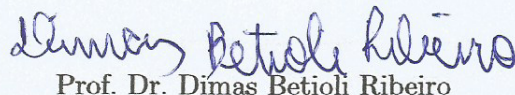


Dissertation presented to the Instituto Tecnológico de Aeronáutica, in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program of Aeronautical Infrastructure Engineering, Field of Airport Infrastructure.

Lucas Orbolato Carvalho

**APPLICATION OF MACHINE LEARNING
TECHNIQUES FOR SOIL CLASSIFICATION
FROM CPTU**

Dissertation approved in its final version by signatories below:



Prof. Dr. Dimas Betioli Ribeiro

Advisor

Prof. Dr. Pedro Teixeira Lacava

Pro-rector of Graduate Courses

Campo Montenegro
São José dos Campos, SP - Brazil
2018

Cataloging-in Publication Data
Documentation and Information Division

Carvalho, Lucas Orbolato

Application of Machine Learning techniques for soil classification from CPTu / Lucas Orbolato Carvalho.

São José dos Campos, 2018.

101f.

Dissertation of Master of Science – Course of Aeronautical Infrastructure Engineering. Area of Airport Infrastructure – Instituto Tecnológico de Aeronáutica, 2018. Advisor: Prof. Dr. Dimas Betioli Ribeiro.

1. Cone penetration test. 2. Soil classification. 3. Machine learning. 4. K-nearest neighbor. 5. Decision tree. 6. Random forest. 7. Artificial neural network. 8. Support vector machines. 9. Multiple model predictor. 10. K-means. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

CARVALHO, Lucas Orbolato. **Application of Machine Learning techniques for soil classification from CPTu**. 2018. 101f. Dissertation of Master of Science – Instituto Tecnológico de Aeronáutica, São José dos Campos.

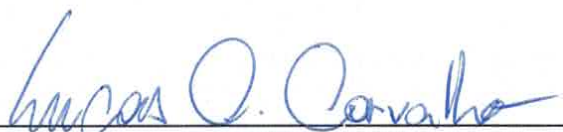
CESSION OF RIGHTS

AUTHOR'S NAME: Lucas Orbolato Carvalho

PUBLICATION TITLE: Application of Machine Learning techniques for soil classification from CPTu.

PUBLICATION KIND/YEAR: Dissertation / 2018

It is granted to Instituto Tecnológico de Aeronáutica permission to reproduce copies of this dissertation and to only loan or to sell copies for academic and scientific purposes. The author reserves other publication rights and no part of this dissertation can be reproduced without the authorization of the author.



Lucas Orbolato Carvalho

Rua Filomeno Antônio Cândido, 231, Parque Higienópolis
19053-605 – Presidente Prudente – SP

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR SOIL CLASSIFICATION FROM CPTU

Lucas Orbolato Carvalho

Thesis Committee Composition:

Prof. Dr. Francisco Alex Correia Monteiro	President	-	ITA
Prof. Dr. Dimas Betioli Ribeiro	Advisor	-	ITA
Prof. Dr. Paulo Scarano Hems	Member	-	ITA
Prof. Dr. Heraldo Luiz Giacheti	Member	-	UNESP
Prof. Dr. José Antônio Schiavon	Substitute	-	ITA
Prof. Dr. Cristina de Hollanda Cavalcanti Tsuha	Substitute	-	USP

To God for the opportunity, to my family for the emotional support and to my advisor for guiding my steps throughout this journey.

Acknowledgments

To Peter K. Robertson, Paul W. Mayne, Company of the Metropolitan of São Paulo, Renato P. da Cunha and Heraldo L. Giacheti 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.

*"If I have seen farther than others,
it is because I stood on the shoulders of giants."*

— SIR ISAAC NEWTON

Resumo

Métodos clássicos de classificação de solos usando dados do ensaio de penetração de cone (CPT) propõem uma abordagem gráfica, exigindo a representação dos dados com apenas duas variáveis a fim de criar representações planas. Buscando superar esta restrição, o objetivo fundamental deste trabalho é desenvolver análises multi-dimensionais usando diferentes algoritmos de Aprendizado de Máquina (AM). Uma metodologia geral e reproduzível é desenvolvida e aplicada em até cinco dimensões, incluindo dados de CPT brutos ou normalizados como atributos contínuos e a idade geológica do solo como um atributo discreto. O conjunto de dados utilizado inclui 179 sondagens de CPT adquiridas de diferentes fontes. O problema preditivo é abordado sob duas perspectivas: influenciada pela granulometria do solo (ISG) ou focada no comportamento do solo (FSB). Procedimentos de estatística descritiva assim como técnicas de AM são aplicadas aos dados para o seu pré-processamento, incluindo a limpeza de dados, o balanceamento e a transformação dos dados. Diferentes técnicas de AM são calibradas, estudadas e comparadas com testes estatísticos e empregando o procedimento *10-fold* de validação cruzada. Estudos iniciais de seleção e análise de importância de atributos são conduzidos com técnicas baseadas em distância. Estes estudos mostraram que o uso de dados brutos de CPT com a profundidade inclusa é suficiente para as técnicas distinguirem entre os tipos de solo com uma acurácia satisfatória, em torno de 90%, e que a profundidade acrescenta informação relevante à classificação. Foi também observado que a inclusão da idade geológica como atributo de entrada discreto contribui com informação relevante. Na segunda parte do estudo, novas técnicas preditivas de AM são testadas e suas particularidades exploradas. Entre todas as técnicas preditivas de AM testadas isoladamente, as simbólicas foram a de melhor desempenho e menor sensibilidade a variações da metodologia. As técnicas são então combinadas para gerar um modelo múltiplo preditivo (MMP), cujo desempenho superou o das demais técnicas segundo os testes estatísticos. Por fim, técnicas de aprendizagem não supervisionada são empregadas para gerar novas classes a partir de atributos não normalizados e as Redes Neurais Artificiais são utilizadas para gerar um modelo matemático analítico simples de classificação preditiva. Os agrupamentos são então testados com cinco sondagens de ensaio de penetração padrão (SPT) pareadas com CPT.

Abstract

Classical methods of soil classification from cone penetration test (CPT) data propose a graphical approach. It requires representing data with just two parameters in order to create plane representations. To overcome this restriction, the fundamental objective of this work is to develop multi-dimensional analyses using different machine learning (ML) algorithms, including supervised and unsupervised techniques. A general methodology for the predictive supervised approach is proposed and applied up to five dimensions, including raw, non-normalized and normalized CPT data as continuous inputs and geological age as a discrete one. The dataset is composed by 179 CPT soundings from several world locations. Two popular chart-based classification methods are reproduced and studied in the predictive supervised approach: one influenced by soil granulometry (ISG) and one focused on soil behavior (FSB). Descriptive statistical procedures as well as ML techniques are used for data preprocessing, including data cleaning, balancing and transformation. Different ML techniques are calibrated, studied and compared with statistical tests and employing the 10-fold cross-validation procedure. Within the first studies, the input features relevance is evaluated with distance-based techniques. These first results have shown that using non-normalized CPT data with depth included is enough for the techniques to properly distinguish between different soil types with a satisfactory accuracy, around 90%, and that depth introduce relevant information for the task of soil classification from CPT data. It was also observed that the inclusion of the geological age as a discrete input can introduce relevant information. In the second part of the study, other ML techniques are tested and their particularities are explored. The symbolic techniques were those with the best performance and lowest sensitivity to methodology variations. The techniques are then combined into a multiple model predictor (MMP), whose performance overcame the one of the other ML techniques according to statistical tests. Finally, unsupervised techniques are employed to generate new classes from non-normalized inputs with two different approaches: a general one and a case-based one. So, an ANN is trained to create a mathematical model that allows the new method reproduction with simple tools as spreadsheets. The clustering results are then compared to the classical methods and to standard penetration test (SPT) data paired with CPT.

List of Figures

FIGURE 2.1 – Excess pore pressure.	32
FIGURE 2.2 – $Q_{tn} \times F_r$ chart from Robertson (1991).	34
FIGURE 2.3 – $Q_{tn} \times B_q$ chart from Robertson (1991).	35
FIGURE 2.4 – $Q_{tn} \times F_r$ chart from Robertson (2016).	36
FIGURE 2.5 – $Q_{tn} \times U_2$ chart from Robertson (2016).	37
FIGURE 3.1 – Full methodology description with all modeling procedures summarized.	38
FIGURE 3.2 – Box-plots of features values for ISG classes.	42
FIGURE 3.3 – Box-plots of features values for FSB classes.	43
FIGURE 3.4 – Main and geological datasets histograms: (a) for ISG classes in the main dataset, (b) for FSB classes in the main dataset, (c) for ISG classes in the geological dataset and (d) for FSB classes in the geological dataset.	45
FIGURE 3.5 – Balancing example: imbalanced dataset on the left and balanced on the right.	46
FIGURE 3.6 – Correlation matrices for all input features.	47
FIGURE 3.7 – Tropical dataset histograms: (a) for ISG classes in the tropical dataset and (b) for FSB classes in the tropical dataset.	48
FIGURE 3.8 – Clustering dataset histograms: (a) for ISG classes in the clustering dataset and (b) for FSB classes in the clustering dataset.	48
FIGURE 3.9 – Histogram for SPT observations in the dataset with CPT paired with SPT.	49
FIGURE 3.10 – Log-scaling effect example in the objects distribution within the $Q_{tn} \times F_r$ space: (a) without log-scaling and (b) with log-scaling. . .	52

FIGURE 3.11 – Summarized methodology for data preprocessing.	53
FIGURE 4.1 – Comparison between DWKNN, DT, RF, ANN, SVM e MMP using the Friedman statistical test.	69
FIGURE 4.2 – Comparison between DWKNN, DT, RF, ANN, SVM and MMP without FSB class 0 by the Friedman statistical test.	70
FIGURE 4.3 – Comparison between DWKNN, DT, RF, ANN, SVM and MMP by the Friedman statistical test.	72
FIGURE A.1 – ANN neuron.	90
FIGURE A.2 – Multilayer neural network.	91
FIGURE B.1 – Soil profiles given by: (a) CPeT-IT v2.0.2.5 and (b) MMP.	95

List of Tables

TABLE 3.1 – Dataset from P. K. Robertson.	39
TABLE 3.2 – Dataset from P. W. Mayne.	39
TABLE 3.3 – Geological dataset.	40
TABLE 4.1 – Comparing Q_{t1} to Q_{tn} (KNN predictive results for ISG).	56
TABLE 4.2 – Best predictive results for ISG.	56
TABLE 4.3 – Best predictive results for FSB.	57
TABLE 4.4 – Comparing input features f_s and R_f	58
TABLE 4.5 – Combinations with and without input feature CG for FSB.	58
TABLE 4.6 – Results obtained with DWKNN.	60
TABLE 4.7 – Results obtained with DT.	61
TABLE 4.8 – Results obtained with RF.	63
TABLE 4.9 – Results obtained with ANN without log-scaling.	64
TABLE 4.10 – Results obtained with ANN without log-scaling and without the FSB class 0.	64
TABLE 4.11 – Results obtained with ANN using log-scaling.	65
TABLE 4.12 – Results obtained with ANN using log-scaling and without the FSB class 0.	65
TABLE 4.13 – Results obtained with SVM without log-scaling.	66
TABLE 4.14 – Results obtained with SVM without log-scaling and without the FSB class 0.	66
TABLE 4.15 – Results obtained with SVM using log-scaling.	67
TABLE 4.16 – Results obtained with SVM using log-scaling without the FSB class 0.	67
TABLE 4.17 – Results obtained with MMP.	68

TABLE 4.18 – Results obtained with MMP without the FSB class 0.	70
TABLE 4.19 – Results for all techniques.	71
TABLE 4.20 – Results using z , q_c and f_s as input features.	71
TABLE 4.21 – Clusters centers.	73
TABLE 4.22 – Comparison between the k-means clustering and the ISG classes. . .	73
TABLE 4.23 – Comparison between the k-means clustering and the FSB classes. . .	73
TABLE 4.24 – Comparison between ISG classes and the SPT observations.	74
TABLE 4.25 – Comparison between FSB classes and the SPT observations.	74
TABLE 4.26 – Comparison between the k-means clustering and the SPT observations.	75
TABLE 4.27 – Contribution of each soil type for the BC.	75
TABLE 4.28 – Case-based clusters centers.	80
TABLE 4.29 – Comparison between the case-based x-means clustering and the ISG classes.	81
TABLE 4.30 – Comparison between the case-based x-means clustering and the FSB classes.	81
TABLE 4.31 – Comparison between the case-based x-means clustering and the SPT observations.	81
TABLE 4.32 – Contribution (%) of each soil granulometrical type for each behavior.	82

List of Abbreviations and Acronyms

ANN	Artificial neural network
CG	Geological classification
CPT	Cone penetration test (used to refer both to CPT and CPTu)
DT	Decision tree
DWKNN	Distance weighted k -nearest neighbors
ENN	Edit nearest neighbor
FSB	Classification method focused on soil behavior only
ISG	Classification method influenced by soil granulometry
KKM	Kernel k-means
KXM	Kernel x-means
MA	Mean accuracy
ML	Machine learning
MLP	Multi-layer perceptron
MMP	Multiple model predictor
RF	Random forest
SD	Standard deviation
SMOTE	Synthetic minority over-sampling technique
SPT	Standard penetration test
SVM	Support vector machines
USCS	Unified Soil Classification System

List of Symbols

z	Depth
q_c	Raw or uncorrected cone resistance
q_t	Total or corrected cone resistance
Q_{t1}	Normalized cone resistance of Robertson (1991)
Q_{tn}	Normalized cone resistance of Robertson (2016)
f_s	Lateral friction
R_f	Friction ratio
F_r	Normalized friction ratio
u_0	Equilibrium pore pressure
u_2	Pore pressured measured behind the cone tip
B_q	Normalized excess pore pressure of Robertson (1991)
U_2	Normalized excess pore pressure of Robertson (2016)
σ'_{v0}	Overburden effective pressure
σ_{v0}	Overburden total pressure

Contents

1	INTRODUCTION	17
1.1	Literature review	19
1.2	Motivation	28
1.3	Aims and contributions	29
1.4	Dissertation general structure	29
2	CLASSIFICATION METHODS USED IN CPET-IT	31
2.1	Influenced by soil granulometry (ISG)	31
2.2	Focused on soil behavior (FSB)	33
3	MODELING PROCEDURES	38
3.1	Data description	38
3.1.1	Main and geological datasets	38
3.1.2	Tropical dataset	45
3.1.3	Clustering approach data description	46
3.2	Feature analysis with distance-based algorithms	47
3.3	General analysis with supervised ML techniques	50
3.4	Preprocessing procedure for supervised techniques	51
3.5	Clustering approach methodology	53
3.5.1	Clustering analysis	53
3.5.2	ANN modeling	54
4	MODELING RESULTS AND DISCUSSIONS	55
4.1	Preliminary studies	55

4.1.1	Distance-based input feature analysis	55
4.1.2	Complementary studies	58
4.2	General supervised analyses	59
4.2.1	DWKNN	59
4.2.2	DT	61
4.2.3	RF	62
4.2.4	ANN	63
4.2.5	SVM	66
4.2.6	MMP evaluation	68
4.2.7	MMP application and validation	70
4.2.8	Classification without pore pressure information	71
4.3	Clustering approaches	72
4.3.1	General clustering approach	72
4.3.2	Case-based clustering approach	80
5	CONCLUSIONS AND FURTHER WORK	84
	ANNEX A – ML TECHNIQUES	87
A.1	KNN and DWKNN	87
A.2	DT and RF	89
A.3	ANN	90
A.4	SVM	91
A.5	Kernel k-means	93
	ANNEX B – EXAMPLE OF SUPERVISED APPLICATION	95
	ANNEX C – MMP APPLICATION CODE	96
	BIBLIOGRAPHY	97

1 Introduction

Different strategies may be employed to investigate soil in situ stratigraphic profile. The most popular approach is to perform the SPT (SALGADO, 2006; SCHNAID; ODEBRECHT, 2012), which allows obtaining samples and evaluating soil resistance with relatively low cost. In this experiment, a sampler is nailed into the soil layers to a predefined distance by repeated blows of a standard hammer. The number of blows provides the main parameter measured in this experiment. Two main drawbacks of the SPT when compared to other methods such as the CPT are low precision and discontinuous logging, among other disadvantages (SALGADO, 2006; SCHNAID; ODEBRECHT, 2012). In the CPT, a metallic cylinder headed by a cone is forced into the soil at a constant velocity, usually 2 cm/s, measuring pressure at the cone, friction at the shaft and pore pressure. In the CPT no sample is recovered, which may be considered a major disadvantage compared to the SPT. In practice, one may complement each other and both can be performed for a better understanding of the soil stratigraphic profile in situ.

To compensate CPT lack of sampling, many authors tried developing methodologies to identify soil layers from CPT data. Most methodologies are based on charts in which no specific soil type is obtained, but only a soil behavior profile (CAI *et al.*, 2011). In this context, Robertson charts became popular. They use normalized CPT measurements, namely cone resistance, sleeve resistance and pore pressure, to separate soil behavior into twelve classes. Improvements were proposed for this method, providing recommendations for cone designs (BOGGESE; ROBERTSON, 2011), evaluating the sensitivity to variations of sleeve resistance (ROBERTSON, 2009), and introducing microstructure identification (ROBERTSON, 2016). In spite of these advances, the basic idea remains to produce 2-dimensional figures divided into regions representing soil behaviors, with no major improvement towards this fundament.

Plenty of work in the literature associate CPT values to advanced computational techniques in order to solve different geotechnical problems. One approach that has grown in the last years is the use of ML techniques (BHATTACHARYA; SOLOMATINE, 2006) (see Annex A). In simple words, ML involves using previous recorded experience or some predefined criteria to solve a problem or to improve performance when solving it. Thus, ML algorithms induce a hypothesis capable of solving a problem from a previous

dataset representing problem instances or they can also solve a problem following some decision criteria previously defined to guide their choices. There are several applications for ML within geotechnics, as soil liquefaction classification (BAZIAR; NILIPOUR, 2003), pile capacity prediction (ALKROOSH; NIKRAZ, 2012), soil behavior classification (AREL, 2012), among others. Considering the example of soil behavior classification, for instance, each data point represents a soil test with known type which is described by a tuple of attributes. These are input features, while the output feature is the soil type to be predicted.

An important requirement for ML algorithms is the ability of dealing with imperfect data, like noise, inconsistencies, incompleteness and redundancies. If these problems are not severe, their influence should be minimized by internal mechanisms used by the algorithms in the hypothesis induction. For severe problems on data, preprocessing techniques may also be required (FACELLI *et al.*, 2011).

Learning tasks may be divided into two major groups: predictive and descriptive. In predictive learning, usually the supervised learning paradigm is followed. Within it, each data point has a known output. The goal is to learn a mapping function from the data features to the desired outputs. This function, also named model or generalization hypothesis, can then be used to predict the labels of new instances. A more formal definition would be, given a dataset of pairs of observations $\mathbf{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i)), i = 1, \dots, n\}$, where f is an unknown function mapping the input features, represented by a vector $\mathbf{x}_i$, into a desired output. A ML algorithm learns an approximation $\hat{f}$ capable of predicting the labels for new values of $\mathbf{x}_i$. For classification outputs may be represented as $y_i = f(\mathbf{x}_i) \in \{c_1, \dots, c_n\}$. In this kind of problem, a measure usually adopted to evaluate the technique performance is the accuracy or variations of it, which is a classification success rate, meaning the number of right predictions divided by the total number of tested objects.

In unsupervised learning, there is not known outputs and the aim is to describe or explore the dataset for unknown regularities. An example of unsupervised descriptive learning procedure is the clustering, which consists on grouping the objects based on recognized similarities between them. This work employ both learning paradigms, which are used for solving classification problems using previously known labels given by some classification methods or for producing new classification criteria.

In this context, a general and reproducible methodology for application of ML techniques is then proposed. It is also made a rigorous analysis of input feature relevance with distance-based and symbolic algorithms. It enables identifying the more important features for soil classification from CPT data and possible replacements to simplify this task or to produce new classification methods. Initially, it is included only depth, cone resistance, lateral friction and pore pressure. These three last features are taken

non-normalized and normalized (ROBERTSON, 1990; ROBERTSON, 2016). In the end, the geological age is included as a discrete input called CG and its contribution is assessed.

Furthermore, several predictive supervised ML techniques are trained with the proposed methodology to classify soil from CPT data, aiming to replicate classification systems accessed with a student version of the CPeT-IT v2.0.2.5 software (IOANNIDES; ROBERTSON, 2016). First, CPeT-IT is used to classify all examples from two datasets, one composed by 111 CPT soundings taken in different countries and the other composed by 62 CPT soundings taken in the city of São Paulo, Brazil. Next, these examples are used to train the following ML techniques within a 10-fold cross validation procedure: DWKNN, boosted DT, RF, ANN, SVM and MMP, which is a combination of the previous. Training is made not only with the original inputs required by CPeT-IT, but also with different input feature combinations in order to investigate and discuss geotechnical aspects of soil classification. One of the analyzed aspects is the possibility of replacing the normalized parameters by the non-normalized or the raw CPT or CPTu parameters. It is also investigated the inclusion of the geological age as an input feature and the relevance of different input features with the distance-based approach and symbolic ML techniques.

Regarding the unsupervised ML approach, another 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. A statistical investigation is performed to define the number of clusters, concluding that it would be best generating 9 soil classes. These classes are compared to SPT samples and to Robertson classification methods (ROBERTSON, 1991; ROBERTSON, 2016) obtained with a student version of the CPeT-IT v2.0.2.5 software (IOANNIDES; ROBERTSON, 2016). An alternative case-based approach is also presented using the kernel x-means algorithm, which was found to be effective in previous work (ROGERS *et al.*, 2017). It is shown that both proposed soil classification methods can be accurately 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 and using alternative CPT equipment to classify soil. These ANN models are considered the main contribution of this work.

1.1 Literature review

The more usual classification standard applied for general purposes is the USCS, which is based on granulometry and plasticity. Nevertheless, this method have its restraints, such as requiring the extraction of soil samples to be subjected to laboratory tests. The

drawbacks of this approach are the unfeasibility of extracting undisturbed samples, the discontinuous extraction, the time delay required to get the results as well as the complex and not automatized procedures. One of the applications of the CPT is to determine the soil stratigraphic profile. It provides a continuous and reliable measurement of parameters that can be used as inputs to classify soil layers based on some classification system. Therefore, it can be more desirable to relate soil classification directly to CPT parameters. However, the CPT parameters are more related to the soil behavior, what is more of interest of engineers, than to the soil composition, as the USCS parameters. Thus, there should be one different classification system which reflects this change of principle from compositional to behavioral.

One of the first classification methods based on CPT measurements was a chart using the uncorrected cone resistance by lateral friction (BEGEMANN, 1965). The chart contained curves allowing to find the percentage of fines and the soil type, but it was still based on granulometry. Further work stated that it should be considered a soil behavior classification system and that the friction ratio should substitute lateral friction to give better results (DOUGLAS; OLSEN, 1981). Next, some pore pressure-based charts were proposed. Following these and others progresses, Robertson *et al.* (1986) proposed an objective classification method that had become a landmark in soil classification. This method presents two charts, one using the total cone resistance by friction ratio and the other using the total cone resistance by normalized excess pore pressure, which takes into account the overload influence. This method establishes 12 soil behavior classes whose descriptions are close to the ones of the USCS. The advantage of this method is its simplicity due to the usage of almost raw parameters. In spite of that, it was noticed that the same soil could be classified differently depending on depth or overburden pressure.

Thus, normalized cone resistance, friction ratio and excess pore pressure as well as two new charts with these updated parameters were proposed by Robertson (1990) in order to account for depth and overburden. The soil classes were reduced from 12 to 9. Besides, the author pointed the risks of pore pressure measurements and the consequently less reliable outputs of pore pressure-based charts. In addition, Jefferies and Davies (1991) have shown that the normalized cone resistance by excess pore pressure chart of Robertson (1990) was unsatisfactory for offshore soils classification. The explanation given was the effect of back pressure and cavitation of the filter element in deep water offshore soils. To avoid this problem, it was suggested to use only the chart with normalized cone resistance by normalized friction ratio, incorporating within the cone resistance normalization the influence of the normalized excess pore pressure. So Robertson (1991) brings an update to his cone resistance by excess pore pressure chart but maintains the parameters and the two charts initially proposed, explaining that the faults observed were caused by the dilative behavior of highly overconsolidated clays such as those found in deep water soils.

Another contribution was the determination of a classification index to estimate class separation curves of the normalized cone resistance by friction ratio chart of Robertson (1990) by approximating them with circles (JEFFERIES; DAVIES, 1993). It was later improved by introducing a new cone resistance normalization with an exponent of the overburden pressure which is a function of that index (ROBERTSON; WRIDE, 1998).

In a similar way of Jefferies and Davies (1991), 439 soundings from the clay quaternary deposit of the North Sea were tested with the cone resistance by excess pore pressure chart of Robertson (1991), indicating that it could not classify this kind of soil properly (RAMSEY, 2002). Likewise, Schneider *et al.* (2008) confirmed that the chart was not accurate for some types of soils and identified the source of the misclassification. It was noticed that, with increasing normalized cone resistance and decreasing normalized excess pore pressure, the chart output indicates a granular soil but it could not distinguish it from an overconsolidated clay, which also has low values of normalized excess pore pressure and high values of normalized cone resistance. This confusion was explained by the existence of a correlation between these variables. Schneider *et al.* (2008) stated that, even with increasing normalized cone resistance and excess pore pressure, for increasing consolidation ratio, it can happen that the normalized excess pore pressure decreases and misclassification occurs. Based on these observations as well as on some theoretical and empirical correlations, Schneider *et al.* (2008) proposed a new chart with normalized cone resistance by normalized excess pore pressure, in which a new normalization for the excess pore pressure is presented. Later, Schneider *et al.* (2012) also proposed a normalized cone resistance by normalized friction ratio chart, suggesting modifications to the chart of Robertson (1991).

Schneider *et al.* (2012) suggested replacing the circular boundaries by hyperbolic ones with a new index. However, while Robertson (1991) uses nine classes, Schneider *et al.* (2012) uses only five. Furthermore, the classes designations were still attached to granulometry in some way in Robertson (1991), while in Schneider *et al.* (2012) they were more related to the soil behavior. Nonetheless, it still concerns conventional soil behaviors explained by the soil macrostructure, i.e., grain size distribution, but does not tell much about some other influent factors on soil behavior such as cementation, soil aging or overconsolidation. It is true that there were some interpretations given by the mentioned methods about these processes, but none of them discussed or defined them objectively.

