Cone Penetration Testing 2022 – Gottardi & Tonni (eds)

Clustering analysis to improve total unit weight prediction from CPTu

S. Collico & M. Arroyo
Department of Civil and Environmental Engineering (DECA), Universidad Politecnica de Cataluña, Barcelona, Spain

M. DeVincenzi, A. Rodriguez & A. Deu
Igeotest s.l. Figueres, Spain

ABSTRACT: Accurate estimates of soil unit weight are fundamental for correctly post process CPTu data and making use of Soil Behavior Type-based classification systems. Soil-specific and global regressions have been proposed for this purpose. However, soil-specific correlation might pose a problem of pertinence when applied at new sites. On the other hand, global correlations are easy to apply, but generally carry large systematic uncertainties. In this context, this work proposes a data clustering technique applied to geotechnical database aiming to identify hidden linear trends among dimensionless soil unit weight and normalized CPTu parameter according to some unobservable soil classes. Global correlations are then revisited according to such data subdivision aiming to improve accuracy of soil unit weight prediction while reducing transformation uncertainty. A new probabilistic criterion for soil unit weight prediction is also obtained. The potential benefits of the proposed procedure are illustrated with data from a Llobregat delta site (Spain).

1 INTRODUCTION

CPTu interpretation is almost always based on stress normalized cone readings. For instance, soil classification charts (Robertson, 1990; Been & Jefferies, 2006; Schneider et al., 2008) use stress normalized tip resistance, friction and excess pore pressure parameters. Total soil unit weight γ_t is thus a necessary input in any CPTu interpretation exercise.

Soil unit weight is measured on undisturbed samples from boreholes. Such samples are not always available, particularly for granular soils and/or in early stages of site investigation. As a simpler alternative, total soil unit weight can be estimated indirectly from piezocone and seismic piezocone readings. In the last decades several proposals have been presented to achieve this purpose (Mayne, 2009; Mayne et al., 2010; Robertson & Cabal, 2010; Mayne, 2014; Lengkeek, 2018).

These proposals are all based on empirical regression using databases in which cone readings are paired with soil unit weight measurements. A trade-off typically arises between coverage (the scope of the original soil database) and precision (the predictive power of the regression). Global regressions, usually described as applicable to soils with a "normal" or common mineralogy, are widely applicable but less precise than regressions developed only for a certain class of soil (i.e., soil specific). On the other hand, soil specific regressions are more difficult to apply at a particular site because they have less coverage and require a previous soil classification step, which may introduce additional uncertainty. One possible way out of this problem is to develop soil-specific regressions for unit weight using a global database that is segmented into soil classes for regression purposes. How to define those classes?

In this work we employ a Gaussian Mixture Model (a.k.a., GMM) technique to identify hidden classes in a global database described by Mayne, (2014) with the purpose of establish more accurate regressions. GMM have been previously applied to CPTu data analyses (Depina et al., 2016; Krogstad et al., 2018) to identify soil classes for stratigraphic delineation. In those studies, a Bayesian perspective was introduced, as having the possibility of updating the stratigraphic groups as more information was gathered was deemed essential. In the present work a fixed database in considered and therefore the Bayesian updating aspect has been omitted.

A key step of applying such data clustering technique is to express the reference database, including soil unit weight and CPTu data, as a multivariate normal distribution before GMM is applied. To this end we use a methodology laid out by Ching et al., (2014), to rationally account for predictor co-dependence and to establish or revise correlations between different variables.

In what follow, key steps to construct valid multivariate distribution are reported followed by a brief

DOI: 10.1201/9781003308829-45

description of finite Gaussian Mixture Models. In the results section, an example of data subdivision and assignment to identified hidden classes is reported using dimensionless soil unit weight and normalized CPTu parameters. Existing correlations are re-examined using the new soil classes and a new one is proposed.

2 METHODOLOGY

2.1 *Setting up a multivariate distribution*

Ching et al., (2014) and Phoon & Ching, (2018) proposed a systematic cumulative transform procedure to build standard multivariate normal distributions for multivariate databases. Such approach is adopted in this study to treat the Mayne, (2014) database.

