Gaussian Mixture Model Based Soil Classification Using Multiple Cone Penetration Tests

Djamila Bouayad\textsuperscript{1,}\textsuperscript{*}, Julien Baroth\textsuperscript{2} and Christophe Dano\textsuperscript{2}
\textsuperscript{1}Department of Civil Engineering, Bejaia University, Route de Targua Ouzemour, 06000 Bejaia, Algeria
\textsuperscript{2}Grenoble Alpes University, CNRS, Grenoble INP, 3SR, F-38000 Grenoble, France

*Corresponding author email: djamila.bouayad@univ-bejaia.dz

Abstract. This study presents an application of the Gaussian mixture (GM) method for soil classification using multiple cone penetration tests (CPT). Compared to the hard clustering methods, the GM model classifies the CPT data by representing the probability density function of observed variables as a mixture of multivariate normal distributions. A GM model based expectation maximization (EM) algorithm with Bayesian information criterion (BIC) for selecting the optimal number of clusters is developed using six real CPT data performed at Dunkerque site in the north of France. The classification results are compared with the classical CPT based interpretation using the non normalized soil behavior type (SBT) index together with the Robertson chart. The results show that the GM model is able to identify accurately the soil layers. In addition, the combination of all CPTs, rather than considering them separately, may improve the soil layers identification because all the site information is considered.

Keywords: Gaussian mixture model; Cone penetration test; Soil classification; Probabilistic clustering.

1. Introduction
Conce penetration tests (CPT) are widely used for the identification of subsurface stratigraphy and soil classification. Traditional methods based on the charts linking CPT parameters to soil types are the most common in soil classification practice. Several charts were proposed in the literature (Schmertmann 1978; Douglas and Olsen 1981; Robertson et al. 1986; Robertson 1990; Eslami and Fellenius 2004; Robertson 2010) but the charts proposed by Robertson et al. (1986) and Robertson (1990) have become very popular. Recently, Robertson (2010) updated the early Robertson et al. (1986) chart to match the soil behavior types of the normalized chart (Robertson, 1990). Although these charts are very common for soil classification, they provide guidelines to the soil behaviour type and they could not properly identify the true number of layers. Therefore, these charts may involve uncertainties and misclassification of soil type especially in the transition zones and thin layers. In addition, the variability of soil properties, even in an homogenous layer, may lead to uncertainties in soil profiles. Hence, to overcome these limitations, several probabilistic approaches have been developed (Zhang and Tumay 1999; Tumay et al. 2008; Jung et al. 2008; Cetin and Isik 2007; Wang et al. 2013; Kurek et al. 2014, Li et al. 2016; Huang et al. 2018). Most of these studies were based on the normalized CPT parameters and focused on the interpretation of single CPT sounding which may lead to inaccurate soil classification especially for thin layers where little amount of data are recorded. Hence, considering multiple CPT sounding simultaneously allows fully taking into account the CPTs information and identify the entire subsurface region of a project site.
In the present study, a Gaussian mixture (GM) model is developed for soil classification based on multiple CPT sounding data. Compared to hard clustering methods (Facciorusso and Uzielli 2004; Rogiers et al. 2017), the GM model classify the CPT data by representing the probability density function of observed variables as a mixture of multivariate normal densities. The construction of GM model is based on the values of dimensionless cone resistance \((q_c/p_a)\) and the friction ration \((R_f = f_s/q_c\) (in percent), \(f_s\) is the sleeve friction) because they are reasonably represented by a normal distribution. The proposed model is validated using data of six CPTs performed at Dunkerque site in the northern of France, and the classification results are compared with those given by the Robertson (2010) chart based on non-normalized soil behaviour type index. In this case, it is considered that the conditionally independent assumption of the GM method, in which the CPT data points are regarded as independent, does not influence the results because the vertical correlation distance is very small \((\delta < 0.5 \text{ m})\) compared to the soil layer thickness as it was emphasized by Depina et al. (2016) and Wang et al. (2019).