Robertson (2016) established that aged or cemented soils could be called structured soils, meaning that they have a microstructure. Soils without microstructure were called ideals. Robertson (2016) stated that all graphical methods were valid only for ideal soils and not reliable for the structured ones. Moreover, it was proposed a class division that would be purely behavioral. Inspired by Schneider *et al.* (2008) and Schneider *et al.* (2012), the author suggested to identify soils as sands, clays or transitional, subdividing

each of these types into contractive or dilative. Furthermore, a separated class was created for contractive sensitive clays. The normalized cone resistance by normalized friction ratio chart and the normalized cone resistance by normalized excess pore pressure chart were updated from the charts of Schneider *et al.* (2008) and Schneider *et al.* (2012). The cone resistance normalization of Robertson (1990) was replaced by that of Robertson (2016) and some changes were made to include the proposed classes considering previous work (ROBERTSON, 2009).

One more definition given in Robertson (2016) to make the method consistent was a procedure to determine if the soil is structured or ideal. This procedure is based on a factor named small-strain normalized rigidity index, which is a function of the small-strain shear modulus and of the normalized cone resistance. When this index is greater than or equal to 330, the soil is considered structured. The small-strain shear modulus can be calculated from seismic CPT data, which provides the shear wave speed at different depth levels. The steps to calculate the soil unit weight for the overburden pressure, which is in turn needed to calculate the normalized cone resistances, are not explicit in the work of Robertson (2016). Nevertheless, there are some known work in the literature which provide ways to calculate an approximation for the soil unit weight (MAYNE *et al.*, 2010; MAYNE, 2014). The relationship between the small-strain shear modulus, that measures the soil rigidity, and the cone resistance, that measures the soil resistance, gives an important clue about the soil behavior predictability. When soil has high rigidity with low resistance, it is a non conventional soil, affected by cementation or aging processes.

In parallel to these graphical approaches, soil classification using statistics and ML techniques were developed. One of the first applied an ANN model with one hidden layer trained by the back-propagation algorithm to classify soil according to fines content and plasticity parameters (CAL, 1995). In spite of some methodological issues, results were satisfactory and encouraged applying ANN or other machine learning techniques to the soil classification problem. In the same decade, a study that analyzes the task of soil classification based on CPT parameters by the fuzzy logic angle was presented (ZHANG; TUMAY, 1999). This study was founded on the idea that the logic is not binary or discrete but, two or more events can be true at the same time with different probabilities. It means that all soil classes can appear at the same time with some probability associated with each one of them as an output for each input values set. In order to address this issue, in that work it was suggested the division of soils among highly probable mixed soil (HPM), highly probable clayey soil (HPC) and highly probable sandy soil (HPS).

Later, a new and descriptive approach of soil classification was presented (HEGAZY; MAYNE, 2002), exploring the method with a hierarchical clustering method. By testing different techniques, it was determined that the single-link algorithm was the one with best performance. This descriptive algorithm follows the same principle of the predictive

1-NN (see Section A.1). Besides, it was found that the ideal number of classes to get better match with real soil types must be under eight. This conclusion was made by varying the soil classes from 2 to 100. Nonetheless, it was used only normalized cone resistance and excess pore pressure as inputs, claiming that lateral friction measurements were not reliable. Referring to this work, a new study was conducted applying a similar hierarchical clustering technique to the problem but using the mean distance between clusters and considering uncorrected cone resistance and raw lateral friction as inputs. Results were compared with those obtained by the fuzzy method of Zhang and Tumay (1999), showing that they were complementary and compatible with the observations. A number of 25 soundings from Gioia Tauro, Italy, were used in this study.

An important landmark on the application of machine learning techniques to soil classification based on CPT data was the work of Bhattacharya and Solomatine (2006). A full clustering and predictive classification method so-called Constraint Clustering and Classification (CONCC) was proposed. The clustering technique employed, named segmentation, comprises a fuzzy logic principle. First, the dataset is partitioned considering the constraint of contiguity between objects of the same partition. From each segment is taken a subset to represent it, excluding objects near the original border, which contracts it. This procedure is based on the idea that there can be noise near the segment's boundaries and that these intersections have to be removed to create representative class regions. Nonetheless, there were chosen only the uncorrected cone resistance and the friction ratio to make this analysis so that even with this cleaning there was still spatial overlap of segments, making it more difficult to classify objects. This problem was solved increasing the dimensionality in a procedure similar to the one used in SVM, so that the objects could be easily distinguished and the overlap of classes removed. For this purpose, it was applied the boundary energy method, which introduces a new artificial input feature. After results pointed a more disjoint configuration, for each class is given a label by a specialist and finally the data is subjected to different machine learning techniques to be classified. It was applied ANN, DT and SVM. Growing the number of segments, the accuracy of the techniques was harmed, getting under 90% for all techniques for seven classes. The author claimed that the results could be worse for a more diverse dataset, which was evidenced by the better results for sandy and clayey soils and meaningfully lower for other types. This study employed only seven CPT soundings, all from the same site.

The problem of soil classification was also treated under the granulometrical point of view (KURUP; GRIFFIN, 2006). A General Regression Neural Network (GRNN) was applied in order to predict the clay, silt and sand contents from CPT measurements as inputs. The results suggested that the methods of Robertson (1990) and of Zhang and Tumay (1999) were satisfactory, close to 90%, for describing soil granulometry beyond its

behavior. One limitation of this work is that, it used only 12 paired SPT-CPT soundings from only four areas, harming the generality of the model. A progress of Kurup and Griffin (2006) over previous work was the implicit inclusion of depth as an input feature. It was considered four inputs: raw uncorrected cone resistance, friction ratio, total overburden pressure and equilibrium pore pressure. These two last ones are highly influenced by depth and could be practically redundant for ML techniques, but no analysis was made in this way.

Motivated by the issues presented above, an upgrade to the single-link descriptive approach previously presented (HEGAZY; MAYNE, 2002) was proposed, now applying hierarchical clustering techniques with more input features and considering a different cluster distance metric (LIAO; MAYNE, 2007). It was considered as inputs the normalized cone resistance, friction ratio and excess pore pressure (ROBERTSON, 1990) and it was used the centroid of the clusters to compute distances, whereas previously it was considered the lowest distance between clusters to join them (HEGAZY; MAYNE, 2002). Likewise, a number of classes around eight was considered (HEGAZY; MAYNE, 2002). This work established a methodology for single-link clustering analysis with normalized parameters of CPT.

Later, there was a study proposing a fuzzy approach of the soil classification based on CPT focused on soil composition, i.e., granulometry and plasticity, like USCS guidelines (CETIN; OZAN, 2009). The output parameters are the fines content, the plasticity index and the liquid limit. The inputs are a normalized net cone resistance that suits the probabilistic nature of the method and the normalized friction ratio. In the same year, another study which extends the discussions about descriptive classification with clustering techniques was presented (DAS; BASUDHAR, 2009). It tested different clustering algorithms and the results were compared with those obtained with the method of Robertson *et al.* (1986). The clustering techniques tested were the self-organizing maps (SOM), the fuzzy clustering *c*-means, the *k*-means and a hierarchical method. It was concluded that the hierarchical method was worse than the others and that the *k*-means and the Robertson's charts were satisfactorily accurate. Nevertheless, the techniques were applied to each sounding individually so that there was not a variety of classes to properly compare them.

In a more practical computational application and following the supervised predictive approach, Arel (2012) applied to the soil classification from CPT data problem feed-forward multi-layer Perceptron ANN trained with the back-propagation algorithm. It was taken as inputs the normalized cone resistance and the normalized friction ratio (ROBERTSON, 1990) and as outputs the classes of Robertson (1990). The objective was to determine the soil profile of Adapazari, Turkey, based on 117 CPT soundings. The data as well as the space were discretized so that the whole area could be described. The result was an accuracy around 96% in comparison with Robertson (1990). A contribution

presented by this work was an objective methodology for application of ANNs to classify real soils.

A recent study describes the discussions about soil types division bringing different clustering techniques and comparing them to Robertson's charts (ROGIERS *et al.*, 2017). The clustering algorithms tested were the x -means and the model-based clustering. As inputs, it was considered different arrangements of classification index, normalized cone resistance, normalized friction ratio and depth. The tests were made for an aquifer soil and the results showed more compatibility between the classes identified by the clustering techniques, mainly by the model-based clustering, than those proposed by Robertson in comparison to soil samples. Based on these observations, it was proposed a site specific classification approach with clustering techniques instead of applying predefined labels. It was a similar approach to the one presented by Facciorusso and Uzielli (2004), but there it was found that the charts of Robertson *et al.* (1986) could give satisfactory outcomes. What could be ascertained is that the methods of Robertson cannot identify all types of soils, which is an empirical evidence of the allegations of Robertson (2016). Nonetheless, to state that clustering techniques would give better results than the methods of Robertson would require more tests.

It is seen that many authors can be found in literature applying ML techniques to different geotechnical problems and most of them use ANN to predict soil characteristics (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 DT (LIVINGSTON *et al.*, 2008), RF (KOEHESTANI *et al.*, 2015) and SVM (GOH; GOH, 2007). Most studies dedicated to soil classification from CPT data aim to seek for new soil classes using data clustering (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017). Another possible approach, which is relatively unexplored in the literature, is using ML techniques to replicate predefined soil classification systems, like classical soil classification methods based on charts (AREL, 2012). Most work adopting this approach use only ANN (KURUP; GRIFFIN, 2006; AREL, 2012; REALE *et al.*, 2018) and, when more ML techniques are used, applications are restricted to small CPT datasets, with all soundings taken at the same location (BHATTACHARYA; SOLOMATINE, 2006). These limitations preclude taking advantage of the potentialities of the ML techniques to investigate and discuss geotechnical aspects of soil classification. Enabling this type of discussion is one of this work objectives.

As one can conclude, most work apply the inductive unsupervised ML paradigm to define new class separations or to evaluate the classical chart-based methods (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017). Nevertheless, although ML algorithms became available decades ago (COVER; HART, 1967; DUDANI, 1976; BREIMAN

et al., 1984; RUMELHART *et al.*, 1986; BREIMAN, 2001), there are still few studies using these tools for soil classification (AREL, 2012; ROGIERS *et al.*, 2017). Such techniques are more commonly applied to the problem of liquefaction (GOH, 1996; JUANG; CHEN, 1999; JUANG *et al.*, 2003; HANNA *et al.*, 2007; LIVINGSTON *et al.*, 2008; GANDOMI *et al.*, 2013; KOHESTANI *et al.*, 2015), although even these applications can be considered limited because they are usually attained to only two or three input features, do not present any analysis about input feature importance and use restricted datasets, commonly based on soundings collected from the same region. Despite these limitations, studies suggest that class separations proposed in popular chart-based methods can be enough for the soil classification problem (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007; SHAHRI *et al.*, 2015).

Studies that applied clustering techniques to the soil classification problem based on CPT data concluded in general that the classes divisions proposed by Robertson *et al.* (1986), Robertson (1990), Robertson (1991) produce similar and good results (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009). The Robertson's charts are also popular and widely explored in literature and in practice (TUMAY *et al.*, 2008; CAI *et al.*, 2011; SHAHRI *et al.*, 2015; GANJU *et al.*, 2017). The chart of Robertson *et al.* (1986) can be sometimes even more attractive to practical purposes because of its simplicity on using raw parameters. Nevertheless, with the normalizations of Robertson (1990), the depth could be taken indirectly into account inside overburden pressure, softening the shortcoming of not including it as an input feature. There was also a change in Robertson (1991) after Jefferies and Davies (1991) comments in order to cover dilative soils in the normalized cone resistance by normalized excess pore pressure chart. These are the explanations why one of the output sets considered as reference in the present work is the one of Robertson (1991).

Until Robertson (2016), there was not an approach that analyzed the intrinsic constraints of soil classification. Studies usually adopted a less diversified database, even commonly restricted to some specific area (ZHANG; TUMAY, 1999; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; AREL, 2012; ROGIERS *et al.*, 2017). This approach tends to limit the generality of the constructed model or of the obtained conclusions. Furthermore, the classification methods still adopted a granulometrical designation, although they were in general called behavioral. This nomenclature harms interpreting soil behavior and one of the first attempts to change this trend was proposed by Zhang and Tumay (1999), which defined only three soil classes. Likewise, Schneider *et al.* (2008) and Schneider *et al.* (2012) also divided soil types into only sand, clay, with some subdivisions, and transitional, simplifying results interpretation. Detaching the classification method definitely from composition and taking into account the intrinsic limitations of aging and

cementation, Robertson (2016) proposed new classes definitions, input parameters and a way to identify if the output is reliable by determining if the soil is structured or ideal. There is not any published work found that analyzed this method with machine learning or statistical techniques. These are the reasons why this method is also being studied.

For the feature selection and the assessment of the methods of Robertson (1991) and Robertson (2016), predictive machine learning methods based on distance were one of the choices. This kind of approach is found in literature only for unsupervised and descriptive classification problems with clustering techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017). It was not found applications of supervised and predictive distance algorithms. Studies that explored the unsupervised approach had the mainly intent of finding new class division, without proposing new theoretical interpretations, or to test if the Robertson's classes are suited. Nonetheless, results suggest that the distance approach is appropriate (HEGAZY; MAYNE, 2002; LIAO; MAYNE, 2007). Furthermore, it was found that the nearest neighbor clustering algorithm (single-link) is feasible for this type of problem. Among the advantages of this kind of technique, one can identify its simplicity, which makes it easily reproducible, and its interpretable results, although no computational model is constructed. Nonetheless, this absence of a model does not restrict the interpretation of the method, because classification can be interpreted by abstracting a space distribution of the objects. The classical graphical approach is based on dividing space into classes regions and the distance-based methods follow the same principle, with the advantage of allowing a dimensionality expansion. The more separated are the objects, represented as points, the more accurate tend to be the classification method based on spatial distribution. Thus, for the assessment and comparison between the reference methods, distance-based methods are suited.

With respect to the input features, it was already highlighted the relevance of including the depth among them to construct a model closer to reality (ROGIERS *et al.*, 2017). Besides, it is known that soil age is important to distinguish between soil classes because the same soil type with different ages can have unequal resistance (ROBERTSON, 2016). None of the above mentioned studies applied a procedure for feature selection or feature importance assessment or even tested different input features as the geological age. The input features are generally arbitrarily chosen and it is also common the inclusion of only two, even in the studies that employ computational techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; DAS; BASUDHAR, 2009). Only a few consider the three basic CPT parameters, cone resistance, lateral friction and pore pressure (LIAO; MAYNE, 2007). It is known that including few input features affects the class separation capacity, because in a space with lower dimensionality the overlaps between classes regions are more frequent. This limitation was already noticed

in the literature (BHATTACHARYA; SOLOMATINE, 2006).

One more shortcoming found in literature was the lack of a clear methodology using supervised ML techniques to be applied to the soil classification based on CPT measurements. In general, the studies found in literature also explores a few number of supervised ML techniques, with superficial comparisons and investigations about their possible improvements. Moreover, even most work try finding new soil classes with unsupervised ML techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017), most of these studies explore only hierarchical clustering techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; LIAO; MAYNE, 2007) and a reduced number of input features (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMATINE, 2006; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017). Nevertheless, a recent study stated that the x-means algorithm can lead to good results and that depth can be a relevant information (ROGIERS *et al.*, 2017). In spite of these conclusions, to the best knowledge of the author, 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.

1.2 Motivation

Based on the literature review, the following gaps can be found:

- Making in situ experiments in order to complement the existing datasets;
- Making deeper data extraction and gathering so that the method accuracy can be improved and more precise evaluations can be done;
- Analyzing the inclusion and the variation of different input and output features;
- Investigating and comparing soil behavioral approaches without the bidimensional dimensionality restriction;
- Establishing a general and simple methodology that can be reproducible;
- Using diverse datasets so that the results can be generalized;
- Evaluate and compare different existing classification methods, mainly the more recent, by using computational techniques;
- Producing a computational methodology more focused on geotechnics than those

found in literature, evaluating with more theoretical discernment the data, the results and eventual obstacles or improvements that allow the techniques and their parameters a better fitting to the problem;

- Generating classification methods fully based on ML techniques with a simple way to be reproducible with common computational tools and without the need of commercial softwares or difficult implementations.

1.3 Aims and contributions

As a result of the investigations within ML applications, this work achieved the following original aims and contributions:

- Proposing a general methodology for the application of ML techniques for soil classification from CPT data;
- Seeming to be the first to test, apply and compare six ML techniques (DWKNN, DT, RF, ANN, SVM and MMP) in a geotechnical application;
- Testing up to five input features for soil classification with ML including a discontinuous input and using hypothesis tests to compare results;
- Using ML techniques to accurately replicate and compare two chart based classification systems used in CPeT-IT Robertson (1991), Robertson (2016);
- Applying ML techniques for soil classification using a dataset that can be considered representative for the city of São Paulo, Brazil, analyzing the dataset extrapolation;
- Finding an input feature combination that makes possible to obtain the soil classes within a reasonable accuracy using alternative CPT investigations when the pore pressure information is not available;
- Proposing two new easily reproducible classification methods, a general one and a case-based one, using improved ML clustering techniques and ANN to create a mathematical model.

1.4 Dissertation general structure

This Chapter presented a literature review about graphical and computational methods for soil classification from CPT data, the main goals chosen based on this review and some motivations that support them. In Chapter 2, the soil classification methods taken

as reference and studied in this work, ISG and FSB, are briefly explained. Chapter 3 presents a data analysis of the different datasets used through this study and defines the methodology developed both for the supervised and unsupervised approaches, including the R programming software packages used to run the techniques and the statistical tests. In Chapter 4, the results obtained by running this methodology are presented. Finally, in Chapter 5, some conclusions are drawn, highlighting the main contributions of this work and its results, and some possible future investigations are pointed out. In Annex A, the ML techniques used in this work are theoretically introduced, being presented the algorithms main functioning principles and how they are explored. Annex B presents an example of soil profiling using MMP in comparison to CPeT-IT for ISG classes. Finally, in Annex C, the code used to run the MMP employed to obtain the soil profile in Annex B is presented.

2 Classification methods used in CPeT-IT

This Chapter describes the two soil classification methods replicated in this work using ML techniques. Beside the soil classes presented below within each method, for both of them there is an additional one named class 0. When one soil is assigned to class 0, it means that the method was not capable of defining a class to it. Reasons for this uncertainty are given below within each method.

2.1 Influenced by soil granulometry (ISG)

One of the chart-based classification methods replicated in this work is the one proposed by Robertson (1991), which is referred as ISG throughout this text. In this reference, the author intended to include soil behavior within the classification, nonetheless the defined classes refer to granulometrical soil composition. Furthermore, borehole samples were used to compatibilize the following soil classes with real soil types:

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 four basic parameters measured in CPT are depth (z), uncorrected cone resistance (q_c), lateral friction (f_s) and pore pressure in a disturbed state (u_2), usually measured behind the cone tip. In the method proposed by Robertson (1991) these parameters are combined to obtain normalized ones. First, q_c is corrected to discount the water pressure aiding cone penetration, resulting the total cone resistance q_t . Next, the equilibrium pore pressure u_0 is needed to calculate the excess pore pressure $u_2 - u_0$. It can be obtained by drawing a straight line through the u_2 graphic like shown in Figure 2.1.

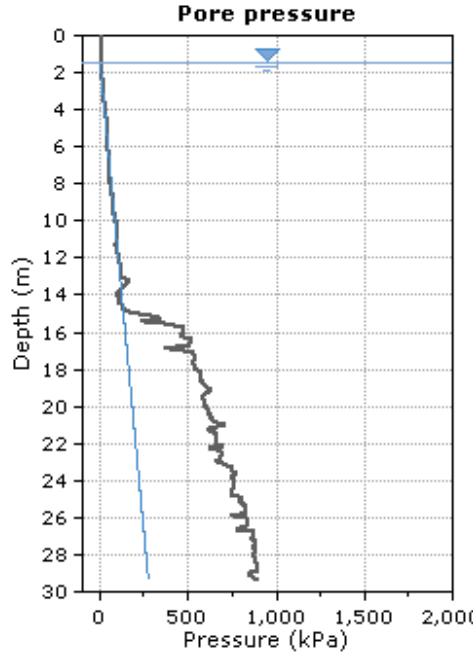


FIGURE 2.1 – Excess pore pressure.

The effective σ'_{v0} and total $\sigma_{v0} = \sigma'_{v0} + u_0$ overburden stresses are then obtained, enabling to calculate the net cone resistance $q_n = q_t - \sigma_{v0}$. In order to eliminate correlations between parameters, in Robertson (1990) it is proposed that q_n should be divided by σ'_{v0} to take into account the overburden and that f_s and $u_2 - u_0$ should be divided by q_n . The result are the following normalizations:

$$Q_{t1} = \frac{q_n}{\sigma'_{v0}} \quad (2.1)$$

$$F_r = \frac{f_s}{q_n} \quad (2.2)$$

$$B_q = \frac{u_2 - u_0}{q_n} \quad (2.3)$$

Later work found that the exponent n of σ'_{v0} in the Q_{t1} expression should be 1 only for pure sands and reduced for mixed soils to the limit of 0.5 for pure clays (ROBERTSON;

WRIDE, 1998). The result is the more general expression:

$$Q_{tn} = \left(\frac{q_n}{p_a} \right) \left(\frac{p_a}{\sigma'_{v0}} \right)^n \quad (2.4)$$

where p_a is the atmospheric pressure taken as reference, with the value of 0.1 MPa. The exponent n can be obtained in an iterative procedure that starts with the following expression:

$$n = 0.381I_c + 0.05(\sigma'_{v0}/p_a) - 0.15 \quad (2.5)$$

The parameter I_c is called classification index and defines circular curves that approximate the class separation curves of the soil classification method. It can be calculated as (ROBERTSON, 2009):

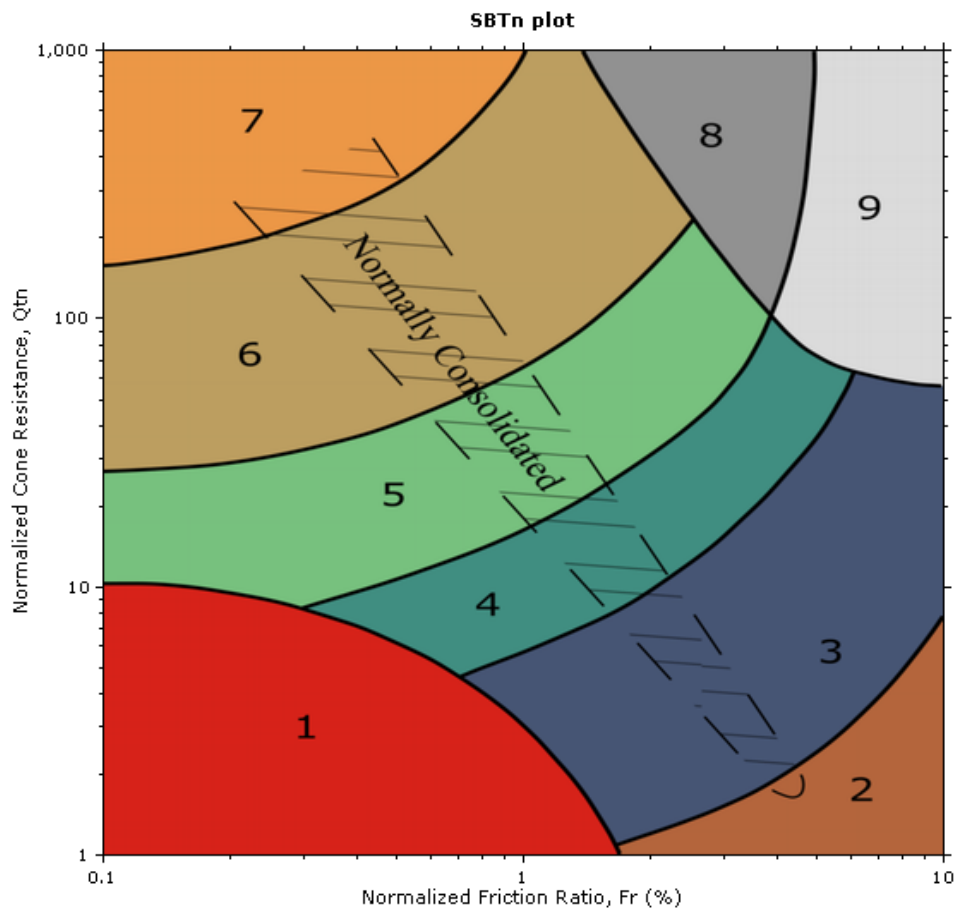
$$I_c = [(3.47 - \log Q_{tn})^2 + (\log F_r + 1.22)^2]^{0.5} \quad (2.6)$$

One can notice that Q_{tn} is required to calculate I_c , which is required to calculate n , which is required to calculate Q_{tn} . This is why an iterative procedure is required.

The charts proposed by Robertson (1991) for soil classification are represented in Figures 2.2 and 2.3. After obtaining raw CPT values and performing all procedures defined above, each soil example can be represented by a point within each of these charts. The area to which the point belongs gives the class of the corresponding soil. In the CPeT-IT software only Figure 2.2 is used, ignoring Figure 2.3. If the obtained point is located outside the Q_{tn} and F_r ranges defined within this chart, the soil class is considered uncertain and class 0 is then assigned.

