Soils and Rocks
www.soilsandrocks.com
An International Journal of Geotechnical and Geoenvironmental Engineering
ISSN 1980-9743
ISSN-e 2675-5475

# Application of kernel k-means and kernel x-means clustering to obtain soil classes from cone penetration test data

L.O. Carvalho[1], D.B. Ribeiro[1]

Article

**Keywords**
Artificial neural network
Cone penetration test
Clustering
Soil classification

**Abstract**
Most methods available in the literature for soil classification from cone penetration test (CPT) data define soil classes using laboratory tests. One disadvantage of this approach is that field soil conditions are difficult to replicate in a lab. The alternative adopted in this work is trying to define soil classes only by the similarity of the CPT measurements, using clustering. This study is the first, to the best knowledge of the authors, to cluster soil classes in a four-dimensional input feature space using measurements directly taken from the CPT experiment. Nine soil classes are produced from a general dataset containing 179 CPT soundings and, in a complementary study, four more specialized classes are obtained from 5 CPT soundings. Artificial neural networks (ANN) are used to produce simple models capable of reproducing both class groups, which are compared with classical soil classifications from the literature and with standard penetration test (SPT) samples. Results show that both general and specialized class groups can be reproduced by ANN although accuracy is better for the latter, reaching a 97.04 % accuracy with a standard deviation of 1.24 %. Furthermore, it is shown that accuracies above 80 % are obtained even if incomplete data is used. This shows that the here proposed soil classes can become an interesting alternative in engineering practice.

## 1. Introduction

The more commonly used soil classification standard is the Unified Soil Classification System, which is based on granulometry and plasticity. Nevertheless, it has disadvantages like the difficulty of extracting undisturbed samples and the time delay required to get the results. On the other hand, the cone penetration test (CPT) allows an accurate measurement of soil parameters, which can be instantaneously used to classify soil layers along a vertical axis. One important issue concerning this classification is its connection to soil behavior in detriment of soil granulometry. In this context, although pioneer work proposing soil classification from CPT data focused only soil granulometry (Begemann, 1965), following studies stated that soil behavior should guide class definitions for being related to the soil load-bearing capacity (Douglas & Olsen, 1981). In later investigations, pore pressure information was included to define soil classes and propose normalizations for the cone resistance and lateral friction to account for the overburden pressure and better separate classes, which produced the well known Robertson charts (Robertson, 1990). A new friction ratio-based chart was later proposed, changing the circular curves of Robertson (1990) by hyperbolic ones (Schneider et al., 2012). Robertson (2016) modified these charts, defining a fully behavioral classification, including also the dilative and contractive behaviors for each of the three soil types.

Most work that use machine learning techniques for classifying soil from CPT data apply clustering to propose new soil classes (Hegazy & Mayne, 2002; Facciorusso & Uzielli, 2004; Bhattacharya & Solomtine, 2006; Liao & Mayne, 2007; Das & Basudhar, 2009; Rogiers et al., 2017; Wang et al., 2019). One limitation of these work is the reduced number of input features included, most times only two. Another limitation is that most work explore only hierarchical clustering techniques (Hegazy & Mayne, 2002; Facciorusso & Uzielli, 2004; Bhattacharya & Solomtine, 2006; Liao & Mayne, 2007). Nevertheless, a recent study stated that including depth as an input can improve cluster-

#Corresponding author. E-mail address: dimas@ita.br.
[1]Departamento de Geotecnia, Divisão de Engenharia Civil, Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brazil.

Submitted on April 29, 2020; Final Acceptance on July 15, 2020; Discussion open until March 31, 2021.
DOI: https://doi.org/10.28927/SR.434607

ing results and that the x-means algorithm can lead to good results (Rogiers et al., 2017). In spite of these conclusions, to the best knowledge of the authors, no work from the literature investigated clustering techniques including all measured CPT parameters. Furthermore, the traditional x-means algorithm implemented with the original k-means can only be used for linearly separable classes.

The kernel k-means algorithm is an iterative clustering technique based on the minimization of the variance inside clusters. It allows objects changing from one cluster to another to reduce the overall variance. The kernel x-means algorithm works running kernel k-means several times, splitting the clusters into new ones in each round. In this context, the objective of this work is to use kernels k-means and kernels x-means to produce soil classification methods using four input features: depth, cone resistance, lateral friction and pore pressure. First, kernel k-means is applied to a dataset composed by 179 CPT soundings, of which 5 have paired SPT soundings, generating 9 soil classes. These classes are compared to SPT samples and to Robertson classification methods (Robertson, 1991, 2016) obtained with a student version of the CPeT-IT v2.0.2.5 software. An alternative specialized approach is also presented using the kernel x-means algorithm, which was found to be effective in previous work (Rogiers et al., 2017). It is shown that both proposed soil classification methods can be replicated by an ANN model, even if the pore pressure is not included as an input. This enables reproducing the obtained methods in simple spreadsheets.

## 2. Classification methods for comparison

The two soil classification methods here used for comparison were developed by Robertson. Only a brief view of their theory is presented here, once they are also used and described in previous work from the authors (Carvalho & Ribeiro, 2019).

### 2.1 Influenced by soil granulometry (ISG)

This method was proposed by Robertson (1991) and its classes descriptions allude to granulometry:

1. Sensitive, fine grained
2. Organic soils - peats
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 sand to silty sand
7. Gravelly sand to sand
8. Very stiff sand to clayey sand
9. Very stiff, fine grained

The normalized parameters used for classification are:

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \quad (1)$$

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}} \quad (2)$$

$$Q_{tn} = \left( \frac{q_t - \sigma_{v0}}{p_a} \right) \left( \frac{p_a}{\sigma'_{v0}} \right)^n \quad (3)$$

where $q_t$ is the total cone resistance, which is a correction of the raw cone resistance $q_c$. $f_s$ is the lateral friction, $u_2$ is the pore pressure measured behind the cone tip, $u_0$ is the hydrostatic pore pressure, $\sigma_{v0}$ is the total overburden stress and $\sigma_{v0}$' is the effective overburden stress. $n$ is given by

$$n = 0.381 I_c + 0.05 \left( \frac{\sigma'_{v0}}{p_a} \right) - 0.15 \quad (4)$$

where $p_a = 0.1$ MPa is a reference pressure and $I_c$ is defined as (Robertson, 2009):

$$I_c = [(3.47 - \log Q_{tn})^2 + (\log F_r + 1.22)^2]^{0.5} \quad (5)$$

The charts of the ISG method are shown in Figure 1 and Figure 2.

