Title
Unsupervised machine learning for detecting soil layer boundaries from cone penetration test data

Authors
Hudson, Kenneth S
Ulmer, Kristin J
Zimmaro, Paolo
et al.

Publication Date
2023

DOI
10.1002/eqe.3961

Peer reviewed

11Equation Section 1Unsupervised Machine Learning for Detecting Soil Layer Boundaries from Cone Penetration Test Data

By Kenneth S. Hudson[1], Kristin J. Ulmer[2], Paolo Zimmaro[3], Steven L. Kramer[4], Jonathan P. Stewart[1], and Scott J. Brandenberg[1]

1. Civil and Environmental Engineering Department, University of California, Los Angeles.
2. Southwest Research Institute.
3. Environmental Engineering Department, University of Calabria, Italy and Civil and Environmental Engineering Department, University of California, Los Angeles.
4. Civil and Environmental Engineering Department, University of Washington.

Abstract

Cone penetration test (CPT) data contains detailed stratigraphic information that is useful in a wide variety of applications. Separating a CPT profile into discrete layers is an important part of many analyses such as critical layer selection in liquefaction triggering analysis, effective stress seismic ground response analysis, analysis of pile shaft and tip resistance, and soil-pile interaction analysis. The discretization of the profile into layers is often done manually, relying on the judgment of the analyst. This manual approach is cumbersome for datasets that include large numbers of CPT profiles [such as the Next Generation Liquefaction (NGL) database and the New Zealand Geotechnical Database] and it may not be consistent or repeatable because different analysts may discretize a given CPT log in different ways. To overcome these difficulties, we present an approach to automatically divide a CPT profile into discrete layers. Automated layer detection is performed using an unsupervised machine learning technique called agglomerative clustering in combination with two cost functions to identify an optimal number of layers. The algorithm is illustrated using CPT profiles from the NGL database, where the approach is being used in the development of liquefaction triggering and manifestation models. Although the algorithm shows promise for replicating our judgment regarding layering, we recommend visual review of the layering produced by the algorithm to check for reasonableness given the site geology and intended use of the CPT data.

Introduction

Cone penetration test (CPT) data is one of the most valuable resources for subsurface characterization by geotechnical engineers. CPT data is used in a large variety of applications from identifying soil types to estimating static and dynamic shear strength of soil. By typically sampling at 1 cm intervals, an individual CPT test may contain thousands of data points, which provide essentially continuous profiles of tip resistance (q_c), sleeve friction (f_s), and sometimes pore pressure (u_2) over the length of the CPT profile. Most geotechnical engineering applications require grouping the CPT data within the site's stratigraphic profile into a discrete number of layers of consistent soil type and behavior. Examples include liquefaction triggering evaluation, including identification of a critical layer, ground response analysis to evaluate earthquake site response, evaluation of the axial and lateral capacity of deep foundations, and many others.

Selection of layers is often based on the judgment of an engineer or geologist with the goal being to select layers that have similar geologic origin and soil properties but are distinct from the materials above and below them. The number and thickness of layers selected to represent the profile depends on the intended application. This process is subjective and hence unrepeatable when based entirely on analyst judgment because different analysts (or the same analyst at a different time) may choose different layer boundaries. Additionally, manual layer selection becomes inefficient when large numbers of profiles require interpretation. Therefore, the engineering community needs an algorithm that can efficiently assign layers to CPT profiles with repeatable, objective results, thereby removing bias that can be introduced by a sole analyst or small group of analysts. The aim of this paper is to describe and propose such an algorithm based on an unsupervised machine learning procedure that, along with a small number of existing alternate approaches (described in the remainder of this section), enables robust analysis of CPT profiles.

Existing Layering Algorithms

A number of techniques have been developed to automate identification of simplified profiles from CPT data. For example, Wang et al.[1,2,3] and Cao et al.[4] developed a Bayesian approach to assign layer boundaries and assign a probability that soil within a particular layer falls within a soil behavior type category. Ching et al.[5] developed a procedure that utilizes the wavelet transform method to distinguish sudden changes in CPT tip resistance from smaller amplitude changes due to within-layer soil variability. These methods are rather complicated, require a significant number of calculations, and only consider one parameter (soil behavior type or tip resistance). Cao et al.[4] proposed a Bayesian identification method based on the soil behavior type index, I_c[6]. Ntritsos and Cubrinovski[7] developed an algorithm that minimizes the within-layer coefficient of variation of q_{c1Ncs} and I_c for the purpose of developing finite element meshes for one-dimensional ground response analysis. Their method is conceptually and computationally simpler than many previous methods and was shown to produce similar results to analyzing the full profile with respect to liquefaction potential. Ntritsos and Cubrinovski[7] caution that the algorithm may result in fictitious layers at layer boundaries and indicate that their algorithm is not intended to replace engineering judgment. Molina-Gómez et al.[8] more recently utilized a multivariate hierarchical clustering approach to identify stratigraphic layers at a site in the Tagus River Valley where gel push sampling was performed in combination with CPT testing to confirm soil types. Layers need not be vertically continuous in their algorithm (e.g., a layer may have another layer within it). They suggest that their algorithm is well-suited to identifying layers at other experimental sites. In addition to identifying layers based on a vertical profile, some of the methods (e.g., Wang et al.[2,3]) assess the lateral spatial variation of stratigraphy within a site where multiple CPT soundings and/or boring logs are available. We recognize that automated lithology detection of rock strata based on geophysical data has been studied by statistical, wavelet, and, more recently, machine learning procedures in the petroleum exploration industry but will not be discussed here as it is a significantly different application compared to our method.

Motivation for Automated Layer Identification Algorithm

A motivation for the work described in this paper was the need to create discretized representations of individual CPT profiles at sites in the Next Generation Liquefaction (NGL) database[9,10]. Such profiles are required for the development of new liquefaction triggering and manifestation models. Our algorithm was developed independently from, and concurrently with, the methods by Ntritsos and Cubrinovski[7] and Molina-Gómez et al.[8] and bears some similarities to both methods, as well as having some advantages. It is similar to the method by Ntritsos and Cubrinovski[7] in that it seeks a set of layers that reduces the within-layer variance. It differs from their method in that it uses unsupervised machine learning rather than prescribed rules for assigning layers, which is advantageous since the algorithms are widely available in Python packages. Our method is similar to that of Molina-Gómez et al.[8] in that it utilizes an unsupervised machine learning technique, hierarchical clustering, to identify layers. However, it differs from their method in two important respects. First, our algorithm requires that layers be vertically contiguous, whereas theirs allows for non-contiguous layers. Second, our algorithm selects the optimal number of layers based on a cost function that is unbiased with respect to the maximum penetration depth, whereas theirs utilizes an automated algorithm to select the number of layers. We show herein that the automated method they adopted results in bias wherein thicker layers are identified for deeper profiles, and thinner layers for shallower profiles, whereas our algorithm is unbiased with respect to the total depth of the CPT sounding. Our proposed method only considers vertical layering and does not consider horizontal spatial variability because it is intended to be used at a single CPT location. A reduced dataset that contains only locations with multiple CPT soundings in proximity would be required to extend the method for horizontal interpretation, which is beyond the scope of this paper.

We consider the existence of multiple automated algorithms to provide a beneficial measure of epistemic uncertainty, which is important for quantification of overall uncertainty in engineering analyses. No single algorithm will best suit the needs of all users and all applications; therefore, it is useful for different algorithms to utilize a range of different approaches to quantify uncertainties related to layer identification decisions. The following sections present some details about CPT measurements and data analyses and introduce, describe, and provide examples on how to apply our proposed layer detection algorithm.