Essentially the steps involved are:

1. For a given set of selected observations $\underline{\Omega}$ (e.g., $\underline{\Omega} = [\Omega_1, \Omega_2]$ with Ω_1= [nx1] and n number of observations) define marginal distributions Ω_i for each of the component variates in the database;
2. Transform the different marginals into standardized normal distributions $\mathbf{X}_i$
3. Assume that transformed observations $\underline{\mathbf{X}} = [\mathbf{X}_1, \mathbf{X}_2]$ follow a multivariate normal distribution $f(\underline{\mathbf{X}})$;
4. Obtain a random sample from $f(\underline{\mathbf{X}})$, and transform it back to the non-normalized space, obtaining the simulated set $\underline{\Omega}_{sim}$
5. Check by inspection the overlap between $\underline{\Omega}_{sim}$ and $\underline{\Omega}$.

A key step in the procedure is the definition of marginal distributions for the different variates in the database (in this case dimensionless soil unit weight and normalized CPTu parameters). Lognormal distributions have been frequently used in geotechnical applications (Phoon & Kulhawy, 1999). The lognormal is included in the Johnson system of distributions (Ching et al., 2014), which has three main families: the lognormal system S_L, the bounded system S_B and unbounded system S_U. Choosing between the whole system of Johnson distributions offers versatility while maintaining the interesting property of having analytical expressions to transform into standard normal variables and back (Ching et al. 2014).

The selection of an appropriate Johnson distribution for a particular dataset is based on the first four moments of a sampling distribution (George & Ramachandran, 2011). Examples of Johnson distributions fitted to different variates of the Mayne, (2014) global database are given in (Figure 1).

2.2 *Multivariate as a gaussian mixture*

The multivariate distribution that represents the database might be conceived as a combination of several

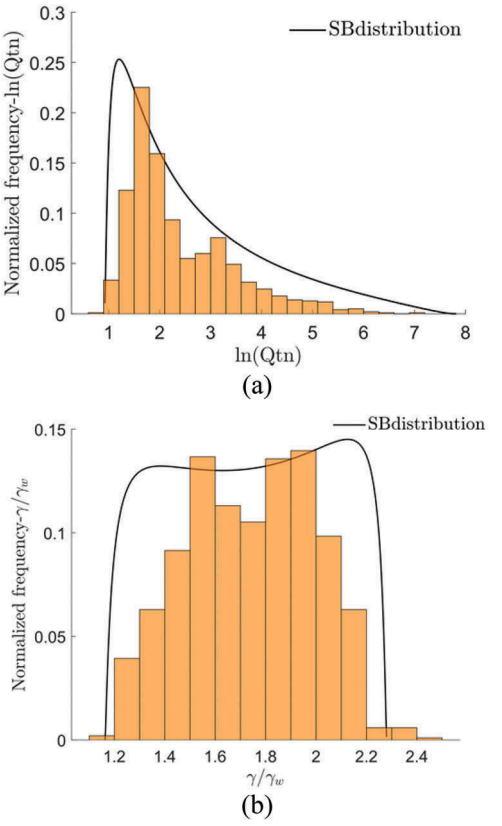

Figure 1. Fitting Johnson probability density function to a) ln (Q_{tn}); b) γ_t/γ_w data of Mayne, (2014) database.

underlying components (the hidden classes that we are trying to identify). This is formalized using a Gaussian Mixture Model, (Depina et al., 2016; McLachlan & Peel, 2000), that expresses the transformed -standardized- multivariate distribution as a linear combination of K multivariate gaussians:

$$f(\underline{x_i}|\zeta) = \sum_{j=1}^{K} \pi_j \phi_j(\underline{x_i}|\Theta_j) \qquad (1)$$

Each ϕ_j represents a component of the mixture and is a multivariate normal probability density function with the same structure as $f(\underline{\mathbf{X}})$. Therefore, each component has its own statistical parameters, denoted by Θ_j In our case they include a vector of means and a covariance matrix (i.e., μ_j, $\underline{\underline{\sum}}j$). The collection of all the Θ_j is denoted by Θ. Each component has a weight π_j (proportion of data assigned to ϕ_j); weights are chosen so that they add up to one and are collected in a vector $\underline{\Pi} = (\pi_1,.., \pi_K)$. $\zeta =$ collection of all the unknown parameters of the mixture model (i.e., $[\Theta; \underline{\Pi}]$).