### 2.2 Focused on soil behavior (FSB)

This method presented by Robertson (2016) is considered fully behavioral and proposes the following classes:

1. CCS: Clay-like - Contractive - Sensitive
2. CC: Clay-like - Contractive
3. CD: Clay-like - Dilative
4. TC: Transitional - Contractive
5. TD: Transitional - Dilative
6. SC: Sand-like - Contractive
7. SD: Sand-like - Dilative

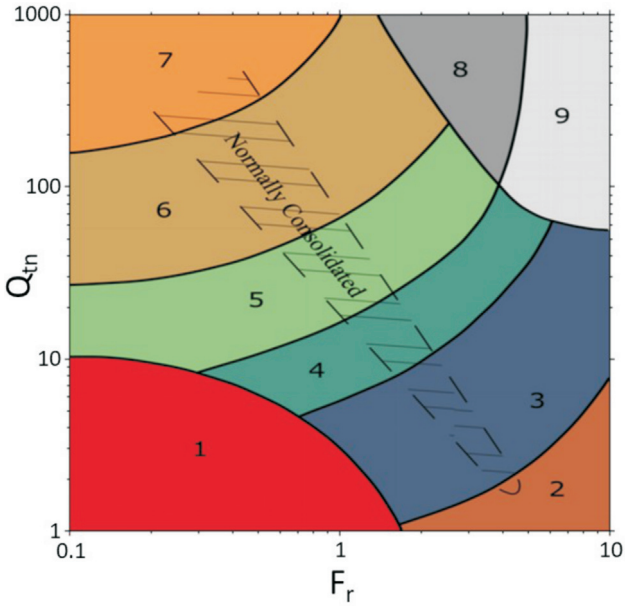


**Figure 1.** $Q_{tn} \times F_r$ chart from Robertson (1991).

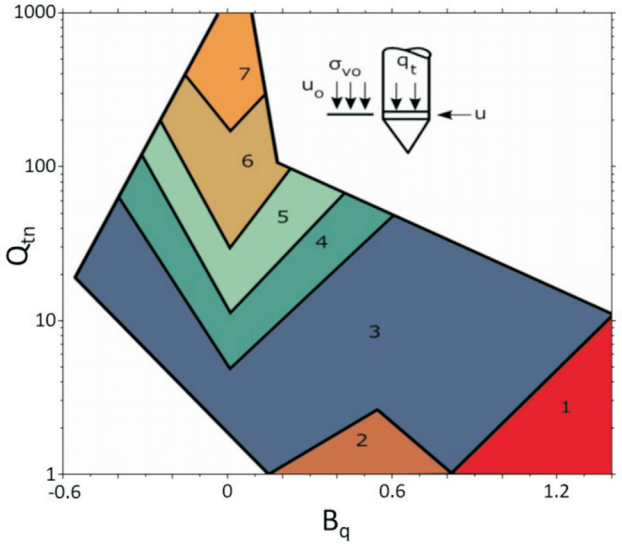


**Figure 2.** $Q_{tn} \times B_q$ chart from Robertson (1991).

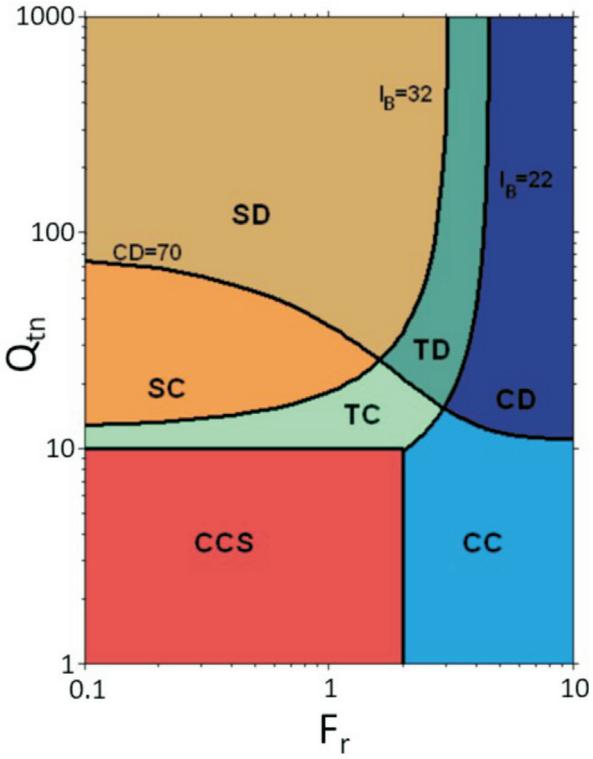


**Figure 3.** $Q_{tn} \times F_r$ chart from Robertson (2016).

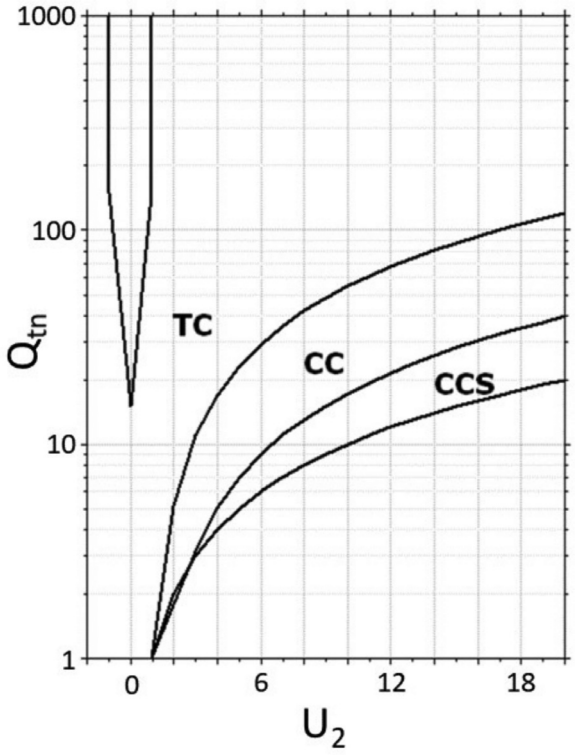


**Figure 4.** $Q_{tn} \times U_2$ chart from Robertson (2016).

It uses the charts presented in Figure 3 and Figure 4 (Schneider et al., 2008, 2012), where $U_2$ is given by:

$$U_2 = \frac{u_2 - u_0}{\sigma'_{v0}} \tag{6}$$

## 3. Machine learning tools

### 3.1 Kernel k-means

The kernel k-means algorithm is a modification of the k-means algorithm, which groups the instances by partition, with a fixed number $k$ of clusters. It is an iterative clustering technique based on the optimization of a clustering criterion, the mean squared error. For each iteration, differently from the hierarchical clustering, the objects can change from one cluster to another to reduce the error. The error is a measure of the variance inside the clusters, which has to be minimized. The mean squared error $E$ is then given by the sum of the variances inside clusters for the $k$ clusters as follows:

$$E = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} d(\mathbf{x}_i, \mathbf{x}^{(j)})^2 \tag{7}$$

where $d(\mathbf{x}_i, \mathbf{x}^{(j)})$ is the distance between the object $\mathbf{x}_i$ and the cluster centroid $\mathbf{x}^{(j)}$.

