INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Lucas Orbolato Carvalho

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR SOIL CLASSIFICATION FROM CPT DATA

Final Paper
2018

Course of Civil-Aeronautical Engineering

CDU 681.3:16

Lucas Orbolato Carvalho

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR SOIL CLASSIFICATION FROM CPT DATA

Advisor

Prof. Dr. Dimas Betioli Ribeiro (ITA)

CIVIL-AERONAUTICAL ENGINEERING

São José dos Campos
Instituto Tecnológico de Aeronáutica

2018

Cataloging-in Publication Data
Documentation and Information Division

Carvalho, Lucas Orbolato
Application of Machine Learning techniques for soil classification from CPT data / Lucas Orbolato Carvalho.
São José dos Campos, 2018.
68f.

Final paper (Undergraduation study) – Course of Civil-Aeronautical Engineering– 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. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

CARVALHO, Lucas Orbolato. **Application of Machine Learning techniques for soil classification from CPT data**. 2018. 68f. Final paper (Undergraduation study) – 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 CPT data.
PUBLICATION KIND/YEAR: Final paper (Undergraduation study) / 2018

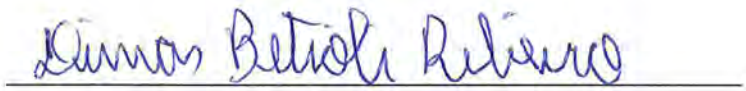
Lucas Orbolato Carvalho
Rua República do Iraque, 80
12.216-540 – São José dos Campos–SP

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR SOIL CLASSIFICATION FROM CPT DATA

This publication was accepted like Final Work of Undergraduation Study

Lucas Orbolato Carvalho

Author



Dimas Betioli Ribeiro (ITA)

Advisor

Prof. Dr. Eliseu Lucena Neto
Course Coordinator of Civil-Aeronautical Engineering

São José dos Campos: November 14, 2018.

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

Acknowledgments

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

"If I have seen farther than others,
it is because I stood on the shoulders of giants."
— Sir Isaac Newton

Resumo

O problema de classificação de solos com dados de ensaio de penetração de cone (*Cone Penetration Test* – CPT) é usualmente tratado com soluções bidimensionais tais como gráficos ou, menos frequentemente, abordagens de Aprendizado de Máquina (*Machine Learning* – ML) em um espaço de dimensionalidade restrita. Para evitar esta restrição, neste trabalho é feita uma análise multi-dimensional dos dados de CPT para a classificação de solos usando algoritmos simbólicos e baseados em distância. Os algoritmos simbólicos são capazes de realizar uma análise de relevância e uma seleção dos atributos internamente, permitindo estimar a importância dos atributos. Estes algoritmos são empregados a fim de avaliar a relevância de cada atributo segundo diferentes critérios e analisar seu desempenho considerando até cinco características, incluindo atributos brutos e normalizados de CPT como entradas contínuas e a idade geológica como discreta. O conjunto de dados utilizado é composto de 111 sondagens provenientes de diferentes locais do planeta. As técnicas simbólicas, nomeadamente árvores de decisão impulsionadas (DT) e florestas aleatórias (RF), são aplicadas ao problema, estudadas e comparadas usando o procedimento 10-*fold* de validação cruzada. Dois métodos de classificação são considerados: um influenciado pela granulometria do solo (ISG) e outro focado no comportamento do solo (FSB). Uma metodologia geral para a classificação de solos usando técnicas de ML é descrita e seguida. Ela envolve procedimentos de estatística descritiva e outras técnicas de ML para o pré-processamento dos dados, incluindo a transformação, a limpeza e o balanceamento dos dados. As técnicas são também comparadas com o algoritmo do vizinho mais próximo ponderado pela distância pela função Gaussiana (DWNN). As comparações são feitas por meio de testes estatísticos de hipóteses. Os resultados mostram que as árvores de decisão impulsionadas e as florestas aleatórias possuem desempenho equivalente e que ambas têm melhor desempenho que o DWNN. A análise de importância dos atributos mostra que a profundidade e a idade geológica introduzem informação relevante para a classificação de solos e que os atributos brutos incluindo a profundidade podem ser suficientes para o desempenho da tarefa.