2.2 Focused on soil behavior (FSB)

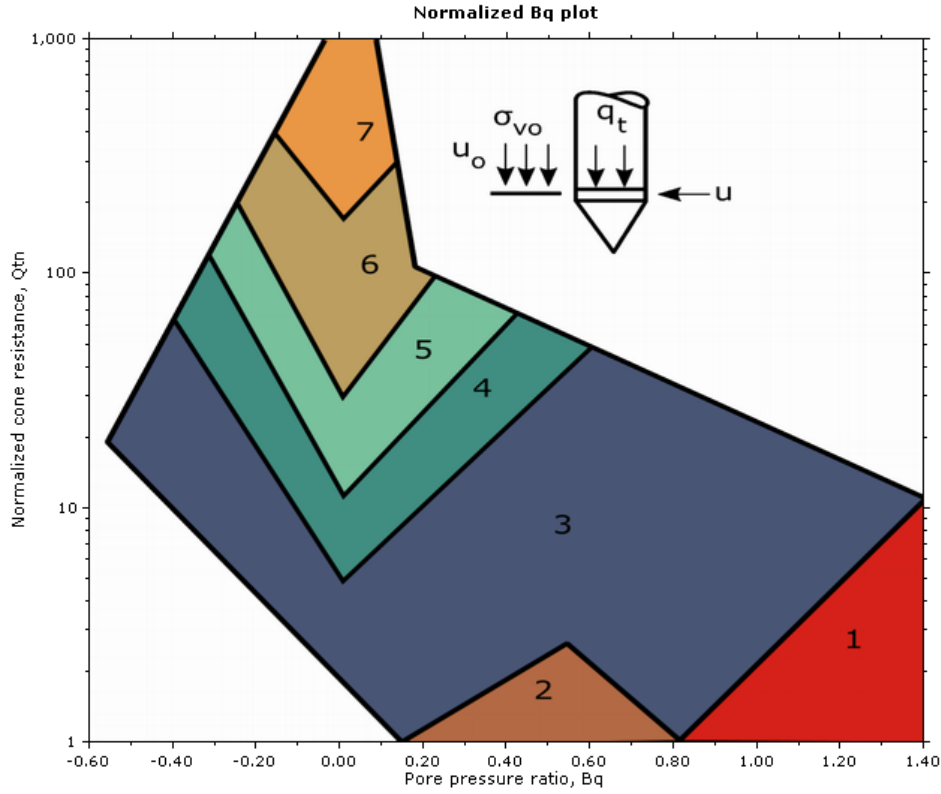
The second soil classification method replicated in this work is the one proposed in (ROBERTSON, 2016), which is referred as FSB throughout this text. This method can be considered fully based on the soil behavior because the only explicit agreement between the proposed classifications and the real soil types refers to sands and clays, which behavior is well defined and clearly separable. Sands usually have high strength, low compressibility and high permeability, which can be easily distinguished from low strength, high compressibility and low permeability usually observed for clays (ROBERTSON, 2016). This lead the author to three basic divisions: clay-like, sand-like and transitional behavior. Each of these classes was subdivided as dilative or a contractive behavior, according

FIGURE 2.2 – $Q_{tn} \times F_r$ chart from Robertson (1991).

to their consolidation state. A separated class was created for contractive clays that are sensitive to disturbance. This procedure generated seven possible classes of soil behavior, which are:

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

The author also detected that B_q is strongly negatively correlated to Q_{t1} , which was explained by the way q_t is used to calculate these normalizations. One consequence of

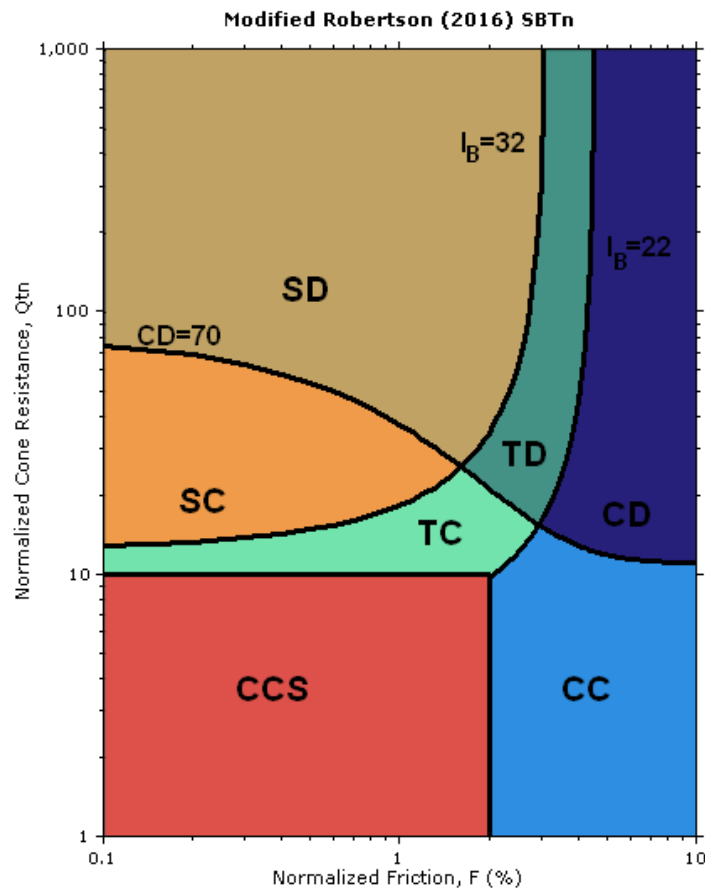
FIGURE 2.3 – $Q_{tn} \times B_q$ chart from Robertson (1991).

this correlation is that, highly overconsolidated clays become indistinguishable from very dense sands (SCHNEIDER *et al.*, 2008). To solve this problem, a new normalized excess pore pressure was proposed:

$$U_2 = B_q \cdot Q_{t1} = \frac{u_2 - u_0}{\sigma'_{v0}} \quad (2.7)$$

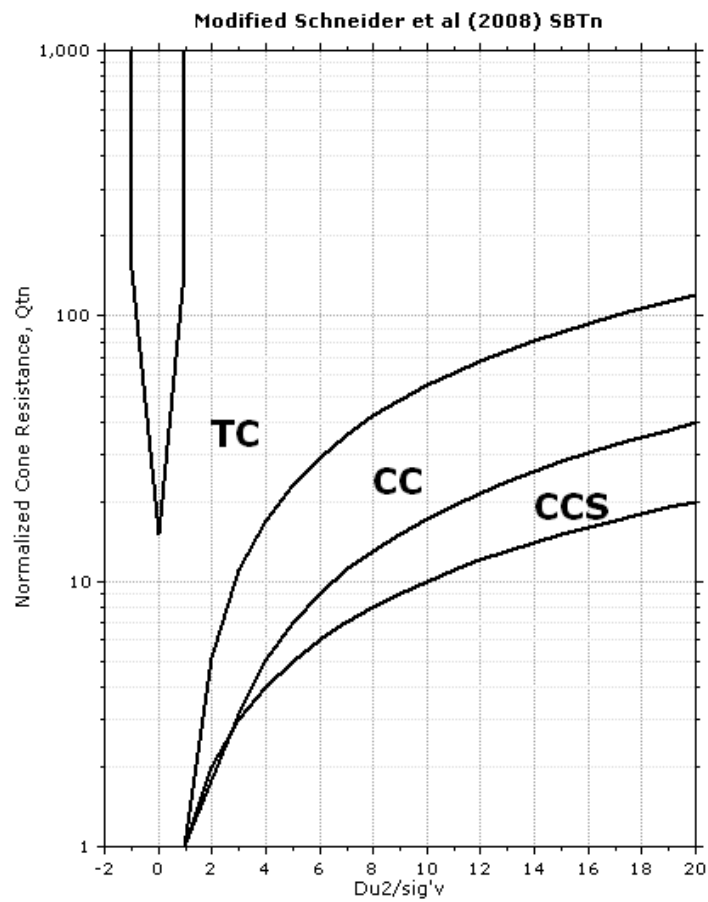
In addition to that, two new charts are proposed, one using F_r and Q_{tn} and the other using U_2 and Q_{tn} . These charts, based on the ones proposed in (SCHNEIDER *et al.*, 2008; SCHNEIDER *et al.*, 2012), are shown in Figures 2.4 and 2.5. The chart that uses U_2 is very similar to the one proposed in Schneider *et al.* (2008), while the chart with F_r uses the hyperbolic curves presented in Schneider *et al.* (2012). New curves are also added to the F_r chart to separate dilative and contractive behaviors, as well as for separating the contractive sensitive behavior.

The procedures for obtaining the soil class with this method are very similar to those employed for the ISG method. The values of Q_{tn} , F_r and U_2 are calculated from raw CPT measurements, using the procedures presented above. This enables obtaining one point in each chart, which gives the soil class according to the area wherein the point is located. In this case, the CPeT-IT software uses both charts and if they do not agree the soil example is assigned to class 0. In addition to that, soil receives class 0 if the point is

FIGURE 2.4 – $Q_{tn} \times F_r$ chart from Robertson (2016).

located outside the ranges of Q_{tn} , F_r and U_2 presented within the charts and if a modified normalized small-strain rigidity index is greater than 330.

In Robertson (2016) it is highlighted that the proposed method is not accurate for aged or cemented soils, which contain microstructure.

FIGURE 2.5 – $Q_{tn} \times U_2$ chart from Robertson (2016).

3 Modeling procedures

In this Chapter, it is first made a data description, presenting the data used through this work and analyzing it by some statistical measures, in order to support the preprocessing procedures proposed for the supervised studies. Following, the modeling methodologies developed for the different performed analyses are described and explained. A summarized diagram of the full methodology, including the supervised and the unsupervised approaches, is shown in Figure 3.1.

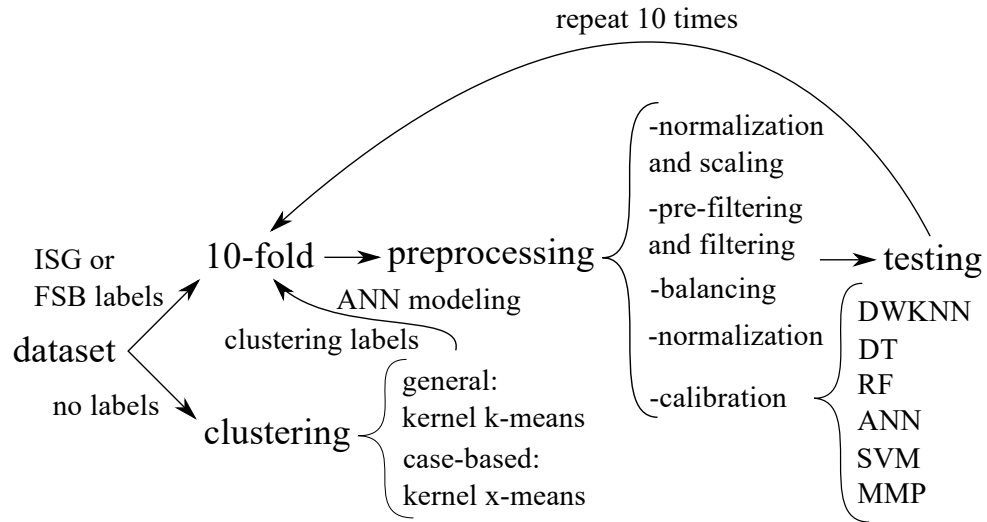


FIGURE 3.1 – Full methodology description with all modeling procedures summarized.

3.1 Data description

3.1.1 Main and geological datasets

The here named main dataset is composed by 111 CPT soundings. An amount of 38 of these soundings were provided by Professor P. K. Robertson and are described in Table 3.1, while 73 were made available online by Professor P. W. Mayne ¹ and are described in Table 3.2.

¹<http://geosystems.ce.gatech.edu/Faculty/Mayne/Research/index.html>

TABLE 3.1 – Dataset from P. K. Robertson.

General soil type	Location	Number of soundings
Mixed Soils	Canada	3
	Italy	1
	USA	6
	Switzerland	1
Soft Clay	UK	1
	Australia	1
	Norway	1
	USA	3
	Canada	2
	Sweden	2
	North Sea	1
	Very soft offshore	1
Soft Rock	USA	4
Stiff Clay	UK	3
	USA	4
	Italy	1
	France	1
	Ireland	1
	Alaska (USA)	1
Total		38

TABLE 3.2 – Dataset from P. W. Mayne.

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

TABLE 3.3 – Geological dataset.

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, UK	Holocene	2
	Burswood, 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

In this work, we intent to investigate the possibility of including information about soil age for soil classification. This interest arose from the difficulty related in the literature for classifying aged soil (ROBERTSON, 2016). One way of measuring soil age is by the geological epoch when it was formed. This information was available only for some of the 38 CPT soundings provided by Robertson, being provided in reference (ROBERTSON, 2016). In this work, soil age is treated as a discrete, ordered and numerical class of geology (CG), with low values meaning new soil and high values meaning old soil. The soundings with their respective geological age and its corresponding CG are given in Table 3.3. The dataset composed by these soundings from Robertson accompanied with soil age information is here called geological dataset.

One problem concerning the use of ML techniques is outliers, which are inputs with discrepant values. One way of evaluating the presence of outliers is by producing box-plots

for all input features, as presented in Figures 3.2 and 3.3 within the main dataset and for ISG and FSB inputs, respectively. In this figures, black dots represent values that are not within the range $[Q_1 - 1.5IQ, Q_3 + 1.5IQ]$, where Q_1 is the first quartile, Q_3 is the third quartile and IQ is the interquartile range. One can notice a high number of black dots for most box-plots, each one representing a potential outlier.

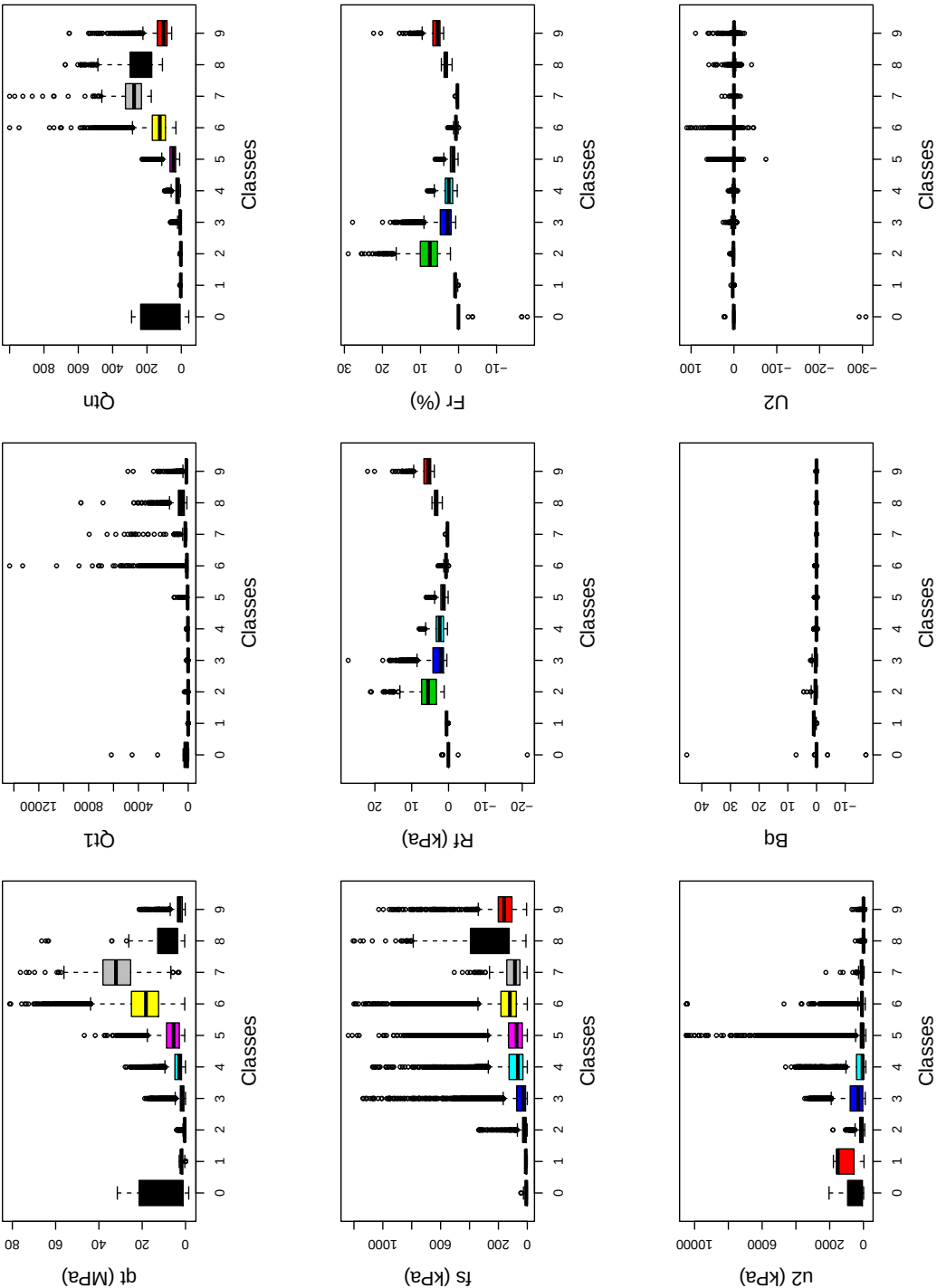


FIGURE 3.2 – Box-plots of features values for ISG classes.

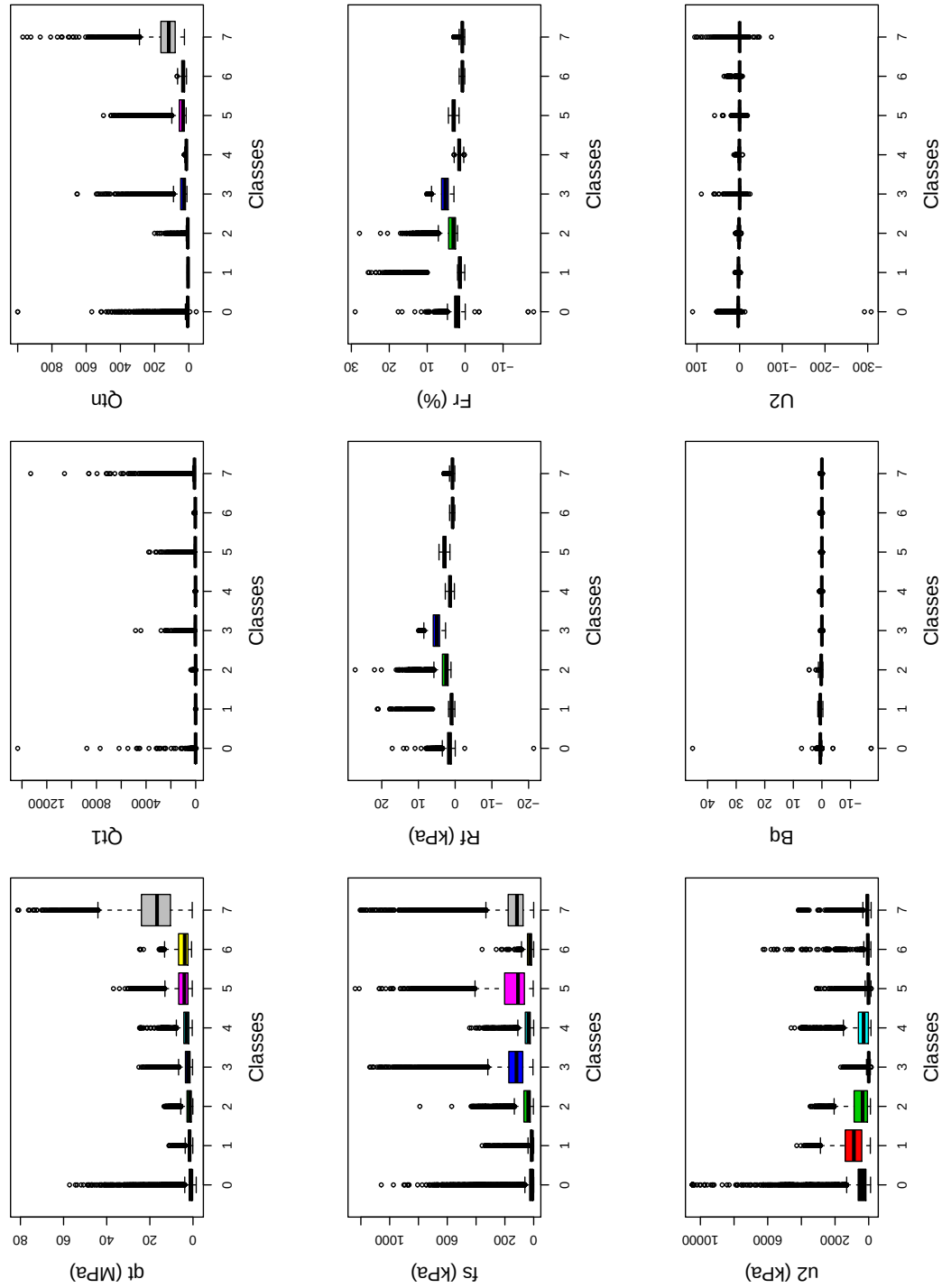


FIGURE 3.3 – Box-plots of features values for FSB classes.

Preliminary tests have shown that removing all these potential outliers harm the accuracy of the ML techniques, meaning that the information provided by the datasets is being severely reduced. To solve this problem, in this work each potential outlier is subjected to the Edit Nearest Neighbor technique before definitely removing it. This technique compares the class of the potential outlier to the one of its nearest neighbor within the input feature space, removing it only if both classes do not match (WILSON, 1972) (see Section A.1).

Another problem that can harm the accuracy of ML techniques in classification problems is imbalance between classes. To investigate this issue, Figure 3.4 presents histograms showing objects distribution among classes for the ISG (see Section 2.1) and the FSB (see Section 2.2) methods, considering the main and the geological datasets. One can notice that classes are strongly imbalanced specially for ISG classes, as illustrated in Figures 3.4a and 3.4c. Imbalance can be treated using techniques to be presented below. Nonetheless, preliminary tests have shown that, even with this treatment, the number of objects assigned to class 0 within the ISG method was not sufficient to represent this class among the others. Therefore, all objects assigned to class 0 within the ISG method were removed from the datasets. Furthermore, as ISG classes are too imbalanced within the geological dataset, analysis with this dataset are here restricted only to the FSB soil classification method.

Treating imbalance involves increasing minority classes and decreasing majority classes, so that all classes result with the same number of objects. The number here adopted is the maximum between 1000 and twice the number of elements in the minority class. A random sampling was applied to reduce majority classes, considering that CPT data contains several redundancies due to many measurements taken within each soil layer. To increment classes, it is proposed to estimate each new object from $d + 1$ existing objects, where d is the dimensionality of the input feature space. This strategy is based on the possibility of creating a d -dimensional simplex with existing objects at its vertices and the new one at its centroid. Figure 3.5 shows an example of this simplex balancing effect. In this Figure, Fr_p means $F_r(\%)$, with the subscripted “p” in Fr_p indicating percentage.

When using ML techniques, it is important to evaluate statistical correlations between input features to auxiliate investigating best input feature combinations. Correlation matrices were produced to address this issue, as presented in Figure 3.6. The two matrices at the top refer to the main dataset, while the two at the bottom refer to the geological dataset. One can observe that f_s and q_t are strongly positively correlated, confirming literature allegations (DOUGLAS; OLSEN, 1981; ROBERTSON *et al.*, 1986). Another confirmed literature allegation is that Q_{t1} and Q_{tn} are negatively correlated to B_q (SCHNEIDER *et al.*, 2008). Thus, strong positive correlation can be observed between Q_{t1} and Q_{tn} and no significant correlation between CG and any other input feature.

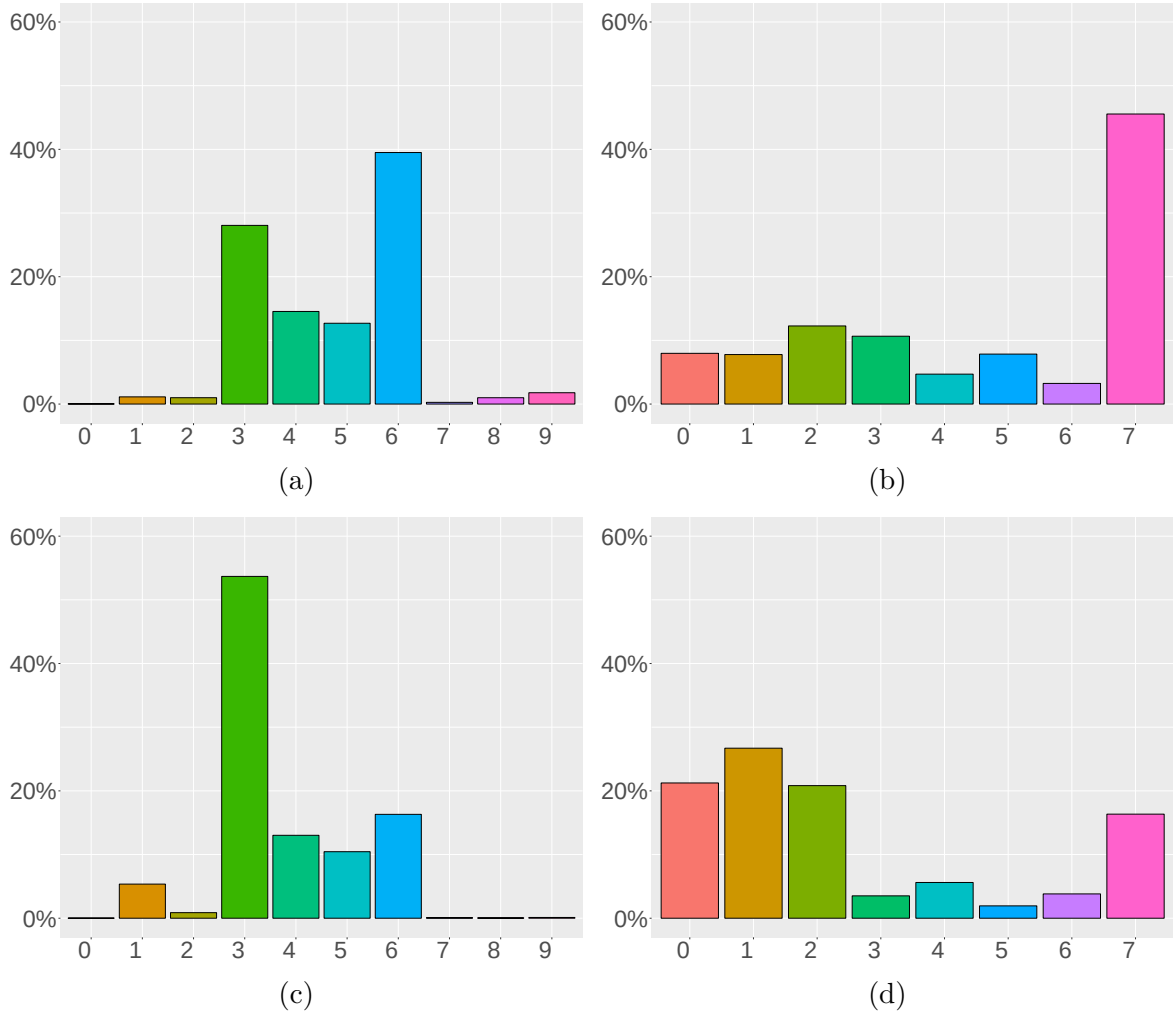


FIGURE 3.4 – Main and geological datasets histograms: (a) for ISG classes in the main dataset, (b) for FSB classes in the main dataset, (c) for ISG classes in the geological dataset and (d) for FSB classes in the geological dataset.

3.1.2 Tropical dataset

Another dataset used in this work is composed by 62 CPT soundings from the metropolitan region of São Paulo, Brazil, being here named tropical dataset. These soundings were executed within in situ investigations for the construction of the subway of São Paulo, being provided by the Company of the Metropolitan of São Paulo under a confidentiality term. Measurements were taken at each 2 cm of depth and, after removing clearly inconsistent data, remained 41381 soil examples.

Most procedures performed for the main and geological datasets, presented in Section 3.1.1, were also performed for the tropical dataset. The only difference is related to data imbalance, which can be evaluated for the tropical dataset by observing the histograms presented in Figure 3.7.