The algorithm does the following steps:

1. The first $k$ centroids are randomly chosen
2. Each object is included in the group whose centroid is closer
3. A new centroid is then defined for each group in order to minimize the mean squared error
4. Steps (2) and (3) are repeated until conversion is observed, within a predefined error margin.

The most used similarity measure is the Euclidean distance, which requires data normalization in order to avoid distortions due to data scale. The main advantage of k-means is its linear complexity, but its main disadvantages include the possibility of converging to local optimum and

being applicable only to linearly separable classes. Other weaknesses that can compromise analysis are its sensitivity to initialization, the possibility of generating imbalanced clusters and the need of previously fixing *k*. One simple alternative to search for the best *k* and avoid local minimums is running the algorithm several times, varying *k* and the initialization. This procedure is adopted in this work.

One way to deal with classes that are not linearly separable is using a function to map the data from the original feature space into a higher dimensionality feature space wherein the objects are linearly separable. Nevertheless, non-linear transformation and high dimensionality are required to guarantee linear separability. Most work that make use of this approach do not define the function directly, but only a kernel function, which is sufficient to obtain the Euclidean distance. The Gaussian kernel adopted in this work is exp ($-\sigma\|\mathbf{x}_i - \mathbf{x}_j\|^2$), where $\mathbf{x}_i$ and $\mathbf{x}_j$ are points within input feature space and σ is the only calibration parameter required, which can be estimated from the data as the median of $\|\mathbf{x}_i - \mathbf{x}_j\|^2$.

### 3.2 Artificial neural networks

Artificial neural networks (ANN) are based on the brain functioning, with a structure constituted by processing units called neurons, which are connected by weighted signals called synapses. The first artificial neuron model, called Perceptron, was proposed by McCulloch & Pitts (1943). Its practical applicability was formalized with the work of Rosenblatt (1957).

In a Perceptron neuron, an object **x** receives *n* signals (inputs), which are weighted by a vector **w**. After these weighted inputs are gathered, an excitatory threshold or bias θ is discounted, producing a net signal *u*. This net signal is then subjected to an activation function g to produce an output signal $y = g(u) = g(\mathbf{w}.\mathbf{x} - \theta)$. This process is illustrated in Fig. 5. In this work, the sigmoid function is used for activation, which is presented below.

$$g(u) = \frac{1}{1 + e^{-\lambda u}} \tag{8}$$

where λ is a parameter to be calibrated. Data normalization is required, rescaling each input feature to the range [0,1]. One limitation of this model is that it can only be used for linearly separable classes. Non-linear cases require using multi-layer neural networks, which can be trained with the back-propagation algorithm (Rumelhart et al., 1986). Figure 6 represents the structure of this model, wherein each neuron is a Perceptron. According to the universal approximation theorem (Hornik et al., 1989), an ANN with one hidden layer is sufficient to replicate any continuous function. Thus, two hidden layers are enough to replicate even discontinuous functions.

Once there are infinite possibilities for an ANN model, restrictions must be defined to limit the number of calibration tests. The sigmoid function was fixed based on previous experience of the authors, the number of neurons for each layer was limited to double the number of classes and the number of layers was limited to 2. These decisions about architecture were based on the universal approximation theorem. Readers interested in further discussions about this issue are referred to Carvalho et al. (2019).

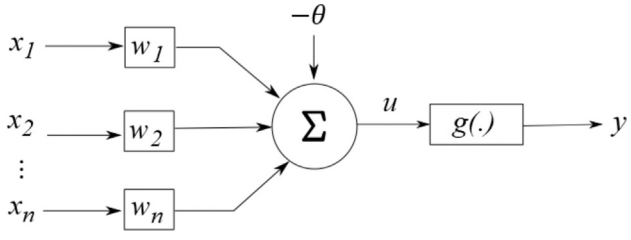


**Figure 5.** Perceptron neuron.

## 4. Methodology

### 4.1 Used datasets

Two datasets are used in this work, one named Full dataset and the other named Specific dataset. The objective is to demonstrate that more homogeneous datasets lead to ANN models with better accuracy. The Full dataset is composed by measurements taken within 179 CPT soundings, which are briefly described below:

- 38 taken in several countries and provided by Professor Peter Robertson. See Carvalho & Ribeiro (2019);
- 73 were taken in the USA and made available online by Professor Paul Mayne. See Carvalho & Ribeiro (2019);
- 1 was taken in Vancouver, Canada and provided by Professor Renato da Cunha. See Cunha (1994);
- 5 were taken in Brazil paired with SPT soundings and provided by Professor Heraldo Giacheti. See reference Ide (2009).
- 62 were taken in Brazil and provided by the São Paulo Metropolitan Trains Company, São Paulo, Brazil.

The 179 CPT soundings produced 130966 examples for the machine learning techniques, each example consisting on a CPT measurement taken at a specific depth. Figures 7a and 7b show histograms for the objects distribution

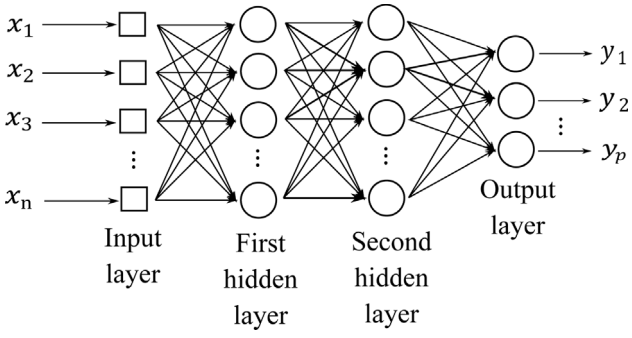


**Figure 6.** Multilayer neural network.

of the Full dataset among the ISG and FSB classes, respectively. Even though there is an imbalance, the minority classes for the ISG and FSB methods have 381 and 5136 objects, respectively. Preliminary tests have shown that this is enough to represent these minority classes among the majority ones.

The Specific dataset is a subset of the Full dataset and is composed by the measurements taken within the 5 CPT soundings provided by Professor Heraldo Giacheti. The paired SPT soundings provided 2847 soil samples, which are here divided into three classes, sands (60,2 % of samples), silts (16,1 % of samples) and clays (23,7 % of samples).

One of the objectives of this work is comparing these three SPT classes to the ones of the ISG method, of the FSB method and also to the ones here obtained by clustering.