Depending on the values that ζ finally takes each observation $\underline{x_i}$ would have a certain probability p_{ij} of belonging to a particular component ϕ_j.

Gaussian Mixture analysis consists of estimating the most probable ζ, $\hat{\zeta}$. This is generally done through the Expectation-Maximization (EM) algorithm (Samé et al., 2011; Huang et al., 2017; Liu et al., 2019) which alternates an expectation step in which observations are assigned exclusively to a particular gaussian component and a maximization step, in which the log-likelihood function for the incomplete dataset is maximized.

The concept of GMM could be also formulated in a Bayesian framework, by integrating prior knowledge p (ζ) and observations $\underline{\Omega}$ to obtain posterior estimates of ζ. Due to lack of previous studies, this work only applies GMM. For more detailed information about the Bayesian formulation, analysis about optimal number of hidden classes to consider and most plausible normalized CPTu parameters to be considered, the reader is referred to Collico, (2021).

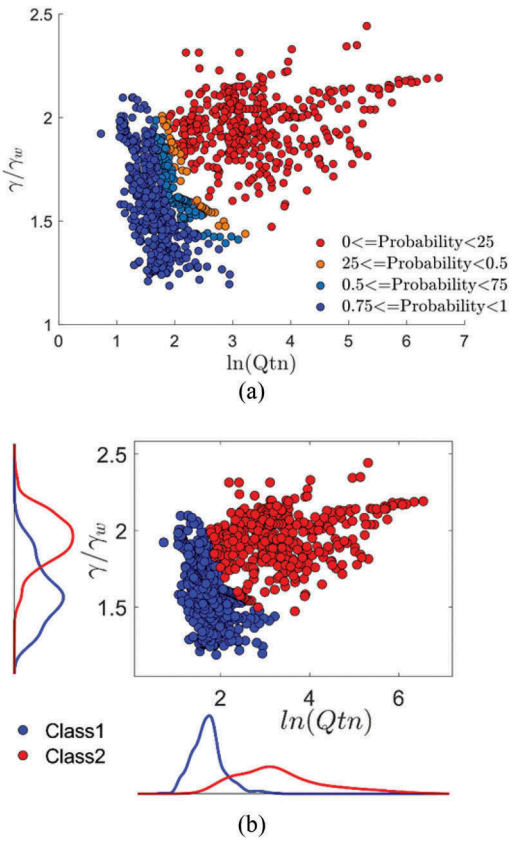

Figure 2. A) Assigned probability of belonging to component 1 of the GMM b) Scatter and marginal distributions of observations for the two hidden classes identified.

3 RESULTS

3.1 *Cluster definition and analysis*

There are six variates in the Mayne, (2014) database. To illustrate the methodology proposed in a simple setting, we consider only two of them, namely dimensionless unit weight γ_t/γ_w and normalized tip resistance, $ln(Q_{tn})$. After fitting a bivariate distribution to all the data pairs $\gamma_t/\gamma_w - ln(Q_{tn})$ we run the expectation-maximization algorithm to identify a GMM with two components.

An example illustrating the probabilities of data belonging to a particular component of the Gaussian mixture is given in Figure 2a. Clustering is based on such probabilities. A simple choice is to assign data to the component in which they have the largest probability of belonging. For two components this is equivalent to enforcing a probability threshold of 0.5 as clustering criteria. The result of doing so in this case is reported in Figure 2b, while statistics for the two clusters in the $\underline{\Omega}$ space are reported in Table 1.

To understand the meaning of this newly identified soil classes, the clustered data is plotted in Soil Behavior Type charts (Robertson, (2016); Schneider et al. (2008) (Figure 3). Results show that the first hidden class identified is dominated by Clay-Like-Contractive soils ($C-C$, Figure 3a), while the second hidden correlation identified is represented by a wider range of conventional soil types.

Table 1. Covariance matrix and mean vector of γ_t/γ_w-$ln(Q_{tn})$ hidden classes identified.