Cone Penetration Test

The CPT probe measures tip resistance, sleeve friction, and sometimes pore pressure (Fig. 1) (e.g., Robertson[6], Lunne et al.[11]). A hydraulic press pushes the cone into the ground generally at a rate of 2 cm/s. The cone tip resistance, q_c, is equal to the measured force on the cone tip divided by the cone area, A_c, and the sleeve friction, f_s, is the force acting on the friction sleeve divided by the surface area of the sleeve. Commonly A_c is 10 to 15 cm^2, and the cone tip angle is $\alpha = 60°$. The most common location for pore pressure measurement is between the cone tip and the friction sleeve, which is deemed the u_2 location. When pore pressure is measured, the corrected tip resistance is computed as,

$$q_t = q_c + u_2(1\text{-}a) \quad (1)$$

where a is the net area ratio of tip, usually between 0.6 and 0.8 depending on cone design. Eq. 1 accounts for the influence of water pressure acting downward behind the cone tip on the measured tip resistance. Measurements are generally recorded at 1 cm intervals.

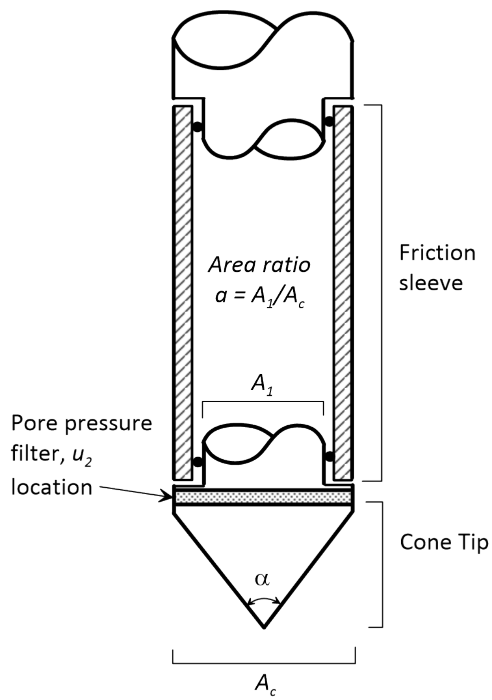


Figure 1. Cross-section schematic of cone penetration test probe.

Various quantities are often computed from CPT measurements, and utilized to identify soil characteristics. Cone tip resistance and sleeve friction increase with depth in uniform soil due to increasing effective stress with tip resistance generally being high relative to sleeve friction in coarse-grained soils and vice versa in fine-grained soils. To assess fundamental soil properties, the normalized cone resistance, Q_{tn}, defined by Eq. 2 is typically used, where σ_{vo} is in-situ vertical total stress, σ_{vo}' is in-situ vertical effective stress, p_a is atmospheric pressure (101.325 kPa), and n is an exponent that defines the soil-type-dependent relationship between σ_{vo}' and q_t. Furthermore, normalized sleeve friction, F_r, is defined by Eq. 3. These dimensionless quantities are combined to define the soil behavior type index, I_c[6], defined by Eq. 4. The exponent n depends on I_c as defined by Eq. 5, and Eqs. 3, 4, and 5 therefore form an implicit system of equations that is solved by iteration.

$$Q_{tn} = \left(\frac{q_t - \sigma_{vo}}{p_a}\right)\left(\frac{p_a}{\sigma_{vo}'}\right)^n \quad (2)$$

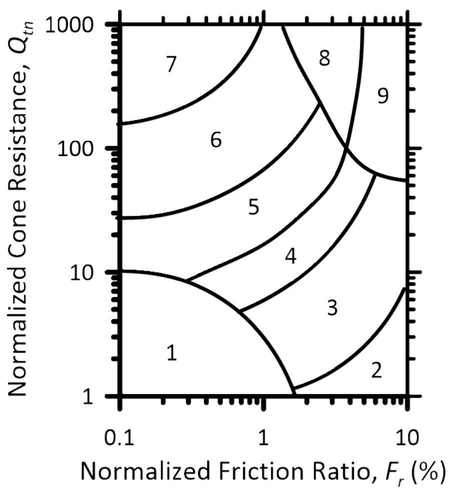
$$F_r = \frac{f_s}{q_t - \sigma_{vo}} \times 100\% \quad (3)$$

$$I_c = \sqrt{(3.47 - \log Q_{tn})^2 + (\log F_r + 1.22)^2} \quad (4)$$

$$n = 0.381 I_c + 0.05\left(\frac{\sigma_{vo}'}{p_a}\right) - 0.15 \quad (5)$$

Soil Behavior Type

Robertson[6] (1990) found that soil behavior type can be classified based on contours of Q_{tn} vs. F_r shown in Fig. 2. Soils that cluster within SBT_n zones 2 through 7 are separated by contours that approximately follow the range of I_c values in Fig. 2, and exhibit soil behavior type that increases in coarseness from SBT_n=2 (Organic soils and clay) to SBT_n = 7 (gravelly sand to dense sand). Sensitive fine-grained soils, very stiff sand to clayey sand, and stiff fine-grained soils do not have a unique I_c range associated with their behavior type.



I_c	SBT_n Zone (Robertson 1990)	Soil Behavior Type
	1	Sensitive Fine Grained
>3.60	2	Organic soils – clay
2.95 – 3.60	3	Clays – silty clay to clay
2.60 – 2.95	4	Silt mixtures – clayey silt to silty clay
2.05 – 2.60	5	Sand mixtures – silty sand to sandy silt
1.31 – 2.05	6	Sands – clean sand to silty sand
<1.31	7	Gravelly sand to dense sand
	8	Very stiff sand to clayey sand
	9	Stiff fine-grained

Figure 2. Soil behavior type based on CPT measurements.

Thin Layer and Transition Zone Effects

Due to its physical dimensions, the CPT probe averages out soil properties within a zone of influence near the cone tip. As a result, the cone may render measurements near layer interfaces that imply incorrect soil behavior type. For example, as a cone transitions out of a stiff sand layer with SBT_n = 6 into a soft underlying clay layer with SBT_n = 3, there will likely be a transition zone in which SBT_n = 4 and 5 will be measured even though silty soil does not exist in these zones. Furthermore, when a cone is advanced through a thin sand layer sandwiched between two softer clay layers, the tip resistance measured at the center of the sand may be lower than the resistance that would be measured in a uniform profile of the same sand.

A number of algorithms have been developed to identify transition zones, where an interface between different types of soil result in CPT measurements that may not accurately reflect the soil at that depth. For example, CPeT-IT[12] provides an algorithm for identifying interface zones, along with the SectionMaker software for assigning layers within a cross-section based on CPT measurements. Boulanger and DeJong[13] developed an inverse-filtering algorithm to recover the "true" CPT soil properties from the measured properties by accounting for the influence of the layered profile on the CPT measurements. Their algorithm tends to increase the tip resistance in stiff layers near the boundaries with softer layers, and to a lesser degree it also decreases the tip resistance in soft layers near the boundaries with stiff layers. Other authors have pointed out the limitations of the Boulanger and DeJong[13] algorithm[14] and have begun introducing refined algorithms[15].