### 4.2 Clustering analysis

Two separated studies are performed, one using the Full dataset and the other using the Specific dataset. Both of them are divided into two phases: clustering analysis and ANN modeling. First, the objects are grouped by the kernel k-means algorithm. For this step, the four measured CPT parameters are used to compose the original feature space: depth $z$ (m), raw cone resistance $q_c$ (MPa), lateral friction $f_s$ (kPa) and pore pressure measured behind the cone tip $u_2$ (kPa). Using these inputs instead of normalizations such as $Q_t$, $B_q$ and $F_r$ avoids reducing information within the dataset. Thus, a previous work from the authors suggests that dismissing this type of normalizations makes sense for soil classification (Carvalho & Ribeiro, 2019). For both approaches the Gaussian kernel, which is calibrated by the median of the distance between points, is used to map the objects into a higher dimension feature space (see Section 3.1).

For the Full dataset, the procedure adopted to define the number of classes was manually varying this number and adopting the one with the lowest total variance inside clusters. This procedure lead to 9 classes, as described in Section 5.1. For the Specific dataset, the kernel x-means algorithm was employed. One basic version of this algorithm consists in running the kernel k-means several times from $k = 2$ and splitting the clusters into two new clusters in each round while a parameter called Bayesian Information Criterion is improved (Pelleg & Moore, 2000). Once this parameter gets any worse, the algorithm stops. The result for this case was 4 classes, as presented in Section 5.2.

After obtaining the clusters, they are compared to ISG classes, to FSB classes and to the three SPT classes defined in Section 4.1.

### 4.3 ANN modeling

In this work, ANN models are created to replicate soil classification systems obtained by clustering. The 10-fold cross-validation procedure (Stone, 1974) is employed to evaluate the predictive performance of the ANN models, as illustrated in Fig. 8. This procedure was adopted to avoid overfitting and to calculate a standard deviation of the accuracies obtained within the 10 iterations, which is an important information to be presented together with the mean accuracy.

The procedure starts dividing the dataset in 10 folds of the same size. At each step, one of the 10 folds is randomly selected and separated from the other 9. These 9 folds are then used for training, while the one kept apart is used for testing, obtaining an accuracy. Selection is made without reposition, allowing all folds to be tested after 10 steps. The mean and standard deviation of the obtained accuracies represent the predictive performance of the ANN model.

Notice that all soil samples received a class within the clustering procedure described in Section 4.2, making possible to check all predictions given by the ANN algorithm. Recall $R_i$ is defined as the number of right predictions for one class $i$ divided by its number of examples $n_i$:

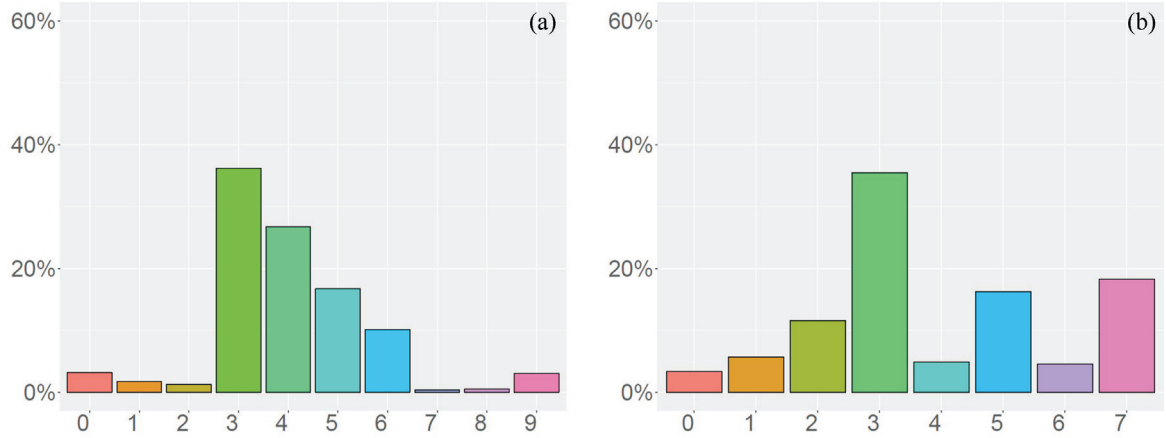


**Figure 7.** Histograms for the Full dataset: (a) distribution for ISG classes and (b) distribution for FSB classes.

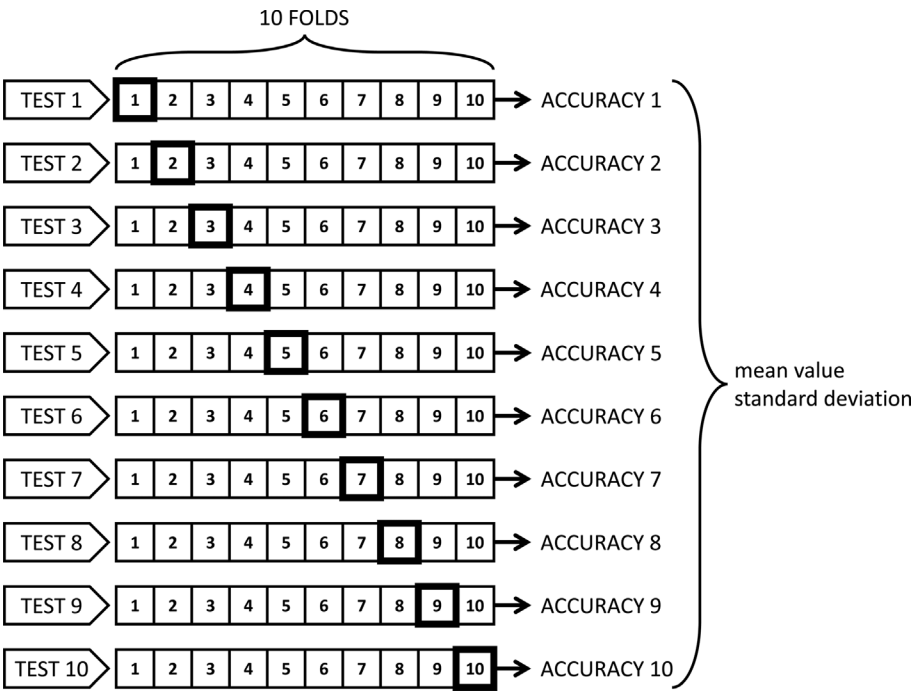


**Figure 8.** 10-fold cross validation.

$$R_i = \frac{1}{n_i}\sum_{j=1}^{n_i} I_{ij} \quad (9)$$

where $I_{ij} = 1$ if the model made a right prediction and $I_{ij} = 0$ otherwise. In this work, the mean recall is used as performance measure and, from this point of the text, referred simply as accuracy $A$ for a sake of clarity. For $c$ classes, it is obtained as

$$A = \frac{1}{c}\sum_{i=1}^{c} R_i \quad (10)$$

Preprocessing procedures are used within the 10-fold cross validation procedure to improve the predictive performance of the ANN algorithms. Once these procedures are described in previous work from the authors (Carvalho et al., 2019), they are here omitted for conciseness.

