Soil Classification System from Cone Penetration Test Data Applying Distance-Based Machine Learning Algorithms

L.O. Carvalho, D.B. Ribeiro

Abstract. Most work from the literature dedicated to soil classification systems from cone penetration test (CPT) data are based on simple two-dimensional charts. One alternative approach is using machine learning (ML) to produce new soil classification systems or to reproduce existing ones. The available studies within this research field can be considered limited, once most of them do not include more than two inputs within their analysis and are applicable only to specific regions. In this context, the aim of this work is to use distance-based ML techniques to replicate two chart-based methods from the literature. Up to five input feature combinations are tested, with the objective of discussing geotechnical aspects of soil classification systems. Results are compared using the statistical test of Friedman with the post-hoc statistics of Nemenyi and the signed-rank statistical test of Wilcoxon. The used dataset can be considered diversified because it contains 111 CPT soundings from several countries. Results show that the used ML techniques maintain reasonable accuracy when inputs are substituted and when incomplete data is used, which can lead to cost reduction in real engineering projects. It is important to notice that these observations would not be possible by using the replicated soil classification systems alone.

Keywords: cone penetration test, distance-based algorithms, machine learning, soil classification system.

1. Introduction

Most available systems for soil classification from CPT data use two-dimensional charts divided into regions which represent different soil types. Initially, these charts were based on soil type (grain size and plasticity) and used raw CPT data, like cone resistance and lateral friction (Begemann, 1965). Nonetheless, later studies produced better classification methods by focusing on soil behavior and by proposing normalizations for CPT data (Douglas, 1981; Robertson *et al.*, 1986). Some popular classification methods make use of two charts instead of one, combining three normalized variables in pairs. In this context, some work propose normalizations to include the influence of depth and overburden (Robertson, 1990). Nevertheless, these methods are not accurate for offshore soils (Jefferies & Davies, 1991) due to the dilative behavior of highly overconsolidated clays commonly found in deep water soils (Robertson, 1991). This limitation is supported by experimental data (Jefferies & Davies, 1991; Ramsey, 2002). Thus, these methods fail to distinguish stiff or dense granular soils from overconsolidated clay (Schneider *et al.*, 2008). Different normalizations and charts were proposed to address this problem (Schneider *et al.*, 2008; Schneider *et al.*, 2012). Nonetheless, Robertson (2016) affirms that soil classification systems that use charts may not be reliable for structured soils, meaning aged or cemented, like some offshore soils. He also recommends to consider soils structured if a modified normalized small-strain rigidity index K_G^* is above 330, although some geotechnical judgment is required.

Another possible approach for soil classification systems from CPT data is based on the use of statistics and ML techniques. Most authors interested in solving general geotechnical problems use artificial neural networks (ANN) to predict values of interest such as soil parameters (Goh, 1995; Goh, 1996; Schaap *et al.*, 1998; Juang & Chen, 1999; Kumar *et al.*, 2000; Juang *et al.*, 2002; Juang *et al.*, 2003; Hanna *et al.*, 2007). Nevertheless, one can find work using support vector machines (SVM) (Goh & Goh, 2007), decision trees (DT) (Livingston *et al.*, 2008) and random forests (RF) (Kohestani *et al.*, 2015). For soil classification systems there are two main approaches, one is replicating existing soil classification systems and the other is trying to propose new ones. Most work in this research field are dedicated to the latter approach, using data clustering (Hegazy & Mayne, 2002; Facciorusso & Uzielli, 2004; Liao & Mayne, 2007; Das & Basudhar, 2009; Rogiers *et al.*, 2017). Usually, among the few work that investigate replicating existing soil classification systems such as Robertson charts (Arel, 2012), the only ML technique tested is ANN (Kurup & Griffin, 2006; Reale *et al.*, 2018). Nonetheless, there is a study that compares different ML techniques when replicating existing systems for soil classification (Bhattacharya & Solomatine, 2006), although the used

Lucas Orbolato Carvalho, M.Sc., Departamento de Geotecnia, Divisão de Engenharia Civil, Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brazil, e-mail: lucasorbol.carvalho@gmail.com.
Dimas Betioli Ribeiro, Associate Professor, Departamento de Geotecnia, Divisão de Engenharia Civil, Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brazil, e-mail: dimas@ita.br.
Submitted on March 12, 2018; Final Acceptance on July 8, 2019; Discussion open until December 31, 2019.
DOI: 10.28927/SR.422167

dataset is restricted to few CPT soundings which are all taken from the same location. Other work related to classifying soil with ML are Bilski & Rabarijoely (2009), Rao *et al.* (2016) and Chandan & Thakur (2018).

In this work, two chart-based soil classification systems proposed by Robertson (1991) and Robertson (2016) are replicated using distance-based ML techniques. These techniques were elected among other options for their simplicity and because there is a lack in the literature for this type of approach. The objective is to investigate and discuss geotechnical aspects of soil classification systems that can not be disclosed by using the original Robertson methods. First, the stratigraphic profiles of 111 CPT soundings taken within several countries are obtained using a student version of CPeT-IT v2.0.2.5 software (Ioannides & Robertson, 2016), which employs Robertson charts in a soil classification system. Next, the so-called k-nearest neighbor (KNN) and distance-weighted nearest-neighbor (DWNN) ML techniques are used to replicate Robertson (1991) and Robertson (2016) charts. For each ML technique and each classification method, 33 input feature combinations are tested and all results are compared using the Friedman statistical test (Friedman, 1937) with the Nemenyi post-hoc statistics (Nemenyi, 1963) and the Wilcoxon statistical test (Wilcoxon, 1945). The proposed discussions produced several original contributions, like showing that:

1. Distance-based ML techniques are capable of reproducing Robertson soil classification systems with good accuracy;
2. Reasonable accuracy can be obtained without normalizations proposed in the literature for the CPT data;
3. Including soil age as an input feature contributes for distinguishing between soil classes.

2. Soil Classification Systems

In this work, two soil classification systems available within a student version of CPeT-IT software are replicated using distance-based ML techniques. The objective of this section is to present the theory that sustains each of these methods.

2.1. Influenced by soil type (IST)

The first replicated method is based on the work of Robertson (1991). Although it was idealized to be oriented towards a behavioral classification, the labels assigned to classes are inspired by conventional soil type classes, showing even some compatibility with real soil types (Kurup & Griffin, 2006). For this reason, this method is here considered influenced by soil type, being referred to as IST throughout this text. It adopts nine possible soils types, within which two are said to be heavily overconsolidated or cemented. The IST classes are in Table 1.