2. Robertson Classification Method

The traditional interpretation of CPT measurements are generally accomplished using the Robertson et al. (1986) and Robertson (1990) charts. The early chart proposed by Robertson et al. (1986) is based on the corrected cone resistance \(q_t\) (or cone resistance, \(q_c\) and friction ration \((R_f)\) and has 12 soil types, while the Robertson (1990) chart uses CPT parameters normalized in terms of effective stress and has 9 soil types. Robertson (2010) updated the early chart in terms of the dimensionless cone resistance \((q_c/p_a)\) and the friction ration \((R_f)\) to match the 9 soil types of Robertson (1990) chart given in table 1. The advantage of non-normalized chart is that it can be used in real time to evaluate soil behaviour type during the CPT while the normalized charts require information on soil unit weight and ground water conditions (Robertson, 2010). According to Robertson (2010), there is little difference between the normalized and non-normalized charts based soil identification when the in situ effective vertical stress is between 50 kPa and 150 kPa. Robertson (2010) also proposed a new non-normalized soil behaviour type index \((I_{SBT})\), representing the boundaries between the 9 zones showed in figure 1. \(I_{SBT}\) is given by the following equation:

\[
I_{SBT} = \left[\left(3.47 - \log(q_c/p_a)\right)^2 + \left(\log R_f + 1.22\right)^2\right]^{1/2} \tag{1}
\]

Table 1. Soil behaviour types (Robertson, 2010).

| Zone | Soil behaviour type (SBT) |
|------|--------------------------|
| 1    | Sensitive fine grained    |
| 2    | Clay- organic soil       |
| 3    | Clays: clay to silty clay|
| 4    | Silt mixtures: clayey silt to silty clay|
| 5    | Sand mixtures: silty sand to sandy silt|
| 6    | Sands: clean sands to silty sands|
| 7    | Dense sand to gravelly sand|
| 8    | Stiff sand to clayey sand*|
| 9    | Stiff fine grained*       |

* Overconsolidated or cemented.
3. Gaussian Mixture Method for CPT Classification

The Gaussian mixture (GM) method is used to classify the CPT data into an unknown number of soil clusters by representing the probability density function of observed variables as a mixture of multivariate normal densities. The resulted clusters are the components of the mixture model. The construction of GM model is based on the values of \((q_c/p_a)\) and \((R_f)\) as they are reasonably represented by a Gaussian distribution. Given a set of CPT observations:

\[
x_i = \left\{ R_{fi}, (qc/pa)_i \right\}; \; i = 1, ..., N
\]

A Gaussian mixture model with \(k\) components is defined as:

\[
f(x_i) = \sum_{j=1}^{k} w_j \phi(x_i | \mu_j, \Sigma_j)
\]

Where \(w_j\) are the mixture weights with \(\sum_{j=1}^{k} w_j = 1\), \(\phi\) is a 2 dimensional multivariate normal probability density function, \(\mu_j\) and \(\Sigma_j\) are respectively the mean vector and the covariance matrix of the \(j\)th mixture component.

The model parameters \((w, \mu, \Sigma)\) are estimated using an iterative Expectation Maximization (EM) algorithm (Bishop, 2006), which assigns posterior probabilities to each cluster with respect to each observation. Then, each point is assigned to the cluster corresponding to the highest posterior probability. To set the initial values to model parameters as required by the EM algorithm, Agglomerative Hierarchical Clustering (AHC) method was used as an alternative to the random initialization that may involve overlapping of the generated clusters. The AHC method groups data by creating a tree of hierarchical clusters based on Wards algorithm using the minimum variance criterion for computing the distance between clusters.