CPT Corrections

Although the layer identification algorithm presented here is general and could be used for many different CPT applications, our specific focus is the evaluation of liquefaction. CPT is a preferred tool for characterizing site conditions for liquefaction analysis due to its repeatability and the nearly continuous profile that it provides. Both corrected cone tip resistance and liquefaction resistance depend fundamentally on soil density and fines content. However, the dependencies are different, which requires adjustments to the measured cone tip resistance to render a quantity that relates more directly to liquefaction resistance. Namely, corrections are applied to account for the influence of σ_{vo}' and fines content, FC. The overburden- and fines-corrected cone tip resistance, q_{c1Ncs}, is defined in Eq. 6, where the overburden correction factor, C_N, is defined by Eq. 7[16]. The fines correction in Eq. 6 is intended to account both for the reduced stiffness and strength of sandy soils containing fines (which affect tip resistances) and the effects of fines on the cyclic resistance of the soil to liquefaction triggering. Liquefaction triggering relationships typically utilize q_{c1Ncs} to define cyclic liquefaction resistance of sand-like soils (e.g., Moss et al.[17], Boulanger and Idriss[16]).

$$q_{c1Ncs} = C_N \frac{q_t}{p_a} + \left(11.9 + \frac{C_N}{p_a}\frac{q_t}{14.6}\right)\exp\left[1.63 - \frac{9.7}{FC+2} - \left(\frac{15.7}{FC+2}\right)^2\right] \tag{6}$$

$$C_N = \left(\frac{p_a}{\sigma_{vo}'}\right)^{n_f} \leq 1.7 \tag{7}$$

$$n_f = 1.338 - 0.249\left(q_{c1Ncs}\right)^{0.264} \tag{8}$$

Example CPT Profile

An example CPT profile, UC-4, obtained at Moss Landing (California) near Sandholdt Road is shown in Fig. 3. This site exhibited severe manifestations of liquefaction due to the 1989 **M**6.9 Loma Prieta earthquake[18,19]. The CPT profile shows that this site consists of alternating layers of fine-grained and coarse-grained materials. Note that coarser-grained materials with lower I_c tend to have higher q_c and q_{c1Ncs}. Furthermore, the averaging of cone penetration tip resistance near layer boundaries is evident, for example at a depth near 6m. The inverse-filtered CPT data[13] have sharper edges due to being corrected for layer transition effects. We consider the inverse-filtered profiles to provide a more accurate representation of the true soil properties, and utilize the inverse-filtered profiles for the remainder of this paper.

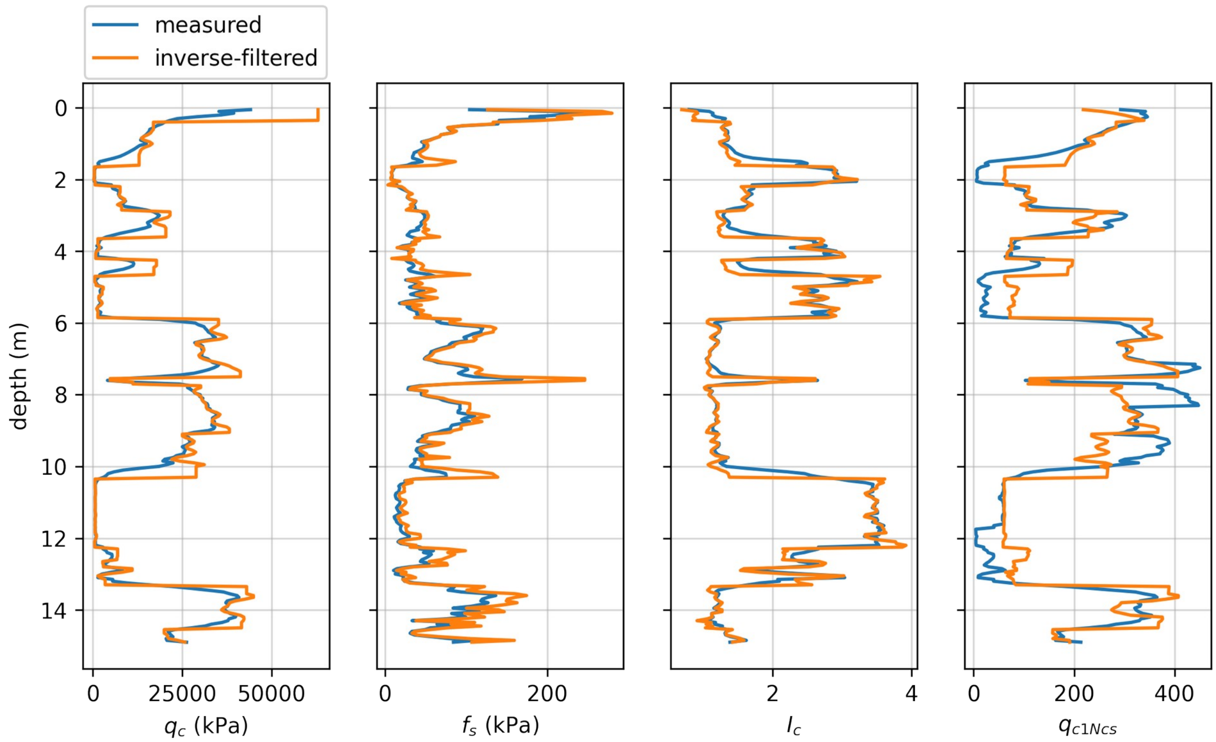


Figure 3. Cone penetration test data for UC-4 at the Moss Landing site near Sandholdt Road, which exhibited liquefaction manifestations due to the 1989 **M**6.9 Loma Prieta earthquake[18,19].

Layer Identification Algorithm

In this section, we first summarize main features of the theoretical framework behind the tools used to produce our layer identification algorithm. We then provide a description of how it was implemented and details on how it should be used. Clustering or cluster analysis is an unsupervised machine learning approach that categorizes data based on common attributes[20]. K-means clustering categorizes data based on the aggregate distance between the data point and the centroid of each cluster, where distance is measured in the parameter space of the variables included in the clustering algorithm[21,22]. Gaussian mixture models assign probabilities that each data point belongs within each cluster based on the cluster statistics, and may be thought of as an extension of K-means clustering that also considers covariance among variables. The number of clusters is provided as an input parameter, and the algorithm assigns data to clusters in a manner that minimizes the sum of within-cluster variance. We perform K-means and Gaussian mixture model clustering using standardized values for cone tip resistance and soil behavior type index defined by Eqs. 9 and 10, where μ_q, σ_q, μ_{Ic}, and σ_{Ic} are the mean and standard deviation of q_{c1Ncs} and I_c for the entire profile, respectively.

$$\hat{q}_{c1Ncs} = \frac{q_{c1Ncs} - \mu_q}{\sigma_q} \quad (9)$$

$$\hat{I}_c = \frac{I_c - \mu_{Ic}}{\sigma_{Ic}} \quad (10)$$

Standardizing the data prior to clustering is important, particularly when the parameter space contains variables of different units and significantly different ranges. Without standardization, variables with higher numerical values may be inadvertently weighted more heavily than variables with smaller numerical values in the distance calculation. For example, q_{c1Ncs} for liquefaction applications generally varies from about 50 to 300, while I_c varies only from about 1.0 to 3.5.

K-Means and Gaussian Mixture Model Results

Fig. 4 shows results for K-means and Gaussian mixture model clustering each with 16 clusters. Calculations were performed using the Python package Scikit-learn[23] with default input parameters. Both algorithms group data into clusters that are close to each other in $\hat{q}_{c1Ncs} - \hat{I}_c$ space, thereby showing promise for grouping data based on similarities in soil composition. The algorithms exhibit subtle differences in their clustering of the data, with the Gaussian mixture model resulting in differently shaped clusters than K-means in some cases (Fig. 4). These approaches to clustering data are similar in concept to the soil behavior type assignments by Robertson[6] in that soils in different regions in $\hat{q}_{c1Ncs} - \hat{I}_c$ space are expected to exhibit different soil behavior type. However, the SBT_n regions defined by Robertson[6] are fixed in Q_{tn}-F_r space, whereas the clusters are determined simply by proximity to other data points. We selected $\hat{q}_{c1Ncs} - \hat{I}_c$ as the clustering parameters rather than Q_{tn} and F_r because the former is more relevant for liquefaction assessments.