Abstract

The soil classification problem with cone penetration test (CPT) data is usually treated with bidimensional solutions such as charts or, less often, machine learning (ML) approaches in a dimensionally restricted feature space. To avoid this restriction, a multidimensional analysis of CPT data for soil classification is here performed by using k-nearest neighbors (KNN) and machine learning symbolic algorithms. The symbolic algorithms are able to do an inner input features relevance analysis and feature selection, calculating the features importance. These algorithms are employed in order to evaluate each input feature importance by different criteria and to analyze their performance considering up to five features including raw and normalized CPT inputs as continuous inputs and soil age as a discrete one. The dataset used is composed by 111 soundings from different locations around the world. The symbolic techniques, namely boosted decision trees (DT) and random forests (RF), are applied to the problem, studied and compared using a 10-fold cross-validation procedure. Two classification methods are considered: one influenced by soil granulometry (ISG) and the other focused on soil behaviour only (FSB). A general methodology for soil classification using ML techniques is described and followed. It covers descriptive statistical procedures and other ML techniques for data preprocessing, including data transformation, cleaning and balancing. The symbolic techniques are compared with the Gaussian distance-weighted nearest neighbor technique (DWNN). The comparisons are made with statistical hypothesis tests. The results shows that RF and boosted DT have equivalent performance and that they both perform better than the DWNN. The features importance analysis indicates that depth and soil age introduce relevant information for soil classification and that the raw inputs including depth can be enough to perform the task.

List of Figures

List of Tables

List of Abbreviations and Acronyms

CPT - Cone penetration test
ML - Machine learning
KNN - k-nearest neighbor
DT - Decision trees
RF - Random forests
ANN - Artificial neural networks
SVM - Support vector machines
ISG - Methodology influenced by soil granulometry
FSB - Methodology focused on soil behavior only
DWNN - Distance-weighted nearest neighbor
ENN - Edit nearest neighbor
CG - Geological age
SMOTE - Synthetic minority over-sampling technique

List of Symbols

z	Depth
q_c	Raw 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

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. The more usual classification standard applied for general purposes is the Unified Soil Classification System (USCS), which is based on granulometry and plasticity. Nevertheless, this method have 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 to get the results as well as the complex and not automatized procedures. Therefore, it can be more desirable to relate soil classification directly to CPT parameters. However, these parameters are more related to the soil behavior, what is more of interest of engineers, than to the soil composition, as the USCS parameters, so that there should be a different classification system which reflects this change of basis.

One of the pioneers classification methods based on CPT measurements was a chart using the uncorrected cone resistance by lateral friction (BEGEMANN, 1965). There were drawn curves that could allow to find the percentage of fines and the soil type, but it was still based on granulometry. It was later stated that there should be considered a soil behavior classification system and that the friction ratio should substitute lateral friction to give better results (DOUGLAS; OLSEN, 1981). Further, it was proposed some charts based on pore pressure. Following these and others advances, an objective classification method that would become a mark in soil classification was proposed in Robertson *et al.* (1986). 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 overburden influence. This method establishes 12 soil behavior classes whose descriptions are close to the USCS ones. One of the advantages 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 the depth, which produces an overburden pressure due to the soil above.

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. However, it was shown in Jefferies and Davies (1991) that the normalized cone resistance by excess pore pressure chart of Robertson (1990) was incapable of identifying some offshore soils. The explanation given was the effect of back pressure and cavitation of the filter element in deep water offshore soils. Then, it was suggested to use only the chart with normalized cone resistance by normalized friction ratio wherein, instead of the normalized cone resistance, it should be used a combination of this feature with the normalized excess pore pressure, to comprise its information inside the new feature. So, in Robertson (1991), an update was brought to his cone resistance by excess pore pressure chart, but maintaining the parameters previously proposed and explaining that the faults observed were caused by the dilative behavior of highly overconsolidated clays such as those found in deep water soils.

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, in Schneider *et al.* (2008), it was confirmed that the chart was not accurate for some kind of soils and then the source of the misclassification was identified. 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 highly overconsolidated clay, which has also 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. In Schneider *et al.* (2008), it was stated that even with increasing normalized cone resistance and absolute 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, a new chart with normalized cone resistance by normalized excess pore pressure was proposed in Schneider *et al.* (2008), in which a new normalization for the excess pore pressure is also presented. Later, Schneider *et al.* (2012) develops a normalized cone resistance by normalized friction ratio chart too, suggesting modifications to the chart of Robertson (1991).