The most important difference to be observed is that, for this case, the number of

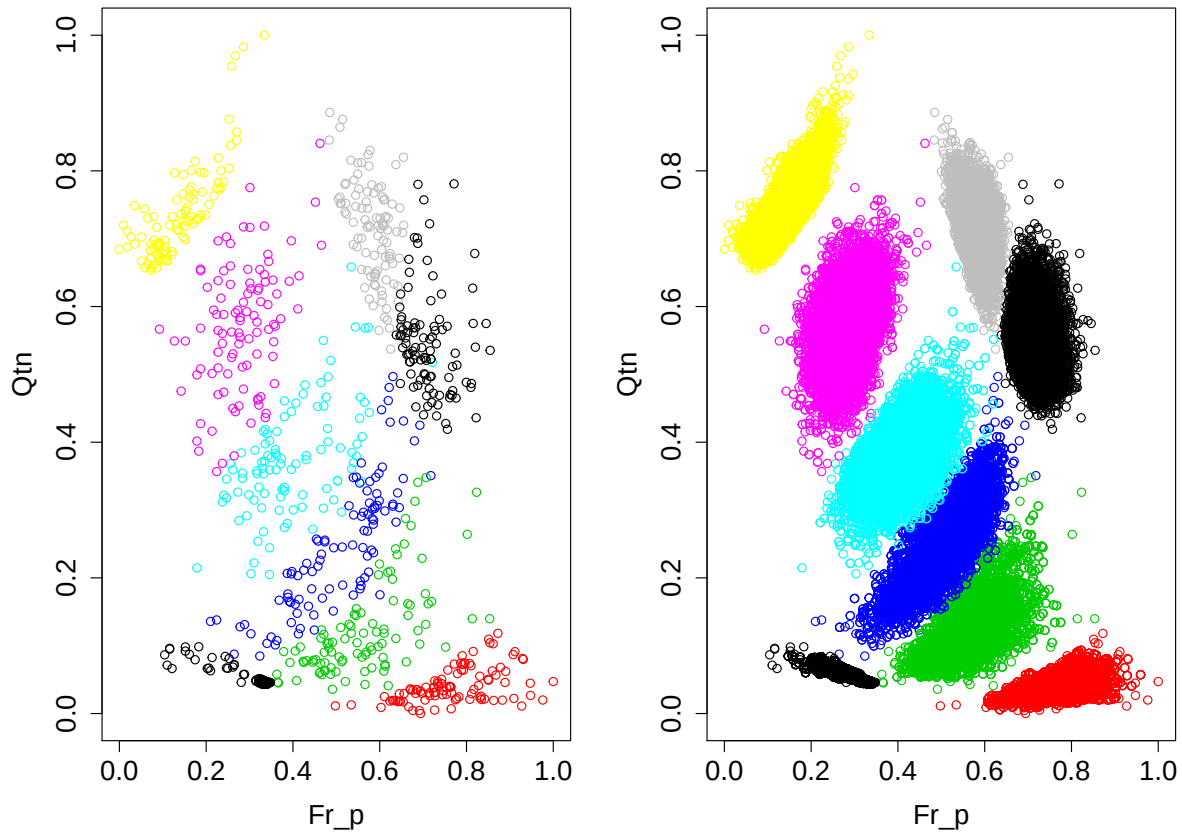


FIGURE 3.5 – Balancing example: imbalanced dataset on the left and balanced on the right.

objects assigned to ISG class 0 is sufficient to represent this class. Therefore, objects assigned to this class were maintained in the dataset. Objects assigned to class 0 within the ISG and FSB methods were removed from the dataset only for training the ML techniques.

3.1.3 Clustering approach data description

The full dataset used for the general clustering is composed by 179 CPT soundings, of which 38 were taken in several countries and provided by Professor P. K. Robertson, 73 were taken in the USA and were made available online by Professor P. W. Mayne ², 62 were taken in Brazil and provided by the Company of the Metropolitan of São Paulo, one was taken in the USA and given by Professor Renato P. da Cunha and five were taken in Brazil paired with SPT soundings and provided by Professor Heraldo L. Giacheti. The 179 soundings produced 130966 examples for the ML techniques, each example consisting on a CPT measurement taken at a specific depth.

²<http://geosystems.ce.gatech.edu/Faculty/Mayne/Research/index.html>

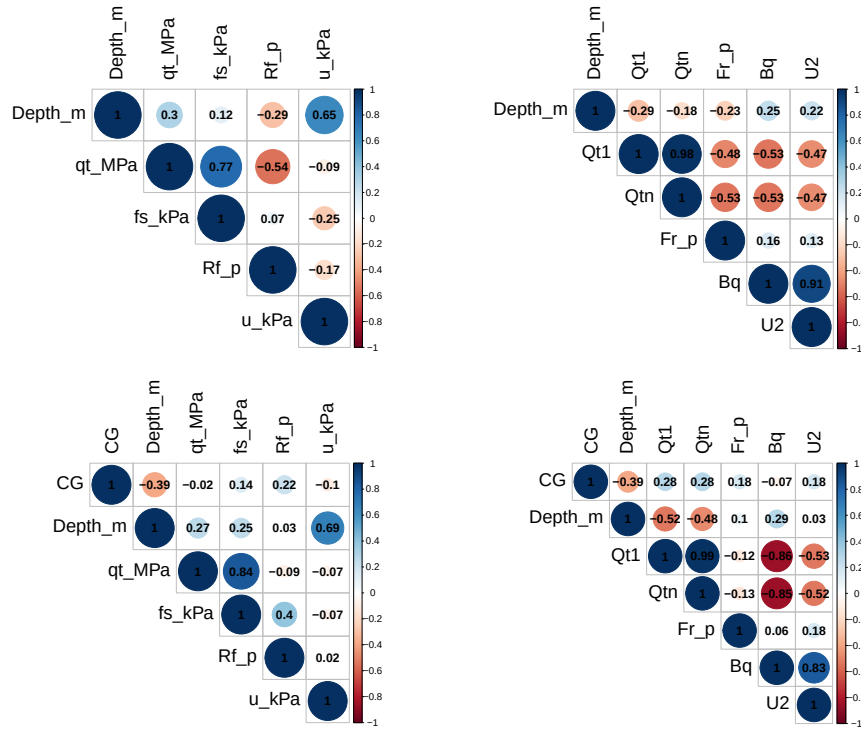


FIGURE 3.6 – Correlation matrices for all input features.

The soundings provided by Professor P. K. Robertson were the ones used by him to develop his soil behavioral classification methods. Those from the Company of the Metropolitan of São Paulo were produced during the construction of the subway of São Paulo and were provided under a confidentiality term.

Figures 3.8a and 3.8b show histograms for the objects distribution 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 paired CPT/SPT soundings were used to compare ISG classes with SPT sample observations. SPT sample observations were divided into sands (1), silts (2) and clays (3). A histogram with object distributions of these soundings among these SPT soil types is shown in Figure 3.9. There are 2847 objects, of which 458 are from the minority class (2), which is enough to represent it.

3.2 Feature analysis with distance-based algorithms

The objective of this Section is to present the used strategy to apply KNN algorithms to classify soil using the outputs given by the program CPeT-IT. Different combinations of features are evaluated, considering normalized and non-normalized inputs. The methodol-

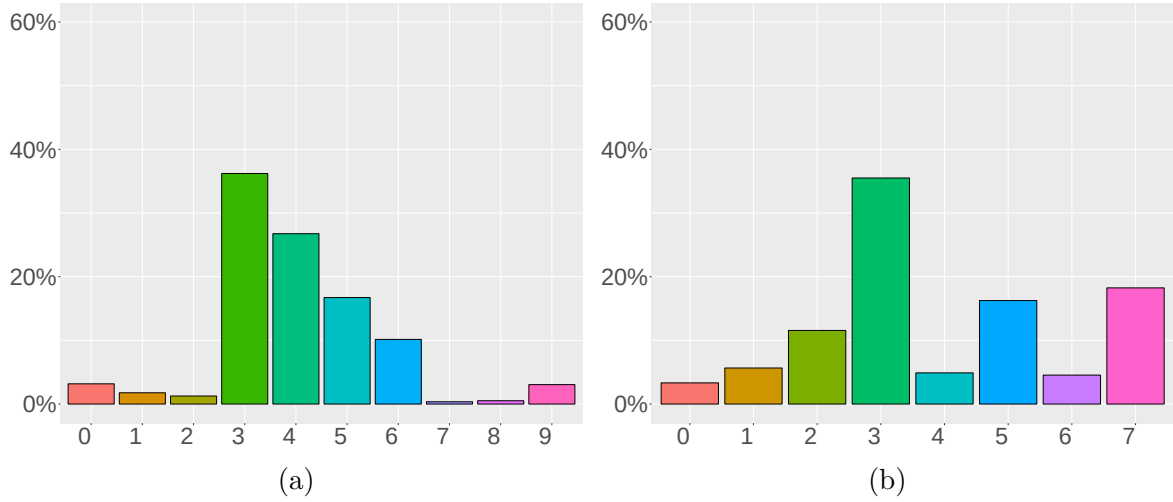


FIGURE 3.7 – Tropical dataset histograms: (a) for ISG classes in the tropical dataset and (b) for FSB classes in the tropical dataset.

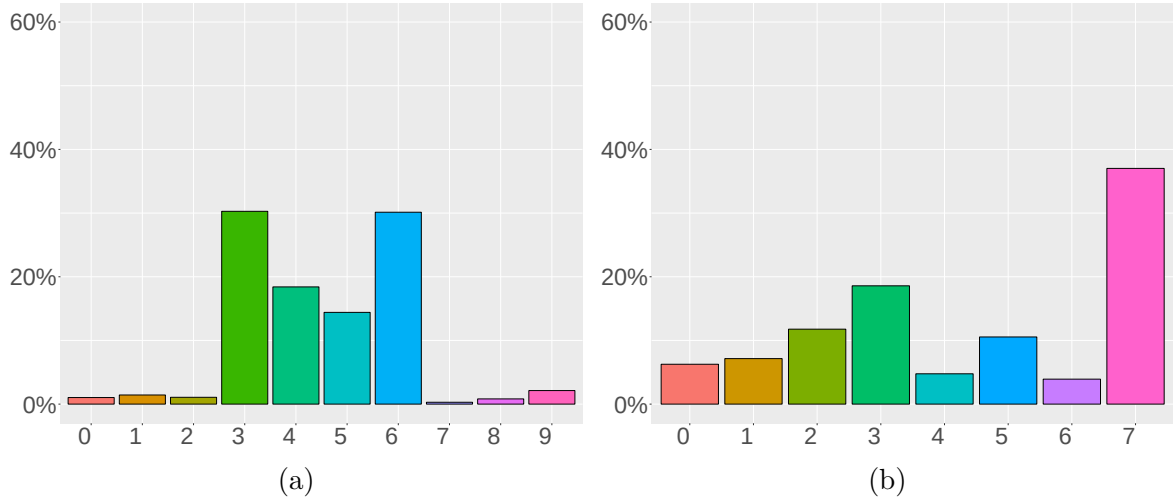


FIGURE 3.8 – Clustering dataset histograms: (a) for ISG classes in the clustering dataset and (b) for FSB classes in the clustering dataset.

ogy used in this work is conceived to be reproducible, applicable to general CPT datasets and flexible to work with different ML techniques. Some procedures can be performed using open source packages available for R programming.

Two algorithms are tested and compared, the classical KNN and the Gaussian DWKNN. Classes from the classification methodologies ISG and FSB are regarded as outputs. For the inputs, three basic sets are defined from which all possible combinations are extracted. One of these sets takes the non-normalized parameters, depth z (m), corrected cone resistance q_t (MPa), lateral friction f_s (kPa) and pore pressure behind the cone tip u_2 (kPa). The other sets replaces these three last indexes by ISG inputs Q_{t1} , F_r (%) and B_q and by FSB inputs Q_{tn} , F_r (%) and U_2 , respectively.

Then, all combinations of two, three and four inputs are taken from these three basic

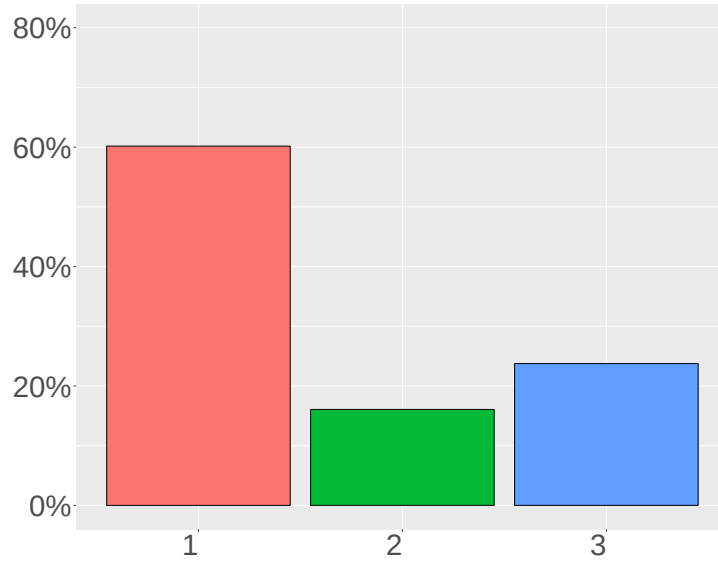


FIGURE 3.9 – Histogram for SPT observations in the dataset with CPT paired with SPT.

sets and each one of them is tested. It means that there are eleven possible combinations for each one of the three input sets. In another test, the raw lateral friction f_s is replaced by R_f for the full non-normalized inputs set, both ISG and FSB methods as well as KNN and DWKNN algorithms to check if the performances are improved, adding four more tested combinations. Finally, the inclusion of geological age is evaluated with the geological dataset considering only FSB outputs and some inputs chosen based on results of the previous experiments.

A 10-fold cross-validation procedure (STONE, 1974) was applied to evaluate classification accuracy. The procedure starts 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 KNN algorithm, starting k with 1 and increasing it while the accuracy measured on the validation dataset rises. Only odd numbers of k are considered.

For each cross-validation round, the average of the accuracies per class are taken. This avoids disregarding minority classes in the performance evaluation. After the 10 folds are used for testing, a mean and standard deviation accuracy performance is computed. For comparing the results of the experiments, the Friedman statistical test (FRIEDMAN, 1937), 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).

3.3 General analysis with supervised ML techniques

The five supervised techniques mentioned in Annex A, as well as the MMP that combines them, were tested with the ISG and FSB methods using the following input feature combinations:

- Basic combinations – for ISG and FSB outputs (main dataset)
 - Raw inputs: z (m), q_t (MPa), f_s (kPa) and u_2 (kPa)
 - Normalized of Robertson (1990): z (m), Q_{t1} (MPa), F_r (%) and B_q
 - Normalized of Robertson (2016): z (m), Q_{tn} (MPa), F_r (%) and U_2
- Geological combinations – only for FSB outputs (geological dataset)
 - Raw inputs: CG , z (m), q_t (MPa), f_s (kPa) and u_2 (kPa)
 - Normalized of Robertson (2016): CG , z (m), Q_{tn} (MPa), F_r (%) and U_2
- Biased combinations (used by CPeT-IT to generate outputs) – for ISG and FSB outputs (main dataset)
 - ISG inputs: Q_{tn} and F_r (%)
 - FSB inputs: Q_{tn} , F_r (%) and U_2

The 10-fold cross-validation procedure described within the previous Section is used to measure the predictive performance of each ML technique. After the calibration process described for each supervised technique in Annex A, they are all tested. Their accuracies are compared using the statistical test of Friedman (1937) with the post-hoc statistics of Nemenyi (1963) and the signed-rank statistical test of Wilcoxon (1945). It is important to highlight that all techniques were trained with the main dataset, using this training for applications at the geological and tropical datasets.

As in the previous analyses, DWKNN was calibrated using only odd numbers for k from 1 and increasing it while the accuracy measured on the validation dataset rises. The command `kknn()` from the R package *kknn* was employed as the DWKNN implementation.

For boosted DT and RF, the only parameter to be calibrated is the number of trees. For DT, the number of trees t starts with the number of classes c and is increased by c until predictive performance drops. On the other hand, for RF, a standard number of 500 trees is adopted. The command `C5.0()` from the R package *C50* was employed as the DT implementation. For RF, the command `randomForest()` from the R package *randomForest* was used.

For ANN, only two hidden layers were considered for its architecture. In this work, three parameters are calibrated for ANN: the number of iterations, the number of neurons in the first hidden layer and the number of neurons in the second hidden layer. The weights are initialized randomly within the range $[-0.3, 0.3]$ and the learning rate is fixed as $\eta = 0.2$. The number of iterations starts at 200 and is increased by 200 until predictive performance drops. This calibration is performed with a preliminary ANN architecture, using a single hidden layer with a number of neurons equal to the number of classes c (see Chapter 2).

The ANN architecture calibration starts with one hidden layer containing one neuron. The number of neurons of this hidden layer is then increased by 1 until predictive performance drops, limited by $2c$. The number of neurons of the one-hidden-layer ANN with best predictive performance is fixed while a second hidden layer with one neuron is added. Then the number of neurons of the second hidden layer is increased by 1 until predictive performance drops, limited to $2c$. The architecture leading to best predictive performance along this full procedure is adopted, including all ANN tested with one and two hidden layers. It was adopted the MLP ANN implementation of the command *mlp()* from the R package *RSNNS*.

For SVM, it is here proposed to nest the calibration of κ inside the calibration of d . This is done varying κ from 1 by 1 for each value of d , which is also varied from 1 by 1, and fixing the (d, κ) pair with best predictive performance. Afterwards, with these best parameters fixed, δ is started with 1 and is increased by 1 until predictive performance drops. For the SVM implementation, it was used the command *ksvm()* from the R package *kernlab*.

Finally, the MMP is validated and the techniques are tested with respect to dataset extrapolation using the tropical dataset for testing and the main dataset for training. An additional test is also done by considering only the CPT parameters without the pore pressure information (alternative CPT data).

3.4 Preprocessing procedure for supervised techniques

A preprocessing procedure is executed within each 10-fold cross-validation step. It includes data transformation for the training, validation and testing sets, and data cleaning and balancing only for the training set.

First step of the preprocessing is a data transformation, rescaling each input feature to the $[0, 1]$ range. In other words, the values v of some feature F are changed into the

values v' according to the following equation:

$$v' = \frac{v - \min(F)}{\max(F) - \min(F)} \quad (3.1)$$

where the minimum and maximum values of F are taken from the training set.

A second data transformation is tested for the ANN, SVM and MMP analyses to impose a logarithmic scale to each input feature. This procedure was adopted because the original charts from Robertson use logarithmic scale and preliminary tests showed that better performance is achieved with this transformation. This can be performed by using the following equation:

$$v'' = \text{sign}(v') \cdot \log_{10} [(10^n - 1) \cdot |v'| + 1]^{(\frac{1}{n})} \quad (3.2)$$

After performing some tests with this expression, it was fixed $n = 2$. Figure 3.10 shows an example of the log-scaling effect.

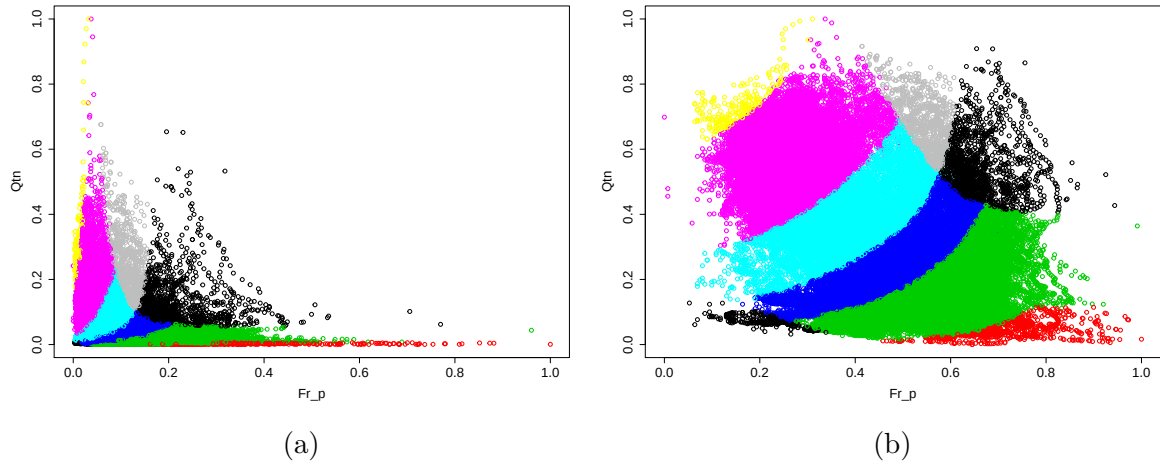


FIGURE 3.10 – Log-scaling effect example in the objects distribution within the $Q_{tn} \times F_r$ space: (a) without log-scaling and (b) with log-scaling.

A pretest that is performed only for DWKNN, DT and RF for each 10-fold cross-validation test is comparing predictive performance for balanced and imbalanced training sets. It is then taken the training set for which best predictive performance was obtained. This procedure is too time-consuming for the ANN and SVM, so the balanced training set was adopted for all cases. This decision was corroborated by preliminary tests.

A summary of the full preprocessing methodology is presented in Figure 3.11.

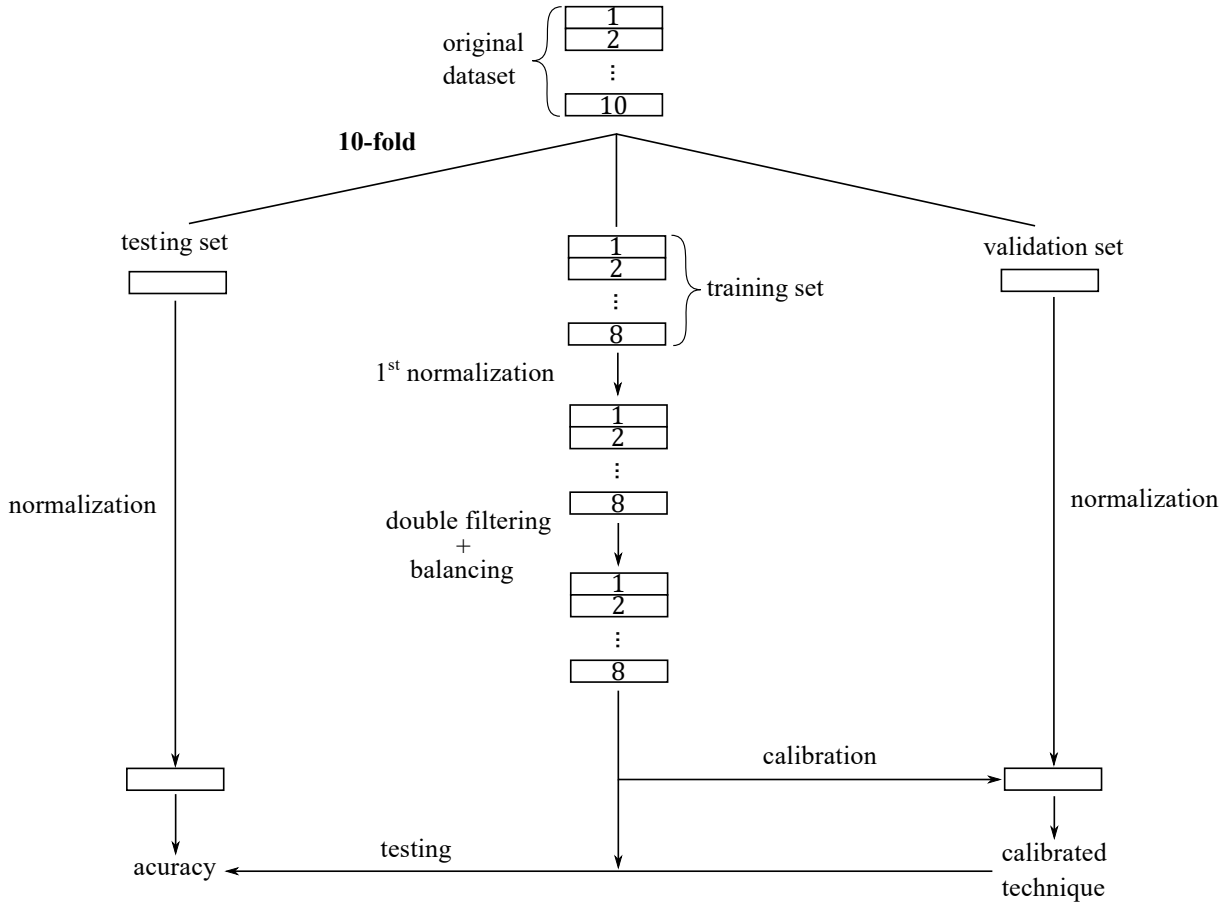


FIGURE 3.11 – Summarized methodology for data preprocessing.

3.5 Clustering approach methodology

3.5.1 Clustering analysis

Both the general and the case-based modeling 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), total cone resistance q_c (MPa), lateral friction f_s (kPa) and pore pressure measured behind the cone tip u_2 (kPa). For both approaches, the Gaussian kernel is used to map the objects into a higher-dimension feature space (see Section A.5).

This kernel was chosen because it depends on only one parameter σ that can be estimated from the data as the median of the squared distances $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ distribution. This is the procedure used by the command *kkmeans()* from the R package *kkmeans*, which is adopted in this work to run the kernel k-means algorithm.

For the general approach, the procedure adopted to define the number of classes k was manually varying the value of k and adopting the one with the lowest total variance inside clusters. On the other hand, for the case-based approach, the kernel x-means algorithm is

employed. This algorithm basically 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 (BIC) is improved (PELLEG; MOORE, 2000). The best clustering results obtained in both cases are then compared to those labels given by the ISG and FSB methods as well as the observations with the SPT for the five soundings for which the CPT soundings have paired SPT.

For this step, only the data normalization is required, rescaling each input feature to the range $[0, 1]$. Next, a logarithm is applied to transform inputs scale and ease the classes separation within the original feature space, which was confirmed in preliminary studies. This transformation can be made with the Eq. (3.2) using also $n = 2$.

As an illustrative example, this proposed log-scaling procedure was applied to the $Q_{tn} \times F_r$ feature space for some CPT soundings to evaluate the effect on class separation. Observing the results shown in Figures 3.10a and 3.10b, one can conclude that classes become better separated with log-scaling.

3.5.2 ANN modeling

In this work, ANN models are created to replicate soil classification systems obtained by clustering. These models are trained with the same dataset used to cluster the soil classes and also with a reduced dataset, in which only the u_2 parameter is omitted. To produce accurate ANN models the training process requires data preprocessing, which can be performed using the methodology presented in Figure 3.11, following the same procedure presented for the supervised approach and adopting the same performance measure, the mean recall. The 10-fold cross-validation procedure is employed within the training process. Once different ANN architectures can be obtained for different testing folds, the final model is defined as the most frequent one.

4 Modeling results and discussions

4.1 Preliminary studies

In this Section, the results achieved by the methodology in Section 3.2 are presented and discussed.

4.1.1 Distance-based input feature analysis

A total of 132 classification results were obtained in the analysis to be presented in this Section: 2 classification outputs (ISG described in Section 2.1 and FSB described in Section 2.2) $\times$ 33 combinations of input features $\times$ 2 types of 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 features are dimensionless.

In the pre-test, the use of the the balanced training set prevailed in most cases, with accuracy gains up to 20%. Besides, in the few cases when the unbalanced set was chosen, accuracy gains never reached 2%. This allows concluding that balancing the training set in all cases would be reasonable for this dataset.