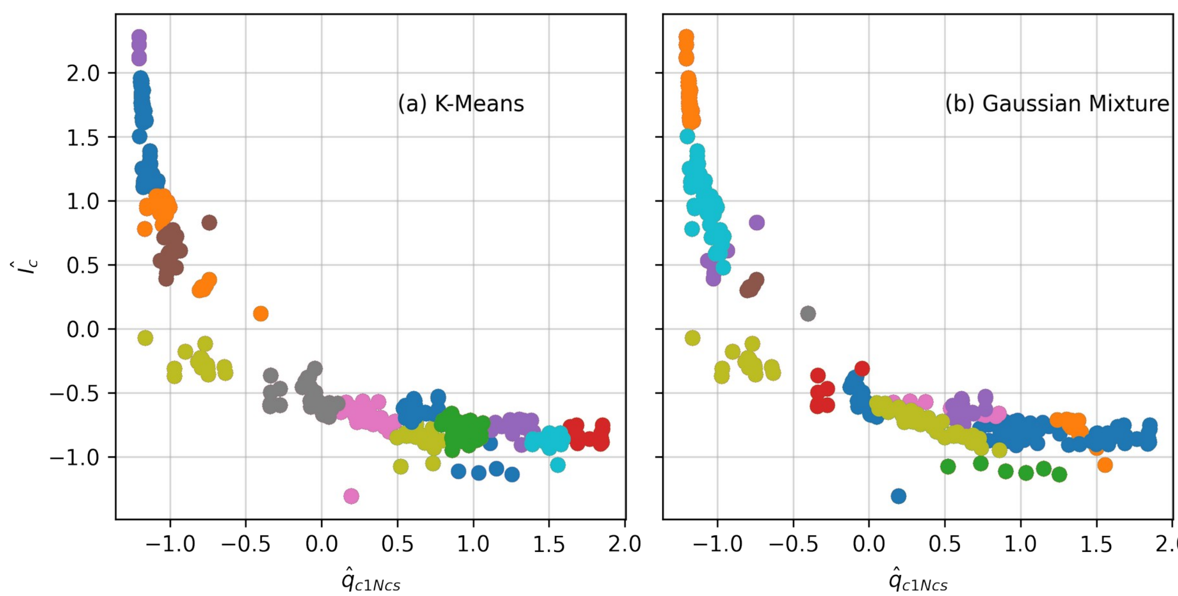


Figure 4. Clustering algorithm results for the UC-4 CPT profile using (a) K-means and (b) Gaussian mixture modeling.

As shown in the profiles in Fig. 5 for the K-means clustering algorithm, these algorithms do not cluster the data into spatially contiguous layers (e.g., the green colored cluster occurs over the depth intervals 6.5-7.5 m, 7.9-9.0 m, and 13.8-14.2 m). The reason is that these algorithms cluster data based only on their similarities in $\hat{q}_{c1Ncs} - \hat{I}_c$ space, and do not consider the fact that the data are hierarchically ordered based on depth. The Gaussian mixture model, which is not shown in Fig. 5 for brevity, produces similar results in that the clusters are not vertically contiguous.

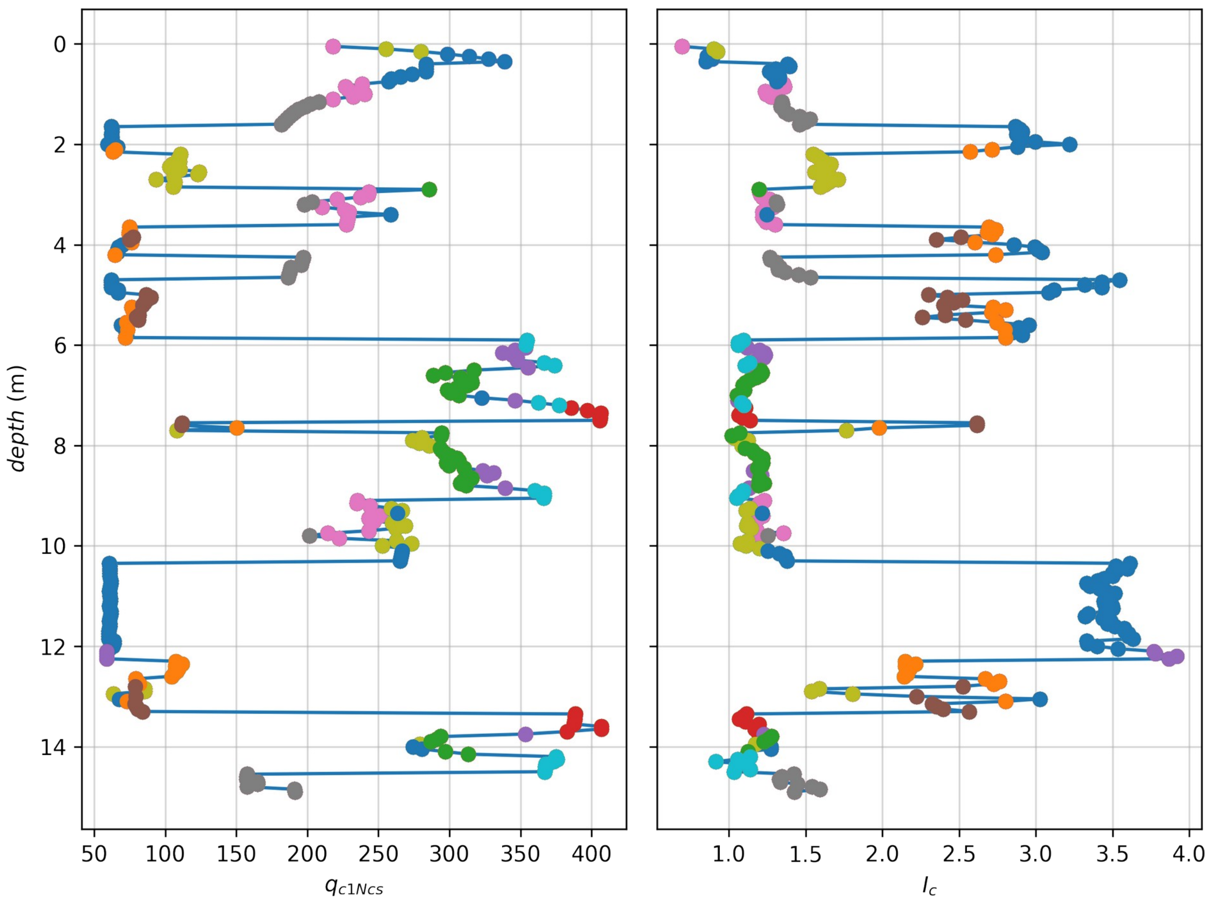


Figure 5. CPT profiles for UC-4 based on K-means clustering. Common coloration indicates that depths are associated with the same cluster, e.g. pink intervals at 1, 3, and 9.7 m depths.

Agglomerative Clustering

We turn to agglomerative clustering, which is a form of hierarchical clustering that groups data based on a cascading "tree" of clusters computed using distances between points[20], to produce clusters that form vertically contiguous layers. A nearest-neighbor matrix is provided to the clustering algorithm to specify which points are permitted to be considered when assigning clusters. For sequentially ordered data such as CPT data, the nearest neighbor matrix is tri-diagonal with ones on the diagonal and the two adjacent diagonals, and zeros elsewhere. This differs from the approach of Molina-Gómez et al.[8] that used a more fully populated nearest neighbor matrix. In our approach, a particular data point is constrained to belong to the same cluster as the point above and the point below (or both) or to constitute its own cluster, but it cannot belong to the same cluster as a distant neighbor unless all of the points in between are part of the same cluster. The algorithm then clusters data by minimizing the collective within-cluster variance for the total number of clusters specified. The resulting data are plotted in Figure 6 for the UC-4 CPT profile using a total of 16 clusters. In this case, the clusters are organized into vertically contiguous layers in a manner that reflects their depositional sequence and is similar to how layers might be assigned using human judgment. In this respect, our approach differs from that of Molina-Gómez et al.[8] which permits clusters to be vertically non-contiguous.