Class	Mean		Covariance		
1	μ_{γ_t/γ_w}	$\mu_{ln(Q_{tn})}$		γ_t/γ_w	$ln(Q_{tn})$
	1.59	1.71	γ_t/γ_w	0.04	-0.036
			$ln(Q_{tn})$	-0.036	0.125
2	μ_{γ_t/γ_w}	$\mu_{ln(Q_{tn})}$		γ_t/γ_w	$ln(Q_{tn})$
	1.95	3.32	γ_t/γ_w	0.025	0.04
			$ln(Q_{tn})$	0.04	0.94

3.2 *Generating cluster-based correlations*

The clustered data can be used to generate new correlations. Such correlations need not be based on the same restricted subset of the global database that was used to generate the clustering. For instance, we use here correlations that follow a template proposed by Mayne, et al., (2010) (i.e., γ_t=a+ b·log(z) + c·log(f_s)+ d·log(q_t)). Results in terms of coefficient of determination, R^2 and regression standard error, σ_{ε_T} are reported in Table 2, while regression coefficients are reported in Table 3. The standard error, after GMM subdivision, is at least 13% smaller than the one of global database. The coefficient of determination $R^2$is also smaller for the clusters than for the global database. These effects are particularly strong for Hidden class 2.

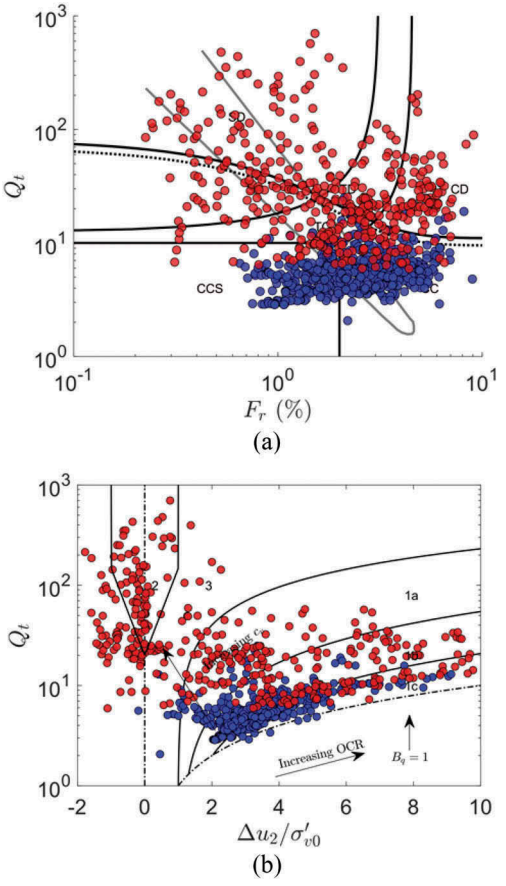

Figure 3. Scatter plot of data belonging to each hidden cluster identified on SBT charts: a) Robertson, (2016); b) Schneider et al., (2008).

Table 2. Coefficient of determination and regression standard error before and after database subdivision.

	R^2	σ_{ε_T}
Global	0.66	1.44
Hidden class1	0.6	1.25
Hidden class 2	0.39	1.22

However, those comparisons are somewhat misleading, as the statistics are computed using different observations. The improvement of predictive strength after clustering is more clearly identified through a cross-validation procedure. To this end we run a simulation exercise in which we randomly selected 85% of the data in each cluster and used them to fit cluster specific correlations, as well as a global one in which no cluster distinction was made. Then we applied those correlations, to the remaining 15% of the dataset -the validation data. The sum of squared residuals $||SS_{res}||$ was computed at each trial, for each one of the correlations (global, cluster 1, cluster 2). 500 such simulations were performed. Results, reported in Figure 4 for both Hidden classes, highlight the benefit of the GMM as the distribution of error norms is clearly shifted towards lower values.