Another contribution was the determination of a classification index to estimate classes separation curves of the normalized cone resistance by friction ratio chart of Robertson (1991) by approximating them with circles (JEFFERIES; DAVIES, 1993). This 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; ROBERTSON, 2009). In Schneider *et al.* (2012), it was suggested replacing the circular boundaries by hyperbolic ones with a new index. However, while in Robertson (1991)

nine classes are considered, in Schneider *et al.* (2012) just five are used. Furthermore, the classes designations in Robertson (1991) were still attached to granulometry in some way. It concerns the soil macrostructure, i.e., grain size distribution, but does not tell much about some other influent factors such as cementation or soil aging. It is true that there were some interpretations given by the mentioned methods about these processes, but none of them discussed or defined them adequately.

In Robertson (2016), it was established that aged or cemented soils could be called structured soils, meaning that they have a microstructure. Soils without this microstructure are called ideals. In Robertson (2016), it is also stated that all graphical methods were valid just for ideal soils and not reliable for 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 sand-like, clay-like or transitional and each of these by contractive, dilative or even sensitive for some contractive clays. The normalized cone resistance by normalized friction ratio and the normalized cone resistance by normalized excess pore pressure charts were updated from the charts of Schneider *et al.* (2008) and Schneider *et al.* (2012), replacing the cone resistance normalization of Robertson (1990) by the one of Robertson (2016) and making some changes to include the proposed classes considering previous work (ROBERTSON, 2009).

One more definition given to make the method consistent was a procedure to determine if the soil is structured or ideal. It is based on a factor named small-strain normalized rigidity index, which is a function of the small-strain shear modulus and normalized cone resistance. When this index is greater than or equal to 330, the soil is considered structured. The steps to calculate the small-strain shear modulus and the soil unit weight for the overburden pressure, which is in turn needed to calculate the normalized cone resistances, are not explicit in Robertson (2016). Nevertheless, there are other work in the literature that gives a way to estimate the small-strain shear modulus (ROBERTSON, 2009) and to calculate an approximation for the soil unit weight (LUNNE *et al.*, 1997; MAYNE *et al.*, 2010; MAYNE, 2014).

In parallel with these graphical approaches, there were developed some alternative ways to treat the soil classification using statistics and machine learning (ML) techniques. One of the firsts applied an artificial neural network (ANN) model with one hidden layer trained by the back-propagation algorithm to classify soils according to fines content and plasticity parameters (CAL, 1995). In spite of some methodological issues, the results were satisfactory and encouraged applying ANN or other machine learning techniques to the soil classification problem. In the same decade, comes up a work that analyzes the task of soil classification based on CPT parameters by the fuzzy logic angle (ZHANG; TUMAY, 1999). This is founded on the idea that the logic is not binary or discrete but more than one event 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. Consequently, it was suggested the division of soils among highly probable mixed soil (HPM), highly probable clayey soil (HPC) and highly probable sandy soil (HPS).

Some time later, a new and descriptive approach of soil classification was presented (HEGAZY; MAYNE, 2002). The problem was explored by a hierarchical clustering method. By testing different techniques, it was determined that the single-link algorithm was the best to the task. This descriptive algorithm follows the same principle of the predictive 1-NN. Besides, it was found that the ideal number of classes to get better match with the real soil types must be under eight. This conclusion was made by varying it from 2 to 100. However, there were 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, called average-link metric, and considering uncorrected cone resistance and raw lateral friction as inputs (FACCIORUSSO; UZIELLI, 2004). The results were compared with those obtained by the fuzzy method of Zhang and Tumay (1999), showing that these methods are complementary and that they are both compatible with the observations. A number of 25 soundings from Gioia Tauro, Italy, were used in this study.

An important mark on the application of machine learning techniques to soil classification based on CPT data was the work of Bhattacharya and Solomtine (2006). A full clustering and predictive classification method called Constraint Clustering and Classification (CONCC) was proposed. The clustering technique employed, named segmentation, comprises a fuzzy logic consideration. 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 and contracting it. It follows the idea that there can be noise near the segments boundaries and that these intersections have to be removed to create representative classes regions. Nonetheless, there were chosen just the uncorrected cone resistance and the friction ratio to make this analysis so that even with this cleaning there was still a critical spatial overlap of segments, making it more difficult to classify the objects later. This problem was solved increasing the dimensionality so that the objects could be easily distinguished and the overlap of classes could be removed. For this purpose, it was applied the boundary energy method, which introduces a new artificial input attribute. The results pointed a more disjoint configuration. Next, for each class is given a label by a specialist. Finally, the data is subjected to different machine learning techniques to be classified. It was applied ANN, decision trees (DT) and support vector machines (SVM). Growing the number of segments, the accuracy of the techniques was harmed, getting under 90% for all tech-

niques for seven classes. The results could be worse for a more diverse dataset, which is evidenced by the better results for sandy and clayey soils and meaningfully lower for other types. Moreover, the study employed just seven CPT soundings of only one field. Possible upgrades could be diversifying the dataset and studying including more input features.