Table 4.1 presents the mean accuracy and its standard deviation obtained for every experiment that used features Q_{t1} and Q_{tn} with KNN and ISG outputs. Although Q_{t1} and Q_{tn} are strongly correlated, as can be seen in Figure 3.6, the accuracy obtained in combinations that use Q_{tn} are higher than those obtained in combinations that use Q_{t1} . This indicates that using Q_{tn} as an input should be preferred over using Q_{t1} .

Combinations that presented best performance with the classical KNN technique applied to ISG outputs are presented in Table 4.2. Once the first KNN combination uses the original ISG inputs and outputs, it was expected that it would lead to the highest mean accuracy among all. Nonetheless, results of the Friedmans statistical test with the Nemenyi post-hoc statistics show an statistical equivalence between the first two KNN combinations in Table 4.2. Thus, the last two KNN combinations shown in Table 4.2 can be considered of lower performance. This shows that including more features among the

TABLE 4.1 – Comparing Q_{t1} to Q_{tn} (KNN predictive results for ISG).

Inputs	Elected k	MA (%)	SD
$Q_{t1} z$	3	69.00	1.22
$Q_{tn} z$	3	72.47	1.00
$Q_{t1} F_r$	5	89.71	1.06
$Q_{tn} F_r$	1	96.71	0.59
$Q_{t1} B_q$	5	57.63	1.37
$Q_{tn} U_2$	7	64.79	1.96
$Q_{t1} z F_r$	1	84.72	1.44
$Q_{tn} z F_r$	1	93.45	0.76
$Q_{t1} z B_q$	1	73.78	1.02
$Q_{tn} z U_2$	1	77.40	1.32
$Q_{t1} F_r B_q$	3	87.74	1.86
$Q_{tn} F_r U_2$	1	94.77	1.18
$Q_{t1} z F_r B_q$	1	89.21	1.22
$Q_{tn} z F_r U_2$	1	93.06	1.00

TABLE 4.2 – Best predictive results for ISG.

Technique	Inputs	Elected k	MA (%)	SD
KNN	$Q_{tn} F_r$	1	96.71	0.59
	$Q_{tn} z F_r$	1	93.45	0.76
	$Q_{tn} F_r U_2$	1	94.77	1.18
	$Q_{tn} z F_r U_2$	1	93.06	1.00
	$z q_t f_s$	1	89.76	0.68
	$z q_t f_s u_2$	1	90.16	0.71
DWKNN	$Q_{tn} F_r$	1	96.71	0.59
	$Q_{tn} z F_r$	1	93.45	0.76
	$Q_{tn} F_r U_2$	1	94.81	1.17
	$Q_{tn} z F_r U_2$	1	93.16	0.80
	$z q_t f_s$	1	89.72	0.65
	$z q_t f_s u_2$	1	90.16	0.71

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 DWKNN technique applied to ISG outputs, which are also presented in Table 4.2. One can observe that results are very close to those of KNN, reinforcing the same conclusions. It is important to notice that KNN and DWKNN algorithms become the same for $k = 1$, which was the best k elected in all cases of Table 4.2. The small differences in results are due to numerical instabilities.

Considering now the classical KNN technique applied to FSB outputs, the best feature combinations are presented in Table 4.3. In this case, the Friedman statistical test with the Nemenyi post-hoc statistics show that the last two feature combinations are equivalent and statistically better than the first two. One can observe that, as expected, the best

TABLE 4.3 – Best predictive results for FSB.

Technique	Inputs	Elected k	MA (%)	SD
KNN	$Q_{tn} F_r$	7	88.76	0.47
	$Q_{tn} z F_r$	1	91.91	0.34
	$Q_{tn} F_r U_2$	3	92.96	0.49
	$Q_{tn} z F_r U_2$	1	93.82	0.57
	$z q_t f_s$	1	88.39	0.65
	$z q_t f_s u_2$	1	90.37	0.53
DWKNN	$Q_{tn} F_r$	7	88.89	0.45
	$Q_{tn} z F_r$	1	91.91	0.34
	$Q_{tn} F_r U_2$	1	93.03	0.39
	$Q_{tn} z F_r U_2$	1	93.82	0.57
	$z q_t f_s$	1	88.39	0.65
	$z q_t f_s u_2$	1	90.37	0.53

combination for this case include all three original FSB inputs, named Q_{tn} , F_r and U_2 . Nevertheless, associating depth with these features contributed to improve performance, even with the biasing due to the way by which the outputs were generated.

In the end, the feature combinations that produced best performance for the Gaussian DWKNN technique applied to FSB outputs are also presented in Table 4.3. One can observe that the results are very close to the ones of KNN, reinforcing that using original FSB inputs leads to good accuracy and that including depth among these features contributes to improve performance.

Concerning more general observations, with respect to the non-normalized inputs, good performance can be observed when they are associated with depth. For ISG and both classification techniques, for example, accuracy gets around 70% when q_t and f_s are used as input features, but rises close to 90% when z is included. Non-normalized inputs resulted better performance than using the normalizations of ISG, but worse than using the FSB normalizations, which are biased. These observations suggest that using only raw CPT data for soil classification, including depth, would be feasible, although confirming this hypothesis would require further studies.

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

TABLE 4.4 – Comparing input features f_s and R_f .

Output	Input	MA (%)	SD
ISG	f_s	90.16	0.71
	R_f	90.88	0.99
FSB	f_s	90.37	0.53
	R_f	91.96	0.43

TABLE 4.5 – Combinations with and without input feature CG for FSB.

Technique	Input	MA (%)	SD
KNN	$Q_{tn} F_r U_2$	92.00	0.96
	$Q_{tn} F_r U_2 CG$	94.47	0.93
	$z q_t f_s u_2$	89.61	1.37
	$z q_t f_s u_2 CG$	91.06	1.08
DWKNN	$Q_{tn} F_r U_2$	92.17	1.11
	$Q_{tn} F_r U_2 CG$	94.44	0.95
	$z q_t f_s u_2$	89.61	1.37
	$z q_t f_s u_2 CG$	91.06	1.08

4.1.2 Complementary studies

The first proposed complementary study is to evaluate if replacing f_s by R_f contributes to better differentiate soil classes, as affirmed by some authors (DOUGLAS; OLSEN, 1981; ROBERTSON, 1990). The evaluation is based on the experiments presented in Table 4.4, for which best performance was obtained for $k = 1$ in all 10-fold testing cases.

One can notice that relevant improvements occur only when FSB outputs are used. For ISG outputs, performance can be considered equivalent.

In the second complementary study, it is evaluated the impact in performance due to the inclusion of the geological age (CG) as an input feature. The geological dataset is used considering as inputs the non-normalized features and the FSB input features. The FSB outputs are considered and best performance was achieved for $k = 1$ in all cases. The results shown in Table 4.5 were obtained applying the classical KNN technique and the Gaussian DWKNN technique subjected to the calibration procedure described in Section 3.2.

A relevant increase in performance can be observed in all cases when the geological age was included. This difference was confirmed using the Wilcoxon statistical test, which obtained a p -value of 0.20%. This difference can be associated with the class 0 of the FSB method, which is probably composed mostly by structured soils. Once such soils are aged or cemented, including the geological age helps distinguishing them from the others.

In the third complementary study, the objective is evaluating if examples of FSB class 0 form a reasonably homogeneous region within the input feature space. The study

is made by removing this class from the complete dataset and observing the incurring change in classification performance.

An increasing predictive performance by class 0 removal would evidence that the objects of this class are mixed with objects of other classes in the feature space. In other words, it would indicate that these objects do not form a reasonably homogeneous region. On the other hand, if the predictive performance is maintained or reduced, one may conclude that class 0 objects are not disturbing the classification technique results. This means that they probably form a reasonably homogeneous region in the feature input space.

The classification tool elected for this evaluation was the DWKNN with Gaussian weighting, performing a 10-fold cross-validation to compare performance with and without the use of objects of class 0. The geological dataset was elected for this test to reduce computational cost, once it can be considered sufficiently diversified for this task.

The results obtained with the use of class 0 considering non-normalized inputs was a mean accuracy of 89.82% with a standard deviation of 1.35%. Without this class, the mean accuracy lowered to 88.47% with a higher standard deviation of 1.44%. Once performance increases if class 0 objects are included, there is evidence that this class forms a reasonable homogeneous region in the non-normalized feature space. However, considering the normalized inputs of Robertson (2016), the predictive performance decreases with the class 0 inclusion. Nonetheless, this decrement is not compromising, being the mean accuracy around 92% with this class and 95% without it. The conclusion is that this class may or may not indicate some specific soil type depending on which feature space is considered, but it can be initially included in the analyses because it does not harm significantly the predictive performance.

4.2 General supervised analyses

In this Section, the results achieved by the methodology in Sections 3.3 and 3.4 are presented and discussed.

4.2.1 DWKNN

Results obtained with the DWKNN are shown in Table 4.6. The first and second columns contain all used inputs and outputs respectively, as described in Section 3.3. The third and fourth columns contain the values here used to compare performance, represented by the mean accuracy of the 10-fold cross-validation procedure and the corresponding standard deviation. The highest and the lowest MA obtained are highlighted

TABLE 4.6 – Results obtained with DWKNN.

Input	Output	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	90.23	0.66
	FSB	90.28	0.48
$z \ F_r \ Q_{t1} \ B_q$	ISG	89.40	0.90
	FSB	88.82	1.62
$z \ F_r \ Q_{tn} \ U_2$	ISG	93.13	0.70
	FSB	93.77	0.57
$F_r \ Q_{tn}$	ISG	96.58	0.59
$F_r \ Q_{tn} \ U_2$	FSB	93.06	0.34
$z \ f_s \ q_t \ u_2 \ CG$	FSB	91.03	1.04
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	94.73	0.73

in bold.

The performances above 90% obtained for z , f_s , q_t and u_2 show that these raw input features could be used for soil classification, even with the output biasing of the here used datasets. Nevertheless, complementary tests showed that accuracy is not reasonable if z is not included.

For the ISG classification method best performance is achieved for the biased case (F_r and Q_{tn}), which was expected. Nonetheless, for FSB, including CG and z among biased inputs enhances performance. This suggests that increasing feature space dimensionality can improve separation between classes.

To complement this analysis, the confusion matrix for FSB with biased inputs (F_r , Q_{tn} and U_2) is evaluated. This matrix is obtained calculating the mean of the 10-fold cross-validation confusion matrices. Lines correspond to predicted values for classes 0 to 7 and columns correspond to reference values for these same classes (obtained with the CPeT-IT). The resulting matrix is:

$$\mathbf{C} = \begin{bmatrix} 487 & 53 & 43.3 & 2.4 & 4.5 & 2.9 & 0.5 & 3.6 \\ 47.9 & 605.7 & 4.7 & 0 & 2.3 & 0 & 0 & 0 \\ 57 & 3.1 & 988.4 & 11.2 & 3.9 & 0.2 & 0 & 0 \\ 15.6 & 0 & 9.9 & 890.1 & 0 & 7 & 0 & 0 \\ 10.3 & 3.3 & 2.7 & 0 & 385.8 & 3 & 2.4 & 0.1 \\ 23.4 & 0 & 0.3 & 7.5 & 3 & 652.3 & 0 & 11.3 \\ 4.8 & 0 & 0 & 0 & 2.4 & 0 & 270.2 & 6.1 \\ 35.4 & 0 & 0 & 0.2 & 0.1 & 5.8 & 4.2 & 3875.1 \end{bmatrix} \quad (4.1)$$

Observing the ratio between correct (at the main diagonal) and wrong (outside the diagonal) predictions, one can conclude that class 0 is the main source of confusion. An additional test was made removing class 0 and repeating this experiment, the result was a

TABLE 4.7 – Results obtained with DT.

Input	Output	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	91.71	0.76
	FSB	91.32	0.47
$z \ F_r \ Q_{t1} \ B_q$	ISG	95.81	0.60
	FSB	96.40	0.22
$z \ F_r \ Q_{tn} \ U_2$	ISG	97.60	0.43
	FSB	97.31	0.22
$F_r \ Q_{tn}$	ISG	96.97	0.57
$F_r \ Q_{tn} \ U_2$	FSB	94.69	0.25
$z \ f_s \ q_t \ u_2 \ CG$	FSB	91.66	0.90
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	97.01	0.84

MA of 97.61% with a SD of 0.30%. Although a higher MA was obtained for this specific case the same was not observed for other feature combinations, leading to the conclusion that maintaining class 0 does not severely compromises performance for DWKNN.

4.2.2 DT

Results obtained with DT are shown in Table 4.7. Comparing DWKNN to DT for all tested input combinations using the test of Wilcoxon shows statistical prevalence of DT over DWKNN with a p -value of 0.20%. One can observe that all input combinations that produced MA above 97% include the normalizations proposed by Robertson (2016) (F_r , Q_{tn} and U_2) combined with z . Raw input combinations produced accuracies above 90%, reinforcing the conclusion draw for DWKNN that these raw inputs could be used for soil classification.

The importance of each input feature for all combinations was evaluated by usage and by split. The feature importance is measured by split by counting the number of nodes associated to the feature. On the other hand, the feature importance is measured by usage by counting the percentage of objects that use the feature to be classified. Even though all features seem to have approximately the same importance when evaluated by usage, some relevant differences were observed when comparing features by split.

Pore pressure u_2 was identified as less important than z , q_t and f_s , including when combined with CG . This can be explained by fact that u_2 is not used to calculate F_r or Q_{tn} (see Section 2.1), which are the main parameters used for both ISG and FSB classification methods. The same explanation suits the low importance observed for B_q when normalizations proposed by Robertson (1990) are used and the low importance observed for U_2 when normalizations proposed by Robertson (2016) are used.

For every feature combination containing the pair Q_{tn} and F_r they are the most im-

portant, as expected. It can be seen that z is the most important for raw inputs with FSB outputs, with or without geological age, and for the normalizations proposed by Robertson (1990) for ISG outputs. It is important to highlight that it was not expected that z would be more important than F_r in this last case.

The same investigation proposed within the Section 4.2.1 is here repeated for DT, evaluating the mean confusion matrix corresponding the case of biased inputs associated with FSB. This matrix is presented below:

$$\mathbf{C} = \begin{bmatrix} 501.6 & 37.2 & 33.8 & 1.4 & 3.7 & 2.2 & 0 & 7.6 \\ 39.5 & 627.8 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 51.2 & 0.1 & 1010.9 & 1.5 & 1.6 & 0 & 0 & 0 \\ 16 & 0 & 2.2 & 902.3 & 0.1 & 3.4 & 0 & 0 \\ 11.9 & 0 & 2 & 0 & 393.3 & 1.3 & 1.7 & 0 \\ 22.1 & 0 & 0 & 6.2 & 1.9 & 662.5 & 0 & 12.1 \\ 5.6 & 0 & 0 & 0 & 1.2 & 0 & 272.9 & 11.9 \\ 33.5 & 0 & 0 & 0 & 0.2 & 1.8 & 2.7 & 3864.6 \end{bmatrix} \quad (4.2)$$

One can notice that class 0 also confuses this technique. Removing this class, it is obtained a mean accuracy of 99.20% with a standard deviation of 0.18%. Tests with all other feature combinations showed that removing class 0 contributes to improve performance, except when CG is included as an input.

4.2.3 RF

A number of 500 trees were used to produce the results presented in Table 4.8. The statistical test of Friedman identified that DWKNN, DT and RF have different performance with a p -value of 0.06%. The statistical test of Wilcoxon showed that RF is equivalent to DT with a p -value of 43.16%. Nevertheless, DT and RF are superior to DWKNN with the p -value of 0.20%.

All input combinations that include F_r and Q_{tn} as inputs resulted MA above 97%, except for biased inputs and FSB outputs. Raw inputs produced MA above 90% for all cases, as obtained for DWKNN and DT. Introducing CG as an input did not produce significant performance changes.

RF was also employed to evaluate input feature importances, using the mean decrease Gini as a reference. It was observed that, for raw inputs, u_2 is the least important. The combined pairs q_t and f_s , Q_{t1} and F_r and Q_{tn} and F_r are clearly more important than any other input features. The importance of z was reduced for RF, when compared to DT.

TABLE 4.8 – Results obtained with RF.

Input	Output	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	91.53	0.69
	FSB	91.43	0.41
$z \ F_r \ Q_{t1} \ B_q$	ISG	96.09	0.52
	FSB	96.38	0.31
$z \ F_r \ Q_{tn} \ U_2$	ISG	97.44	0.40
	FSB	97.27	0.24
$F_r \ Q_{tn}$	ISG	97.31	0.48
$F_r \ Q_{tn} \ U_2$	FSB	94.63	0.19
$z \ f_s \ q_t \ u_2 \ CG$	FSB	91.78	1.21
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	97.31	0.58

As performed for DWKNN and DT, the mean confusion matrix corresponding the case of biased inputs associated with FSB is here considered for RF. This matrix is:

$$\mathbf{C} = \begin{bmatrix} 506.9 & 36.4 & 34.8 & 1.2 & 3.1 & 2.3 & 0.1 & 10.4 \\ 38.3 & 628.3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 48.2 & 0.4 & 1007.3 & 2.2 & 1 & 0.1 & 0 & 0 \\ 15.8 & 0 & 3.6 & 899.9 & 0 & 3.6 & 0 & 0 \\ 12.1 & 0 & 2.6 & 0.1 & 394.9 & 2.1 & 2.1 & 0.1 \\ 24.6 & 0 & 0 & 8 & 2 & 662.6 & 0.1 & 40.3 \\ 5.6 & 0 & 0 & 0 & 1 & 0 & 273.7 & 36.4 \\ 29.9 & 0 & 0 & 0 & 0 & 0.5 & 1.3 & 3809 \end{bmatrix} \quad (4.3)$$

One can observe that class 0 confuses the technique. Removing the FSB class 0 it is obtained $MA = 98.82\%$ and $SD = 0.14\%$, which implies performance improvement. After repeating all other experiments with the FSB class 0 removal, it was concluded that, in general, there is a significant performance gain. DT was again compared to RF after class 0 was removed with the statistical test of Wilcoxon, concluding that they remain equivalent with a p -value of 18.75%.

4.2.4 ANN

The results for the mean accuracies and their standard deviations obtained with ANN without log-scaling are presented in Table 4.9, where n is the number of neurons in the first hidden layer and m is the number of neurons in the second hidden layer. These values are calibrated within the 10-fold cross-validation procedure described in Section 3.3, choosing the model architecture with best predictive performance. One can observe that, every time the same inputs are used, the MA obtained with of ISG is higher than

TABLE 4.9 – Results obtained with ANN without log-scaling.

Input	Output	n	m	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	16	13	88.36	1.71
	FSB	13	16	84.66	1.22
$z \ F_r \ Q_{t1} \ B_q$	ISG	12	18	89.55	2.22
	FSB	13	10	80.61	6.56
$z \ F_r \ Q_{tn} \ U_2$	ISG	15	15	95.24	1.08
	FSB	14	13	85.92	3.43
$F_r \ Q_{tn}$	ISG	18	18	96.04	1.34
$F_r \ Q_{tn} \ U_2$	FSB	9	15	85.16	4.57
$z \ f_s \ q_t \ u_2 \ CG$	FSB	16	8	84.56	1.84
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	15	15	89.98	3.33

TABLE 4.10 – Results obtained with ANN without log-scaling and without the FSB class 0.

Input	n	m	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	11	14	87.43	1.24
$z \ F_r \ Q_{t1} \ B_q$	13	11	89.64	3.55
$z \ F_r \ Q_{tn} \ U_2$	13	13	95.84	1.61
$F_r \ Q_{tn} \ U_2$	14	6	95.39	2.18
$z \ f_s \ q_t \ u_2 \ CG$	10	6	86.65	1.63
$z \ F_r \ Q_{tn} \ U_2 \ CG$	11	5	94.96	1.55

the one obtained with FSB.

The confusion matrix C_{FSB} for the FSB biased inputs reveals that its lower performance is mainly due to class 0. This matrix is given below:

$$\mathbf{C}_{FSB} = \begin{bmatrix} 337.2 & 119.9 & 88.2 & 1.2 & 2.2 & 1.2 & 0.1 & 12.4 \\ 103.3 & 488.8 & 5.2 & 0 & 8.4 & 0 & 0 & 0 \\ 110.2 & 24.9 & 886.2 & 17.1 & 4.4 & 0.2 & 0 & 0 \\ 18.9 & 0.4 & 48.8 & 877.3 & 0.2 & 13.4 & 0 & 0.1 \\ 24.6 & 31.1 & 19.2 & 0 & 366.5 & 5.2 & 7.5 & 0.1 \\ 36.4 & 0 & 1.7 & 15.6 & 13.3 & 649.6 & 0.9 & 55.4 \\ 6.4 & 0 & 0 & 0 & 7 & 0 & 257.8 & 61.2 \\ 44.4 & 0 & 0 & 0.2 & 0 & 1.6 & 11 & 3767 \end{bmatrix}$$

Results shown in Table 4.10 were obtained by removing FSB class 0 from the dataset and repeating only the experiments that make use of FSB outputs. Comparing to the corresponding results from Table 4.9, one can observe that performance was significantly increased for all cases.

To evaluate if the log-scaling impacts predictive performance, all experiments from Tables 4.9 and 4.10 were repeated with the inclusion of log-scaling. The result can be

TABLE 4.11 – Results obtained with ANN using log-scaling.

Input	Output	n	m	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	16	10	91.57	1.22
	FSB	14	15	85.88	1.03
$z \ F_r \ Q_{t1} \ B_q$	ISG	13	9	93.59	0.59
	FSB	12	16	84.39	3.33
$z \ F_r \ Q_{tn} \ U_2$	ISG	14	0	96.56	0.49
	FSB	16	13	86.18	3.44
$F_r \ Q_{tn}$	ISG	15	13	96.48	0.84
$F_r \ Q_{tn} \ U_2$	FSB	8	13	82.85	1.09
$z \ f_s \ q_t \ u_2 \ CG$	FSB	16	10	87.79	2.20
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	14	15	89.96	1.99

TABLE 4.12 – Results obtained with ANN using log-scaling and without the FSB class 0.

Input	n	m	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	12	12	88.56	0.82
$z \ F_r \ Q_{t1} \ B_q$	14	13	94.18	1.92
$z \ F_r \ Q_{tn} \ U_2$	7	0	97.45	0.79
$F_r \ Q_{tn} \ U_2$	12	8	97.08	0.17
$z \ f_s \ q_t \ u_2 \ CG$	14	0	89.63	1.40
$z \ F_r \ Q_{tn} \ U_2 \ CG$	8	0	96.40	1.53

seen in Tables 4.11 and 4.12.

Comparing Table 4.9 to Table 4.11 and Table 4.10 to Table 4.12, one can observe that log-scaling improved predictive performance for most cases. One can also notice that $m = 0$ for some cases, which indicates that the calibrating procedure lead to an architecture with only one hidden layer.

All experiments that used only the raw inputs z , f_s , q_t and u_2 resulted MA above 84%, which can be considered a reasonable predictive performance within geotechnical applications. This suggests that these inputs are sufficient for the model to distinguish between different soil types.

Any of the produced models obtained within each calibration procedure can be reproduced using the weights obtained for every connection between neurons, combined to the sigmoid functions described in Section A.3. Considering $\mathbf{W}_k$ a matrix containing the weights between layers $k - 1$ and k , Θ_k the vector of biases of the layer k , $\mathbf{s}_0$ a vector containing inputs and g its sigmoid activation function, one can write:

$$\mathbf{s}_k = g \left(\mathbf{W}_k^T \cdot \mathbf{s}_{k-1} + \Theta_k \right) \quad (4.4)$$

This expression enables reproducing any of the trained models with a simple spread-

TABLE 4.13 – Results obtained with SVM without log-scaling.

Input	Output	d	δ	κ	MA (%)	SD (%)
$z f_s q_t u_2$	ISG	6	1	9	91.75	1.65
	FSB	10	2	4	85.44	0.54
$z F_r Q_{t1} B_q$	ISG	7	2	9	94.44	1.08
	FSB	10	2	4	90.46	0.55
$z F_r Q_{tn} U_2$	ISG	6	2	10	97.31	0.55
	FSB	11	1	3	92.81	0.44
$F_r Q_{tn}$	ISG	6	2	12	97.65	0.22
$F_r Q_{tn} U_2$	FSB	12	1	3	89.63	0.46
$z f_s q_t u_2 CG$	FSB	7	2	7	86.50	1.84
$z F_r Q_{tn} U_2 CG$	FSB	7	2	7	94.54	0.94

TABLE 4.14 – Results obtained with SVM without log-scaling and without the FSB class 0.

Input	d	δ	κ	MA (%)	SD (%)
$z f_s q_t u_2$	9	1	4	88.67	0.53
$z F_r Q_{t1} B_q$	8	3	8	97.23	0.47
$z F_r Q_{tn} U_2$	9	2	5	99.31	0.13
$F_r Q_{tn} U_2$	11	2	4	99.54	0.16
$z f_s q_t u_2 CG$	5	1	8	89.77	1.15
$z F_r Q_{tn} U_2 CG$	5	1	7	97.24	0.82

sheet, obtaining the soil classification by comparing the final values obtained by each neuron of the last layer. The neuron with the highest output indicates the soil type obtained by the model. The matrices containing the weights of the models were omitted from this text for a sake of conciseness.

4.2.5 SVM

The objective of this application is repeating all experiments of the previous section with SVM and compare its the predictive performance with ANN. Preliminary tests using Gaussian, Laplacian and Polynomial kernels have shown that the performance of the last is significantly superior to the performance of the other two for the datasets used in this work. Therefore, only results using polynomial kernels are presented throughout this section.

Applying SVM to the same conditions described for Table 4.9 resulted the values presented in Table 4.13. In this case, three parameters need calibration within each experiment, d , δ and κ . They refer to the polynomial kernel, as described in Section A.4. As observed for ANN, the predictive performance with ISG overcomes the predictive performance with FSB for the same inputs.

TABLE 4.15 – Results obtained with SVM using log-scaling.

Input	Output	d	δ	κ	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	6	2	4	92.63	0.63
	FSB	9	1	3	87.57	0.30
$z \ F_r \ Q_{t1} \ B_q$	ISG	6	2	6	96.47	0.67
	FSB	9	1	3	91.24	0.65
$z \ F_r \ Q_{tn} \ U_2$	ISG	5	1	8	97.97	0.33
	FSB	10	1	2	92.77	0.63
$F_r \ Q_{tn}$	ISG	7	2	8	98.03	0.20
$F_r \ Q_{tn} \ U_2$	FSB	9	2	4	89.24	0.54
$z \ f_s \ q_t \ u_2 \ CG$	FSB	6	2	6	89.75	1.20
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	7	1	5	94.44	0.86

TABLE 4.16 – Results obtained with SVM using log-scaling without the FSB class 0.