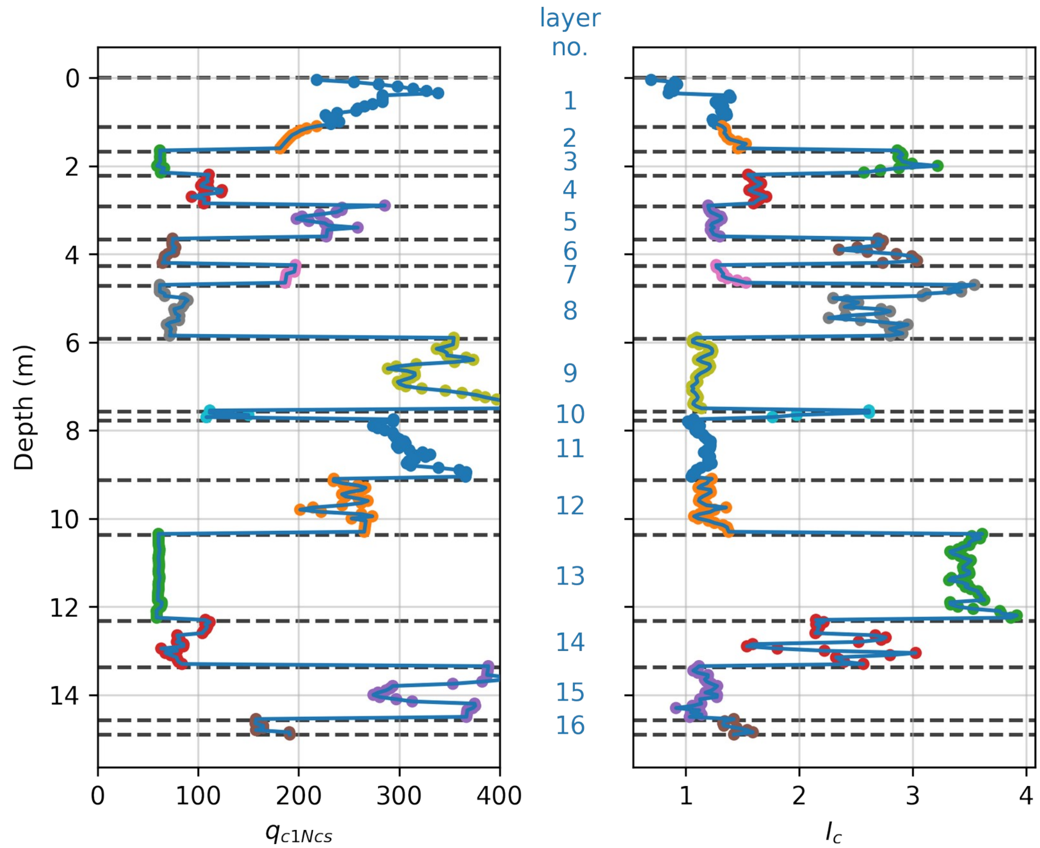


Figure 6. CPT profiles for UC-4 using agglomerative clustering with tri-diagonal nearest neighbor matrix.

Number of Layers

A crucial consideration in the clustering algorithm is selection of an appropriate number of clusters (i.e., layers). In the preceding examples we have manually set the number of clusters as 16. Here we seek an algorithm capable of selecting the optimal number of clusters, which is expected to vary depending on profile depth and complexity. The goal is to separate the CPT data into contiguous layers with similar soil properties using the fewest clusters possible. The optimal number of clusters is therefore subjective, and different analysts would likely select different numbers of layers for a given CPT profile. Our goal is therefore to identify a method for automatically assigning the number of layers in a manner that captures the stratigraphic details important for liquefaction evaluations.

Distortion Score

In agglomerative clustering, a distortion score, J_D, is often utilized to identify the optimal number of clusters, and is defined for the two-standardized-variable case considered here in Eq. 11,

$$J_D = \frac{\sum_{i=1}^{N}\left[\left(\hat{q}_{c1Ncs_i} - \mu_{\hat{q}_i}\right)^2 + \left(\hat{I}_{c_i} - \mu_{\hat{I}c_i}\right)^2\right]}{\sum_{i=1}^{N}\left[\hat{q}_{c1Ncs_i}{}^2 + \hat{I}_{c_i}{}^2\right]} \tag{11}$$

where μ_{q_i} and $\mu_{I_{c_i}}$ are the mean values of $\hat{q}_{c1Ncs}$ and $\hat{I}_c$, respectively, for the i^{th} cluster (i.e., subscript i is the index for clusters and identifies values of these parameters for each individual cluster), and N is the total number of data points in the profile. Note that J_D decreases as the number of clusters, K, increases, and by definition is equal to zero when $K=N$ because every point would constitute its own cluster and the numerator would be zero. The optimal number of clusters therefore cannot be computed by minimizing the distortion score, but rather is a compromise between reducing the distortion score while retaining the smallest possible number of clusters that adequately categorizes the data.

Thickness-Dependent Cost Function and Combined Cost Function

We define a cost function, J_T, that penalizes the average layer thickness within a profile using Eq. 12.

$$J_T = 0.2\left(\frac{0.5m}{t_{avg}}\right)^3 \tag{12}$$

The average thickness is defined as $t_{avg} = z_{max}/K$, where z_{max} is generally the total depth of the CPT profile. Note that predrilling is sometimes necessary for CPT profiles, in which case the first depth at which data is recorded is nonzero. In those cases, z_{max} is the difference between the deepest and shallowest CPT measurement. The purpose of Eq. 12 is to penalize selection of a high value of K if it results in average layer thicknesses that are too small to be considered geotechnically significant. Based on inspections and analyses of hundreds of CPT profiles in the NGL database, we believe that 0.5 m is a fairly thin stratum, and we set the coefficients in Eq. 12 such that $J_T = 0.2$ for this condition. The cubic form of Eq. 12 was adjusted until the achieved average layer thickness accorded well with our judgment. A combined cost function is then defined in Eq. 13, where w_D and w_T are weights assigned to the components of the cost function. We herein utilize $w_D = w_T = 1.0$, but these weights can be adjusted based on user judgment in a site- or region-specific manner.,

$$J = w_D J_D + w_T J_T \tag{13}$$

Elbow and min(J) Methods

We consider two methods for utilizing the distortion score and the combined cost function to select the optimal number of layers. First, the "elbow" method graphically interprets a plot of J_D vs. K, which has a negative curvature over the full range of K, but flattens as K increases (Figure 7). The optimum value of K (9 in the case of Figure 7) is identified on the basis of curvature having decreased to a sufficiently low level, which is subjective. As such, the elbow method is based only on J_D and not on J_T. We utilize the Yellowbrick[24] Python package to implement the elbow method which identifies the point of maximum curvature of the J_D vs. K curve and assigns that as the optimum number of layers. The silhouette method[24] is also often utilized to identify the optimal number of clusters. This method is based on a so-called "silhouette" value that measures the similarity of data points within a cluster compared to other clusters. We found it to produce similar results to the elbow method. Thus, results from this method are not reported in Figure 7. Molina-Gómez et al.[8] utilize the silhouette method to define the number of clusters in their algorithm.