The initial inputs used by CPeT-IT to classify soil with the IST method are raw CPT data, named cone resistance q_c (MPa), lateral friction f_s (kPa), pore pressure measured behind the cone tip u_2 (kPa) and depth z (m). These values are used to obtain the input features originally considered by Robertson (1990), named normalized cone resistance Q_{t1} (Eq. 1), normalized friction ratio F_r (Eq. 2) and normalized excess pore pressure B_q (Eq. 3). The cone resistance normalization was later updated to Q_{tn} (Schneider *et al.*, 2008), resulting in the charts presented in Figs. 1a and 1b. Beside the nine classes predicted within these charts, an additional class 0 is used for misclassified soils.

To obtain the normalized values, first the raw cone resistance q_c is replaced by the total cone resistance q_t, to compute the pore pressure assisting cone penetration. Next step is estimating the soil unit weight γ (kN/m^3) (Lunne *et al.*, 2002; Mayne *et al.*, 2010; Mayne, 2014), which is used to obtain the total overburden pressure σ_{v0} (kPa) and the effective overburden pressure σ_{v0}' (kPa). If the water table is not known, it can be estimated by fitting a straight line in the chart $z \times u_2$ (Fig. 2) when a drained penetration is observed. The water table depth is then used to compute the equilibrium pore pressure u_0, which is used to determine the excess pore pressure $u_2 - u_0$.

Given these estimations, the following normalizations are obtained:

$$Q_{t1} = \frac{q_t - \sigma_{v0}}{\sigma'_{v0}} \tag{1}$$

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \tag{2}$$

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}} \tag{3}$$

Nevertheless, work from the literature state that the exponent n of σ_{v0}' (n = 1 in Eq. 1) should vary from 0.5 for sands to 1 for clays (Zhang *et al.*, 2002). To calculate n, one can consider its correlation with the classification index I_c (Robertson, 2009):

$$I_c = [(3.47 - \log Q_{tq})^2 + (\log F_r + 1.22)2]^{0.5} \tag{4}$$

The normalized cone resistance Q_{tn} is then given by:

Table 1 - IST classes.

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

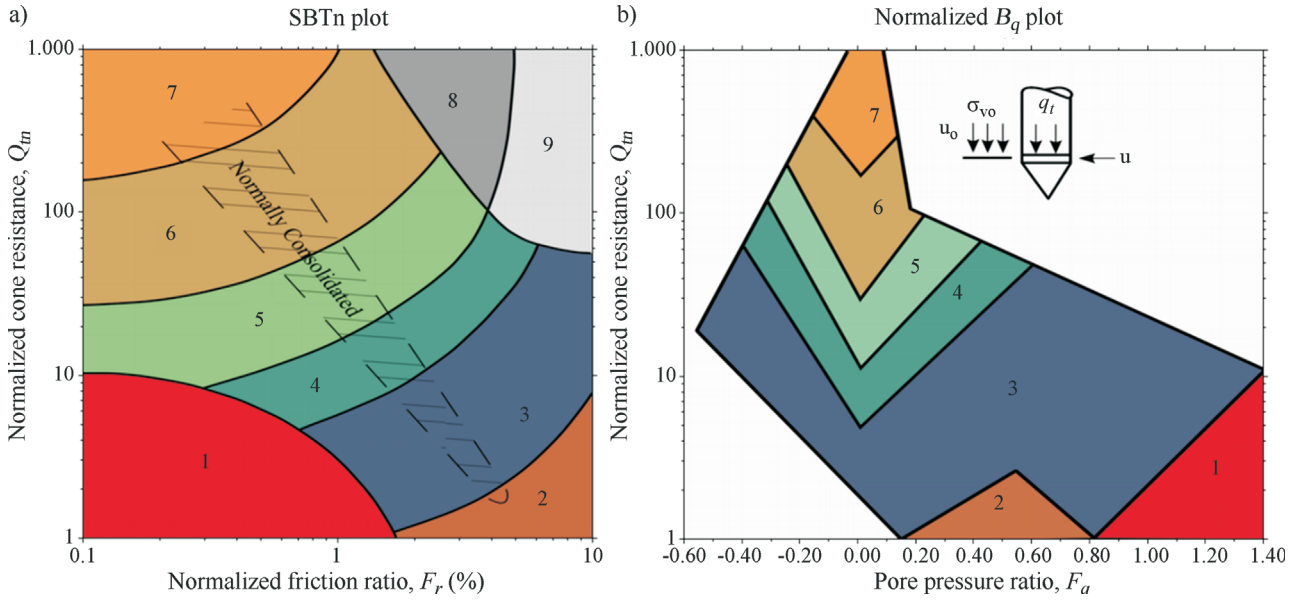

Figure 1 - a) $Q_{tn} \times F_r$ chart from Robertson (1991) updated by Robertson (2009). b) $Q_{tn} \times B_q$ chart from Robertson (1991) updated by Robertson (2009).

$$Q_{tn} = \left(\frac{q_t - \sigma_{v0}}{p_a} \right) \left(\frac{p_a}{\sigma'_{v0}} \right)^n \quad (5)$$

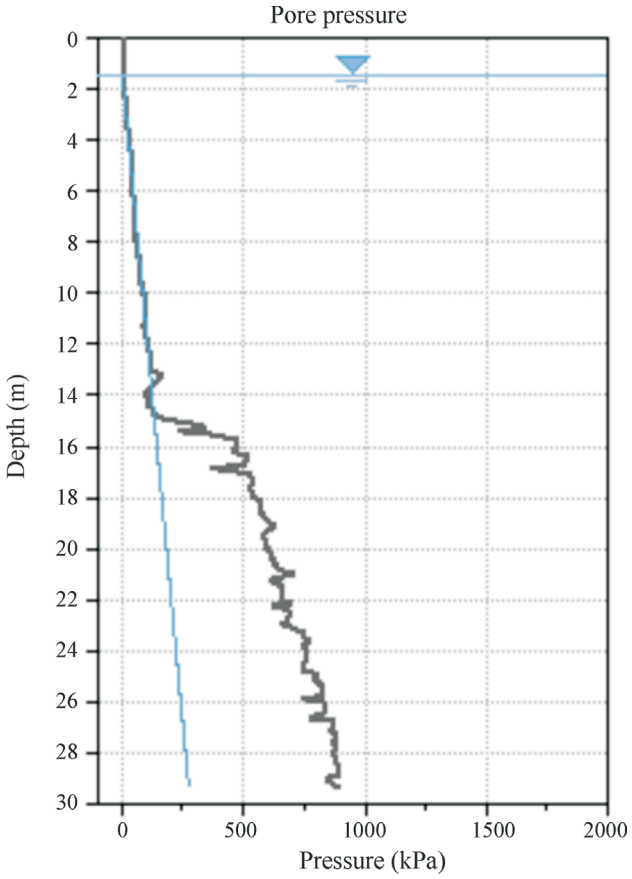

Figure 2 - Excess pore pressure.