Table 3. Coefficient of the regression for linear form of Mayne (2010) after BMA subdivision.

	a	b	c	d
Global	8.78	-0.67	2.24	1.457
Class1	-0.78	-4.2	7.77	-0.57
Class2	15.34	0.052	0.02	1.94

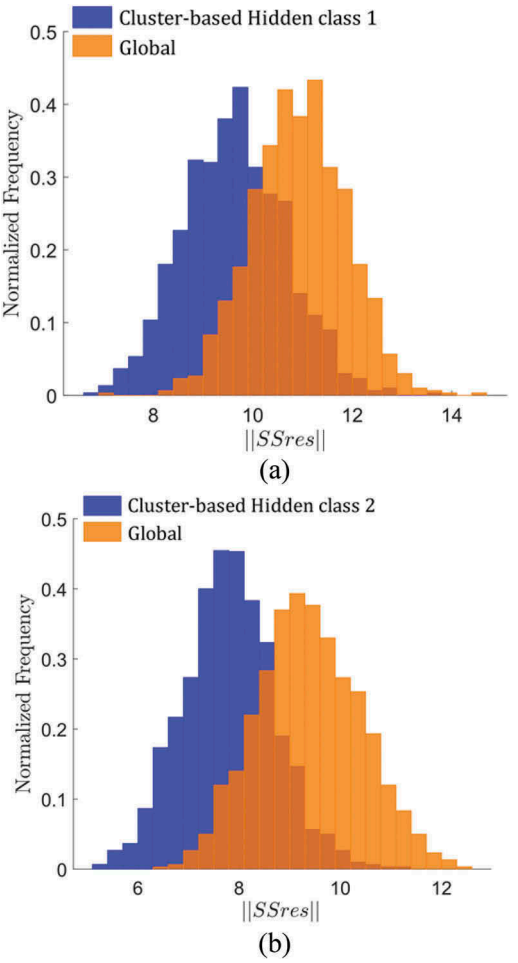

Figure 4. Cross validation of revisited regression and comparison with global literature correlation for a) Hidden class1. B) Hidden class 2.

3.3 *Application to new sites: Decision boundary*

To assign CPTu observations at a new site to a cluster, we use the (non- normalized) Robertson, (1990) SBT chart. Plotting the clustered data in that chart we apply discriminant analysis (Ghojogh & Crowley, 2019) to establish a user-friendly separation criteria for new CPTu observations. The decision boundary (Figure 5) obtained takes the shape of a quadratic in Robertson (1990) chart. This boundary line between the two clusters has the following expression:

$$2.37 + \left[0.2\ln(R_f) - 1.58\ln(q_t)\right]^2 \pm \sigma_{boundary} = 0 \quad (2)$$

with $\sigma_{boundary}$ systematic uncertainty associated with decision boundary ($\sigma_{boundary} = 0.13$).

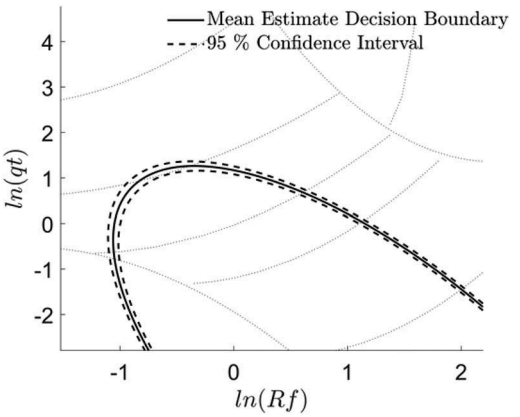

Figure 5. Mean estimate and 95% Confidence Interval of decision boundary.

3.4 *Illustrative example*

The cluster-based correlations just developed are now applied to results of a particular CPTu campaign performed at a Llobregat site (SW Barcelona-Spain). Several infrastructures have been developed around this site during the last decades, requiring extensive onshore-nearshore geotechnical investigation. Along with in-situ investigation, laboratory tests were performed on selected sub-samples from undisturbed Shelby tubes recovered every 5m in each borehole. For this study we consider data from 20 CPTu and 44 total soil unit weight measurements. More detail on this campaign is given by Deu et al., (2021).

All the CPTu data were plotted on Robertson graph, showing that the site is dominated by silty soils. Then each CPTu observation was assigned to class 1 or 2 according to the mean value of the decision boundary (Figure 6a). Most of the data were assigned to Hidden class 2.

The cluster-based correlation was applied to the 20 CPTu data and the results are plotted in (Figure 6b), indicating the mean and spread of the predicted values. The same is done using the global correlation, obtained using the whole database, without clustering. Summary results are also presented in Table 4. They include the

Table 4. Statistics of mean value of total unit weight prediction.