Input	d	δ	κ	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	7	1	5	90.05	0.38
$z \ F_r \ Q_{t1} \ B_q$	8	1	6	98.24	0.21
$z \ F_r \ Q_{tn} \ U_2$	8	1	5	99.47	0.12
$F_r \ Q_{tn} \ U_2$	8	1	6	99.66	0.08
$z \ f_s \ q_t \ u_2 \ CG$	5	1	4	91.06	1.08
$z \ F_r \ Q_{tn} \ U_2 \ CG$	4	1	4	97.71	0.92

Evaluating the confusion matrix C_{FSB} for the FSB biased inputs enables identifying class 0 as the main source of confusion again, which motivated producing Table 4.14. This matrix is given below:

$$\mathbf{C}_{FSB} = \begin{bmatrix} 371.8 & 109 & 140.4 & 7.1 & 5.2 & 2.2 & 0.4 & 18.4 \\ 118.8 & 553.4 & 4.9 & 0 & 0.1 & 0 & 0 & 0 \\ 88.4 & 2.5 & 894.7 & 6.4 & 0.5 & 0 & 0 & 0 \\ 17.5 & 0 & 8.6 & 894.9 & 0 & 1 & 0 & 0 \\ 19.5 & 0.2 & 0.7 & 0 & 394.4 & 0.7 & 0.5 & 0 \\ 24 & 0 & 0 & 3 & 0.8 & 667.2 & 0 & 14.5 \\ 6.1 & 0 & 0 & 0 & 1 & 0 & 276.2 & 13.4 \\ 35.3 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 3849.9 \end{bmatrix}$$

Comparing Tables 4.13 and 4.14 enables concluding that significant improvement is achieved for the predictive performance when class 0 is suppressed. Furthermore, SVM obtains MA close to 100% with SD close to zero in some cases. Next step is including the log-scaling, which produced the results shown in Tables 4.15 and 4.16.

The improvements in predictive performance observed for SVM when log-scaling is included are similar to those observed for ANN. We highlight that even MA values close to 100% in Table 4.14 are increased in Table 4.16.

TABLE 4.17 – Results obtained with MMP.

Input	Output	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	ISG	93.17	0.71
	FSB	91.82	0.37
$z \ F_r \ Q_{t1} \ B_q$	ISG	96.64	0.60
	FSB	96.15	0.30
$z \ F_r \ Q_{tn} \ U_2$	ISG	98.25	0.34
	FSB	97.08	0.21
$F_r \ Q_{tn}$	ISG	97.95	0.24
$F_r \ Q_{tn} \ U_2$	FSB	94.66	0.18
$z \ f_s \ q_t \ u_2 \ CG$	FSB	93.23	0.82
$z \ F_r \ Q_{tn} \ U_2 \ CG$	FSB	97.42	0.54

Comparing the result obtained with ANN for each experiment to the corresponding one with SVM, one can observe that the MA obtained with SVM is higher for all cases. Furthermore, results obtained with SVM reinforce that using only raw input features for classification, named z , f_s , q_t and u_2 , is enough to differentiate classes in feature space.

A quantity of 625 objects were obtained as support vectors for the ISG method and 9388 support vectors were obtained for the FSB method.

4.2.6 MMP evaluation

The results obtained with the MMP are presented in Table 4.17, with each line representing a test or computational experiment. First column presents the used input features, according to the strategy defined in Section 3.3. The second column presets the replicated method, ISG (Section 2.1) or FSB (Section 2.2). The third and fourth columns present the predictive performance obtained within each experiment, with MA meaning mean accuracy and SD referring to accuracy standard deviation.

One can observe that good predictive performance was achieved with the MMP for all experiments, with $MA > 90\%$. This values are specially interesting for the raw input features, which shows that it is possible to reach accurate soil classification without the data transformations proposed by Robertson. It is also interesting to observe that the highest MA values were not obtained for the biased combinations, as expected. Using z , F_r , Q_{tn} and U_2 overwhelmed the predictive performance of the biased combinations for both ISG and FSB classification methods. In addition to that, including CG among these four inputs resulted even better accuracy for replicating FSB. The same is observed when including CG among raw inputs. This suggests that depth and geological age bring relevant information for soil classification, even though they are not directly used in ISG and FSB methods.

Results presented in previous sections have shown that objects assigned to the class 0 of the FSB method harm the predictive performance of ANN and SVM. Nevertheless, this accuracy loss was not observed for the MMP, which presented the highest MA in most experiments. For a more rigorous comparison between all five ML techniques employed plus the MMP, the Friedman test was employed. Its result is represented in Figure 4.1. In addition to that, the signed-rank statistical test of Wilcoxon have shown that the MMP is equivalent to the boosted DT with a p -value of 10.55% and to the RF with a p -value of 23.24%, and that SVM is equivalent to the DWKNN with a p -value of 92.19%.

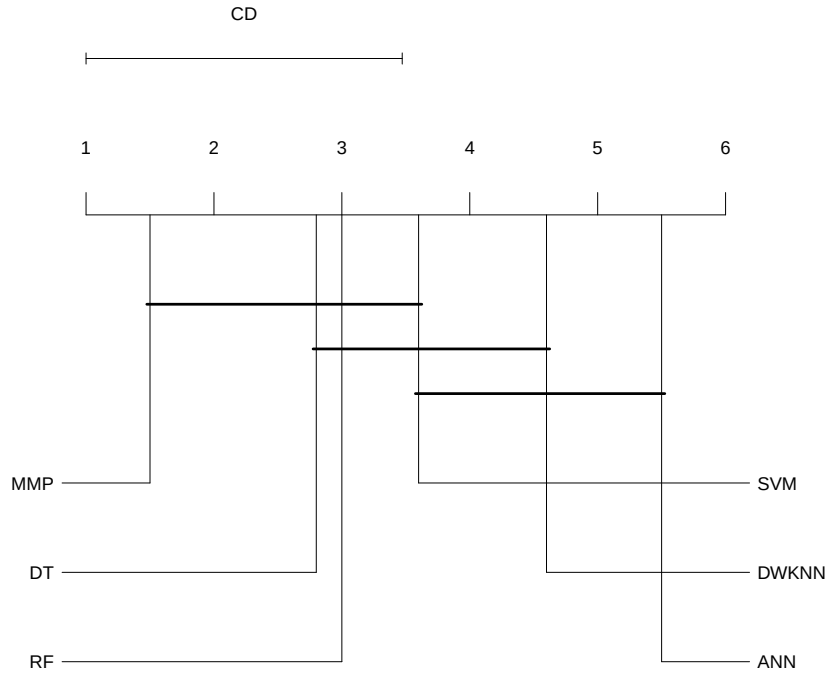


FIGURE 4.1 – Comparison between DWKNN, DT, RF, ANN, SVM e MMP using the Friedman statistical test.

In order to quantify the impact of the objects assigned to the class 0 of the FSB method in accuracy, additional experiments were performed removing these objects from the training and test sets, resulting the values presented in Table 4.18. Once it is expected that these objects would impact only results for the FSB method, results obtained for the ISG method were omitted.

One can observe that a higher MA is achieved for most cases, including values close to 100%. This suggests that objects assigned to the class 0 of the FSB method do not form an homogeneous region within input feature space. It is important to notice that an opposite evidence was obtained when only DWNN was considered, in Section 4.2.1. All techniques were again compared using the Friedman statistical test, which is illustrated in Figure 4.2. The signed-rank statistical test of Wilcoxon showed that MMP is superior or equivalent to DT and RF with a p -value of 15.63%.

TABLE 4.18 – Results obtained with MMP without the FSB class 0.

Input	MA (%)	SD (%)
$z \ f_s \ q_t \ u_2$	92.62	0.39
$z \ F_r \ Q_{t1} \ B_q$	98.55	0.18
$z \ F_r \ Q_{tn} \ U_2$	99.41	0.15
$F_r \ Q_{tn} \ U_2$	99.60	0.12
$z \ f_s \ q_t \ u_2 \ CG$	92.37	1.19
$z \ F_r \ Q_{tn} \ U_2 \ CG$	98.35	0.63

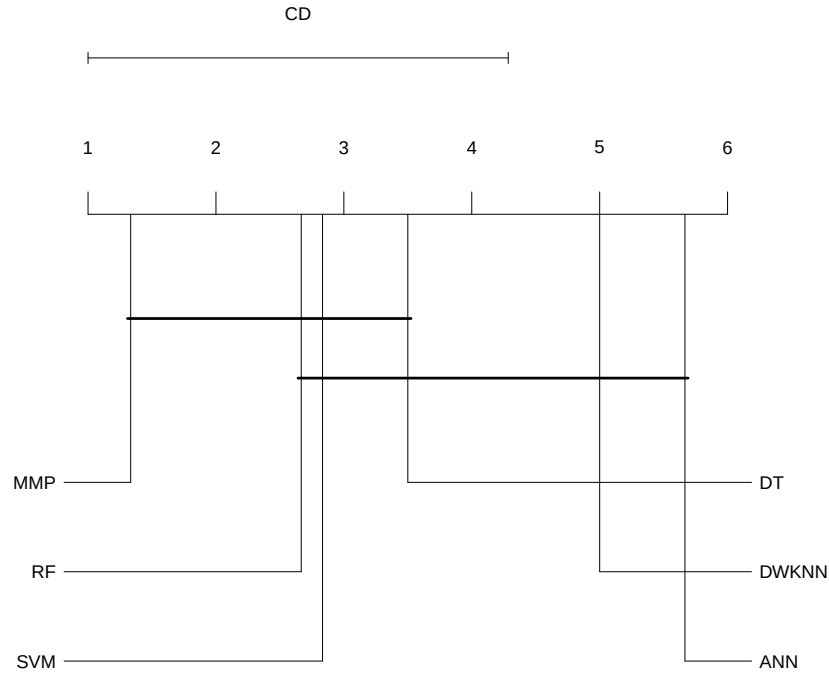


FIGURE 4.2 – Comparison between DWKNN, DT, RF, ANN, SVM and MMP without FSB class 0 by the Friedman statistical test.

4.2.7 MMP application and validation

The results obtained with all techniques using the main dataset (see Section 3.1.1) as training set and the subway dataset (see Section 3.1.2) as test set are shown in Table 4.19. The traditional accuracy was adopted as performance measure. One can observe that, even though the multiple model predictor is not the best-performing technique for all testing combinations, its performance is always at least close to the best one. This shows that MMP is stable, while big performance variations can be observed for all other techniques. This is confirmed by the Friedman statistical test, as presented in Figure 4.3. If the mean recall is considered instead of the traditional accuracy, the MMP keeps the best general performance.

TABLE 4.19 – Results for all techniques.

Outputs	Inputs	DWKNN	DT	RF	ANN	SVM	MMP
ISG	z, q_t, f_s, u_2	59.22	68.7	62.92	74.68	77.2	71.29
	z, Q_{t1}, F_r, B_q	74.16	88.03	88.27	85.55	85.84	87.94
	z, Q_{tn}, F_r, U_2	81.45	89.94	89.8	92.98	89.86	91.5
	Q_{tn}, F_r	89.86	88.2	94.59	93.57	93.84	93.61
FSB	z, q_t, f_s, u_2	71.13	79.5	79.49	82.82	82.68	82.01
	z, Q_{t1}, F_r, B_q	80.36	92.98	92.76	88.68	92.86	92.99
	z, Q_{tn}, F_r, U_2	91.24	95.87	95.78	92.41	95.78	95.87
	Q_{tn}, F_r, U_2	95.33	95.47	94.21	92.5	96.13	95.94

TABLE 4.20 – Results using z, q_c and f_s as input features.

Technique	ISG		FSB		FSB (no 0)	
	MA	SD	MA	SD	MA	SD
DWKNN	90.30	0.65	86.79	0.47	88.12	0.72
DT	90.13	0.56	87.35	0.45	88.99	0.40
RF	90.31	0.60	87.84	0.35	89.37	0.32
ANN	90.18	0.85	82.86	0.64	85.82	1.08
SVM	91.04	0.70	82.44	0.48	85.58	0.82
MMP	91.83	0.56	87.58	0.32	89.15	0.30

4.2.8 Classification without pore pressure information

Additional experiments were performed to verify if the friction penetrometer without pore pressure measurements could provide enough information for an accurate soil classification for general purposes. It has to be highlighted that the pore pressure information is important when the cone resistance correction is significant in comparison to its value, what occurs for soft clays, which are soils of interest when using CPT. For this reason, it is recommended to use all information available for classification, leaving this simplified application only for cases for which more information is not available or it is not reliable.

Therefore, all techniques plus the MMP were tested with the main dataset using only z, q_c and f_s as input features, resulting the values presented in Table 4.20. The MA and SD were obtained for replicating the ISG method, the FSB method with the objects assigned to the class 0 of the FSB method and for the FSB method without these objects, resulting three experiments for each technique.

One can notice that all techniques achieved accuracy higher than 90% for the ISG method, which can be considered good for the purpose of soil profiling. Although lower accuracies were obtained for the FSB method, values can also be considered reasonable, specially when objects assigned to the class 0 are removed. These results show that soil can be classified within reasonable accuracy with CPT data that do not include the pore pressure measurements.

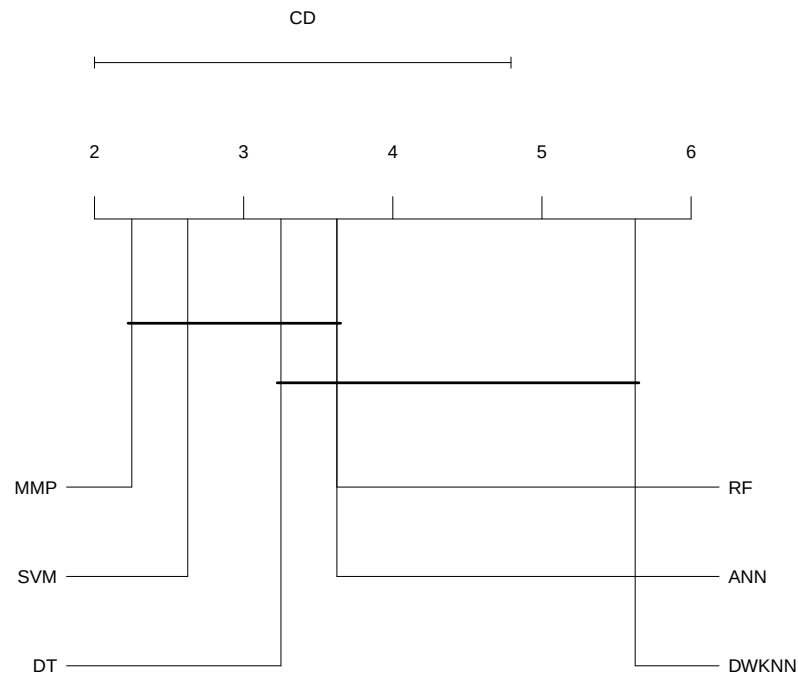


FIGURE 4.3 – Comparison between DWKNN, DT, RF, ANN, SVM and MMP by the Friedman statistical test.

4.3 Clustering approaches

In this Section, the results achieved by the methodology in Section 3.5 are presented and discussed.

4.3.1 General clustering approach

The kernel k-means algorithm was applied with k varying from 7 to 10 with the full dataset and all CPT non-normalized parameters: z (m), q_c (MPa), f_s (kPa) and u_2 (kPa). The lowest total internal cluster variance was obtained for $k = 9$ behavioral classes (BC). The clusters centers are presented in Table 4.21.

In Tables 4.22 and 4.23 the clustering results are compared to the ISG and FSB labels, respectively, which were obtained using the CPeT-IT software. In these tables, lines represent clustering classes and columns represent chart-based methods. One can observe a subtle similarity between clustering results and the chart-based methods. For ISG classes 1, 2, 7, 8 and 9, there is a concentration of the objects in some specific clustering classes and similar correspondences also occur for FSB classes 1, 2, 3, 4 and 6.

The confusion matrix presented in Table 4.24 was produced to compare ISG classes obtained via CPT with the sample observations obtained via SPT sampling. We were

TABLE 4.21 – 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

TABLE 4.22 – Comparison between the k-means clustering and the ISG classes.

Class	1	2	3	4	5	6	7	8	9
1	0	222	11747	8746	3708	642	1	41	695
2	533	691	9789	4943	3527	866	3	0	4
3	6	64	1946	1477	2418	4195	16	0	0
4	0	0	17	103	476	10943	87	6	3
5	0	1	14	155	2270	15901	170	0	0
6	0	1	1785	2642	2159	4411	7	168	383
7	0	60	3649	2468	2395	1924	91	855	1660
8	1131	133	9826	2751	753	18	0	0	0
9	207	230	879	822	1163	566	6	2	42

TABLE 4.23 – Comparison between the k-means clustering and the FSB classes.

Class	0	1	2	3	4	5	6	7
1	146	287	1876	14235	245	5999	13	3001
2	2418	2353	5349	2893	1943	1929	1757	2034
3	239	370	1703	111	1049	426	1840	4564
4	90	0	12	40	17	179	5	11292
5	34	2	18	2	83	105	417	17858
6	118	2	85	3082	34	2422	4	5810
7	1031	502	2355	3226	848	1987	65	3133
8	2813	5408	3780	253	1563	572	204	210
9	1310	435	237	488	454	193	831	577

TABLE 4.24 – Comparison between ISG classes and the SPT observations.

Class	1	2	3
2	8	0	16
3	239	131	162
4	294	65	141
5	584	160	193
6	580	102	161
7	1	0	0
8	0	0	4
9	1	0	0

TABLE 4.25 – Comparison between FSB classes and the SPT observations.

Class	1	2	3
0	7	0	1
1	3	0	17
2	47	6	68
3	244	151	130
4	85	17	26
5	222	40	100
6	128	55	60
7	976	189	275

unable to identify any correspondence between these two soil classifications, which seem to be completely uncorrelated. FSB classes were also compared to SPT sampling, resulting values presented in Table 4.25. A clearer correspondence is found, with those soil types identified by the FSB method matching reasonably with those observed via SPT sampling.

The clustering results were also compared with SPT sample observations, resulting Table 4.26. It is important to highlight that the SPT sand class (1) includes silty sands, clayey sands and fine sands mixed with silt or clay. Therefore, some clayey or silty behavior can be observed for this class. Likewise, the SPT silt (2) and clay (3) classes have some mixed soil types that may not behave like expected for the main soil type identified within the visual-tactile observation. When Tables 4.24 and 4.26 are compared, one can observe that clustering classes are more correlated to SPT samples than ISG classes. Once SPT classes were defined considering soil types with clearly distinguishable behaviors, this result suggests that ISG classes are more influenced by soil granulometry than the ones here obtained by clustering.

To complement the analysis, each line in Table 4.26 was divided by its sum, resulting the values presented in Table 4.27. One can observe in this table that the BC 5, for example, is composed by sand and clay in a proportion of approximately 79% to 21%. Further studies are needed to explain why some classes, like BC 3, are balanced between sand and clay samples. One way to investigate this issue would be considering the clusters

TABLE 4.26 – Comparison between the k-means clustering and the SPT observations.

Class	1	2	3
1	374	140	189
2	515	181	141
3	207	61	168
4	11	0	17
5	283	0	74
6	117	0	52
7	48	63	10
8	44	0	26
9	113	13	0

TABLE 4.27 – Contribution of each soil type for the BC.

Class	1	2	3
1	53.20	19.91	26.88
2	61.53	21.62	16.85
3	47.48	13.99	38.53
4	39.29	0.00	60.71
5	79.27	0.00	20.73
6	69.23	0.00	30.77
7	39.67	52.07	8.26
8	62.86	0.00	37.14
9	89.68	10.32	0.00

centers presented in Table 4.21, which can provide clues about soil stiffness and strength.

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 (mean recall) of 89.35% with a standard deviation of 0.40%, which can be considered reasonable for soil classification. The correspondent weight matrices and bias vectors are presented below. These parameters define an ANN architecture with only one hidden

layer with 18 neurons.

$$\mathbf{W}_1^T = \begin{bmatrix} -37.52 & -6.73 & -63.68 & -0.86 \\ -32.35 & -32.11 & -11.07 & -6.07 \\ -5.77 & -4.42 & 23.97 & 28.13 \\ 16.19 & 13.54 & -22.40 & 5.78 \\ 18.21 & -10.62 & 4.36 & -11.41 \\ -3.25 & -52.91 & 22.30 & 9.76 \\ -48.02 & 5.77 & 10.28 & -4.93 \\ -30.92 & 22.56 & -4.56 & -57.45 \\ 19.06 & 10.05 & -0.89 & 21.55 \\ 30.78 & 23.89 & 16.36 & 17.76 \\ -0.60 & 14.52 & 43.70 & -5.50 \\ -15.83 & -47.19 & -16.83 & -7.09 \\ 17.61 & -40.56 & -2.47 & 9.86 \\ 14.31 & 25.76 & -41.13 & 18.37 \\ -27.97 & -3.40 & 22.12 & 6.23 \\ -29.83 & 2.67 & -14.79 & 0.73 \\ -9.44 & -16.09 & -38.84 & 1.03 \\ 11.12 & 23.61 & 26.66 & -6.29 \end{bmatrix}, \Theta_1 = \begin{bmatrix} 17.18 \\ 29.27 \\ -17.27 \\ -12.90 \\ -8.51 \\ 11.71 \\ 4.30 \\ 33.86 \\ -23.10 \\ -51.37 \\ -20.97 \\ 43.08 \\ 2.99 \\ -4.45 \\ -1.07 \\ 13.27 \\ 21.54 \\ -35.21 \end{bmatrix} \quad (4.5)$$

$$\mathbf{W}_2 = \begin{bmatrix}
-9.07 & -10.03 & -3.54 & -3.56 & -8.00 & -2.55 & -5.56 & 5.57 & 14.64 \\
-0.68 & 7.25 & -8.44 & -4.19 & -7.54 & -5.37 & -0.41 & -3.25 & 5.31 \\
-2.75 & -14.35 & 3.39 & -2.14 & -4.59 & 2.54 & 5.41 & -3.46 & -11.20 \\
-9.82 & -4.26 & -0.84 & -3.63 & 3.74 & -12.72 & 0.18 & 5.94 & -3.72 \\
-7.34 & -25.18 & 5.53 & -8.15 & -6.80 & 3.96 & 10.67 & -5.95 & -7.38 \\
1.15 & -4.67 & -1.63 & -17.89 & -3.28 & -2.10 & -2.28 & 4.73 & 6.43 \\
-5.88 & -16.88 & -10.16 & -13.45 & -9.10 & -9.10 & 4.16 & -3.29 & 4.97 \\
12.16 & 5.29 & 11.82 & 7.06 & 2.65 & 5.65 & -6.58 & -8.27 & 3.80 \\
-10.32 & -20.29 & 0.50 & 3.28 & -1.03 & 0.51 & -2.01 & -0.37 & -9.03 \\
-6.10 & -4.64 & -9.86 & 5.80 & -5.62 & -9.33 & 4.77 & -15.09 & -3.46 \\
3.31 & -18.45 & -6.68 & -1.98 & 8.73 & 1.99 & 1.95 & -0.66 & -6.28 \\
8.24 & 2.40 & 5.10 & -8.06 & -5.17 & -6.51 & -3.83 & 3.37 & -1.32 \\
-3.46 & 5.36 & 1.46 & -7.96 & -9.77 & 2.20 & 0.46 & -0.14 & -7.08 \\
-8.07 & -3.55 & 5.02 & 0.69 & 6.39 & -4.22 & 0.78 & 1.47 & 4.90 \\
-1.76 & 10.70 & -14.67 & -4.51 & -9.86 & 0.82 & 4.27 & -3.43 & -7.43 \\
-3.83 & -4.71 & 3.95 & -3.56 & -3.19 & 3.08 & 10.31 & -8.66 & 3.38 \\
-26.20 & 0.52 & -4.46 & -5.03 & -25.74 & -2.74 & -1.85 & 0.88 & -6.32 \\
-3.97 & -5.02 & -11.86 & 6.00 & -6.79 & 3.07 & -7.23 & -1.99 & -2.79
\end{bmatrix}, \quad (4.6)$$

$$\Theta_2 = \begin{bmatrix}
-16.58 \\
-6.41 \\
-15.32 \\
-13.37 \\
-12.12 \\
-6.12 \\
-0.64 \\
-4.30 \\
-11.33
\end{bmatrix}$$