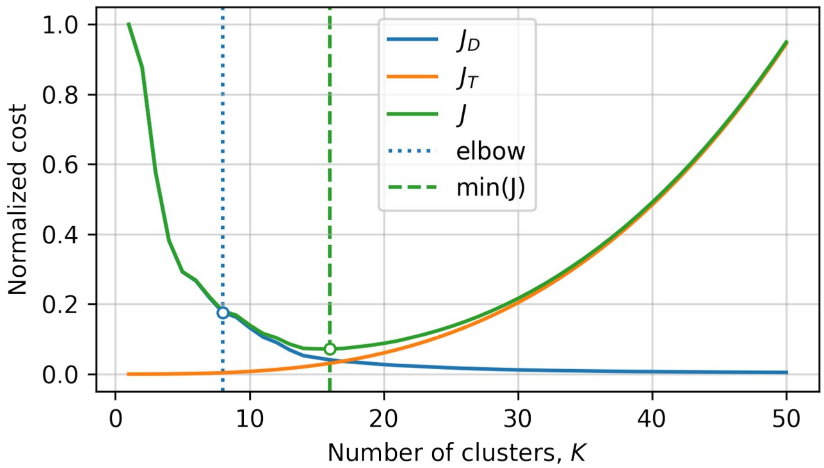


Figure 7. Cost functions and layer selection for CPT profile UC-4.

We also apply an alternative method in which K is selected as the point where J (from Eq. 13) is minimized. For this reason, we call this the min(J) method. The combined cost function is minimized for K = 16 clusters for the example of CPT UC-4 in Figure 7.

Profiles of 16 and 9 layers are shown in Fig. 8, where (a) and (b) have 16 layers by using the min(J) method, whereas (c) and (d) have 9 layers by using the elbow method. The primary differences between these two profiles are in layers number 3, 4, and 6 for the 9-layer profile. These layers clearly contain within-layer regions that are vertically contiguous with different q_{c1Ncs} and I_c values (e.g., the layer for the 2.2-3.8 m depth range), yet they are clustered together in the 9-layer profile. By contrast, they are separated into different layers in the 16-layer profile. The 16-layer profile accords better with our judgment, and similar observations observed across diverse profiles with a wide range of depths (as described in the next section) causes us to prefer use of the min(J) approach over the elbow method when selecting the number of layers. We recognize that a different curvature threshold in the application of the elbow method would have produced a different number of layers and, possibly, a solution that accords better with our judgment. However, the superiority of the min(J) method is related to the fact that it is based on layer thickness, which is a physically meaningful quantity, whereas the gradient of J_T vs. K used in the elbow and silhouette methods does not have a clear physical meaning.

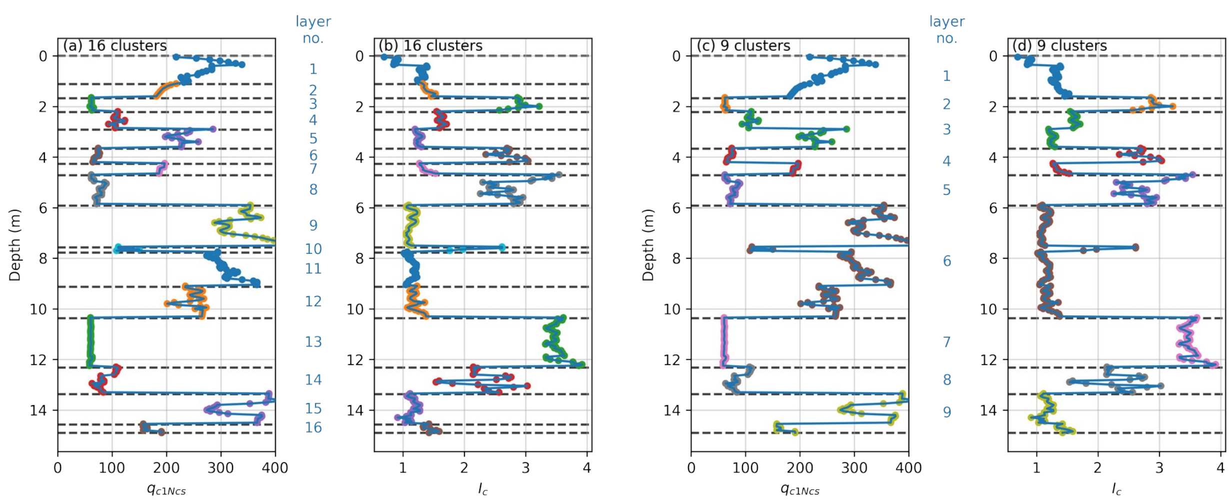


Figure 8. Profiles of q_{c1Ncs} and I_c with 16 layers by using the min(J) method (a and b) and 9 layers by using the elbow method (c and d).

Calculations for Many CPT Profiles

Calculations of the optimal numbers of layers were performed for a total of 272 CPT profiles contained in the NGL database[9,10]. Both the elbow method and the min(J) method were utilized to select the optimal number of layers. We expect that t_{avg} should be independent of z_{max} because t_{avg} depends upon vertical heterogeneity of the soil profile, which is controlled by the geological processes that formed the soil deposit, whereas z_{max} arises from a decision controlled by the objectives of the site investigation. For example, z_{max} may be higher for a site investigation for a pile-supported tall building with a corresponding deep zone of influence than for a single-story building supported by spread footings with a corresponding shallow zone of influence.

Values of t_{avg} vs. z_{max} are plotted in Fig. 9. The elbow method exhibits a strong positive correlation in which t_{avg} increases essentially linearly with z_{max}. This is an undesirable outcome since we anticipate t_{avg} to be independent of z_{max}. By contrast, values of t_{avg} are essentially independent of z_{max} using the min(J) method, particularly for values of z_{max} > 12m. For liquefaction triggering evaluation, profiles shorter than about 15m may miss layers that could potentially liquefy and produce surface manifestation. In this regard, the slight bias in the min(J) method for shallow profiles has little practical impact.

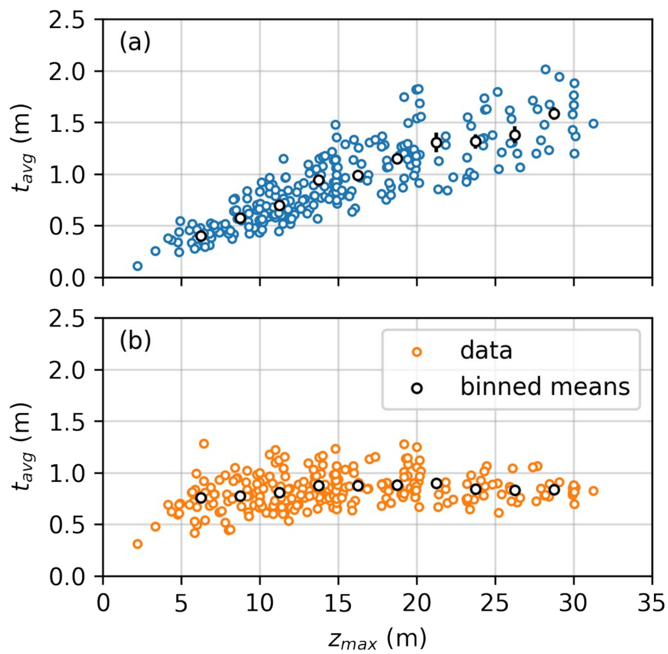


Figure 9. Average layer thickness, t_{avg}, versus total CPT profile length, z_{max} for (a) elbow method and (b) min(J) method

The influence of maximum depth on average layer thickness is further explored in Fig. 10, which illustrates normalized cost versus number of clusters for (a) a shallow profile with z_{max} = 5.1m from CPT_8933 at Site 76 in Edgecumbe, New Zealand, and (b) a deep profile with z_{max} = 31.3m from CPT001 at the Inage site in Urayasu City, Japan (CPT names are those reported in the NGL database). Note that the J_T functions are significantly different for these two profiles because the same average thickness in Eq. 12 produces fewer layers for the shallow profile than for the deep profile. For the shallow profile, the elbow method indicates that 8 sublayers is ideal (t_{avg} = 0.64m), while the min(J) approach provides 7 layers (t_{avg} = 0.73m). These results are very similar. By contrast, for the deep profile, the elbow method indicates that 8 layers is ideal (t_{avg} = 3.9m), while min(J) provides 36 sublayers (t_{avg} = 0.87m). These results are significantly different, and the average layer thickness using the elbow method is too large to capture potential critical layers of sand-like soil with low q_{c1Ncs}.