and the exponent n can be written as:

$$n = 0.381 I_c + 0.05 \left(\frac{\sigma'_{v0}}{p_a} \right) - 0.15 \quad (6)$$

where p_a = 0.1 MPa is a reference pressure.

The CPeT-IT software uses only the $Q_{tn} \times F_r$ chart to generate the soil classification system outputs. Soil is considered misclassified and is labeled with class 0 if the values obtained for Q_{tn} and F_r are not within the ranges presented in this chart.

2.2. Focused on soil behavior only (FSB)

The system proposed by Robertson (2016) establishes a full behavioral-oriented soil classification, which is why it is here considered more focused on soil behavior and named FSB throughout this text. FSB method includes seven classes (Table 2).

One can observe that the three main soil types are sand-like, clay-like and transitional. Each of these soil types is divided into contractive or dilative. A seventh class

Table 2 - FSB 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

is reserved for contractive clays that have high sensitivity to disturbance, which can be related to the friction ratio using the expression $S_t = 7.1/F_r$ (Robertson, 2009). If sensitivity is greater than 3, which corresponds to $F_r < 2\%$, then the clay is considered sensitive. The upper limit for the normalized cone resistance for sensitive clays is defined as 10 because they are soft.

Likewise for the IST system, q_c, f_s, u_2 and z are the initial inputs used by CPeT-IT to classify soil with the FSB system. Nonetheless, in this case the soil classification system is based on the charts shown in Figs. 3 and 4, which use the normalized cone resistance Q_{tn}, the normalized friction ratio F_r and the normalized excess pore pressure U_2 (Schneider *et al.*, 2008) as inputs. The FSB method also includes a class 0 for misclassified soil, which is identified if Q_{tn}, F_r or U_2 are not within the ranges presented in the charts and if the class given by both charts is not the same.

The excess pore pressure normalization U_2 is obtained as:

$$U_2 = \frac{u_2 - u_0}{\sigma'_{v0}} \tag{7}$$

The curves that separate soil classes are inspired by Schneider *et al.* (2008) and Schneider *et al.* (2012). The $Q_{tn} \times F_r$ chart has closely circular curves in the IST method, while in Robertson (2016) the curves have hyperbolic shapes as suggested by Schneider *et al.* (2012). The $Q_{tn} \times U_2$ chart was taken from Schneider *et al.* (2008) with minor changes, containing the classes originally proposed there.

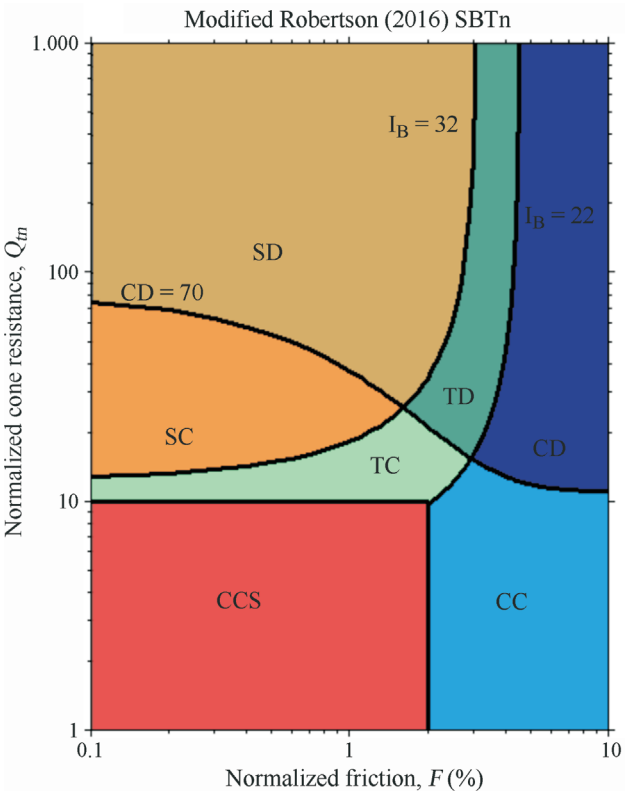

Figure 3 - $Q_{tn} \times F_r$ chart from Robertson (2016).

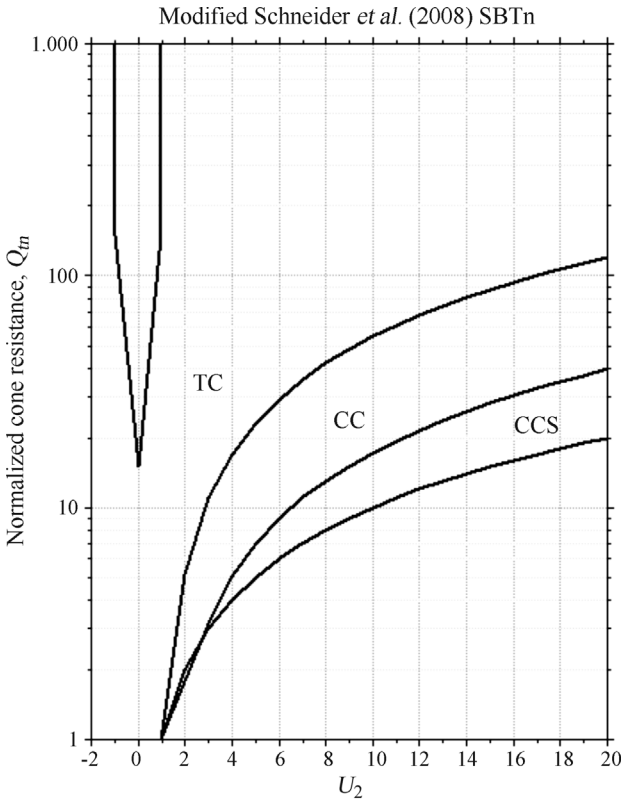

Figure 4 - $Q_{tn} \times U_2$ chart from Robertson (2016).