Another ANN model was trained using only z , q_c and f_s as input features. The objective is using the raw parameters given by the CPT without a pore pressure sensor, to evaluate if this classification method could be replicated using alternative CPT equipment. The resultant model presented an accuracy (mean recall) of 84.47% with a standard deviation of 0.30%, which is also reasonable for soil classification. In this case, the weight

matrices and their respective bias vectors are the following:

$$\mathbf{W}_1^T = \begin{bmatrix} 36.43 & -21.81 & 10.09 & -48.9 \\ 14.58 & -32.74 & 37.42 & -45.72 \\ -41.07 & -27.8 & -19.71 & -7 \\ 44.46 & 0.37 & 4.55 & 20.71 \\ -44.39 & 46.84 & -19.31 & 10.8 \\ -12.93 & -29.47 & 8.4 & 41.45 \\ -11.33 & -42.14 & 19.7 & -10.5 \\ -28.38 & -31.23 & -1.79 & 38.29 \\ -58.39 & -2.15 & -6 & 7.89 \\ -18.16 & 47.53 & -32.66 & -27.07 \\ -29.1 & 24.8 & 17.51 & -8.61 \\ 22.38 & -7.67 & 38.99 & 8.33 \\ 5.54 & -19.75 & -26.34 & -23.35 \\ 1.77 & 21 & -33.22 & 26.82 \\ -30.06 & -28.69 & -1.89 & -10.11 \\ 23.66 & 45.23 & -20.54 & 0.57 \\ -23.39 & 0.73 & -10.51 & -42.96 \\ 44.43 & -13.25 & 17.56 & 26.13 \end{bmatrix}, \Theta_1 = \begin{bmatrix} -16.68 \\ 1.19 \\ 14.24 \\ -17.06 \\ 14.00 \\ 41.71 \\ -10.59 \\ -0.85 \\ 36.26 \\ -22.90 \\ 9.97 \\ 33.68 \\ -11.52 \\ -3.16 \\ 16.54 \\ 12.09 \end{bmatrix} \quad (4.7)$$

$$\mathbf{W}_2 = \begin{bmatrix} -11.15 & 1.69 & 4.54 & 4.16 & 9.74 & -9.49 & 8.37 & -0.59 & -5.03 & -12.33 & -4.33 & 2.18 & 1.36 & -11.2 & -6.28 & -4.67 & 0.08 & -7.17 \\ -11.22 & -21.1 & 2.7 & -0.06 & -0.41 & -4.14 & -3.98 & 15.96 & 4.49 & 15.66 & 5.67 & -0.9 & 6.39 & 14.44 & -6.25 & -2.58 & 5.03 & -2.8 \\ 6.28 & -4.77 & 5.16 & -4.2 & 3.47 & -0.64 & -1.09 & -10.85 & 3.64 & 0.85 & -18.36 & -7.2 & -4.97 & -0.68 & -1.95 & -10.98 & -0.96 & 2.72 \\ -8.17 & -0.56 & 6.23 & -15.1 & -12.12 & -12.96 & -1.79 & -5.1 & -7.51 & 3.85 & 6.93 & -6.14 & -1.89 & -20.33 & 3.3 & 6 & -17.07 & -7.91 \\ 12.6 & -3.55 & 1.26 & -12.08 & -3.75 & 1.03 & -0.21 & -5.33 & -0.02 & -8.68 & 0.8 & -3.15 & -5.26 & -10.54 & -2.88 & -17.29 & -0.41 & 1.56 \\ 1.54 & 1.07 & 3.04 & -3.27 & -7.61 & 13.25 & -7.57 & -4.27 & 14.14 & -6.33 & -15.04 & -4.15 & 1.5 & 0.25 & 1.97 & 6.67 & 4.37 & 6.25 \\ 13.55 & -5.18 & 6.78 & -3.98 & 9.49 & -8.26 & -1.79 & -7.18 & 5.51 & 7.97 & -3.76 & -9.86 & -0.23 & 15.84 & -0.69 & 3.58 & 3.85 & -9.76 \\ 14.17 & 2.89 & -6.53 & 8.72 & -2.33 & 0.85 & -19.42 & -1.4 & 2.73 & -8.32 & -3.28 & -0.14 & -0.49 & -8.36 & -1.4 & 2.51 & -5.8 & 5.81 \\ 8.6 & -7.39 & 7.62 & -3.42 & 0.33 & 5.49 & -5.82 & 4.07 & -9.79 & -10.97 & -4.58 & -6.63 & 5.8 & 7.69 & 8.34 & -12.23 & 12.63 & -6.85 \\ -3.85 & -3.82 & 1.53 & 1.21 & -2.15 & 6.03 & 11.85 & -5.02 & -15.98 & -10.69 & -9.12 & 8.52 & -0.84 & -6.17 & -0.79 & -7.34 & 0.07 & -10.6 \\ -3.98 & 1.07 & -9.19 & 16.31 & -0.41 & 10.64 & 3.76 & 6.82 & -12.64 & 1.22 & -6.65 & 12.74 & 13.1 & 2.42 & 3.24 & 3.55 & 7.48 & 4.71 \\ -4.31 & 11.44 & -2.51 & 5.38 & -3.63 & -6.21 & -0.42 & 17.23 & 3.33 & 4.4 & 10.68 & 3.52 & 8.51 & 0.71 & -4.99 & 2.6 & -2.03 & 9.86 \\ -1.45 & -2.2 & -8.79 & 2.07 & 3.64 & -3.92 & 8.23 & -0.96 & 3.62 & -6.16 & -5.27 & -4.99 & -22.36 & 14.75 & -8.71 & -6.13 & 2.94 & -8.49 \\ 6.13 & 3.47 & -14.74 & 6.24 & -5.51 & 2.26 & -2.44 & 4.26 & -11.64 & -6.61 & -0.43 & 1.56 & -5.46 & -4.49 & 11.69 & 8.87 & -16.18 & 2.92 \\ 4.84 & -12.44 & 3.66 & -1.33 & 0.79 & -1.3 & -5 & -4.98 & 4.92 & 6.52 & -7.72 & -4.98 & -5.7 & 10.95 & -6.24 & -4.01 & 11.74 & -4.28 \\ -12.32 & 8.2 & -6.69 & 9.97 & 5.6 & -5.84 & 8.83 & 1.78 & -12.96 & 7.1 & -5.19 & 12.67 & 0.58 & -7.42 & 9.78 & -3.69 & -6.25 & 0.67 \\ -9.18 & -6.64 & -4.3 & -1.96 & 14.87 & 13.87 & 1.1 & -1.58 & -4.27 & 11.32 & -4.53 & 3.37 & -5.1 & 6.64 & -11.86 & -6.12 & 14.61 & -1.93 \\ -2.85 & 8.13 & 2.37 & 19.88 & -5.91 & 3.76 & -17.13 & -9.71 & 0.79 & 2.74 & 7.64 & 1.75 & -3.86 & -3.49 & 2.05 & 4.59 & -5.69 & 4.96 \end{bmatrix}, \quad (4.8)$$

$$\Theta_2^T = \begin{bmatrix} -9.48 & -6.72 & -9.96 & -10.38 & 0.05 & -29.57 & 6.91 & 7.30 & -4.11 & -3.89 & -4.88 & -7.60 & -0.35 & 4.91 & -5.86 & 4.14 & 3.23 & -7.06 \end{bmatrix}$$

TABLE 4.28 – Case-based 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

$$\mathbf{W}_3 = \begin{bmatrix} -1.46 & -3.3 & 1.11 & 0.38 & -1.99 & -1.09 & 2.37 & 3.68 & 1.09 \\ -0.84 & -1.91 & -1.04 & -3.29 & 2.63 & 2.06 & -4.3 & -1.44 & 3.01 \\ 1.07 & -2.45 & -0.66 & 2.01 & -1.61 & 2.47 & -3.38 & -2.85 & -2.34 \\ -1.7 & -2.43 & -3.28 & -3.66 & -1.31 & 1.84 & 4.01 & -6.12 & -1.54 \\ 3.02 & 2.67 & 2.93 & -3.85 & 0.9 & -3.16 & 4.49 & 2.16 & -3.18 \\ -1.24 & -4.4 & -2.81 & 1.37 & 0.04 & -2.79 & 0.18 & 3.27 & 2.28 \\ 3.49 & 2.59 & -3.84 & -1.39 & 2.49 & -1.34 & -2.92 & -1.25 & -1.38 \\ -4.17 & -6.05 & -3.25 & -2.55 & -4.06 & -2.87 & 4.81 & -3.86 & 2.05 \\ -3.82 & 3.6 & 3.63 & -1.38 & -6.04 & -1.16 & -3.1 & -2.77 & 1.25 \\ -4.71 & 5.22 & -4.96 & -1.99 & -1.95 & -7.66 & 6.59 & -4.19 & -2.47 \\ 0.15 & -0.54 & -1.51 & -1.91 & -0.79 & 1.61 & 1.07 & -1.16 & -1.13 \\ -3.1 & -1.47 & -1.18 & -3.63 & 3.36 & -3.23 & 2.68 & 0.03 & -3.62 \\ -2.14 & -8.26 & -7.75 & -8.17 & 0.79 & -5.51 & 13.02 & -2.91 & -6.99 \\ -4.16 & 4.58 & -2.19 & -0.74 & -2.1 & -1.47 & -3.04 & -1.14 & -2.52 \\ -5.16 & -3.02 & -1.68 & 1.89 & 4.8 & 0.19 & 0.03 & -4.81 & 6.07 \\ -1.73 & -2.94 & 1.59 & -2.38 & -2.03 & 1.67 & 1.56 & -1.56 & 3.51 \\ -4.73 & 4.68 & -3.72 & -3.9 & -3.23 & -1.49 & -5.54 & 6.6 & -2.23 \\ -1.35 & -3.02 & -4.73 & -1.86 & -5.41 & -2.21 & -4.88 & -1.87 & 1.98 \end{bmatrix}, \Theta_3 = \begin{bmatrix} -6.69 \\ -8.14 \\ -4.28 \\ -5.30 \\ -2.00 \\ -6.47 \\ -5.84 \\ -8.57 \\ -5.70 \end{bmatrix} \quad (4.9)$$

These parameters define an ANN architecture with two hidden layers, with 16 neurons in the first one and 18 in the second one.

4.3.2 Case-based clustering approach

One way to improve classification accuracy is obtaining, by clustering, specific soil classes for each region of interest. This approach has the advantage of creating specialized classification methods, that tend to represent better one region separated from the other. 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 were 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 behavioral classes (BC) were found to be the best for the

TABLE 4.29 – Comparison between the case-based x-means clustering and the ISG classes.

Class	2	3	4	5	6	7	8	9
1	0	201	220	320	67	0	0	0
2	0	1	34	158	413	0	4	1
3	24	257	132	296	100	0	0	0
4	0	73	114	163	263	1	0	0

TABLE 4.30 – Comparison between the case-based x-means clustering and the FSB classes.

Class	0	1	2	3	4	5	6	7
1	1	0	33	208	27	180	68	291
2	0	0	0	10	0	59	0	542
3	2	20	80	217	63	63	131	233
4	5	0	8	90	38	60	44	374

considered dataset, with their centers presented in Table 4.28. Crossing results with the ISG and FSB classification methods, Tables 4.29 and 4.30 are obtained, respectively. Only a subtle similarity can be observed.

Crossing the results with the SPT visual-tactile observations, Table 4.31 is obtained. The same result is presented in Table 4.32, considering values relative to the total number of objects within each clustered class. Comparing to the confusion matrices in Tables 4.26 and 4.31, one can conclude that the specialized classification method is capable of better agreement with SPT sampling than the general method. This can be considered an advantage, for uniting the model capability of predicting soil behavior to a correspondence with SPT visual-tactile observations.

In the end, a ANN model was produced in order to reproduce the obtained specialized classification method. The obtained model presented very good predictive performance, with an accuracy (mean recall) 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 case-based approach are more feasible.

TABLE 4.31 – Comparison between the case-based x-means clustering and the SPT observations.

Class	1	2	3
1	423	94	291
2	448	0	163
3	452	134	223
4	389	230	0

TABLE 4.32 – Contribution (%) of each soil granulometrical type for each behavior.

Class	1	2	3
1	52.35	11.63	36.01
2	73.32	0	26.68
3	55.87	16.56	27.56
4	62.84	37.16	0

The weight matrices and their respective bias vectors for this last ANN model are:

$$\mathbf{W}_1^T = \begin{bmatrix} 10.85 & -4.93 & -0.90 & 5.63 \\ -6.63 & -11.79 & -11.30 & -4.61 \\ -23.58 & -3.76 & -5.71 & -17.35 \\ 24.06 & 15.36 & 6.63 & 24.38 \\ -6.34 & -32.29 & -18.20 & -1.50 \end{bmatrix}, \Theta_1 = \begin{bmatrix} -5.71 \\ 19.24 \\ 25.97 \\ -42.46 \\ 35.98 \end{bmatrix} \quad (4.10)$$

$$\mathbf{W}_2 = \begin{bmatrix} -7.51 & 6.63 & -2.41 & -9.80 \\ -7.44 & -2.87 & 7.86 & 6.84 \\ -10.68 & -6.62 & -23.54 & 15.41 \\ 15.99 & 4.71 & -15.83 & -7.36 \\ 19.26 & -16.87 & -5.45 & 1.73 \end{bmatrix}, \Theta_2 = \begin{bmatrix} -14.79 \\ -5.20 \\ 10.22 \\ -7.19 \end{bmatrix} \quad (4.11)$$

These parameters define an ANN architecture with only one hidden layer composed by five neurons.

In this case, it was also evaluated if suppressing pore pressure information harms 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:

$$\mathbf{W}_1^T = \begin{bmatrix} 26.01 & 7.43 & 0.63 \\ -27.72 & 0.3 & 2.88 \\ -11.63 & -17.05 & 2.1 \\ 10.45 & 12.12 & 13.9 \\ 0.37 & 0.52 & -24.2 \\ 19.4 & -24.54 & -12.07 \end{bmatrix}, \Theta_1 = \begin{bmatrix} -22.32 \\ 19.84 \\ 15.86 \\ -23.11 \\ 10.45 \\ 0.07 \end{bmatrix} \quad (4.12)$$

$$\mathbf{W}_2 = \begin{bmatrix} -1.74 & -16.37 & -14.24 & -0.92 & 0.15 \\ -3.11 & 5.96 & 15.79 & -1.36 & -5.08 \\ 6.34 & 2.68 & 6.87 & 2.68 & -16.13 \\ -3.42 & -13.38 & -4.87 & -11.16 & 11.62 \\ 4.09 & 7.71 & 7.26 & -2.25 & -3.56 \\ 9.62 & -5.48 & -8.32 & 8.63 & -17.26 \end{bmatrix}, \Theta_2 = \begin{bmatrix} 0.78 \\ -5.42 \\ 11.46 \\ -0.42 \\ -9.51 \end{bmatrix} \quad (4.13)$$

$$\mathbf{W}_3 = \begin{bmatrix} 3.14 & -6.33 & -3.29 & 0.44 \\ -6.64 & -4.56 & -6.56 & 8.56 \\ -6.35 & -2.3 & 6.44 & 2.5 \\ -4.74 & -2.06 & 4.74 & -0.92 \\ -11.77 & 9.8 & -2.39 & -6.47 \end{bmatrix}, \Theta_3 = \begin{bmatrix} 4.40 \\ -3.50 \\ -4.38 \\ -6.44 \end{bmatrix} \quad (4.14)$$

These parameters define an ANN architecture with two hidden layers, with 6 neurons in the first one and 5 in the second one.

5 Conclusions and further work

A general and optimized methodology for the application of ML techniques for soil classification from CPT data was presented and tested with six techniques: DWKNN, boosted DT, RF, ANN, SVM and MMP, which is a combination of the previous. Different calibration procedures were proposed for each technique and the classification within the MMP procedure was produced by simple majority voting. All techniques were applied initially to 111 CPT soundings which compose the so-called main dataset, testing ten different feature combinations within a 10-fold cross-validation procedure. Data was subjected to a preprocessing procedure within each 10-fold cross-validation step, including data transformation, cleaning and balancing.

In the first feature selection analysis, two distance-based techniques were tested using two classification criteria and three sets of eleven feature combinations, totalizing 132 experiments. It was concluded that including depth as a feature can improve classification for some feature combinations like non-normalized inputs, for which the performances get close to those obtained using normalized inputs, sometimes overcoming them. It suggests that depth could be used in the classification methods or in clustering analyses. Best results were obtained using ISG classes, which suggests they are better separated than the ones from FSB charts, but it is mainly due to the FSB class 0, which has objects that could not be classified by the FSB method. It was also concluded that the Gaussian DWKNN technique achieved better performance than the classical KNN technique.

Next, three complementary studies were presented. In the first, raw lateral friction was replaced by the non-normalized friction ratio considering some selected feature combinations. The Wilcoxon statistical test showed that such substitution slightly improves classification for FSB classes, however no relevant difference was observed for ISG classes. In the second complementary study, the inclusion of geological age as a discrete input was studied for some selected feature combinations. Results have shown that all tested combinations presented significant improvements with the inclusion of geological age when FSB classes were used. In the third complementary study, it was investigated if examples that receive class 0 of FSB charts form a homogeneous region within input feature space. The evidence found in the non-normalized feature space corroborates with this hypothesis, what suggests that 0-class examples represent an specific soil type in the non-normalized

feature space.

One advantage of the used methodology is that it can be considered general, once it was applied to DWKNN, DT and RF with few adaptations and could be easily adapted to other ML techniques. Another advantage is that it enables a rigorous comparison between different ML techniques, using statistical approaches for comparison and avoiding to produce results under unequal conditions. Other authors that use ML for soil classification usually present superficial investigations and, to our best knowledge, we are the first to present studies involving more than three input features in this research field.

The results obtained with DWKNN, DT and RF lead to the following conclusions:

- Although the normalized inputs produce the best performances, the non-normalized inputs achieve accuracies over 90%;
- Both symbolic techniques had equivalent performances and higher than the DWKNN;
- Removing the FSB class 0 completely from the dataset resulted in an accuracy of around 99%;
- Depth was the most important feature for DT, although it was lower for RF.

In the following analysis, SVM had superior predictive performance than ANN for all comparisons. Nevertheless, ANN has the advantage of producing models which can be easily reproduced with spreadsheets by simply combining the calibrated weights with the used activation function. Both SVM and ANN achieve better results when class 0 of the FSB method is removed from the dataset and, in general, they are both benefited by using log-scaling normalizations. Preliminary tests with the SVM showed that, for this dataset, the polynomial kernel calibrated with a certain procedure produce better predictive performance than Gaussian and Laplacian kernels adjusted with a heuristics. ANN and SVM were capable of differentiating soil classes using only non-normalized inputs, which also indicates that these features could be used for soil classifications. It was also observed superior predictive performance for ISG outputs than for FSB outputs.

The presented comparisons using rigorous statistical procedures allows concluding that the predictive performance of the MMP overcomes all other techniques alone. A test was made with the raw parameters given by a friction-cone penetrometer without the pore water pressure sensor, allowing to conclude that this device is not needed for reaching reasonable accuracy for soil classification. In the final analysis, a real application case was simulated using a so-called tropical dataset, which is composed by 62 CPT soundings from the city of São Paulo. In this procedure, the tropical dataset was used as test set and all techniques were trained with the main dataset. The MMP showed a more stable

and higher performance when compared to isolated techniques, which was confirmed by the Friedman statistical test for the traditional accuracy and the mean recall.

This work also explores the kernel k-means and kernel x-means clustering algorithms to group CPT data into different soil behavioral classes. Using a kernel function to modify the k-means algorithm enables evaluating classes that are not linearly separable. Next, ANN was used to create mathematical models which can be easily reproduced. Two different approaches were studied: a general and a case-based one. 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 case-based 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 was applied to 5 soundings for which the CPT soundings were paired with SPT soundings. Results confirm that the case-based model produces more well-defined behavioral classes and a more accurate ANN model. Nevertheless, both models have satisfactory accuracy and interpretability.

One of the contributions of this work is investigating up to five-dimensional input feature spaces for soil classification from CPT data, overcoming the dimensionality limitation of the chart-based methods. It also uses more input features than other work that treat the problem with ML clustering techniques, exploring all CPT information available. It uses a modification of the traditional k-means algorithm, the kernel k-means, turning it more feasible for non-linearly separable classes, improving the clustering results obtained with the original k-means. This technique was also combined with the x-means and a case-based approach was proposed. Another important contribution is the construction of mathematical models with ANN, allowing the reproduction of the obtained classification methods with simple computational tools. It was also performed omitting pore pressure information, considering cases for which it is not possible to take this measurement, and reasonable accuracy is still obtained.

For future work, it is suggested exploring other ML clustering techniques able to identify non-linearly separable classes instead of the kernel k-means. Other tools can also be used to reproduce the classification methods, like DWKNN, DT, RF and SVM. It could be also exploited introducing other input features together with CPT parameters to generate clustered classes like data from other in situ experiments such as the SPT or the flat dilatometer test. Another possible investigations are applying ML techniques for regression problems with different in situ tests data. An example of these investigations in regression problems is studying the application of ML techniques to predict piles load capacity.

Annex A - ML techniques

In this work, supervised ML techniques are used to replicate the soil classification methods described in Sections 2.1 and 2.2. This type of approach requires a dataset containing many examples of correctly classified input features, in this case, CPT values accompanied with the correct soil class. These known examples are used to train each ML technique, which should become capable of predicting the class of unknown examples within a certain accuracy. This Section contains only a brief description of the used supervised ML techniques, further detail about the training and testing procedures is given in Chapter 3. In the MMP, the five techniques described below are used to classify each unknown object and its final class is defined by voting.

Another approach adopted in this work is using descriptive unsupervised ML techniques to create new soil classes, proposing a new classification method that combines unsupervised and supervised ML techniques. This approach involves clustering the examples based only on the input features, providing new classes groupings different from those proposed in the methods of Chapter 2. Next, an ANN model is trained to reproduce these new classification methods. The used unsupervised ML techniques are based on the kernel k-means algorithm, which is explained in Section A.5 and more details of the techniques execution are given in further sections.

A.1 KNN and DWKNN

Distance-based techniques (COVER; HART, 1967; DUDANI, 1976) are based on the hypothesis that objects that are close in the feature space probably describe the same concept. This means that objects from the same class tend to be close from each other, while objects of different classes tend to be more separated, although overlaps between classes may exist. One of the most common metrics used for measuring these distances is the Minkowski metric. Considering a pair $(\mathbf{x}_i, \mathbf{x}_j)$ of objects in a d -dimensional feature space,

the distance between them is given by:

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[p]{\sum_{l=1}^d |x_i^l - x_j^l|^p} \quad (\text{A.1})$$

where it is usually adopted the Euclidean distance, therefore $p = 2$.

Nearest neighbors-based algorithms evaluate the class of unknown examples by measuring their distance to every example with a known class within a dataset. The class of the new example is decided by voting, using the labels of the closest known examples. The simplest form of this algorithm is the 1-NN (COVER; HART, 1967), which considers only one nearest neighbor for the decision. It can be described as follows:

Nearest neighbor pseudo-code

Inputs: training set $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ and
an object to be classified ($\mathbf{x}_t$) using a distance function $d(\mathbf{x}_i, \mathbf{x}_t)$

Output: y_t

$d_{min} \leftarrow +\infty$

for each $i \in 1, \dots, n$ **do**

if $d(\mathbf{x}_i, \mathbf{x}_t) < d_{min}$ **then**

$d_{min} \leftarrow d(\mathbf{x}_i, \mathbf{x}_t)$

$idx \leftarrow i$

end

$y_t = y_{idx}$

return y_t

In this case, the prediction for a new example is simply equal to that of its nearest neighbor. This algorithm can be easily extended to include two or more neighbors in a voting procedure to decide the prediction, considering a suitable tiebreaker. When an arbitrary number of k neighbors are considered, the method is called k nearest neighbor (KNN). This leads to the task to decide which is the best k . An overestimated k increases computational cost and tends to favor majority classes. On the other hand, always using $k = 1$ can ignore important information from other neighbors. In practice, the safer solution is to test several values for k and elect the one with best performance on known data.

One alternative procedure that can improve this method is to weight the vote of each neighbor by the inverse of its distance from the unknown example. This is the essence of the distance-weighted nearest neighbor algorithms (DWKNN) (DUDANI, 1976). One

specific way for defining these weights is by using Gaussian weighting, which is defined by the following expression (HECHENBICHLER; SCHLIEP, 2004):

$$w(d) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2} \quad (\text{A.2})$$

where d is the distance value.

A.2 DT and RF

Several algorithms can be employed to build a DT (HUNT; MACILROY, 1976; BREIMAN *et al.*, 1984; QUINLAN, 1986; QUINLAN, 1993). One simple definition is that DT are structured by direct acyclic graphs as a common tree structure, containing decision nodes and leaf nodes. The objects that reach a decision node are distributed between other possible nodes, according to a conditional test applied to an input feature. When an object reaches a leaf-node, a class is assigned to it.

Binary trees are often used, meaning that all decision nodes distribute objects to only two other nodes. The conditional test used by each decision node should be calibrated to maximize the tree capacity of distinguishing between possible classes. This capacity can be calculated by comparing the two new datasets to the previous one and evaluating the reduction of class mixtures. In other words, by using an impurity measure to verify if the class impurity has decreased. Impurity can be defined as:

$$i(t) = \phi(p_1, p_2, \dots, p_k) \quad (\text{A.3})$$

where t is the decision node whose impurity is measured, ϕ is some impurity function, $p_i, i = 1, 2, \dots, k$ is the proportion of the class i in the dataset and k is the number of classes.

Thus, if S is a conditional test that distributes objects between left L and right R son-nodes, then the impurity decrease is given by:

$$\delta(S) = \phi(p_1, \dots, p_k) - P_L \cdot \phi(p_{1,L}, \dots, p_{k,L}) - P_R \cdot \phi(p_{1,R}, \dots, p_{k,R}) \quad (\text{A.4})$$

where P_L and P_R are the probabilities of an object going to the left and right son-nodes, respectively.

An advantage of the DT is its capacity to make an inner input feature selection, due to the way it is constructed. Nevertheless, one disadvantage of DT is that it tends to produce overfitting, which is becoming too specialized on the training set. It means that,

although very good predictive performance is observed for the training set, performance drops abruptly for slight changes within the tested dataset. For DT, overfitting can be avoided with two main procedures. One is by using boosted DT, which is DT combined to the bootstrap method. This method consists on creating multiple trees trained with samples from the original training set, aiming to increase the probability of choosing objects difficult to classify. Another procedure to avoid overfitting is by using random forests, which consists on creating multiple DT with the bootstrap aggregating method, also called bagging (BREIMAN, 2001). The two main differences between boosted DT and RF is that, for the latter, training samples are taken randomly and input feature samples are taken to create different trees.

A.3 ANN

ANN are based on the brain structure and work only with numeric features. Their fundamental units, the neurons, communicate to each other using weighted signals that belong to the interval $[0, 1]$, so the output of an neuron can be an input to another one.

The first artificial neuron model proposed in the literature, represented in Figure A.1 (MCCULLOCH; PITTS, 1943), is called a perceptron neuron. It receives input signals x_i , which are weighted with w_i . The resulting signals are then aggregated by a function Σ and an excitation threshold is discounted, resulting a signal u . Finally, over the net signal u , an activation function g is applied, generating the output signal y .

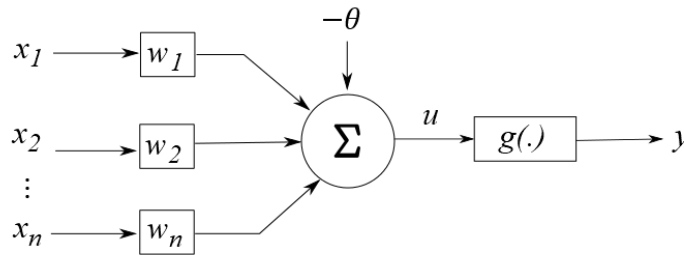


FIGURE A.1 – ANN neuron.

where $x_1, x_2, \dots, x_n$ are the input signals, $w_1, w_2, \dots, w_n$ are the signal weights, Σ is the aggregation function, θ is the activation threshold, u is the net total input signal, g is the activation function and y is the output signal.

The activation function usually adopted in classification problems is the sigmoid, which can be defined as:

$$g(u) = \frac{1}{1 + e^{-\lambda u}} \quad (\text{A.5})$$

Multi-layer perceptron neural networks are used in this work, which are feed-forward networks trained with the back-propagation algorithm (RUMELHART *et al.*, 1986). The Figure A.2 represents its general architecture. The sigmoid function is used to activate neurons within the hidden and output layers. The output layer has one neuron representing each class, totalizing c neurons. Each of these neurons provides a value within the interval $[0, 1]$ and the neuron that gives the highest value defines the final classification.

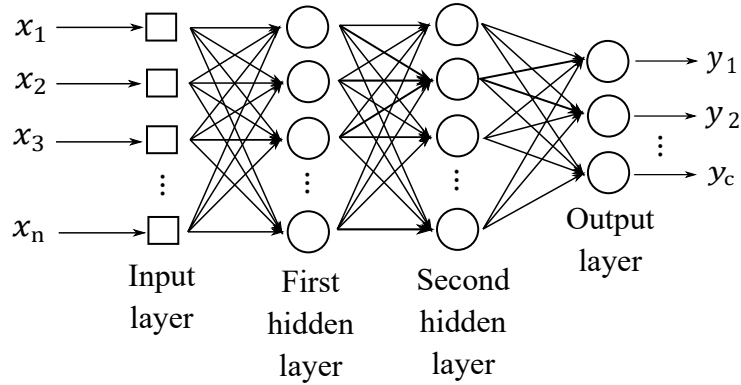


FIGURE A.2 – Multilayer neural network.