The problem of soil classification was also treated under the granulometrical point of view by ML techniques (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 Tumay *et al.* (2008) are reasonable for describing soil granulometry beyond its behavior. Again, it was used just 12 paired SPT-CPT soundings from only four areas, harming the generality of the model. A progress over the previous works was the implicit inclusion of depth as an input parameter. It was considered four inputs: raw uncorrected cone resistance, friction ratio, total overburden pressure and equilibrium pore pressure. These two last ones takes depth into account and seems to be redundant, but no analysis was made in this way. It could be included the pore pressure measurement behind the cone tip too in order to bring more information to the model.

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

Later on, 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 takes into account the probabilistic nature of the method and the normalized friction ratio. In the same year, there is a research which extends the discussions about descriptive classification with clustering techniques (DAS; BASUDHAR, 2009). It was 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 the k-means and the charts of

Robertson *et al.* (1986) were satisfactorily accurate. However, the techniques were applied to each sounding individually so that there wasn't a variety of classes to properly compare them.

In the way of a more practical computational application and following the supervised predictive approach, in Arel (2012), it was applied a multi-layer perceptron (MLP) ANN trained with the back-propagation algorithm to the problem taking as inputs the normalized cone resistance and the friction ratio (ROBERTSON, 1990) and as outputs the classes of Robertson (1990). The aim 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 ANN for $3D$ soil profiling.

A recent work resumes the discussions about soil types division bringing different clustering techniques and comparing them to the charts of Robertson (ROGIERS *et al.*, 2017). The clustering algorithms tested were the x-means and the model-based clustering. As inputs, different arrangements of classification index, normalized cone resistance, normalized friction ratio and depth were considered. The tests were made over an aquifer soil. The results have shown more compatibility between the classes identified by the clustering techniques, mainly by the model-based clustering, than those proposed by Robertson. Based on this, the proposal was a site specific classification approach with clustering techniques instead of applying predefined labels. It is similar to what was done in Facciorusso and Uzielli (2004), but there it was found that the charts of Robertson *et al.* (1986) could give satisfactory results. What could be ascertained is that the methods of Robertson can not identify all kind of soils, which is an empirical evidence of the allegations in Robertson (2016). On the other hand, to state that clustering techniques would always give better results than the methods of Robertson or that this is not suited for classification in general cases, more tests should be done with a more diversified dataset.

Work that apply clustering techniques to the soil classification problem based on CPT data conclude, 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; SOLOMTINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009). The charts of Robertson 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 charts of Robertson *et al.* (1986) can be sometimes even more attractive to practical purposes because of its simplicity on using raw parameters. However, 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, then, a change in Robertson (1991) after the comments in

Jefferies and Davies (1991) in order to cover dilative soils in the pore pressure-based chart. These are the explanations why one of the outputs set considered in this work is the one of Robertson (1991).

Until Robertson (2016), there wasn't an approach that analyzed the intrinsic constraints of soil classification. Studies usually adopted a less diversified dataset, even commonly restricted to some specific area (ZHANG; TUMAY, 1999; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMTINE, 2006; AREL, 2012; ROGIERS *et al.*, 2017). It ends up limiting the generality of the model constructed or of the conclusions obtained. Furthermore, until then, the classification methods used to adopt a granulometrical designation, although they were in general called behavioral. This nomenclature hinders to interpret the soil behavior. One of the first attempts to solve this problem was proposed by Zhang and Tumay (1999), which defined just three soil classes. Likewise, in Schneider *et al.* (2008) and in Schneider *et al.* (2012), soil types were also divided into only sand, clay, with some subdivisions, and transitional, simplifying the 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. No published work was found analyzing this method with machine learning or statistical techniques. These are the reasons why this method is being here studied.