## 5. Results and discussion

### 5.1 Clustering analysis with the full dataset

To produce the results presented within this section, the kernel k-means algorithm was applied. $k$ was varied from 7 to 10, using the Full dataset and all CPT original measurements: $z$ (m), $q_c$ (MPa), $f_s$ (kPa) and $u_2$ (kPa). The model with $k = 9$ was the one with the lowest total internal cluster variance, therefore it is the only one here presented. The 9 clusters, each one representing a soil class, have centers which coordinates are presented in Table 1.

In Tables 2 and 3 the clustering results are compared to ISG and FSB classes, respectively. Lines represent clustering classes and columns represent chart-based methods.

**Table 1.** Clusters centers.

| Class | $z$ (m) | $q_c$ (MPa) | $f_s$ (kPa) | $u_2$ (kPa) |
|---|---|---|---|---|
| 1 | 44.16 | 32.33 | 662.38 | 1864.94 |
| 2 | 38.10 | 21.35 | 262.94 | 2296.68 |
| 3 | 53.57 | 35.98 | 369.36 | 2391.77 |
| 4 | 67.26 | 63.86 | 787.28 | 2835.83 |
| 5 | 57.50 | 53.01 | 573.79 | 2490.12 |
| 6 | 54.30 | 51.00 | 835.36 | 1964.86 |
| 7 | 44.93 | 35.75 | 612.18 | 4017.64 |
| 8 | 68.29 | 24.19 | 278.07 | 4931.17 |
| 9 | 23.91 | 20.88 | 165.71 | 2036.20 |

Each value is a percentage of soil samples that were assigned to a clustering class (line) and also to a specific ISG or FSB class (column). ISG class 0 is omitted from Table 2 due to its low representative among the used examples.

Observing Tables 2 and 3, the following interpretations were produced for the 9 cluster classes:

- Classes 1 and 2: They present similar distributions among ISG classes, with a predominance of clay behavior (ISG classes 3 and 4). This predominance is also observed within FSB classes (1, 2 and 3), although the cluster classes appear to become different.
- Class 3: ISG classes 5 and 6, which represent sand behavior, compose 65 % of this cluster class. Similar per-

**Table 2.** Comparing cluster classes to ISG classes (%).

| | Sensitive | Organic | Clays | Clayey silt | Sand mixtures | Sands | Gravelly sand | Stiff to clayey sand | Stiff fine grained |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 0 | 1 | 46 | 34 | 14 | 2 | 0 | 0 | 3 |
| 2 | 3 | 3 | 48 | 24 | 17 | 4 | 0 | 0 | 0 |
| 3 | 0 | 1 | 19 | 15 | 24 | 41 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 4 | 94 | 1 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 12 | 86 | 1 | 0 | 0 |
| 6 | 0 | 0 | 15 | 23 | 19 | 38 | 0 | 1 | 3 |
| 7 | 0 | 0 | 28 | 19 | 18 | 15 | 1 | 7 | 13 |
| 8 | 8 | 1 | 67 | 19 | 5 | 0 | 0 | 0 | 0 |
| 9 | 5 | 6 | 22 | 21 | 30 | 14 | 0 | 0 | 1 |

**Table 3.** Comparing cluster classes to FSB classes (%).

| | | CCS | CC | CD | TC | TD | SC | SD |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 1 | 1 | 7 | 55 | 1 | 23 | 0 | 12 |
| 2 | 12 | 11 | 26 | 14 | 9 | 9 | 8 | 10 |
| 3 | 2 | 4 | 17 | 1 | 10 | 4 | 18 | 44 |
| 4 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 97 |
| 5 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 96 |
| 6 | 1 | 0 | 1 | 27 | 0 | 21 | 0 | 50 |
| 7 | 8 | 4 | 18 | 25 | 6 | 15 | 0 | 24 |
| 8 | 19 | 37 | 26 | 2 | 11 | 4 | 1 | 1 |
| 9 | 29 | 10 | 5 | 11 | 10 | 4 | 18 | 13 |

centage is obtained if FSB classes 6 and 7 are added, which also represent sand behavior.

- Classes 4 and 5: These classes clearly represent sand behavior, with high percentages assigned to ISG class 6 and FSB class 7. Their similarity suggests merging them together.
- Classes 6 and 7: Once the behavior of these classes is well distributed among ISG and FSB classes, they are here considered transitional. In other words, behavior that cannot be clearly distinguished between sand and clay.
- Class 8: This class is strongly identified with clay behavior, with 86 % of ISG classes 3 and 4 and 65 % of FSB classes 1, 2 and 3.
- Class 9: Its behavior is also distributed among ISG and FSB classes, being here considered transitional.

Table 4 was produced to compare ISG classes (columns) with the sample observations obtained via SPT sampling (lines). Numbers represent percentages, similarly to the previous tables, and some ISG classes are omitted for being underrepresented with samples. As defined in Section 3, SPT classes represent sand, silt and clay. Moving from ISG classes 3 to 6, one can observe an increase of sand and decrease of clay, which is coherent with their names given in Section 2.1. An analogous analysis is proposed with Table 5, comparing FSB classes (columns) to SPT (lines). The correspondence to the FSB class names given in Section 2.2 is not clear, except for FSB classes 3 and 7. This suggests that FSB is less sensitive to soil granulometry than ISG.

The clustering results were also compared to the SPT sample observations, resulting Table 6. Cluster classes 3 and 8 contain relevant parts of sand and clay, being here identified as transitional. Class 4 is the only one with predominance of clay and the other can be identified with sand. These observations do not match the ones provided by the comparisons to the ISG and FSB methods, showing that

**Table 4.** Comparison between SPT observations and ISG classes (%).

| | Clays | Clayey silt | Sand mixtures | Sands |
|---|---|---|---|---|
| | 3 | 4 | 5 | 6 |
| Sand | 45 | 59 | 62 | 69 |
| Silt | 25 | 13 | 17 | 12 |
| Clay | 30 | 28 | 21 | 19 |

**Table 5.** Comparison between SPT observations and FSB classes (%).

| | CC | CD | TC | TD | SC | SD |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| Sand | 39 | 46 | 66 | 61 | 53 | 68 |
| Silt | 5 | 29 | 13 | 11 | 23 | 13 |
| Clay | 56 | 25 | 20 | 28 | 25 | 19 |

**Table 6.** Comparison between the k-means clustering and the SPT observations (%).

| | Sand | Silt | Clay |
|---|---|---|---|
| 1 | 53 | 20 | 27 |
| 2 | 62 | 22 | 17 |
| 3 | 47 | 14 | 39 |
| 4 | 39 | 0 | 61 |
| 5 | 79 | 0 | 21 |
| 6 | 69 | 0 | 31 |
| 7 | 40 | 52 | 8 |
| 8 | 63 | 0 | 37 |
| 9 | 90 | 10 | 0 |

soil granulometry alone is not enough to explain its mechanical behavior.