	μ_{γ_t}	$\sigma_{\mu_{\gamma_t}}$
Global correlation	17.47	0.48
Cluster-based correlation	18.1	0.42
Laboratory	19.45	0.66

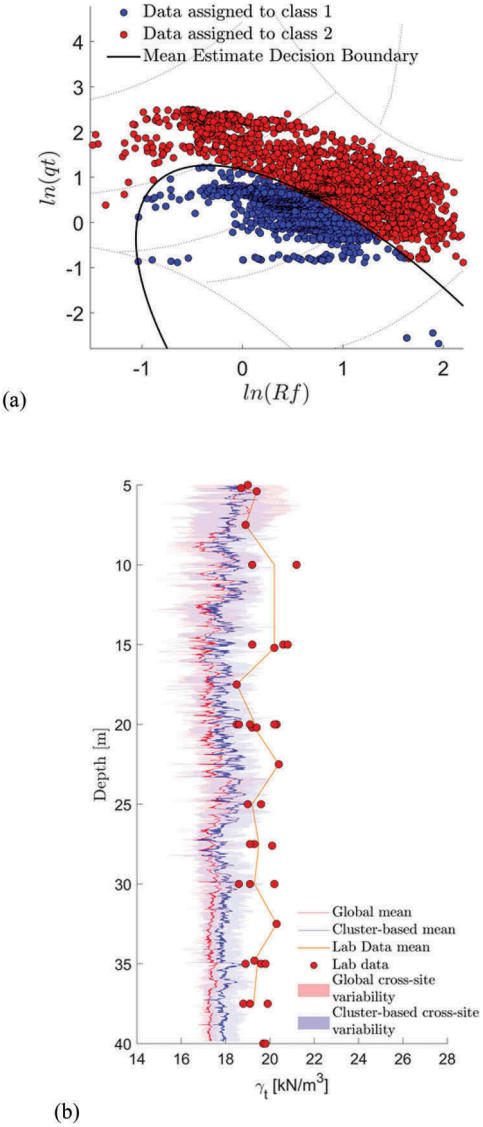

Figure 6. A) CPTu data assigned to class 1 and class 2 according to the proposed criterion. b) Profile of total unit weight prediction for cluster-based and global correlations and laboratory observations.

overall predicted mean estimate as well as the standard deviation of the means evaluated at different depths.

The cluster-based correlation clearly exhibits an improvement of total unit weight prediction at the site (Figure 5b), when compared global correlation. It has a lower dispersion and its mean is closer to that of the laboratory measurements. It is clear, however, that even the cluster-based correlation underestimates the laboratory mean. There are several possible explanations for this discrepancy. One is that samples of silty soils generally tend to be densified upon sampling (Lim et al. 2018). The other is that the reference database is somewhat scarce in the silt area were most datapoints in the example lie (compare Figures 6a and 3a).

4 CONCLUSION

This study describes a first attempt to apply GMM clustering to geotechnical database.

The clustering technique has been applied in a space parametrized by dimensionless soil unit weight and normalized cone tip resistance, assuming the existence of two unobserved classes, selected at 50% probability. One of emerging hidden classes can be associated to Clay-Like-Contractive class, while the second is representative of a wide range of soil types. Assignation of new data to the hidden classes is based on a discriminant line generated on non-normalized SBT charts (Robertson, 1990). CPTu data subdivision according to such classes has been shown to increase the predictive strength of correlations. This has been illustrated using a correlation template from Mayne et al., (2010) in which unit weight is predicted from shaft friction, depth and tip resistance.

Beyond the particular application explored here, the technique presented here may prove useful to improve accuracy and precision of empirical CPTu based correlations established using global databases.

REFERENCES

Ching, J., Phoon, K., & Chen, C. (2014). Modeling piezocone cone penetration (CPTU) parameters of clays as a multivariate normal distribution. 91(July 2012), 77–91.

Collico, S., (2021). Geotechnical characterization of marine sediments via statistical analysis (Unpublished doctoral dissertation). Universidad Politècnica de Cataluña (UPC), Barcelona, Spain.