3.1. Optimal Number of Classes Using BIC

The number of classes (or clusters) in the GM model may not be known a priori. The optimal number of clusters is generally determined using the Bayesian information criterion (BIC) which is defined as follows (Kurek et al., 2014):

\[
BIC = -2ln(L) + Mln(N)
\]

Where \(L\) is the maximized likelihood of the parameters in the model, \(M\) is the number of parameters in the GM model and \(N\) is the number of data points. When the BIC does not improve any further by clustering, the optimal number of clusters is reached (Kurek et al., 2014).

The data \((R_f, q_c/p_a)\) of all CPTs are fitted to different Gaussian mixture models with different number of clusters \((k=1\) to \(5)\) and the corresponding BIC are calculated for each model using MATLAB.
statistical toolbox (R2014a). The model that gives the minimum BIC is considered as the model with the optimal number of clusters. The interpretation of obtained clusters and the identification of soil type corresponding to each cluster is done using the Robertson (2010) chart.

4. Soil Classification at Dunkerque Site Based on Multiple CPT Data

4.1. Site Description and Soil Classification Using Robertson (2010) Chart

The study site is located in the north of France near to the port of Dunkerque. The soil characterization at the Dunkerque site, which included multiple CPT tests, boreholes and laboratory testing, is well described by Jardine et al. (2005). The soil profile consists of 3 m of dense to very dense hydraulic marine sand fill overlying a dense marine sand, which is locally crossed by thin silty layers (at depths between 8 m and 9 m, between 14 m and 15 m). The sand become very dense up to a depth of 9 m. The level of the water table was detected at a depth varying between 4 m and 4.7 m.

The present study is concerned with six cone penetration tests (CPTu) performed at a depth varying between 10 m to 18 m with a vertical measurement interval of 0.01 m. All the measured data for the six CPT are displayed in the Robertson (2010) chart as shown in figure 2. It can be seen that the data points are mainly located in zone 6 classified as clean sand to silty sand (according to table 1) with some data points falling in zone 7 (dense sand). A few scattered data points occupy the zones 3, 4 and 5. These results are in agreement with the observations of boreholes and laboratory tests that revealed a dense to very dense sand (Jardine and Standing, 2012). However, it is hard to separate the silty sand, the dense sand and very dense sand using the Robertson (2010) chart alone. Hence, we will use the GM method to identify the different soil behaviour types.

5. Application of the Gaussian Mixture Method

We propose now to cluster the CPTs measurements using the GM method and find the optimal number of soil classes based on the BIC. The data \((R_1, q_c/p_a)\) from the six CPTs are fitted to different Gaussian mixture models with different number of classes \((k=1 \text{ to } 5)\) and the corresponding BIC is calculated for each model using MATLAB statistical toolbox (R2014a).

Figure 3 shows the BIC calculated for different models with the number of classes varying from \(k=1 \text{ to } 5\). It can be seen that there is no improvement in the values of the BIC for \(k=4\) classes. Therefore, the model with four classes is considered as optimal. The projection of these classes on the Robertson (2010) chart is shown in figure 4. It can be seen that the most representative class \((C1)\) occupy the center of zone 6, the class \(C2\) occupy the upper part with some points in the zone 7. The class \(C3\), composed of few points, falls in the zone 3 and the last class \((C4)\) occupy the zone 5 with few points in the lower part of zone 6. These classes \((C1 \text{ to } C4)\) can be identified as clean dense sand, very dense sand, clay and silty sand respectively. As shown in the Robertson (2010) chart, the GM model provides a clear classification; the model is able to separate the dense and very dense behavior of the soil of type 6. In addition, the proposed model is able to identify the two thin layers corresponding to clay and silty sand soils.

Figure 3. BIC values versus the number of clusters \((k=1 \text{ to } 5)\).

Figure 4. GM classification results for \(k=4\) shown in the Robertson chart.
This is an advantage of the GM model over the existing unsupervised learning approach (Wang et al. 2019) when all the CPT data is used simultaneously.