To better illustrate the cluster classes obtained, a case study is presented in Fig. 9. A 29.3 m sounding from the USA was used, being classified using the ISG classes (Fig. 9a), the FSB classes (Fig. 9b) and the cluster classes presented in this section (Fig. 9c). The name of the classes is the same adopted in Tables 2 and 3 and colors are used independently for each classification method.

The last step of this analysis is applying ANN to produce a model capable of reproducing the obtained classification method. This procedure resulted a model with an accuracy of 89.35 % with a standard deviation of 0.40 %, corresponding to an architecture with only one hidden layer with 18 neurons.

Another ANN model was trained using only $z$, $q_c$ and $f_s$ as input features. The objective is verifying if CPT equipment without a pore pressure filter can provide enough information to approximate the method. The resultant model presented an accuracy of 84.47 % with a standard deviation of 0.30 %, corresponding to an architecture with two hidden layers, the first with 16 neurons and the second with 18 neurons.

The weight matrices and bias vectors produced for the ANN models of this section are here omitted for conciseness. Readers interested in this information are advised to contact the authors.

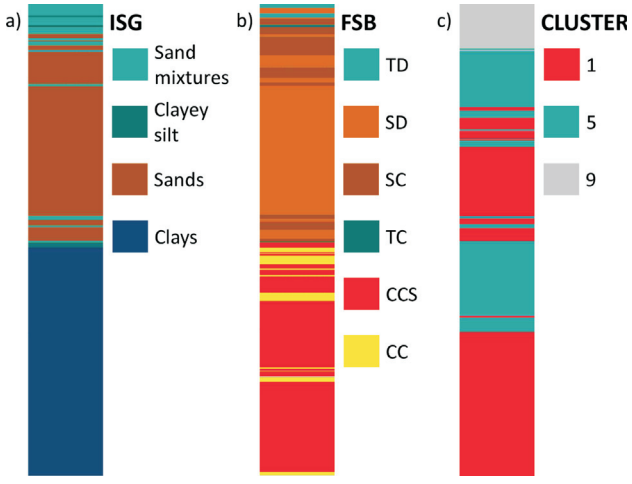


**Figure 9.** Comparing cluster classes to ISG and FSB classes: (a) distribution for ISG classes, (b) distribution for FSB classes and (c) distribution for cluster classes.

### 5.2 Specialized approach

Using CPT data from only 5 soundings, all from the same site, tends to improve classification accuracy. Nonetheless, the obtained model becomes limited to the soil types measured within these 5 soundings. For that reason, these clusters are here considered more specialized than those obtained in the previous section. This strategy is here investigated using the kernel x-means algorithm instead of varying manually the number of classes, which enables maintaining the minimum total internal cluster variance as a performance measure. This allows comparing different results given by this algorithm in cases wherein a high variation of the number of classes $k$ is observed.

Only 5 CPT soundings are used to obtain the specialized classification method by clustering, all taken from the same location and paired with SPT soundings. With the kernel x-means algorithm, 4 classes were found to be the best for the considered dataset, with their centers presented in Table 7.

Crossing results with the ISG and FSB classification methods and to SPT soundings, Tables 8, 9 and 10 are obtained, respectively. As in the previous section, values represent percentages of soil assigned to a cluster class (line) and also to a reference method class (column).

**Table 7.** Specialized clusters centers.

| Class | $z$ (m) | $q_c$ (MPa) | $f_s$ (kPa) | $u_2$ (kPa) |
|---|---|---|---|---|
| 1 | 11.61 | 18.77 | 304.56 | 828.22 |
| 2 | 12.24 | 29.70 | 394.80 | 842.16 |
| 3 | 10.23 | 15.94 | 270.72 | 605.18 |
| 4 | 6.83 | 15.54 | 214.32 | 423.96 |

**Table 8.** Comparison between the specialized x-means clustering and the ISG classes (%).

| | Organic | Clays | Clayey silt | Sand mixtures | Sands |
|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 |
| 1 | 0 | 25 | 27 | 40 | 8 |
| 2 | 0 | 0 | 6 | 26 | 68 |
| 3 | 3 | 32 | 16 | 37 | 12 |
| 4 | 0 | 12 | 19 | 27 | 43 |

**Table 9.** Comparison between the specialized x-means clustering and the FSB classes (%).

| | CCS | CC | CD | TC | TD | SC | SD |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 0 | 4 | 26 | 3 | 22 | 8 | 36 |
| 2 | 0 | 0 | 2 | 0 | 10 | 0 | 89 |
| 3 | 2 | 10 | 27 | 8 | 8 | 16 | 29 |
| 4 | 0 | 1 | 15 | 6 | 10 | 7 | 60 |

**Table 10.** Contribution of each soil granulometrical type for each behavior (%).

| | Sand | Silt | Clay |
|---|---|---|---|
| 1 | 52 | 12 | 36 |
| 2 | 73 | 0 | 27 |
| 3 | 56 | 17 | 28 |
| 4 | 63 | 37 | 0 |

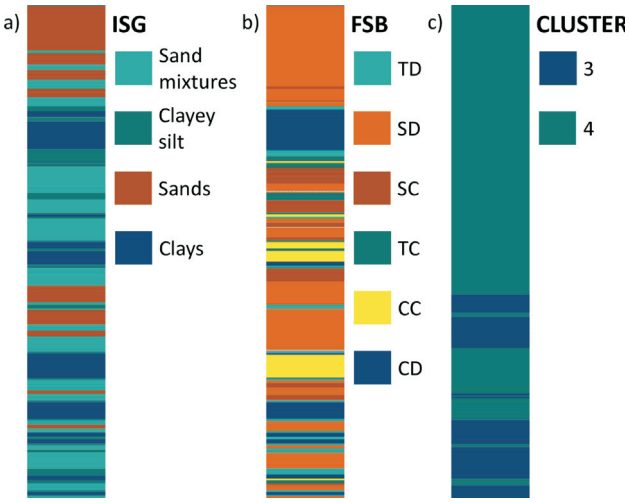


**Figure 10.** Specialized cluster classes compared to ISG and FSB classes: (a) distribution for ISG classes, (b) distribution for FSB classes and (c) distribution for specialized cluster classes.

For this case, some agreement can be observed for the soil type of the cluster classes when compared to ISG, FSB and SPT. Cluster class 1 shows a subtle predominance of sand over clay when compared to the ISG, which is also observed for FSB and SPT. The predominance of sand is clearer for cluster class 2, specially comparing to FSB. Cluster class 3 seems to confuse the ISG and FSB methods, although it can be identified as sand considering SPT alone. Finally, cluster class 4 can be also identified as sand, although such correlation is weaker than the one observed for cluster class 2.