For the feature selection and the assessment of the methods of Robertson (1991), Robertson (2016), predictive machine learning methods based on distance and symbolic techniques were chosen. The symbolic techniques include an inner input feature importance evaluation and selection. These techniques are also robust for general ML applications, are less sensitive to outliers, can extrapolate data and are invariable for linear data transformations. Moreover, these techniques were poorly explored in the literature (BHATTACHARYA; SOLOMTINE, 2006). It explains why this kind of technique was chosen for this work.

The distance-based approach is, in turn, found in literature only for unsupervised and descriptive classification problems with clustering techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMTINE, 2006; LIAO; MAYNE, 2007; DAS; BASUDHAR, 2009; ROGIERS *et al.*, 2017). No applications were found applying supervised and predictive distance algorithms. Work that explored the unsupervised approach had the mainly intent of finding a new class division, without proposing new theoretical interpretations, or to testify if the classes of Robertson are suited. Nonetheless, the results suggest that the distance approach is appropriate (HEGAZY; MAYNE, 2002; LIAO; MAYNE, 2007) and that the nearest neighbor clustering algorithm (single-link) is feasible to face the problem. Among the advantages of this kind of technique, one can mention its simplicity,

manking it handy and reproducible, and its interpretable results, although there is not a construction of a computational model, i.e., an estimator that needs only one training to be defined. In other words, this absence of a model doesn't restrict the interpretation of the method, because classification can be interpreted by abstracting a space distribution of the objects. The classical graphical approach are founded on dividing space into classes regions. The distance-based methods follows the same principle, but it allows a dimensionality expansion. The more separated are the objects, represented as points, the more accurate tend to be the classification based on spatial distribution. Thus, for the assessment and comparison between the methods of Robertson, distance-based methods are suited.

With respect to the input features, it was already detached the relevance of including the depth as one of them to construct a model closer to reality (ROGIERS *et al.*, 2017). Besides, it is known that the soil age is important to distinguish it, because the same soil type with different ages can have unequal resistance (ROBERTSON, 2016). In the mentioned work there is no application of a procedure for feature selection or a features importance assessment or even the attempt of some new inclusions of input features as geological age. The input features are generally arbitrarily chosen and it is also common the inclusion of just two, even in work that employ computational techniques (HEGAZY; MAYNE, 2002; FACCIORUSSO; UZIELLI, 2004; BHATTACHARYA; SOLOMTINE, 2006; DAS; BASUDHAR, 2009). Just a few consider the three basic CPT parameters, cone resistance, lateral friction and pore pressure (LIAO; MAYNE, 2007). It affects the classes separation capacity, because in a space with lower dimensionality the overlaps between classes regions are more critical. This limitation was already noticed in the literature (BHATTACHARYA; SOLOMTINE, 2006), but a solution that creates an artificial input to ease the classification task was proposed, the boundary energy method.

One more drawback found in literature was the lack of a clear methodology using supervised machine learning techniques for soil classification based on CPT parameters. In this context, one of the objectives of this work is to propose a general and reproducible methodology for application of machine techniques to the problem. It is also made a rigorous analysis of features relevance with distance-based and symbolic algorithms. This allows choosing the more important features for the task of soil classification based on the spatial distribution of the objects. Initially, only depth, cone resistance, lateral friction and pore pressure are included. These three last features are considered raw and normalized (ROBERTSON, 1990; ROBERTSON, 2016). The relevance of an attribute is evaluated by its contribution to the models prediction capacity or with some information measures. In the end, the geological age is included as a discrete input called classification of geology (CG) and its contribution is assessed.

2 Soil classification methods

2.1 Influenced by soil granulometry (ISG)

The soil classification method proposed by Robertson (1990) was idealized to be oriented to the soil mechanical behavior. However, the labels assigned to the classes and the number of classes divisions are inspired by conventional granulometrical classes, showing even some compatibility with real soil types (KURUP; GRIFFIN, 2006). It adopts as possible soils types the following: 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 last ones are said to be heavily overconsolidated or cemented.

The raw parameters given by CPT are the uncorrected cone resistance q_c, the lateral friction f_s, the pore pressure usually measured behind the cone tip u_2 and depth z. The input features originally considered by Robertson (1990) are the normalized cone resistance Q_{t1}, the normalized friction ratio F_r and the normalized excess pore pressure B_q. The separation curves between classes were later modified to comprise deep water offshore soils that could be dilative (JEFFERIES; DAVIES, 1991; ROBERTSON, 1991) and the cone resistance normalization was replaced by Q_{tn} (ROBERTSON, 2009). It results in the charts presented in Figures 2.1 and 2.2.