3. Distance-Based Techniques

In this work, distance-based ML techniques are used to replicate the soil classification systems described in Section 2. These ML techniques have the advantage of using an approach similar to the chart-based methods to be replicated, representing soil examples as points in a space composed by the input features. It also uses the hypothesis that, if two soil examples produce close points, they are similar. One way of measuring the distance between points is with the Euclidean metric. Considering a pair (x_i, x_j) of objects in a d-dimensional feature space, the distance between them is given by:

$$dist(x_i, x_j) = \sqrt[p]{\sum_{l=1}^{d} \left| x_i^l - x_j^l \right|^2} \tag{8}$$

The distance-based ML algorithms used in this work predict the class of an unknown example using a dataset of examples whose classes are known. The simplest strategy is detecting which known example produces a point that is the nearest neighbour of the point that represents the unknown example. It is then assigned to the unknown example the same class of its nearest neighbor (Cover & Hart, 1967).

It is also possible to use an arbitrary number k of nearest neighbors and decide the class of the unknown example by voting, which corresponds to the k-nearest neighbors (KNN) technique. Tests can be performed to calibrate which k leads to best predictive performance. In this work, only odd values of k are tested starting from one, increasing *k* until decreasing predictive performance is observed.

It is also possible to weight the votes, so that closer neighbors are more valued than farther ones. In this case, the technique is named distance-weighted nearest neighbor (DWNN) (Dudani, 1976). One specific way for defining these weights is by using Gaussian weighting, which is defined by the following expression (Hechenbichler & Schliep, 2004):

$$w(dist) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}dist^2} \quad (9)$$

where *dist* is the distance value. In this work, the KNN and the DWNN with Gaussian weighting are used and compared to replicate the soil classification systems presented in Section 2.

4. Methodology

4.1. Datasets description

The ML programs used in this work require a dataset of known examples to predict new examples. This dataset can be formatted as a table, where each line represents a different soil example. Input features are represented as columns and the last column contains the output feature. Table 3 presents a sample with 10 soil examples (lines), within a 0.45 m soil layer. In this sample, the inputs are raw CPT data and the output is the corresponding IST soil class, obtained using the CPeT-IT software. This program is also used to produce other input-output combinations for the ML techniques, as described with more detail in Section 4.3.

Thirty eight of all CPT soundings used to compose the datasets were sent directly by Professor P.K. Robertson.

Table 3 - Sample of soil examples.

Inputs				Output
z (m)	q_c (MPa)	f_s (kPa)	u_2 (kPa)	IST Class
13.00	16.93	55.84	120.32	6
13.05	16.53	46.32	124.47	6
13.10	10.14	36.69	129.95	6
13.15	7.13	24.76	146.66	6
13.20	4.92	22.80	158.69	6
13.25	3.90	21.47	163.42	5
13.30	3.28	21.46	159.58	5
13.35	2.73	23.89	153.03	5
13.40	3.70	33.80	148.37	5

They are the same ones used in Robertson (2016) to produce the FSB method, which is described in Section 2.2. Once detailed information about these soundings can be found in the original reference, only a brief description about them is presented in Table 4.

The first column of Table 4 gives a general description of the soil types within the CPT soundings. The second identifies where soundings were taken and the third gives the geological age when the soil was deposited. The last column presents a discrete ordered variable named "class of geology" (CG), considering the most recent age 1 and the other numbered sequentially to the oldest. The information from these 38 soundings plus the variable CG compose the here named geological dataset. The objective of including CG, as an input feature in some of the studies presented in Section 5, is investigating if information about geological age can help differentiate one soil class from the other.

Another 73 CPT soundings were obtained from the website of Professor P.W. Mayne, whose information is summarized in Table 5. Further detail about the soundings can be found on the website. All these soundings were taken within the United Sates of America and more specific information about location is presented in the table. Information about geological age was not available for these soundings, so they are not included in studies that make use of the variable CG. These soundings grouped with the ones sent by Robertson compose the here named complete dataset, totalizing 111 CPT soundings. All CPT data used in this dataset were taken in intervals of 2 to 5 cm, the pore pressure was measured behind the cone tip (u_2) and the raw cone tip resistance q_c was corrected to q_t using CPeT-IT.

4.2. Data preprocessing

In this work, all CPT data were used to classify soil using the CPeT-IT software, which was later replicated using the methods described in Section 3. The accuracy of the final results depends on the quality of the used datasets, which can be improved with data preprocessing.

The first problem is that distance-based ML techniques are sensitive to data scale. When the distance between points is calculated, the importance of input features that vary within large ranges tends to be emphasized, while the ones with low variation tend to be ignored. The solution adopted here is normalizing all input features to the interval [0, 1].

Another issue is that data taken within CPT soundings can contain noise, which is here defined as any variable becoming severely different from what it was supposed to be. Noise can have several causes, like sensor errors, formatting problems and human mistakes. The main noise types are missing data and outliers, which are here defined as distorted or corrupted values. CPeT-IT is unable to classify most noisy examples, assigning class 0 in both IST and FSB methods or no class whatsoever. Once the ML

Table 4 - Geological dataset (Robertson, 2016) (Classification in terms of geology age – GC).

General soil type	Identification	Geological age	CG
Mixed Soils	UBC, Canada	Holocene	2
	Venice Lagoon, Italy	Holocene	2
	Ford Center, USA	Pleistocene	4
	San Francisco, USA	Late Pleistocene	3
	Tailings, USA	Recent	1
	UBC KIDD, Canada	Holocene	2
	UBC KIDD, Canada (2)	Holocene	2
Soft Clay	Bothkennar, RU	Holocene	2
	Burswoord, Perth, Australia	Holocene	2
	Onsoy, Norway	Holocene	2
	Amherst, USA	Late Pleistocene	3
	San Francisco Bay, USA	Holocene	2
	San Francisco Bay, USA (2)	Holocene	2
Soft Rock			
	Newport Beach, USA	Miocene	5
	LA Downtown, USA	Miocene	5
	Newport Beach, USA (2)	Miocene	5
Stiff Clay			
	Madingley, UK	Cretaceous	6
	Houston, USA	Pleistocene	4

Table 5 - Number of CPTs and test location from P.W. Mayne database (acquired in years 2000 – 2003).

