, R.A. 1987. *Communication Systems Analysis and Design*. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Slifker, J.F. and Shapiro, S.S. (1980). The Johnson system: selection and parameter estimation. 841 Technometrics, 22(2), 239-246.

Bishop,C.M. (2006). Pattern Recognition and Machine Learning, Chapter 9: Mixture models and EM Springer, 423-459.