First, the raw cone resistance q_c has to be corrected to take into account the pore pressure aiding penetration, like represented in the right upper corner of Figure 2.2, being turned into the total cone resistance q_t. This correction is more significant for soft soils, which have low q_c and high u_2 values (ROBERTSON, 1990).

However, the same soil type can be subjected to different degrees of consolidation due to the overburden pressure, changing its stiffness and strength. The depth is an information intimately connected to the overburden pressure but in the graphical methods it is not considered because of the constraint of dimensionality. It means that to recognize accurately the soil type, the three basic inputs, cone resistance, lateral friction and pore pressure, have to be modified to take this effect into account.

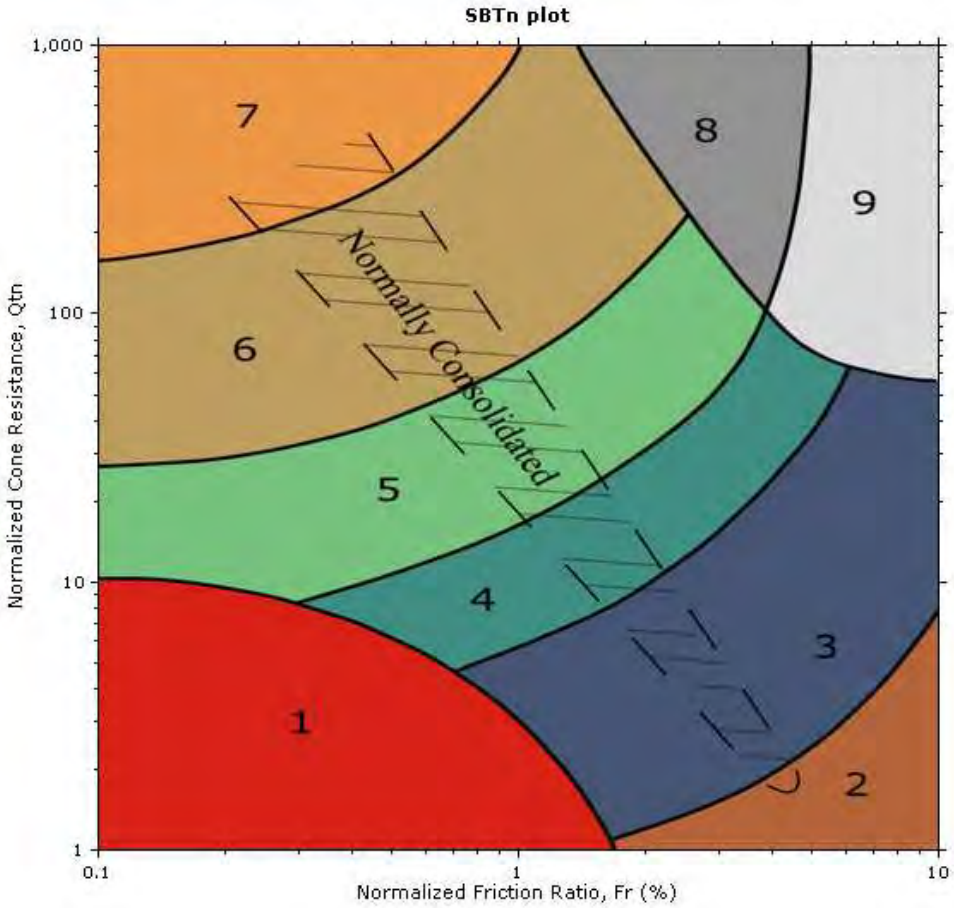


FIGURE 2.1 – $Q_{t1} \times F_r$ chart of Robertson (1990).

To calculate the normalized parameters just the full raw inputs set is used. For this, it is required to estimate the total overburden pressure σ_{v0} and the effective overburden pressure $\sigma'_{v0} = \sigma_{v0} - u_0$. They are obtained by the estimation of the soil unit weight γ (kN/m^3) (LUNNE *et al.*, 1997; MAYNE *et al.*, 2010; MAYNE, 2014). Moreover, the water table depth is needed to compute the equilibrium pore pressure u_0, which is used to determine the excess pore pressure $u_2 - u_0$ and also the effective overburden pressure. If the water table is not known it has to be estimated too. It can be done by fitting a straight line into the chart depth by pore pressure like in the Figure 2.3. This line has to touch the vertical axis close to the same point that the measured pore pressure does and its slope is the water unit weight.