The universal approximation theorem states that an ANN with a single hidden layer is capable of reproducing any continuous function (HORNIK *et al.*, 1989). In addition to that, if two hidden layers are used any function can be reproduced.

A.4 SVM

For a binary linear SVM (VAPNIK, 1998) classifier, only two classes are included and the hyperplane that separates them can be represented with a linear function $\mathbf{w} \cdot \mathbf{x} + b = 0$. The margin between classes can be defined by two parallel hyperplanes, which are represented by the functions $\mathbf{w} \cdot \mathbf{x} + b = -1$ and $\mathbf{w} \cdot \mathbf{x} + b = 1$. If y_i is the class of an object $\mathbf{x}_i$, then

$$y_i = \begin{cases} 1, & \mathbf{w} \cdot \mathbf{x}_i + b \geq 1 \\ -1, & \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \end{cases}$$

The objective of the algorithm is to separate the two classes by maximizing the margin, which is equivalent to finding b^* and $\mathbf{w}^*$ that minimize $\frac{1}{2} \|\mathbf{w}\|^2$. Applying a Lagrangian operator and deriving the expression with respect to $\mathbf{w}$ and b , it is obtained:

$$\text{Maximize}(\alpha_1, \alpha_2, \dots) : \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (\text{A.6})$$

$$\text{Subjected to: } \begin{cases} \alpha_i \geq 0, \forall i = 1, \dots, n \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

where α_i are Lagrangian multipliers. The optimal solutions α_i^* satisfy the following condition:

$$\alpha_i^*(y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - 1) = 0 \quad (\text{A.7})$$

One can notice that $\alpha_i^* \neq 0$ only for objects over one of the parallel hyperplanes that define the margin. Objects that satisfy this condition are called support vectors.

A binary linear SVM is adequate only for problems without outliers and noise, which rarely suits real applications. This limitation can be solved by introducing a tolerance $\xi_i \geq 0$ to smooth out the margin.

Another limitation of binary linear SVM is the inability to deal with non linear problems. One solution is increasing dimensionality by using a function Φ to map the objects to a higher dimensionality feature space. If an adequate mapping function is chosen, objects become linearly separable in this new feature space.

Obtaining Φ directly usually requires unpractical computational cost. Nonetheless, observing the expression of the maximization problem, one may notice that what is needed indeed is not the isolated function Φ , but the scalar product $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. This function is called kernel and usually it is easier to obtain than Φ . There are several types of kernels that can be associated to the SVM algorithm and choosing the best one for a specific problem is usually based on preliminary tests. In this work three kernels were tested, the Gaussian, the Laplacian and the polynomial, which are defined as follows:

$$\text{Polynomial: } (\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)^d \quad (\text{A.8})$$

$$\text{Gaussian: } \exp(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (\text{A.9})$$

$$\text{Laplacian: } \exp(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (\text{A.10})$$

The Gaussian and the Laplacian kernels depend on only one parameter σ , which can be estimated by a heuristics consisting in taking, for example, the median of squared distances $\|x - x'\|^2$ between every pair of objects x and x' . On the other hand, the polynomial kernel requires three calibration parameters: the offset κ , the scale δ and the degree d .

Although all the procedures described above are valid for a binary SVM classifier, it is possible to extend it to multi-class problems by simply combining two or more binary

classifiers. In this procedure all classes must be evaluated in pairs, generating $\binom{c}{2}$ classifiers for c classes.

A.5 Kernel k-means

The kernel k-means algorithm (DHILLON *et al.*, 2005) 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:

$$E = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} d(\mathbf{x}_i, \bar{\mathbf{x}}^{(j)})^2 \quad (\text{A.11})$$

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

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 convergence is observed, within a predefined error margin.

The most used similarity measure is the Euclidian distance, which requires data normalization in order to avoid distortions due to data scale. In this context, the algorithm can be described by the following pseudocode:

k-means algorithm**Inputs:** Dataset $\mathbf{X}_{n \times d}$ and number of clusters k **Outputs:** A partition $\mathbf{C}$ of $\mathbf{X}$ in k clusters**While** a change is observed**For each** object $\mathbf{x}_i$ and cluster C_j **do** Calculate $d(\mathbf{x}_i, \bar{\mathbf{x}}^{(j)})$ **For each** object $\mathbf{x}_i$ **do** Link $\mathbf{x}_i$ to the cluster with the closest centroid**For each** cluster C_j **do** Redefine the centroid $\bar{\mathbf{x}}^{(j)}$ minimizing E

The main advantage of k-means is its linear complexity, but its main disadvantages include the possibility of converging to a local optimum and being applicable only to linearly separable classes. Other weaknesses that can compromise the analysis are its sensitivity to initialization, the possibility of generating imbalanced clusters and the need of a previously fixed k . One simple alternative to search for the best k and avoid local minima is running the algorithm several times, varying k and the random initialization.

One way to deal with classes that are not linearly separable is using a function $\Phi : \mathbf{X} \rightarrow \mathfrak{S}$ to map the data from the original feature space $\mathbf{X}$ into a higher dimensionality feature space $\mathfrak{S}$ wherein the objects are linearly separable. Nevertheless, two conditions have to be guaranteed to ensure linear separability in $\mathfrak{S}$: a non-linear transformation Φ and a high dimensionality of $\mathfrak{S}$. Most work that make use of this approach do not define the function Φ directly, but only the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, which is sufficient to obtain the Euclidian distance. The Gaussian kernel is adopted in this work, which is defined as follows:

$$\text{Gaussian: } \exp(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (\text{A.12})$$

Annex B - Example of supervised application

In order to illustrate the application of ML techniques for soil profiling, the MMP was employed to determine the soil profile according to the ISG method for the sounding provided by Prof. Renato P. da Cunha (see Section 3.1.3) using as training set the main dataset (see Section 3.1.1). The result can be compared to the obtained with CPeT-IT v2.0.2.5. The soil profiles given by those different tools are shown represented in Figure B.1. These results represent a situation with an accuracy (mean recall) of 95.4%. The code used to generate the soil profile given by the MMP is presented in Annex C, wherein the MMP and the preprocessing implementation details are omitted for a sake of conciseness but they can be implemented with the guidelines and R packages described through this work.

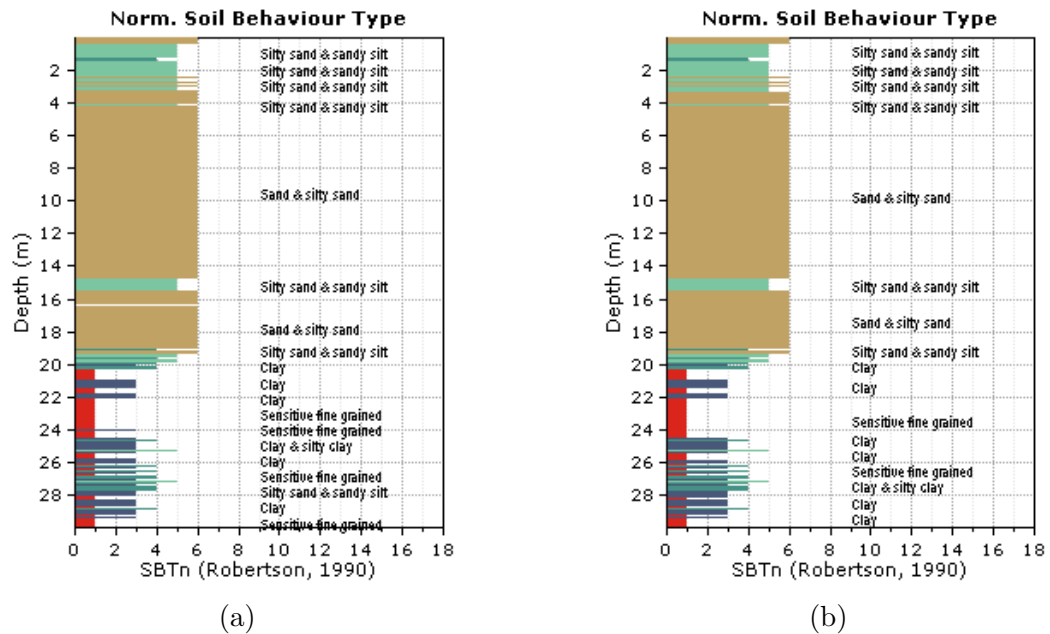


FIGURE B.1 – Soil profiles given by: (a) CPeT-IT v2.0.2.5 and (b) MMP.

Annex C - MMP application code

```
1  # Including functions
2
3  path1 = file.path("C:", "Exemplo", "preproc.R")
4  path2 = file.path("C:", "Exemplo", "mmp.R")
5  path3 = file.path("C:", "Exemplo", "calcAc.R")
6  source(path1)
7  source(path2)
8  source(path3)
9
10 # Reading CSV file and saving data in variables
11
12 path4 = file.path("C:", "Exemplo", "main.csv")
13 path5 = file.path("C:", "Exemplo", "cunha.csv")
14 main = read.table(path4, header = TRUE, sep = ",")
15 cunha = read.table(path5, header = TRUE, sep = ",")
16
17 train=main
18 test=cunha
19
20 # Running applications
21
22 inp=c("Qtn", "Fr_p")
23 out="SBTn"
24
25 info = preproc(train, test, inp, out) # Preprocessing
26 datasets = info[[1]]
27 func_param = info[[2]]
28 minVal = info[[3]]
29 maxVal = info[[4]]
30
31 res = mmp(datasets, inp, out, func_param) # Modeling and predicting
32 lpreds = res[[1]]
33 lconf = res[[2]]
34
35 ac = calcAc(lconf$conf, rzero=1) # Calculating accuracy (mean recall)
```

Bibliography

- ALKROOSH, I.; NIKRAZ, H. Predicting axial capacity of driven piles in cohesive soils using intelligent computing. **Engineering Applications of Artificial Intelligence**, vol. 25, no. 3, pp. 618–627, apr 2012.
- AREL, E. Predicting the spatial distribution of soil profile in Adapazari/Turkey by artificial neural networks using CPT data. **Computers & Geosciences**, vol. 43, pp. 90–100, jun 2012.
- BAZIAR, M. H.; NILIPOUR, N. Evaluation of liquefaction potential using neural-networks and CPT results. **Soil Dynamics and Earthquake Engineering**, vol. 23, no. 7, pp. 631–636, oct 2003.
- BEGEMANN, H. K. S. The Friction Jacket Cone as an Aid in Determining the Soil Profile. **Proceedings of the 6th International Conference on Soil Mechanics and Foundation Engineering**, ICSMFE, Montreal, pp. 17–20, 1965.
- BHATTACHARYA, B.; SOLOMATINE, D. P. Machine learning in soil classification. **Neural Networks**, vol. 19, no. 2, pp. 186–195, mar 2006.
- BOGGESE, R.; ROBERTSON, P. K. CPT for Soft Sediments and Deepwater Investigations. **Offshore Technology Conference**, may 2011.
- BREIMAN, L. Random forests. **Machine Learning**, vol. 45, no. 1, pp. 5–32, Oct 2001. ISSN 1573-0565.
- BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. **Classification and Regression Trees**. New York: Chapman & Hall, 1984. 358 pp. ISBN 0-412-04841-8.
- CAI, G.; LIU, S.; PUPPALA, A. Comparison of CPT charts for soil classification using PCPT data: Example from clay deposits in Jiangsu Province, China. **Engineering Geology**, vol. 121, no. 1-2, pp. 89–96, jul 2011.
- CAL, Y. Soil classification by neural network. **Advances in Engineering Software**, vol. 22, no. 2, pp. 95–97, dec 1995.
- CETIN, K. O.; OZAN, C. CPT-Based Probabilistic Soil Characterization and Classification. **Journal of Geotechnical and Geoenvironmental Engineering**, vol. 135, no. 1, pp. 84–107, jan 2009.
- COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, vol. 13, no. 1, pp. 21–27, jan 1967. ISSN 0018-9448.

- DAS, S. K.; BASUDHAR, P. K. Utilization of self-organizing map and fuzzy clustering for site characterization using piezocone data. **Computers and Geotechnics**, vol. 36, no. 1-2, pp. 241–248, 2009.
- DHILLON, I.; GUAN, Y.; KULIS, B. **A Unified View of Kernel k-means, Spectral Clustering and Graph Partitioning**. Austin, 2005.
- DOUGLAS, B. J.; OLSEN, R. S. Soil Classification Using Electric Cone Penetrometer. **Symposium on Cone Penetration Testing and Experience, Geotechnical Engineering Division**, ASCE, St. Louis, pp. 209–227, 1981.
- DUDANI, S. A. The distance-weighted k-nearest-neighbor rule. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-6, no. 4, pp. 325–327, apr 1976. ISSN 0018-9472.
- FACCIORUSSO, J.; UZIELLI, M. Stratigraphic profiling by cluster analysis and fuzzy soil classification from mechanical cone penetration tests. **Proceedings ISC-2 on Geotechnical and Geophysical Site Characterization**, Porto, sep 2004.
- FACELLI, K.; LORENA, A. C.; GAMA, J. M. P.; CARVALHO, A. C. P. L. F. **Inteligência Artificial – Uma abordagem de Aprendizado de Máquina**. 1st. ed. Rio de Janeiro: LTC Editora, 2011. 394 pp.
- FRIEDMAN, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. **Journal of the American Statistical Association**, Taylor & Francis, vol. 32, no. 200, pp. 675–701, 1937.
- GANDOMI, A. H.; FRIDLINE, M. M.; ROKE, D. A. Decision Tree Approach for Soil Liquefaction Assessment. **The Scientific World Journal**, vol. 2013, 2013.
- GANJU, E.; PREZZI, M.; SALGADO, R. Algorithm for generation of stratigraphic profiles using cone penetration test data. **Computers and Geotechnics**, vol. 90, pp. 73–84, oct 2017.
- GOH, A. T. C. Modeling Soil Correlations Using Neural Networks. **Journal of Computing in Civil Engineering**, vol. 9, no. 4, pp. 275–278, oct 1995.
- GOH, A. T. C. Neural-Network Modeling of CPT Seismic Liquefaction Data. **Journal of Geotechnical Engineering**, vol. 122, no. 1, pp. 70–73, jan 1996.
- GOH, A. T. C.; GOH, S. H. Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. **Computers and Geotechnics**, vol. 34, no. 5, pp. 410–421, jul 2007.
- HANNA, A.; URAL, D.; SAYGILI, G. Neural network model for liquefaction potential in layered soils using turkey and taiwan earthquake data. **Soil Dynamics and Earthquake Engineering**, vol. 27, no. 6, pp. 521–540, jun 2007.
- HECHENBICHLER, K.; SCHLIEP, K. Weighted k -Nearest-Neighbor Techniques and Ordinal Classification. **Sonderforschungsbereich**, vol. 386, no. 399, 2004. Available from Internet: <<http://epub.ub.uni-muenchen.de>>.

- HEGAZY, Y. A.; MAYNE, P. W. Objective Site Characterization Using Clustering of Piezocone Data. **Journal of Geotechnical and Geoenvironmental Engineering**, vol. 128, no. 12, pp. 986–996, dec 2002.
- HORNIK, K.; STINCHOMBE, M.; WHITE, H. Multilayer Feedforward Networks are Universal Approximators. **Soil Science Society of America Journal**, vol. 2, pp. 359–366, 1989.
- HUNT, J. W.; MACILROY, M. D. **An algorithm for differential file comparison**. Murray Hill: Bell Laboratories, 1976.
- IOANNIDES, J. T.; ROBERTSON, P. K. **CPeT-IT v.2.0 – CPT interpretation software**. 2016. Available from Internet: <<https://geologismiki.gr/>>.
- JEFFERIES, M.; DAVIES, M. Use of CPTu to Estimate Equivalent SPT N_{60} . **Geotechnical Testing Journal**, vol. 16, no. 4, pp. 458–468, 1993.
- JEFFERIES, M. G.; DAVIES, M. P. Soil classification by the cone penetration test: Discussion. **Canadian Geotechnical Journal**, vol. 28, no. 1, pp. 173–176, 1991.
- JUANG, C. H.; CHEN, C. J. CPT-Based Liquefaction Evaluation Using Neural Networks. **Computer-Aided Civil and Infrastructure Engineering**, vol. 14, no. 1, pp. 221–229, 1999.
- JUANG, C. H.; LU, P. C.; CHEN, C. J. Predicting Geotechnical Parameters of Sands from CPT Measurements Using Neural Networks. **Computer-Aided Civil and Infrastructure Engineering**, vol. 17, no. 1, pp. 31–42, dec 2002.
- JUANG, C. H.; YUAN, H.; LEE, D.-H.; LIN, P.-S. Simplified Cone Penetration Test-based Method for Evaluating Liquefaction Resistance of Soils. **Journal of Geotechnical Geoenvironmental Engineering**, vol. 129, no. 1, pp. 66–80, jan 2003.
- KOHESTANI, V. R.; HASSANLOURAD, M.; ARDAKANI, A. Evaluation of liquefaction potential based on CPT data using random forest. **Natural Hazards**, vol. 79, no. 2, pp. 1079–1089, jul 2015.
- KUMAR, J. K.; KONNO, M.; YASUDA, N. Subsurface Soil-Geology Interpolation Using Fuzzy Neural Network. **Journal of Geotechnical and Geoenvironmental Engineering**, vol. 126, no. 7, pp. 632–639, jul 2000.
- KURUP, P. U.; GRIFFIN, E. P. Prediction of Soil Composition from CPT Data Using General Regression Neural Network. **Journal of Computing in Civil Engineering**, vol. 20, no. 4, pp. 281–289, 2006.
- LIAO, T.; MAYNE, P. W. Stratigraphic delineation by three-dimensional clustering of piezocone data. **Georisk**, vol. 1, no. 2, pp. 102–119, jun 2007.
- LIVINGSTON, G.; PIANTEDOSI, M.; KURUP, P.; SITHARAM, T. G. Using Decision-Tree Learning to Assess Liquefaction Potential from CPT and V_s . **Geotechnical Earthquake Engineering and Soil Dynamics Congress IV**, Sacramento, may 2008.

- MAYNE, P. W. Interpretation of geotechnical parameters from seismic piezocone tests. **Proceedings of the 3rd International Symposium on Cone Penetration Testing (CPT14)**, Las Vegas, vol. 102, pp. 47–73, jan 2014.
- MAYNE, P. W.; PEUCHEN, J.; BOUWMEESTER, D. Soil unit weight estimated from CPTu in offshore soils. **Frontiers in Offshore Geotechnics II**, Perth, pp. 371–376, 2010.
- MCCULLOCH, W.; PITTS, W. A Logical Calculus of Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biophysics**, vol. 5, no. 4, pp. 115–133, 1943.
- NEMENYI, P. **Distribution-free Multiple Comparisons**. Thesis (PhD) — Princeton University, 1963.
- PELLEG, D.; MOORE, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: **In Proceedings of the 17th International Conference on Machine Learning**. San Francisco: Morgan Kaufmann, 2000. pp. 727–734. ISBN 1-55860-707-2.
- QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, vol. 1, pp. 81–106, 1986.
- QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.
- RAMSEY, N. A calibrated model for the interpretation of cone penetration tests (CPTs) in north sea quaternary soils. **Offshore site investigation and geotechnics 'Diversity and Sustainability' conference**, London, pp. 341–356, jan 2002.
- REALE, C.; GAVIN, K.; LIBRIĆ, L.; JURIĆ-KAĆUNIĆ, D. Automatic classification of fine-grained soils using cpt measurements and artificial neural networks. **Advanced Engineering Informatics**, Elsevier, vol. 36, pp. 207–215, 2018.
- ROBERTSON, P. K. Soil classification using the cone penetration test. **Canadian Geotechnical Journal**, vol. 27, no. 1, pp. 151–158, 1990.
- ROBERTSON, P. K. Soil classification using the cone penetration test: Reply. **Canadian Geotechnical Journal**, vol. 28, no. 1, pp. 176–178, jan 1991.
- ROBERTSON, P. K. Interpretation of cone penetration tests – a unified approach. **Canadian Geotechnical Journal**, vol. 46, no. 11, pp. 1337–1355, nov 2009.
- ROBERTSON, P. K. Cone penetration test (CPT)-based soil behaviour type (SBT) classification system – an update. **Canadian Geotechnical Journal**, vol. 53, no. 12, pp. 1910–1927, jul 2016.
- ROBERTSON, P. K.; CAMPANELLA, R. G.; GILLESPIE, D.; GREIG, J. Use of Piezometer Cone Data. **In Situ '86 Use of In-Situ Testing in Geotechnical Engineering ASCE Specialty Conference**, Blacksburg, jun 1986.
- ROBERTSON, P. K.; WRIDE, C. E. Evaluating cyclic liquefaction potential using the cone penetration test. **Canadian Geotechnical Journal**, vol. 35, no. 3, pp. 442–459, 1998.

- ROGIERS, B.; MALLANTS, D.; BATELAAN, O.; GEDEON, M.; HUYSMANS, M.; DASSARGUES, A. Model-based classification of CPT data and automated lithostratigraphic mapping for high-resolution characterization of a heterogeneous sedimentary aquifer. **PLoS ONE**, vol. 12, no. 5, pp. e0176656, 2017.
- RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagation errors. **Nature**, vol. 323, no. 6088, pp. 533–536, oct 1986.
- SALGADO, R. **The Engineering of Foundations**. 1st. ed. New York: McGraw-Hill Education, 2006. 896 pp.
- SCHAAP, M. G.; LEIJ, F. J.; GENUCHTEN, M. T. V. Neural Network Analysis for Hierarchical Prediction of Soil Hydraulic Properties. **Soil Science Society of America Journal**, vol. 62, no. 4, pp. 847–855, jul 1998. ISSN 0361-5995.
- SCHNAID, F.; ODEBRECHT, E. **Ensaio de campo e suas aplicações à engenharia de fundações**. 2nd. ed. São Paulo: Oficina de Textos, 2012. 224 pp.
- SCHNEIDER, J. A.; HOTSTREAM, J. N.; MAYNE, P. W.; RANDOLPH, M. F. Comparing CPTU $Q - F$ and $Q - \Delta u_2 / \sigma'_{v0}$ soil classification charts. **Géotechnique Letters**, vol. 2, no. 4, pp. 209–215, oct 2012. ISSN 2045-2543.
- SCHNEIDER, J. A.; RANDOLPH, M. F.; WAYNE, P. W.; RAMSEY, N. R. Analysis of Factors Influencing Soil Classification Using Normalized Piezocone Tip Resistance and Pore Pressure Parameters. **Journal of Geotechnical and Geoenvironmental Engineering**, vol. 134, no. 11, pp. 1569–1586, nov 2008.
- SHAHRI, A. A.; MALEHMIR, A.; JUHLIN, C. Soil classification analysis based on piezocone penetration test data – A case study from a quick-clay landslide site in southwestern Sweden. **Engineering Geology**, vol. 189, pp. 32–47, apr 2015.
- STONE, M. Cross-Validatory Choice and Assessment of Statistical Predictions. **Journal of the Royal Statistical Society. Series B (Methodological)**, vol. 36, no. 2, pp. 111–147, 1974.
- TUMAY, M. T.; ABU-FARSAKH, M. Y.; ZHANG, Z. From Theory to Implementation of a CPT-Based Probabilistic and Fuzzy Soil Classification. **From Research to Practice in Geotechnical Engineering Congress**, New Orleans, vol. 13, mar 2008.
- VAPNIK, V. **Statistical learning theory**. 1st. ed. New York: Wiley-Interscience, 1998. 768 pp. ISBN 978-0-471-03003-4.
- WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, vol. 1, no. 6, pp. 80–83, dec 1945.
- WILSON, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-2, no. 3, pp. 408–421, jul 1972.
- ZHANG, Z.; TUMAY, M. T. Statistical to Fuzzy Approach Toward CPT Soil Classification. **Journal of Geotechnical and Geoenvironmental Engineering**, vol. 125, no. 3, pp. 179–186, mar 1999.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO DM	2. DATA 27 de dezembro de 2018	3. DOCUMENTO Nº DCTA/ITA/DM-113/2018	4. Nº DE PÁGINAS 101
5. TÍTULO E SUBTÍTULO: Application of Machine Learning techniques for soil classification from CPTu			
6. AUTOR(ES): Lucas Orbolato Carvalho			
7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES): Instituto Tecnológico de Aeronáutica – ITA			
8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR: Cone penetration test; Soil classification; Machine learning; K-nearest neighbor; Decision tree; Random forest; Artificial neural network; Support vector machine; Multiple model predictor; K-means.			
9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO: Aprendizagem (inteligência artificial); Árvores de decisão; Máquinas de vetores de suporte; Máquinas aprendizes; Algoritmos; Processamento de sinais multidimensionais; Técnicas de previsão, Computação.			
10. APRESENTAÇÃO: (X) Nacional () Internacional ITA, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Engenharia de Infraestrutura Aeronáutica. Área de Infraestrutura Aeroportuária. Orientador: Prof. Dr. Dimas Betioli Ribeiro. Defesa em 20/12/2018. Publicada em 2018.			
11. RESUMO: Classical methods of soil classification from cone penetration test (CPT) data propose a graphical approach. It requires representing data with just two parameters in order to create plane representations. To overcome this restriction, the fundamental objective of this work is to develop multi-dimensional analyses using different machine learning (ML) algorithms, including supervised and unsupervised techniques. A general methodology for the predictive supervised approach is proposed and applied up to five dimensions, including raw, non-normalized and normalized CPT data as continuous inputs and geological age as a discrete one. The dataset is composed by 179 CPT soundings from several world locations. Two popular chart-based classification methods are reproduced and studied in the predictive supervised approach: one influenced by soil granulometry (ISG) and one focused on soil behavior (FSB). Descriptive statistical procedures as well as ML techniques are used for data preprocessing, including data cleaning, balancing and transformation. Different ML techniques are calibrated, studied and compared with statistical tests and employing the 10-fold cross-validation procedure. Within the first studies, the input features relevance is evaluated with distance-based techniques. These first results have shown that using non-normalized CPT data with depth included is enough for the techniques to properly distinguish between different soil types with a satisfactory accuracy, around 90%, and that depth introduce relevant information for the task of soil classification from CPT data. It was also observed that the inclusion of the geological age as a discrete input can introduce relevant information. In the second part of the study, other ML techniques are tested and their particularities are explored. The symbolic techniques were those with the best performance and lowest sensitivity to methodology variations. The techniques are then combined into a multiple model predictor (MMP), whose performance overcame the one of the other ML techniques according to statistical tests. Finally, unsupervised techniques are employed to generate new classes from non-normalized inputs with two different approaches: a general one and a case-based one. So, an ANN is trained to create a mathematical model that allows the new method reproduction with simple tools as spreadsheets. The clustering results are then compared to the classical methods and to standard penetration test (SPT) data paired with CPT.			
12. GRAU DE SIGILO: (X) OSTENSIVO () RESERVADO () SECRETO			