Note that when K=8, J_D is near 0.2 for the shallow profile and near 0.4 for the deep profile. A fundamental limitation of the elbow method is that it considers only the curvature of the cost function, and not the value of the cost function itself.

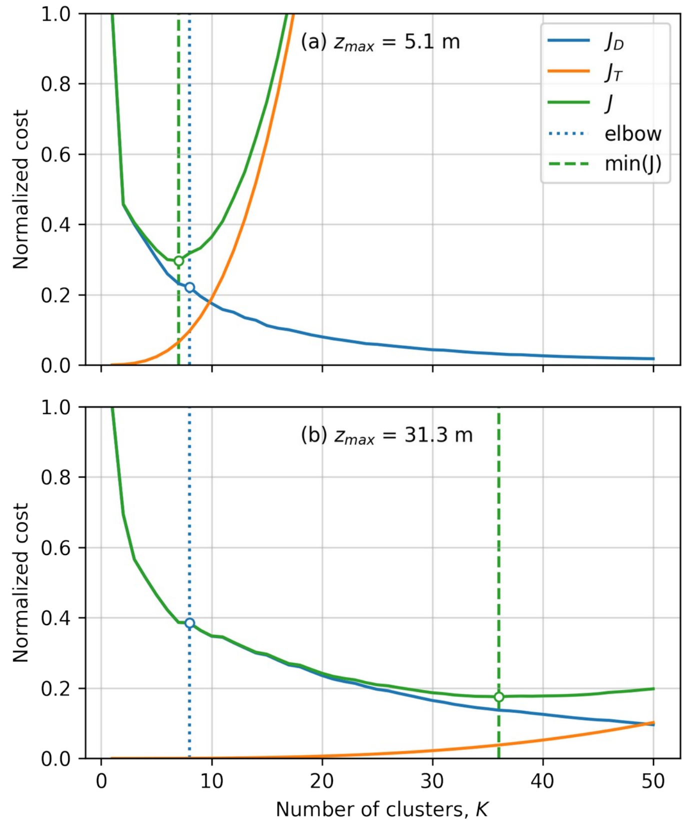


Figure 10. Normalized cost versus number of clusters for (a) a shallow profile with z_{max}=5.1m corresponding to CPT_8933 at Site 76 in Edgecumbe, New Zealand, and (b) a deep profile with z_{max}=31.3m corresponding to CPT001 at the Inage site in Urayasu City, Japan.

The two profiles are illustrated in Fig. 11 with a common depth axis to illustrate the clear differences in the maximum penetration depth. The average layer thicknesses determined using the min(J) method are similar for these two profiles despite the different total depths. Furthermore, it is clear that reducing the number of layers for the deeper site from 36 (using the min(J) method) to only 8 (using the elbow method) would result in significantly higher average layer thickness, and would miss much of the stratigraphic detail within that profile.

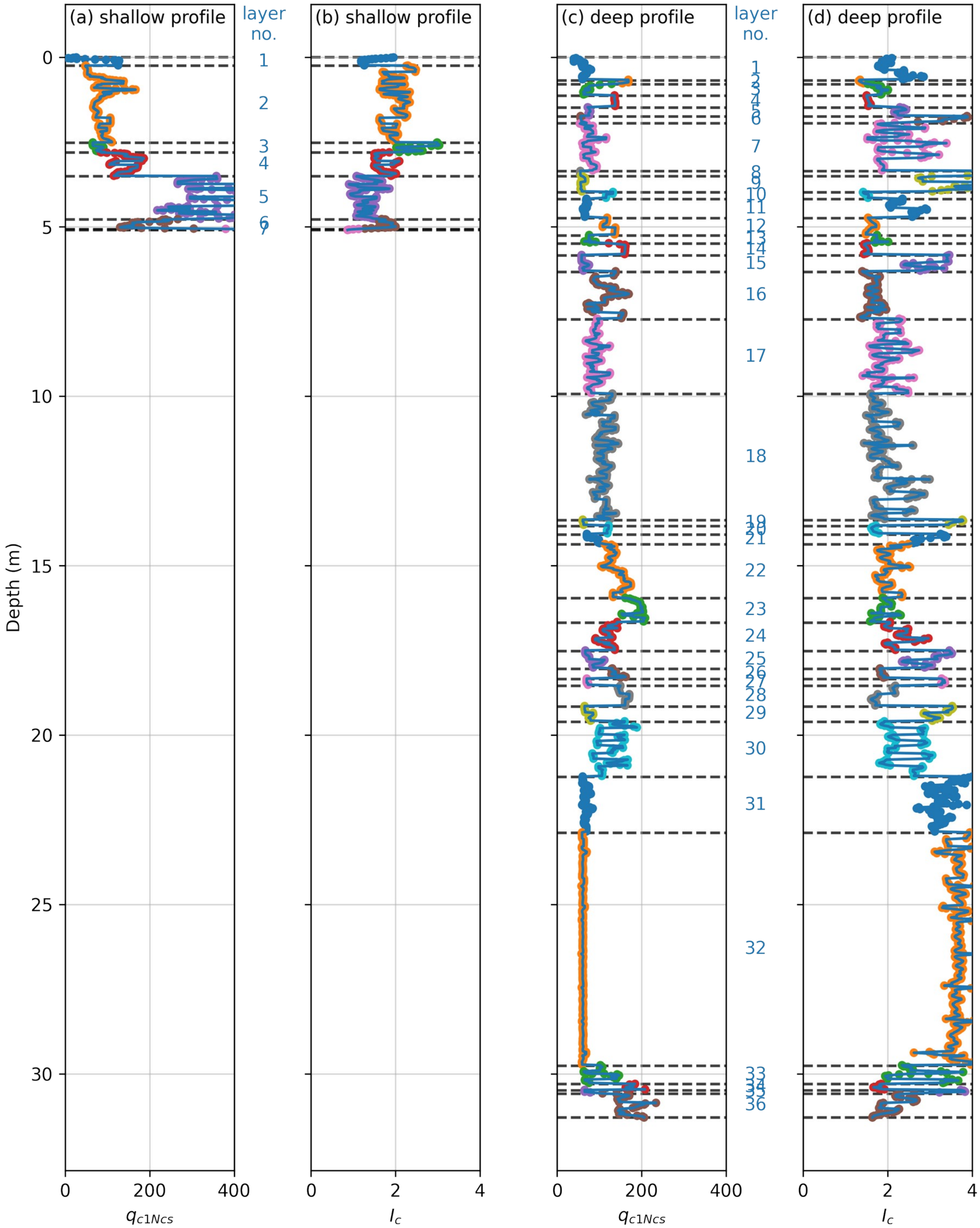


Figure 11. Profiles of q_{c1Ncs} and I_c for (a) and (b) a shallow profile corresponding to CPT_8933 at Site 76 in Edgecumbe, New Zealand, and (c) and (d) a deep profile corresponding to CPT001 at the Inage site in Urayasu City, Japan.