Location	Number of soundings
Gosnell, Arkansas	1
Lenox, Tennessee	1
Memphis, Tennessee	16
Dexter, Missouri	6
Mooring, Tennessee	6
Marked Tree, Arkansas	19
Collierville, Tennessee	1
Meramec, Missouri	4
Opelika, Alabama	4
Wilson, Arkansas	4
Wolf, Wyoming	7
Wyatt, Missouri	4
Total	73

techniques presented in Section 3 are here used to replicate CPeT-IT, these errors tend to be also replicated.

Although it is difficult to completely eliminate noise from the datasets, it is desirable to reduce them as much as possible in order to avoid classification errors. In this work, dataset cleaning was first performed manually, removing the noisy examples that could be easily identified. This procedure was then complemented by an automatic cleaning procedure that makes use of the box-plots of the input features, as illustrated in Fig. 5.

In the box-plot, the base of the rectangle represents the first quartile Q_1 and the top of the rectangle represents the third quartile Q_3. The whiskers above and below the rectangle represent the interval $[Q_1 - 1.5 \times IQ, Q_3 + 1.5 \times IQ]$, where $IQ = Q_3 - Q_1$. Values outside this range (white circles) are identified as potential outliers. Preliminary tests have shown that removing all potential outliers affects accuracy, which indicates that relevant information is being eliminated. To solve this problem, the Edit Nearest Neighbor technique (Wilson, 1972) is used in this work as a second criterion to decide if each potential outlier will be, in fact, removed. This technique compares the potential outlier with its nearest neighbor and removes it only if the classes given by CPeT-IT do not match.

This procedure is illustrated in Fig. 6 for two input features, where the white dot represents the potential outlier and the black dots represent other known examples from the dataset. The numbers close to each dot represent the class assigned by CPeT-IT. One can observe that, in this example, the classes of the potential outlier and its nearest

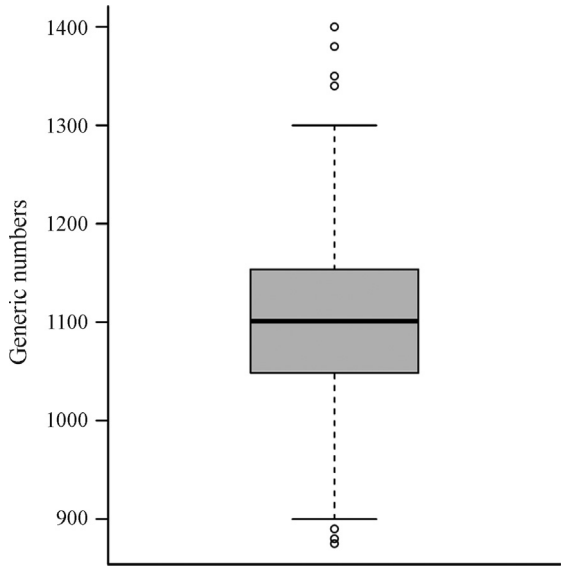

Figure 5 - Box-plot example using generic numbers. The rectangle represents ordinate values within the 1st and 3rd quartiles and the circles represent outliers.

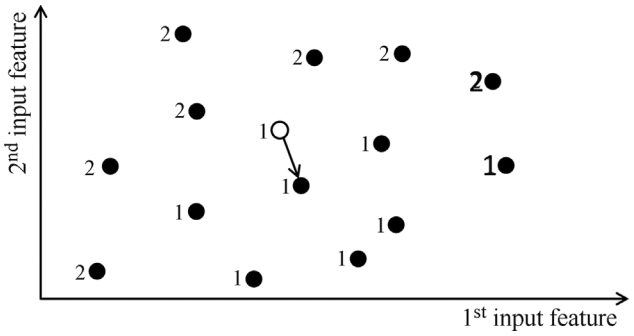

Figure 6 - Edit nearest neighbor technique, with two possible classes (1 and 2) and black dots representing known examples. The unknown example (white) is labeled with the class of its nearest neighbor.

neighbor are the same. This means that this potential outlier will be maintained.

The next issue to be evaluated is if the number of examples within each soil class is balanced, considering both IST and FSB methods. Severe unbalance can compromise the accuracy of distance-based ML techniques because they tend to focus majority classes and ignore minority classes. The distribution of examples among classes can be checked using histograms, as presented in Figs. 7a and 7b for the complete dataset and Figs. 7c and 7d for the geological dataset.

One can observe that the classes are, in fact, imbalanced, which is expected for real CPT soundings. In this work, data imbalance is prevented by eliminating examples of majority classes and creating new artificial examples for minority classes. Preliminary results have shown that random elimination does not affect predictive performance, which can be explained by the fact that CPT data contains redundancies due to several data items being taken within each soil layer.

To create new artificial examples for minority classes, the SMOTE (Chawla *et al.*, 2002) technique was used. For better distribution within the input feature space, it is here proposed to estimate each d-dimensional new artificial object from $d + 1$ original examples. This corresponds to the vertex number of a d-dimensional simplex. The maximum between 1000 and two times the number of elements of the minority class was stipulated as the final number of elements of each class for the balanced dataset. Once class 0 of the IST method could not be well represented within the geological dataset even with the use of SMOTE, examples of this class were completely removed from the geological dataset.

4.3. General strategy

Two ML algorithms are tested and compared, the classical KNN and the Gaussian DWNN, with respect to their capacity for replicating the IST and FSB soil classification systems. This comparison is made using several input feature combinations, including three basic sets:

- First set: depth z (m), corrected cone resistance q_t (MPa), lateral friction f_s (kPa) and pore pressure behind the cone tip u_2 (kPa);
- Second set: depth z (m), normalized cone resistance Q_{t1}, normalized lateral friction F_r (%) and normalized pore pressure B_q;
- Third set: depth z (m), normalized cone resistance Q_{tn}, normalized lateral friction F_r (%) and normalized pore pressure U_2.

The first set contains only non-normalized parameters, the second contains inputs of the IST method combined with depth and the third contains inputs of the FSB method combined with depth. For the main analysis, all combinations of two, three and four input features within each set were tested, although not all of them are presented in Section 5. Additional selected input feature combinations are tested in three complementary studies.