Given these estimations, the cone resistance is initially converted into the net cone resistance $q_n = q_t - \sigma_{v0}$, discounting the overburden aiding penetration. The excess pore pressure $u_2 - u_0$ is taken instead of u_2 and all attributes are divided by the overburden pressure σ'_{v0}. The results are the normalized parameters taking into account the overburden. However, it was initially stated that the lateral friction and the excess pore pressure are both correlated with the cone resistance. Therefore, these parameters are

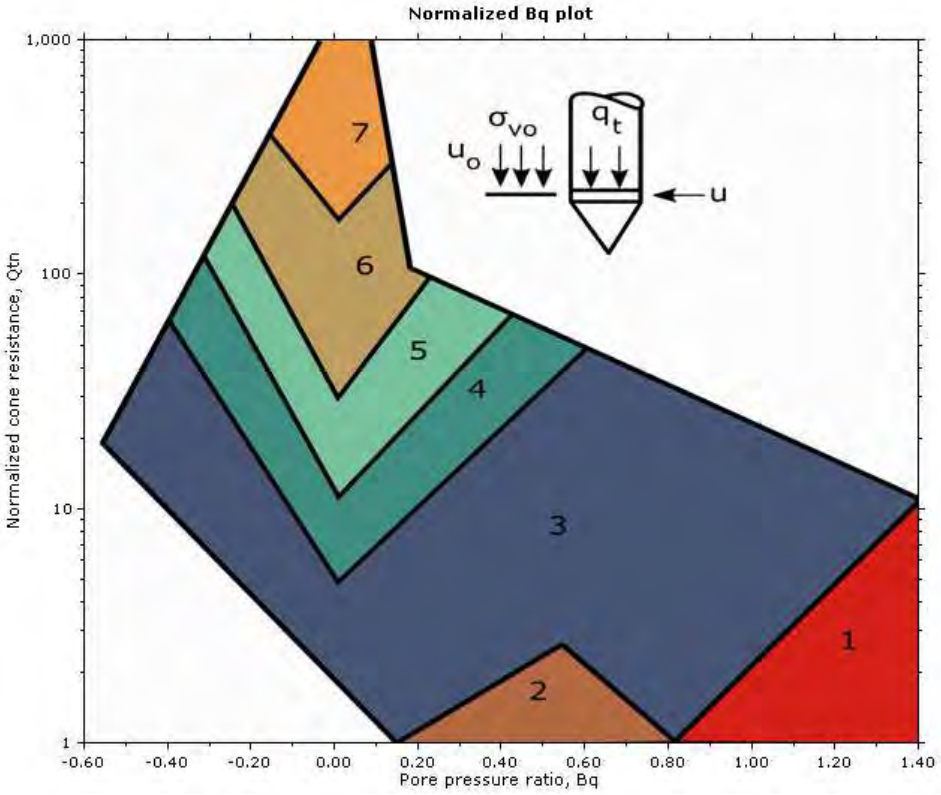


FIGURE 2.2 – $Q_{tn} \times B_q$ chart of Robertson (1991).

finally divided by the cone resistance resulting in the following expressions:

$$Q_{t1} = \frac{q_t - \sigma_{v0}}{\sigma'_{v0}} \tag{2.1}$$

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \tag{2.2}$$

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}} \tag{2.3}$$

Nevertheless, it was assumed the linear dependence between net cone resistance and overburden pressure. It was found that the exponent n of the overburden pressure could vary from 0.5 for sands to 1 for clays (ROBERTSON; WRIDE, 1998; ZHANG *et al.*, 2002). A correlation between this exponent n and the classification index I_c was then proposed (ROBERTSON, 2009):