Conclusions

This study developed an unsupervised machine learning approach for identifying layers from cone penetration test data and selecting the optimal number of layers (or clusters). The clustering parameter space consisted of $\hat{q}_{c1Ncs}$ and $\hat{I}_c$, which are standardized values of the overburden-corrected clean sand equivalent cone tip resistance, q_{c1Ncs}, and the soil behavior type index, I_c. The clustering algorithm utilizes the Scikit learn Python package, which is widely available and easy to implement. We utilize agglomerative clustering with a tridiagonal nearest neighbor matrix to identify vertically contiguous soil layers.

A crucial aspect of the proposed algorithm is selecting the optimal number of clusters. The elbow method, a traditional approach commonly utilized in machine learning, did not perform well for our application because the resulting average thickness of the soil layers was strongly dependent on the maximum depth explored by the CPT. We posit that soil stratigraphy is independent of the maximum depth to which the CPT probe is advanced. To overcome this limitation, we introduced a supplemental cost function that penalizes small average layer thicknesses. The optimal number of clusters is selected at the minimum point of this cost function added to the normalized distortion score. This approach produced an average layer thickness that is essentially independent of maximum depth, which is a desired outcome. Compared with manual assignment of layer boundaries, our method is automated and rapid, and shifts human judgment from a case-by-case basis (which is not repeatable) to selection of input parameters in the clustering algorithm (which is repeatable).

All calculations presented herein were performed on inverse-filtered CPT data rather than on raw recorded CPT data. We believe this is more appropriate because CPT measurements are influenced by soil layering, and the inverse filtering attempts to recover the “true” CPT profile. Although not shown in this paper, we found that the proposed algorithm often grouped transition layers into a single cluster. In this manner, the algorithm may be useful for application to raw measurements as well, provided that the analyst properly accounts for these transition zones for liquefaction evaluation or other applications.

The proposed algorithm provides a convenient means for rapidly developing a tentative layering profile for further engineering evaluation. Modeling parameters were adjusted to accord with our judgment regarding layer assignments. However, the algorithm may do a poor job identifying layers in some situations, and we urge users to review the layering that arises from the algorithm and to exercise their own judgment and available geological knowledge in assigning layers for their particular application before proceeding with calculations. For instance, users could calibrate the J_T function by adjusting the t_{avg} or the weights in Eq. 13 based on their dataset and intended application.

Acknowledgements

Financial support for the NGL project is provided by the U.S. Nuclear Regulatory Commission (NRC) and the U.S. Bureau of Reclamation (USBR) through the Southwest Research Institute (SWRI). Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party’s use, or the results of such use, of any information, apparatus, product, or process disclosed in this paper, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the NRC or USBR.

References

1. Wang, Y., H. Kai, and Z. Cao. (2013) Probabilistic identification of underground soil stratification using cone penetration tests. Canadian Geotechnical Journal. 50(7): 766-776, 10.1139/cgj-2013-0004.
2. Wang, X., Wang, H., Liang, R. Y., & Liu, Y. (2019a). A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data. Engineering geology, 248, 102-116.
3. Wang, H., X. Wang, J.F. Wellmann, and R.Y. Liang (2019b). A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data. Canadian Geotechnical Journal, 56(8), 1184-1205.

4. Cao, Z. J., S. Zheng, D.W. Li, and K.K. Phoon (2019). Bayesian identification of soil stratigraphy based on soil behavior type index. Canadian Geotechnical Journal, 56(4), 570-586.
5. Ching, J., J.S. Wang, H.C. Juan, and C.S. Ku (2015). Cone Penetration Test (CPT)-based stratigraphic profiling using the wavelet transform modulus maxima method. Can. Geotech. J. 52(1) p. 1993-2007, 10.1139/cgj-2015-0027.
6. Robertson, P.K. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal*, 27(1): 151-158, 10.1139/t90-014
7. Ntritsos, N., and M. Cubrinovski (2020). A CPT-based effective stress analysis procedure for liquefaction assessment. *Soil Dynamics and Earthquake Engineering*. 131, 10.1016/j.soildyn.2020.106063.
8. Molina-Gómez, F., D. Cordeiro, C. Ferreira, and A. Viana da Fonseca (2022). "Soil stratigraphy from seismic piezocone data and multivariate clustering in alluvial soil deposits: Experience in the Lower Tagus Valley region." *Cone Penetration Testing 2022*, Gottardi & Tonni (eds), DOI: 10.1201/9781003308829-84
9. Brandenberg S.J., P. Zimmaro, J.P. Stewart, D.Y. Kwak, K.W. Franke, R.E.S. Moss, K.O. Cetin, G. Can, M. Ilgac, J. Stamatakos, T. Weaver, S.L. Kramer (2020). Next generation liquefaction database. Earthquake Spectra, 36(2), p. 939-959, 10.1177/8755293020902477.
10. Zimmaro P, S.J. Brandenberg, J.P. Stewart, et al. (2019) Next-Generation Liquefaction Database. *Next-Generation Liquefaction Consortium*, 10.21222/C2J040
11. Lunne, T., P.K. Robertson, and J.J.M. Powell (1997) Cone Penetration Testing in Geotechnical Practice. Blackie Academic & Professional, London, 312 p.
12. GeoLogismiki (2022) CPeT-IT v.3.0 – CPT interpretation software.
13. Boulanger, R.W., and J.T. DeJong, (2018). Inverse filtering procedure to correct cone penetration data for thin-layer and transition effects. Proc., Cone Penetration Testing. Hicks, Pisano, and Peuchen, eds., Delft University of Technology, The Netherlands, 10.1201/9780429505980-2.
14. Yost, K.M., R.A. Green, S. Upadhyaya, B.W. Maurer, A. Yerro-Colom, E.R. Martin, J. Cooper, (2021) Assessment of the efficacies of correction procedures for multiple thin layer effects on Cone Penetration Tests, *Soil Dynamics and Earthquake Engineering*, 144, doi:10.1016/j.soildyn.2021.106677.
15. Cooper, J., E.R. Martin, K.M. Yost, A. Yerro, R.A. Green, (2022) Robust identification and characterization of thin soil layers in cone penetration data by piecewise layer optimization, *Computers and Geotechnics*, 141, doi:10.1016/j.compgeo.2021.104404
16. Boulanger, R. and I.M. Idriss. (2014). CPT and SPT Based Liquefaction Triggering Procedures. Report No. UCD/CGM-14/01. Center for Geotechnical Modeling. University of California, Davis.
17. Moss, R.E.S., R.B. Seed, R.E. Kayen, J.P. Stewart, A. Der Kiureghian, and K.O. Cetin. (2006) CPT-based Probabilistic and Deterministic Assessment of In Situ Seismic Soil Liquefaction Potential. *Journal of Geotechnical and Geoenvironmental Engineering*. Vol. 132, No. 8. pp. 1,032–1,051
18. Boulanger, R.W., I.M. Idriss, and L.H. Mejia. (1995). Investigation and evaluation of liquefaction related ground displacements at Moss Landing during the 1989 Loma Prieta earthquake. Report No. UCD/CGM-95/02, Center for Geotechnical Modeling, Department of Civil & Environmental Engineering, University of California, Davis, 231.
19. Boulanger, R.W., L.H. Mejia, and I.M. Idriss. (1997). Liquefaction at Moss Landing during Loma Prieta Earthquake. *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE, 123(5): 453-467, 1997. 10.1061/(ASCE)1090-0241(1997)123:5(453)
20. Nielsen, F. (2016). Hierarchical Clustering. 10.1007/978-3-319-21903-5_8.
21. Lloyd, S.P. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.
22. Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*. 21 (3): 768–769.
23. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res*., 12, 2825–2830.
24. Bengfort, B., L. Gray, R. Bilbro, R. Prema, D. Patrick, K. McIntyre, M. Morrison, A. Ojeda, E. Schmierer, A. Morris, S. Molin, and S. Swadik (2022). "Yellowbrick v1.5 (1.5). Zenodo. DOI 10.5281/zenodo.1206239.