An example of the soil segmentation results along the depth is illustrated in figure 5 for the CPT6. The friction ratio ($\tau_0 / q_c$), the cone resistance ($q_c$) and the soil behavior type (SBT) index along depth are shown in the first three columns. The profiles are continuous along the depth with fluctuations at the transition zones. The soil stratification identified by the GM model is shown in the fourth column; the difference with the $I_{SBT}$ profile is the identification of the very dense sand layer at a depth between 1 and 2 meters, between 9 and 11 meters and at a depth between 16 and 18 meters. Particularly, this is not revealed by the $I_{SBT}$ index for CPT 2 and CPT4. The probabilities associated to each cluster along the depth are shown in the last four columns for C1, C2, C3 and C4 respectively. The probabilities are nearly equal to 1, which means that the different clusters are accurately identified by the GM model. Similar results were obtained for the remaining CPTs profiles. The accuracy of the GM results may be partially related to the little fluctuations of CPTs measurements in this case. This result was highlighted by Wang et al. (2018b). In addition, the conditionally independent assumption of the GM method does not influence the results because the vertical correlation distance is very small ($\delta < 0.5$ m) compared to the soil layer thickness as it was emphasized by Depina et al. (2016) and Wang et al. (2019).

**Figure 5.** CPT data, SBT index and GM classification results along depth for the CPT6.
5.1. Comparison between Results Using Multiple and Single CPT Data

To show the soil classification improvements when combining multiple CPT data, we compare the number of optimal clusters calculated for each single CPT based on the BIC with that obtained previously by considering the data of the six CPTs simultaneously. The results in figure 6 show that the optimal number of clusters is equal to two for each single CPT; this indicates a reduction in the number of clusters by twice when compared to the first case (all CPTs). An obvious interpretation is that the GM model for each CPT is not able to identify the thin layers because of the limited number of measurement points. Therefore, the model with multiple CPT data provide a more optimal and accurate soil classification, especially for identifying thin soils layers and distinguishing soil formations with very close properties. The combination of CPT data is seldom in literature; it has been used by Wang et al. (2018b) and proved its efficiency for a probabilistic site characterization and soil classification.

6. Conclusion

In the present paper, a Gaussian mixture (GM) model based probabilistic soil classification was developed using multiple cone penetration tests. The GM model classify the CPT data into an unknown number of classes (soil clusters) by representing the probability density function of observed variables (dimensionless cone resistance \((q_c/p_a)\) and friction ratio \((R_f)\) as a mixture of multivariate normal densities. The model based on expectation maximization algorithm with Bayesian information criterion for the selection of optimal number of clusters was validated using six real CPT data performed at Dunkerque site in the north of France. The classification results showed that the GM model is able to separate the dense and very dense behaviour of the soil type 6 in the Robertson chart unlike the soil behaviour index \((I_{BR})\). In addition, the combination of all CPTs data, rather than taking them separately, allowed the identification of two thin layers corresponding to clay and silty sand soils.

Further studies are still expected to investigate the effect of vertical correlation on the CPTU classification and to take into account the spatial characteristics in a sophisticated machine learning based Bayesian model.

References

[1] Bishop C M 2006. *Pattern recognition and machine learning*. Springer Science & Business Media, LLC, USA.
[2] Cetin K O and Isik N S 2007. Probabilistic assessment of stress normalization for CPT data. *Journal of Geotechnical and Geoenvironmental Engineering* vol 133(7), ASCE, pp 887–897
[3] Ching J, Wang JS, Juang CH, Ku CS 2015b. Cone penetration test (CPT) based stratigraphic profiling using the wavelet transform modulus maxima method. *Canadian Geotechnical Journal* vol 52(12), pp 1993–2007
[4] Depina I, Le T M H, Eiksend G R, Strom P 2016. Cone penetration data classification with Bayesian mixture analysis. *Georisk: Assessment and management of risk for engineered systems and geohazards* vol 10(1), pp 27–41.
[5] Douglas B J and Olsen R S 1981. Soil classification using electric cone penetrometer. *Proceedings of Conference on Cone Penetration Testing and Experience*, ASCE, St. Louis, Missouri, pp 209–227.