In order to generate statistically relevant comparisons, a 10-fold cross-validation procedure (Stone, 1974) was applied to evaluate classification accuracy. The procedure starts by randomly separating the dataset into ten partitions or folds with approximately the same size, maintaining the same proportion between classes observed in the complete dataset. At each cross-validation round, one partition is left for testing, one partition (chosen at random) is chosen as a validation set and the remaining partitions compose the training set. The validation set is used to calibrate the best number of neighbors k to be used in the distance-based algorithms.

For each cross-validation round, the average of the accuracies per class are taken. This avoids disregarding mi-

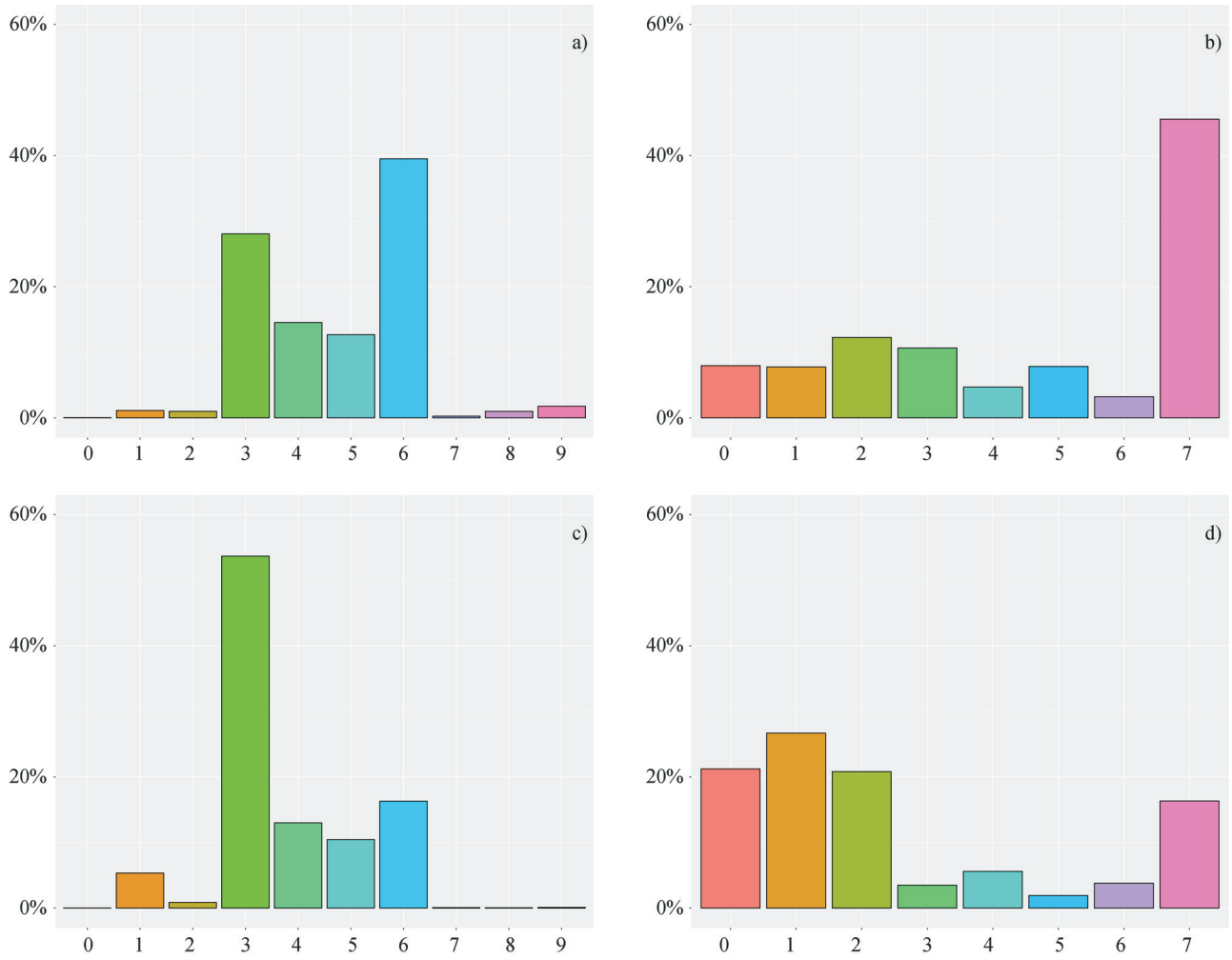

Figure 7 - Histograms. (a) For IST classes and the complete dataset. (b) For FSB classes and the complete dataset. (c) For IST classes and the geological dataset. (d) For FSB classes and the geological dataset.

nority classes in the performance evaluation. After all folds are used for testing, a mean and a standard deviation accuracy performance are computed. For comparing the results of the experiments, the Friedman statistical test (Friedman, 1937) with the Nemenyi post-hoc statistics (Nemenyi, 1963) and the Wilcoxon statistical test (Wilcoxon, 1945) are used, based on the 10 accuracies recorded (per test fold).

5. Results and Discussion

A total of 132 classification results were generated to produce the comparisons presented in this main analysis: 2 replicated classification methods (IST and FSB described in Section 2) × 33 input feature combinations × 2 distance-based classification algorithms. The units used for the input features are z (m), q_t (MPa), f_s (kPa), u_2 (kPa), F_r (%) and the other ones are dimensionless. Each predicted soil class is compared to the one originally given by CPeT-IT to compute accuracy. Tables present the mean and standard deviation accuracy obtained within the 10-fold cross-validation procedure described in Section 4.3.

Combinations that presented best performance with KNN for replicating IST outputs are presented in Table 6. Once the first combination uses the original IST inputs and outputs, it was expected that it would lead to the highest mean accuracy among all. Nonetheless, results of the Friedman statistical test with the Nemenyi post-hoc statistics show a statistical equivalence between the first two combinations in Table 6. Thus, the last two combinations shown in Table 6 can be considered of lower performance. This

Table 6 - Best KNN predictive results for replicating IST.

Inputs	Elected k	Mean	SD
$Q_{tn}\ F_r$	1	96.52	0.57
$Q_{tn}\ F_r\ U_2$	3	94.70	0.96
$Q_{tn}\ z\ F_r$	1	93.49	0.67
$Q_{tn}\ z\ F_r\ U_2$	1	92.54	1.12

shows that including more features among the original ones does not contribute to improve performance in this case.