Depina, I., Le, T. M. H., Eiksund, G., & Strøm, P. (2016). Cone penetration data classification with Bayesian Mixture Analysis. Georisk, 10(1),27–41. https://doi.org/10.1080/17499518.2015.1072637

Deu, A., Marti X., Peña S., Tarragò D., Gens A., & Devincenzi, M. (2021) DMT, CPTU and laboratory tests comparison for soil classification and strength parameters of deltaic soft soils in Barcelona Port, Proceedings of the 6th International Conference on Site Characterization ISC-6 Budapest

George, F., & Ramachandran, K. M. (2011). Estimation of parameters of johnson's system of distributions. Journal of Modern Applied Statistical Methods, 10(2),494–504. https://doi.org/10.22237/jmasm/1320120480

Ghojogh, B., & Crowley, M. (2019). Linear and Quadratic Discriminant Analysis: Tutorial. 4, 1–16. http://arxiv.org/abs/1906.02590

Huang, T., Peng, H., & Zhang, K. (2017). Model selection for Gaussian mixture models. Statistica Sinica, 27 (1),147–169. https://doi.org/10.5705/ss.2014.105

Krogstad, A., Depina, I., & Omre, H. (2018). Cone penetration data classification by Bayesian inversion with a Hidden Markov model. Journal of Physics: Conference Series, 1104(1). https://doi.org/10.1088/1742-6596/1104/1/012015

Lengkeek, H. J. (2019). CPT based unit weight estimation extended to soft organic soils and peat. January 2018.

Lim GT, Pineda JA, Boukpeti NA, Fourie AN, Carraro JA. Experimental assessment of sampling disturbance in calcareous silt. Géotechnique Letters. 2018 Sep;8(3):240–7.

Liu, Y., Ye, L., Qin, H., Ouyang, S., Zhang, Z., & Zhou, J. (2019). Middle and Long-Term Runoff Probabilistic Forecasting Based on Gaussian Mixture Regression. Water Resources Management, 33(5),1785–1799. https://doi.org/10.1007/s11269-019-02221-y

Mayne, P.W., Peuchen, J., & Bouwmeester, D. (2010). Soil unit weight estimation from CPTs. 2nd International Symposium on Cone Penetration Testing, 2 (May),8. https://doi.org/10.1201/b10132-41 10.1201/b10132-41

Mayne, P. W. (2014). Interpretation of geotechnical parameters from seismic piezocone tests. 3rd International Symposium on Cone Penetration Testing (CPT`14), 47–73.

Mayne, Paul W, Coop, M. R., Springman, S. M., & Zornberg, J. G. (2009). Geomaterial behavior and testing. Proceedings of the 17th International Conference on Soil Mechanics and Geotechnical Engineering: The Academia and Practice of Geotechnical Engineering, 5, 2777–2872.

McLachlan, G. J., & Peel, D. (2000). Finite mixture models. John Wiley & Sons.

Phoon, K. K., & Ching, J. (2018). Risk and Reliability in Geotechnical Engineering. In Risk and Reliability in Geotechnical Engineering. https://doi.org/10.1201/b17970

Phoon, Kok Kwang, & Kulhawy, F. H. (1999). Characterization of geotechnical variability. Canadian Geotechnical Journal, 36(4),612–624. https://doi.org/10.1139/t99-038

Robertson, P. K. (1990). Soil classification using the cone penetration test. Canadian Geotechnical Journal, 27 (1),151–158. https://doi.org/10.1139/t90-014

Robertson, P. K. (2016). Cone penetration test (CPT) - based soil behaviour type (SBT) classification system — an update. 1927(July), 1910–1927.

Robertson, P. K., & Cabal, K. L. (2010). Estimating soil unit weight from CPT. May.

Samé, A., Chamroukhi, F., Govaert, G., & Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. Advances in Data Analysis and Classification, 5(4),301–321. https://doi.org/10.1007/s11634-011-0096-5

Schneider, J. A., Randolph, M. F., Mayne, P. W., & Ramsey, N. R. (2008). Analysis of factors influencing soil classification using normalized piezocone tip resistance and pore pressure parameters. Journal of Geotechnical and Geoenvironmental Engineering, 134 (11),1569–1586. https://doi.org/10.1061/(ASCE)1090-0241(2008)134:11(1569)