Comparing these results with the ones of the previous section, one can conclude that specializing classification improves agreement with SPT sampling. This can be considered an advantage, for uniting the model capability of predicting soil behavior to a correspondence with SPT visual-tactile observations.

A case study is also presented for the specialized cluster classes, which can be observed in Fig. 10. This sounding is 12.3 m long and is one of the 5 used to produce the specialized cluster classes used in this section. Class names are the same used in Tables 8 and 9.

In the end, an ANN model was produced in order to reproduce the obtained specialized classification method. The obtained model presented very good predictive performance, with an accuracy of 97.04 % and a standard deviation of 1.24 %. This result can be considered significantly better than the one obtained for the general approach, suggesting that limited extrapolations with the specialized approach are feasible. The weight matrices and their respective bias vectors for this last ANN model are:

$$\mathbf{W}_1^T = \begin{bmatrix} 10.85171 & -4.92947 & -0.90277 & 5.630497 \\ -6.626574 & -11.7869 & -11.2992 & -4.61007 \\ -23.57666 & -3.75753 & -5.7113 & -17.3486 \\ 24.05597 & 15.3592 & 6.625066 & 24.37708 \\ -6.341736 & -32.29 & -18.1987 & -1.49596 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} -5.71002 \\ 19.23977 \\ 25.96866 \\ -42.46379 \\ 35.97992 \end{bmatrix} \tag{11}$$

$$\mathbf{W}_2 = \begin{bmatrix} -7.511974 & 6.628167 & -2.40889 & -9.79977 \\ -7.438816 & -2.86605 & 7.8636 & 6.841357 \\ -10.67961 & -6.61815 & -23.5403 & 15.40886 \\ 15.99294 & 4.706883 & -15.8327 & -7.36243 \\ 19.26499 & -16.8685 & -5.45426 & 1.734535 \end{bmatrix}, \quad \Theta_2 = \begin{bmatrix} -14.7905 \\ -5.20106 \\ 10.22392 \\ -7.19002 \end{bmatrix} \tag{12}$$

One can notice that these matrices correspond to an architecture with only one hidden layer containing five neurons. In this case it was also evaluated if suppressing pore pressure information prejudices predictive performance. The resultant ANN model, that makes use of only $z$, $q_c$ and $f_s$, presented an accuracy of 90.37 % with a standard

deviation of 2.48 %. This accuracy can be considered good within geotechnical engineering problems, with the advantage of enabling the use of alternative CPT equipment. This ANN model uses the following weight matrices and bias vectors:

$$\mathbf{W}_1^T = \begin{bmatrix} 26.01 & 7.43 & 0.63 \\ -27.72 & 0.3 & 2.88 \\ -11.63 & -17.05 & 2.1 \\ 10.45 & 12.12 & 13.9 \\ 0.37 & 0.52 & -24.2 \\ 19.4 & -24.54 & -12.07 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} -22.3225 \\ 19.83825 \\ 15.85711 \\ -23.10527 \\ 10.44827 \\ 0.0736 \end{bmatrix} \tag{13}$$

$$\mathbf{W}_2 = \begin{bmatrix} -1,74 & -16.37 & -14.24 & -0.92 & 0.15 \\ -3.11 & 5.96 & 15.79 & -1.36 & -5.08 \\ 6.34 & 2.68 & 6.87 & 2.68 & -16.13 \\ -3.42 & -13.38 & -4.87 & -11.16 & 11.62 \\ 4.09 & 7.71 & 7.26 & -2.25 & -356 \\ 9.62 & -5.48 & -8.32 & 8.63 & -17.26 \end{bmatrix}, \quad \Theta_2 = \begin{bmatrix} 0.77985 \\ -5.41468 \\ 11.45654 \\ -0.42104 \\ -9.51452 \end{bmatrix} \tag{14}$$

$$\mathbf{W}_3 = \begin{bmatrix} 3.14 & -6.33 & -3.29 & 0.44 \\ -6.64 & -4.56 & -6.56 & 8.56 \\ -6.35 & -2.3 & 6.44 & 2.5 \\ -4.74 & -2.06 & 4.74 & -0.92 \\ -11.77 & 9.8 & -2.39 & -6.47 \end{bmatrix}, \quad \Theta_3 = \begin{bmatrix} 4.40147 \\ -3.49722 \\ -4.38422 \\ -6.43941 \end{bmatrix} \tag{15}$$

These matrices correspond to an ANN architecture with two hidden layers, the first with 6 neurons and the second with 5 neurons.

## 6. Conclusions and recommendations

This work explores the kernel k-means and kernel x-means clustering algorithms to group CPT data into different soil classes. Using a kernel function to modify the k-means algorithm enables evaluating classes that are not linearly separable. Next, ANN are used to create mathematical models which can be easily reproduced. Two different approaches are studied, one is general and the other more specialized. The general approach uses 179 soundings from different sources to develop an ANN model that can be better extrapolated to any new CPT data. On the other hand, the specialized approach requires running the kernel x-means to generate specialized classes for each site investigation, as well as producing a new ANN model. The specialized model is expected to be more accurate for sites with soils similar to those for which it was trained, but it is also expected to be more limited for extrapolation. This approach is applied to 5 soundings for which the CPT soundings were paired with SPT soundings. Results confirm that the specialized model produces more well-defined classes and a more accurate ANN model. The mean accuracy (*MA*) and standard deviation (*SD*) obtained for all ANN models are summarized in Table 11.

**Table 11.** Mean accuracy and standard deviation obtained for all ANN models.

| | Inputs | *MA* (%) | *SD* (%) |
|---|---|---|---|
| Full dataset | $z\ q_c\ f_s\ u_2$ | 89.35 | 0.40 |
| | $z\ q_c\ f_s$ | 84.47 | 0.30 |
| Specific dataset | $z\ q_c\ f_s\ u_2$ | 97.04 | 1.24 |
| | $z\ q_c\ f_s$ | 90.37 | 2.48 |

These values can be considered reasonable when compared to other studies from the literature that used ANN to predict soil classes from CPT data, as Bhattacharya & Solomatine (2006) that achieved 83 % and Kurup & Griffin (2006) that achieved 86 %. Thus, Elkateb et al. (2003) cite a case study that shows that pure engineering judgment can lead to 70 % of poor to bad soil predictions.

One advantage of the here proposed methodology is that the ANN models can be reproduced with spreadsheets by simply combining the calibrated weights with the used activation functions. What makes it different from other methods from the literature is the possibility of approximating the soil classes without pore pressure information, becoming an important alternative for geotechnical engineers in cases that high accuracies are not required. Thus, to the best knowledge of the authors, this is the first study that produces ANN models from tropical soil CPT data, being recommended for projects within tropical countries. Nonetheless, this model can be considered limited to the types of

soil for which the ANN models were trained, which is critical particularly for the specialized approach.