The same comparison is proposed for the combinations that lead to the best performance with the Gaussian DWNN technique for replicating IST outputs, which are presented in Table 7. One can observe that results are very close to those presented in Table 6, reinforcing the same conclusions.

Considering now the classical KNN technique for replicating FSB outputs, the best feature combinations are presented in Table 8. In this case, the Friedman statistical test with the Nemenyi post-hoc statistics show that last two feature combinations are equivalent and statistically better than the first two. One can observe that, as expected, the best combination for this case include all three original FSB inputs, named Q_{tn}, F_r and U_2. However, associating depth to these features contributed to improve performance, even with the biasing due to the way in which the outputs were generated.

In the end, the feature combinations that produced best performance for the Gaussian DWNN technique for replicating FSB outputs are presented in Table 9. One can observe that the results are very close to the ones from Table 8, reinforcing that using original FSB inputs leads to good accuracy and that including depth among these features contributes to improve performance.

Table 7 - Best DWNN predictive results for replicating IST.

Inputs	Elected *k*	Mean	SD
$Q_{tn}\ F_r$	1	96.52	0.57
$Q_{tn}\ F_r\ U_2$	1	94.63	0.98
$Q_{tn}\ z\ F_r$	1	93.49	0.67
$Q_{tn}\ z\ F_r\ U_2$	1	92.63	1.02

Table 8 - Best KNN predictive results for replicating FSB.

Inputs	Elected *k*	Mean	SD
$Q_{tn}\ F_r$	7	88.79	0.40
$Q_{tn}\ z\ F_r$	1	91.86	0.28
$Q_{tn}\ F_r\ U_2$	3	92.97	0.46
$Q_{tn}\ z\ F_r\ U_2$	1	93.83	0.55

Table 9 - Best DWNN predictive results for replicating FSB.

Inputs	Elected *k*	Mean	SD
$Q_{tn}\ F_r$	7	88.90	0.40
$Q_{tn}\ z\ F_r$	1	91.86	0.28
$Q_{tn}\ F_r\ U_2$	1	93.02	0.38
$Q_{tn}\ z\ F_r\ U_2$	1	93.83	0.55

Concerning more general observations, both tested ML techniques presented good performance for replicating both soil classification systems. With respect to the non-normalized inputs, good performance can be observed when they are associated with depth. For IST and both ML techniques, for example, accuracy is around 70% when only q_t and f_s are used as input features, but rises close to 90% when z is included. These observations suggest that proposing a soil classification system that uses only raw CPT data would be feasible if depth is included. Nevertheless, one should notice that confirming this hypothesis would require further studies.

Another general observation concerns evaluating which classification technique is better, comparing the classical KNN and the Gaussian DWNN. The Wilcoxon test was employed for this task adopting a p-value of at most 5%. Comparing all combinations, results show that the Gaussian DWNN presents better predictive performance than the classical KNN.

6. Conclusions and Recommendations

In this work, distance-based ML techniques are used to replicate systems for soil classification from CPT data. It is important to notice that the proposed discussions and obtained conclusions would not be possible by using the original soil classification systems alone, because these original methods do not allow changing input features. It was the flexibility of the ML techniques that made possible to evaluate if raw inputs without normalizations have enough information for reproducing the original methods accurately, for example.

The main advantages of the proposed approach are the ease of applying it to different datasets and little adaptation required for it to be associated with other ML techniques. The use of distance-based techniques can also be considered advantageous for its simplicity, once accurate results were obtained. Thus, the presented method can be considered rigorous compared to other work from the literature that make use of ML applications in geoscience, which do not present a data analysis as detailed as in Section 4.

A total of 132 tests were performed to draw the discussions and conclusions presented and in all of them the mean accuracy is above 85%, which can be considered reasonable within geotechnical applications. Notice the good results obtained using raw parameters, which suggests that would make sense to dismiss some types of data normalization that are proposed in the literature for soil classification systems. Reducing data normalization is advantageous because any data transformation proposed to the original dataset tend to diminish its original information, specially if the original number of input features is reduced. Results presented here are not sufficient to affirm that using raw parameters would lead to greater performance, nonetheless

they can justify future studies about this issue. Other conclusions to be pointed out are:

- Highest accuracies were obtained when using the original IST inputs and outputs;
- Including depth as an input increased accuracy, in most cases;
- Gaussian DWNN is better than the classical KNN, considering the Wilcoxon test with a p-value of at most 5%.

Future studies that can be conducted include applying and comparing different ML techniques to this same problem, discussing other geotechnical issues about soil classification systems that can not be exposed using distance-based techniques. Another possible investigation is applying clustering techniques to the problem, taking advantage of the ease of increasing dimensionality to test several normalized and non-normalized feature combinations. Thus, CPT data can be associated with data from other in situ experiments like the standard penetration test or the flat dilatometer test, exploring the problem with even higher dimensionality.

Acknowledgments

To Peter K. Robertson and Paul W. Mayne for making available the dataset used in this work. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Computer Code Availability

The codes produced to generate all results presented in this paper can be downloaded from the following link: https://github.com/Orbolato/KNN.git.

References

Arel, E. (2012). Predicting the spatial distribution of soil profile in Adapazari/Turkey by artificial neural networks using CPT data. Computers and Geosciences, 43:90-100.

Begemann, H.K.S. (1965). The friction jacket cone as an aid in determining the soil profile. Proc. 6th Int. Conf. on Soil Mech. and Found. Engn., ICSMFE, Montreal, v.1, pp. 8-15.

Bhattacharya, B. & Solomatine, D.P. (2006). Machine learning in soil classification. Neural Networks, 19(2):186-195.

Bilski, P. & Rabarijoely, S. (2009). Automated soil categorization using CPT and DMT investigations. Proc. 2nd Int. Conf. on New Developments in Soil Mechanics and Geotechnical Engineering, Nicosia, North Cyprus, v. 1, pp. 1-8.

Chandan, T.R. (2018). Recent trends of machine learning in soil classification: A review. International Journal of Computational Engineering Research (IJCER), 8(9):25-32.

Chawla, N.V.; Bowyer, K.W.; Hall, L.O. & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321-357.

Cover, T.M. & Hart, P.E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1):21-27.

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.