[6] Eslami A and Fellenius B H 2004. CPT and CPTu data for soil profile interpretation: review of methods and a proposed new approach. *Iranian journal of science and technology* vol 28(B1) pp 69–85.

[7] Facciorusso J and Uzielli M 2004. Stratigraphic profiling by cluster analysis and fuzzy soil classification from mechanical cone penetration tests. *Proceedings ISC2 on Geotechnical and Geophysical Site Characterization*, Millpress, Rotterdam, pp 905–912.

[8] Huang J, Zheng D, Li D, Kelly R, Sloan SW 2018. Probabilistic characterization of two-dimensional soil profile by integrating cone penetration test (CPT) with multi-channel analysis of surface wave (MASW) data. *Canadian Geotechnical Journal* vol 55(8) pp 1168–1181.

[9] Jardine R, Show F, Overy R, Standing J 2005. *ICP design methods for driven piles in sands and clays*. Thomas Telford, London, UK.

[10] Jardine R and Standing J 2012. Field axial cyclic loading experiments on piles driven in sand. *Soils and Foundations* vol 52(4) pp 723–736.

[11] Jung B C, Gardoni P, Biscontin G 2008. Probabilistic soil identification based on cone penetration tests. *Géotechnique* vol 58(7) pp 591–603.

[12] Kurek J, Krug M, Bilski P, Rabargioly S, Swederski B 2014. Automatic estimation of the number of soil profile layers using Bayesian information criterion. *Journal of Communication and Computer* vol 11(7) pp 565–572.

[13] Li J, Cassidy M J, Huang J, Zhang L, Kelly R 2016. Probabilistic identification of soil stratification. *Geotechnique* vol. 66(1) pp 16–26.

[14] Robertson P K 2010. Soil behavior type from the CPT: an update. *2nd International Symposium on Cone Penetration Testing, CPT'10*, Huntington Beach, CA, USA.

[15] Robertson P. K. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal* vol 27(1) pp 151–158.

[16] Robertson P K, Campanella R G, Gillespie D, Greig J 1986. Use of piezometer cone data. *Proceedings of In-situ 86 Conference*, Blacksburg, Geotechnical Special Publication N°6, ASCE, pp 1263–1280.

[17] Rogiers B, Mallants D, Batelaan O, Gedeon M, Huysmans M, Dassargues A 2017. Model based classification of CPT data and automated lithostratigraphic mapping for high resolution characterization of heterogeneous sedimentary aquifer. *PLoS ONE* vol 12(5): e0176656.

[18] Schmertmann J H 1978. *Guidelines for cone test, performance and design*. Federal Highway Administration, Report FHWA-TS-78209, Washington, p 145.

[19] Tumay M T, Abufarsakh M Y, Zhang Z 2008. From theory to implementation of a CPT based probabilistic and fuzzy soil classification, *EJGE*.

[20] Wang Y, Huang K, Cao ZJ 2013. Probabilistic identification of underground soil stratification using cone penetration tests. *Canadian Geotechnical Journal* vol 50(7) pp 766–776.

[21] Wang X, Wang H, Liang R Y., Liu Y 2018a. A semi supervised clustering based approach for stratification identification using borehole and cone penetration test data. *Journal of engineering geology*, DOI: 10.1016/j.jenggeo.2018.11.014.

[22] Wang X, Wang H, Liang R Y, Zhu H Y, Di H 2018b. A hidden Markov random field model based approach for probabilistic site characterisation using multiple cone penetration test data. *Journal of Structural Safety*.

[23] Wang H, Wang X, Wellmann F, Liang R Y 2019. A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data. *Canadian Geotechnical Journal* vol 56(8) pp 1184–1205.

[24] Zhang Z and Tumay MT 1999. Statistical to fuzzy approach toward CPT soil classification. *Journal of Geotechnical and Geoenvironmental Engineering* vol 125(3) pp 179–186.