## Acknowledgments

To Peter K. Robertson, Paul W. Mayne, Company of São Paulo Metropolitan, Renato P. da Cunha and Heraldo L. Giacheti for making available the dataset used in this work.

## References

Begemann, H.K.S. (1965). The friction jacket cone as an aid in determining the soil profile. *Proc. 6th International Conference on Soil Mechanics and Foundation Engineering*, Montreal, Vol. 1, Univ. of Toronto Press, 17-20.

Bhattacharya, B., & Solomatine, D.P. (2006). Machine learning in soil classification. *Neural Networks*, 19(2), 186-195. https://doi.org/10.1016/j.neunet.2006.01.005

Carvalho, L.O., & Ribeiro, D.B. (2019). Soil classification system from cone penetration test data applying distance-based machine learning algorithms. *Soils and Rocks*, 42(2), 167-178. https://doi.org/10.28927/SR.422167

Carvalho, L.O., Ribeiro, D.B., & Monteiro, F.A.C. (2019). Comparing artificial neural networks with support vector machines for soil classification. *Proc. XL Ibero-Latin-American Congress on Computional Methods in Engineering*, Natal, ABMEC, 11-14.

Cunha, R.P. (1994). *Interpretation of selfboring pressuremeter tests in sand* [Doctoral dissertation, The University of British Columbia] University of British Columbia's repository. https://doi.org/10.14288/1.0050418

Das, S.K., & Basudhar, P.K. (2009). Utilization of self-organizing map and fuzzy clustering for site characterization using piezocone data. *Computers and Geotechnics*, 36(1-2), 241-248. https://doi.org/10.1016/j.compgeo.2008.02.005

Douglas, B.J. (1981). Soil classification using electric cone penetrometer. *Proc. Symp. on Cone Penetration Testing and Experience, Geotechnical Engineering Division*, *New York*, ASCE, 209-227.

Elkateb, T., Chalaturnyk, R., & Robertson, P.K. (2003). An overview of soil heterogeneity: quantification and implications on geotechnical field problems. *Canadian Geotechnical Journal*, 40(1), 1-15. https://doi.org/10.1139/t02-090

Facciorusso, J., & Uzielli, M. (2004). Stratigraphic profiling by cluster analysis and fuzzy soil classification from mechanical cone penetration tests. *Proc. ISC-2 on Geotechnical and Geophysical Site Characterization, Porto*. Vol. 1, Millpress, 905-912.

Hegazy, Y.A., & Mayne, P.W. (2002). Objective site characterization using clustering of piezocone data. *Journal of Geotechnical and Geoenvironmental Engineering*, 128(12), 986-996. https://doi.org/10.1061/(ASCE)1090-0241(2002)128:12(986)

Hornik, K., Stinchombe, M., & and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Ide, D.M. (2009). *Geotechnical site characterization and study of an erosion process caused by urban setting* [Master's dissertation, University of São Paulo] University of São Paulo's repository. https://doi.org/10.11606/D.18.2009.tde-22032010-094227

Kurup, P.U., & Griffin, E.P. (2006). Prediction of soil composition from CPT data using general regression neural network. *Journal of Computing in Civil Engineering*, 20(4), 281-289. https://doi.org/10.1061/(ASCE)0887-3801(2006)20:4(281)

Liao, T., & Mayne, P.W. (2007). Stratigraphic delineation by three-dimensional clustering of piezocone data. *Georisk*, 1(2), 102-119. https://doi.org/10.1080/17499510701345175

McCulloch, W., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115-133. https://doi.org/10.1007/BF02478259

Pelleg, D., & Moore, A. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Proc. 17th Int. Conf. Mach. Learning*, *Stanford.* Stanford University, 727-734.

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. (1991). Soil classification using the cone penetration test: Reply. *Canadian Geotechnical Journal*, 28(1), 176-178. https://doi.org/10.1139/t91-024

Robertson, P.K. (2009). Interpretation of cone penetration tests - a unified approach. *Canadian Geotechnical Journal*, 46(11), 1337-1355. https://doi.org/10.1139/T09-065

Robertson, P.K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system - An update. *Canadian Geotechnical Journal*, 53(12), 1910-1927. https://doi.org/10.1139/cgj-2016-0044

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 a heterogeneous sedimentary aquifer. *Plos One*, 12(5), e0176656. https://doi.org/10.1371/journal.pone.0176656

Rosenblatt, F. (1957). The Perceptron: A Perceiving and Recognizing Automation. Technical Report. *Cornell Aeronautical Laboratory, Buffalo*, NY.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagation errors.

*Nature*, 323(9), 340-347. https://doi.org/10.1038/323533a0

Schneider, J.A., Hotstream, J.N., Mayne, P.W., & Randolph, M.F. (2012). Comparing CPTU Q-F and $Q$-$\Delta u_2/\sigma'_{v0}$ soil classification charts. *Geotechnique Letters*, 2(4), 209-215. https://doi.org/10.1680/geolett.12.00044

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)

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111-133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Wang, X., Wang, H., Liang, R.Y., & Liu, Y. (2019). A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data. *Engineering Geology*, 248, 102-116. https://doi.org/10.1016/j.enggeo.2018.11.014

## Internet resources

Prof. Paul Mayne website, http://geosystems.ce.gatech.edu/Faculty/Mayne/Research/index.html, accessed at 02/03/2020.

## List of symbols

$A$: mean recall
$B_q$: normalized excess pore pressure
$c$: number of classes
$d(\mathbf{x}_i, \mathbf{x}^{(j)})$: distance between $\mathbf{x}_i$ and $\mathbf{x}^{(j)}$
$E$: mean squared error
$F_r$: normalized friction ratio
$f_s$: lateral friction
$I_c$: classification index
$I_{ij}$: equals 1 if prediction j of class $i$ is correct, equals 0 otherwise
$k$: number of clusters
$n$: exponent of $\sigma_{v0}'$
$n_i$: number of examples of class $i$
$p_a$: reference pressure
$q_c$: cone resistance
$q_t$: total cone resistance
$Q_{t1}$: normalized cone resistance
$Q_m$: updated normalized cone resistance
$R_i$: recall of class $i$
$u_0$: equilibrium pore pressure
$u_2$: pore pressure measured behind the cone tip
$U_2$: updated normalized excess pore pressure
$\mathbf{w}$: Gaussian weighting
$\mathbf{x}_i$, $\mathbf{x}_j$: points representing objects
$\mathbf{x}^{(j)}$: cluster centroid
$y$, $g$, $u$, $w$, $x$, $\theta$, $\lambda$: parameters of the Perceptron neuron
$z$: depth
$\sigma$: calibration parameter
$\sigma_{v0}$: total overburden pressure
$\sigma_{v0}'$: effective overburden pressure