Douglas, B.J. (1981). Soil classification using electric cone penetrometer. In Symp. on Cone Penetration Testing and Experience, Geotech. Engrg. Div., American Society of Civil Engineers, New York, NY, pp. 209-227.

Dudani, S.A. (1976). The distance-weighted *k*-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(4):325-327.

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. Rotterdam, v. 1, pp. 905-912.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200):675-701.

Goh, A.T. (1995). Modeling soil correlations using neural networks. Journal of Computing in Civil Engineering, 9(4):275-278.

Goh, A.T. (1996). Neural-network modeling of CPT seismic liquefaction data. Journal of Geotechnical Engineering, 122(1):70-73.

Goh, A.T. & Goh, S.H. (2007). Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. Computers and Geotechnics, 34(5):410-421.

Hanna, A.M.; Ural, D. & Saygili, G. (2007). Neural network model for liquefaction potential in soil deposits using Turkey and Taiwan earthquake data. Soil Dynamics and Earthquake Engineering, 27(6):521-540.

Hechenbichler, K. & Schliep, K. (2004). Weighted *k*-nearest-neighbor techniques and ordinal classification. Collaborative Research Center 386, Discussion Paper 399.

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.

Ioannides, J. & Robertson, P.K. (2016). CPeT-IT v.2.0 – CPT interpretation software. URL: https://geologismiki.gr/, accessed on July 26th 2019.

Jefferies, M.G. & Davies, M.P. (1991). Soil classification by the cone penetration test: Discussion. Canadian Geotechnical Journal, 28(1):173-176.

Juang, C.H. & Chen, C.J. (1999). CPT-based liquefaction evaluation using artificial neural networks. Computer-

Aided Civil and Infrastructure Engineering, 14(3):221-229.

Juang, C.H.; Yuan, H.; Lee, D.H. & Lin, P.S. (2003). Simplified cone penetration test-based method for evaluating liquefaction resistance of soils. Journal of Geotechnical and Geoenvironmental Engineering, 129(1):66-80.

Kohestani, V.R.; Hassanlourad, M. & Ardakani, A. (2015). Evaluation of liquefaction potential based on CPT data using random forest. Natural Hazards, 79(2):1079-1089.

Kumar, J.K.; Konno, M. & Yasuda, N. (2000). Subsurface soil-geology interpolation using fuzzy neural network. Journal of Geotechnical and Geoenvironmental Engineering, 126(7):632-639.

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.

Liao, T. & Mayne, P.W. (2007). Stratigraphic delineation by three-dimensional clustering of piezocone data. Georisk, 1(2):102-119.

Livingston, G.; Piantedosi, M.; Kurup, P. & Sitharam, T.G. (2008). Using decision-tree learning to assess liquefaction potential from CPT and V_s. Proc. of the 4th Geotechnical Earthquake Engineering and Soil Dynamics Congress, Sacramento, California, v. 1, pp. 1-10.

Lunne, T.; Robertson, P.K. & Powell, J.J. (2002). Cone Penetration Testing in Geotechnical Practice. Taylor & Francis Group, London and New York.

Mayne, P.W.; Peuchen, J. & Bouwmeester, D. (2010). Soil Unit Weight Estimated from CPTu in Offshore Soils. Gourvenec & White (eds.) Front Offshore Geotech, Taylor & Francis Group, London, pp. 371-376.

Mayne, P.W. (2014). Interpretation of geotechnical parameters from seismic piezocone tests. Proc. of 3rd International Symposium on Cone Penetration Testing, CPT14, Las Vegas, Nevada, v. 1, pp. 1-27.

Nemenyi, P.B. (1963). Distribution-Free Multiple Comparisons. Ph.D. Thesis, Princeton University.

Ramsey, N. (2002). A calibrated model for the interpretation of cone penetration tests (CPTs) in North Sea quaternary soils. Proc. of Offshore Site Investigation and Geotechnics Diversity and Sustainability, Society of Underwater Technology, London, v. 1, pp. 1-16.

Rao, A.; Janhavi, U.; Abhishek, G.N.; Manjunatha & Beham, R.A. (2016). Machine learning in soil classification and crop Detection. International Journal for Scientific Research & Development, 4(1):792-794.

Reale, C.; Gavin, K.; Libric, L. & Juric-Kacunic, D. (2018). Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks. Advanced Engineering Informatics, 36:207-215.

Robertson, P.K.; Campanella, R.G.; Gillespie, D. & Greig, J. (1986). Use of Piezometer Cone Data. Clemence, S.P. (ed.) Use of In Situ Tests in Geotechnical Engineering. American Society of Civil Engineers, New York, pp. 1263-1280.

Robertson, P.K. (1990). Soil classification using the cone penetration test. Canadian Geotechnical Journal, 27(1):151-158.

Robertson, P.K. (1991). Soil classification using the cone penetration test: Reply. Canadian Geotechnical Journal, 28(1):176-178.

Robertson, P.K. (2009). Interpretation of cone penetration tests - a unified approach. Canadian Geotechnical Journal, 46(11):1337-1355.

Robertson, P.K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system - an update. Canadian Geotechnical Journal, 53(12):1910-1927.

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.

Schaap, M.G.; Leij, F.J. & Van Genuchten, M.T. (1998). Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal, 62(4):847-855.

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.

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.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2):111-133.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 1(6):80-83.

Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3):408-421.

Zhang, G.; Robertson, P.K. & Brachman, R.W. (2002). Estimating liquefaction-induced ground settlements from CPT for level ground. Canadian Geotechnical Journal, 39(5):1168-1180.

List of Symbols

B_q: normalized excess pore pressure
CG: class of geology
d: feature space dimensionality
dist: distance between points
F_r: normalized friction ratio
f_s: lateral friction

I_c: classification index
IQ: interquartile range
k: number of nearest neighbors
n: exponent of σ_{v0}'
p_a: reference pressure
Q_1: first quartile
Q_3: third quartile
q_c: cone resistance
q_t: total cone resistance
Q_{t1}: normalized cone resistance
Q_m: updated normalized cone resistance
R_f: friction ratio
SD: standard deviation
S_t: sensitivity
u_0: equilibrium pore pressure
u_2: pore pressure measured behind the cone tip
U_2: updated normalized excess pore pressure
w: Gaussian weighting
x_i, x_j: points representing objects
z: depth
γ: soil unit weigh
σ_{v0}: total overburden pressure
σ_{v0}': effective overburden pressure