$$n = 0.381 I_c + 0.05(\sigma'_{v0}/p_a) - 0.15 \tag{2.4}$$

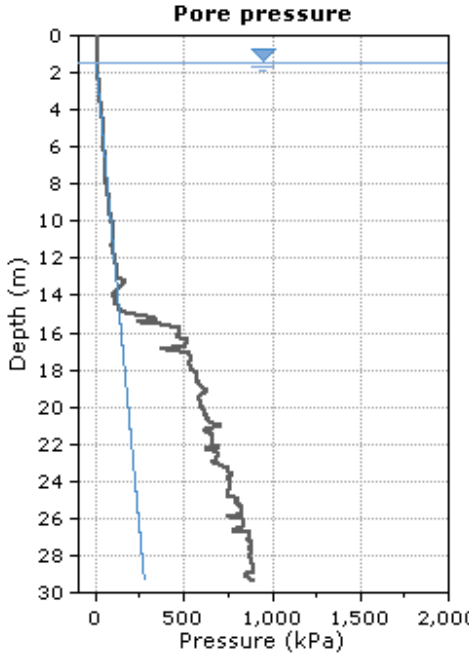


FIGURE 2.3 – Excess pore pressure.

where $p_a = 0.1$ MPa is a reference pressure and I_c is given by Robertson (2009):

$$I_c = \left[(3.47 - \log Q_{tn})^2 + (\log F_r + 1.22)^2\right]^{0.5} \tag{2.5}$$

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

$$Q_{tn} = \left(\frac{q_t - \sigma_{v0}}{p_a}\right)\left(\frac{p_a}{\sigma'_{v0}}\right)^n \tag{2.6}$$

2.2 Focused on soil behavior only (FSB)

The new method proposed by Robertson (2016) leaves the intent to give a more accurate idea of the soil composition and establishing a full behavioral-oriented soil classification. The proposal was to divide soils just in types with well defined behaviors reducing the number of granulometrical references. Then, the soils were first separated into sand-like, clay-like and transitional. It was made considering that sands or coarse-grained soils and clays or fine-grained soils have opposed and well defined behaviors. In general sands have high strength, high permeability and low compressibility, clays have low strength, low permeability and high compressibility. The transitional type was introduced to represent mixed soils or with intermediate granulometry. Afterwards, each one of these classes were divided into contractive or dilative. It was founded on the idea that, despite of soil granulometry, it can have either a contractive or a dilative behavior at big strains or close

to failure. This occurs in penetration, depending on its degree of consolidation, in the case of fine-grained soils, or on its relative density, for coarse-grained soils.

Furthermore, there are some kind of soft clays, which are contractive, that have high sensitivity to disturbance. They are called sensitive, and are identified using a parameter called sensitivity, given by the division of the natural shear strength and the remolded one. As friction ratio is a measure of shear strength and is obtained from CPT taken in a remolded situation, a correlation between sensitivity and friction ratio can be defined from theoretical and empirical assumptions, giving $S_t = 7.1/F_r$ (ROBERTSON, 2009). If the sensitivity is greater than 3.5, then the clay is considered sensitive. In the method of Robertson (2016) it is considered the conservative inferior limit of $S_t = 3$. For this value, the normalized friction ratio F_r is 2%. It was also defined a superior limit to the normalized cone resistance of 10 for sensitive clays because they are soft. Therefore, there are seven classes: 1) CCS: Clay-like – Contractive – Sensitive, 2) CC: Clay-like – Contractive, 3) CD: Clay-like – Dilative, 4) TC: Transitional – Contractive, 5) TD: Transitional – Dilative, 6) SC: Sand-like – Contractive, 7) SD: Sand-like – Dilative. The inputs are the normalized cone resistance Q_{tn} (ROBERTSON; WRIDE, 1998), the normalized friction ratio F_r and the normalized excess pore pressure U_2 (SCHNEIDER *et al.*, 2008). It results the charts shown in Figures 2.4 and 2.5.

The excess pore pressure normalization U_2 is based on the observation that the division of the excess pore pressure by the cone resistance in order to eliminate a correlation between them was in fact creating it (SCHNEIDER *et al.*, 2008). The recommendation was then not to divide one by another so that the normalization of the excess pore pressure should be considered as:

$$U_2 = \frac{u_2 - u_0}{\sigma'_{v0}} \tag{2.7}$$

The curves to separate soil classes were inspired by Schneider *et al.* (2008) and Schneider *et al.* (2012). The $Q_{tn} \times F_r$ chart has closely circular curves in the method of Robertson (1991), whereas in Robertson (2016) it has curves with hyperbolic shapes, as suggested by Schneider *et al.* (2012). The $Q_{tn} \times U_2$ was directly modified from Schneider *et al.* (2008) with minor changes. This is why this chart do not contain all possible classes but basically only the originally established by Schneider *et al.* (2008).

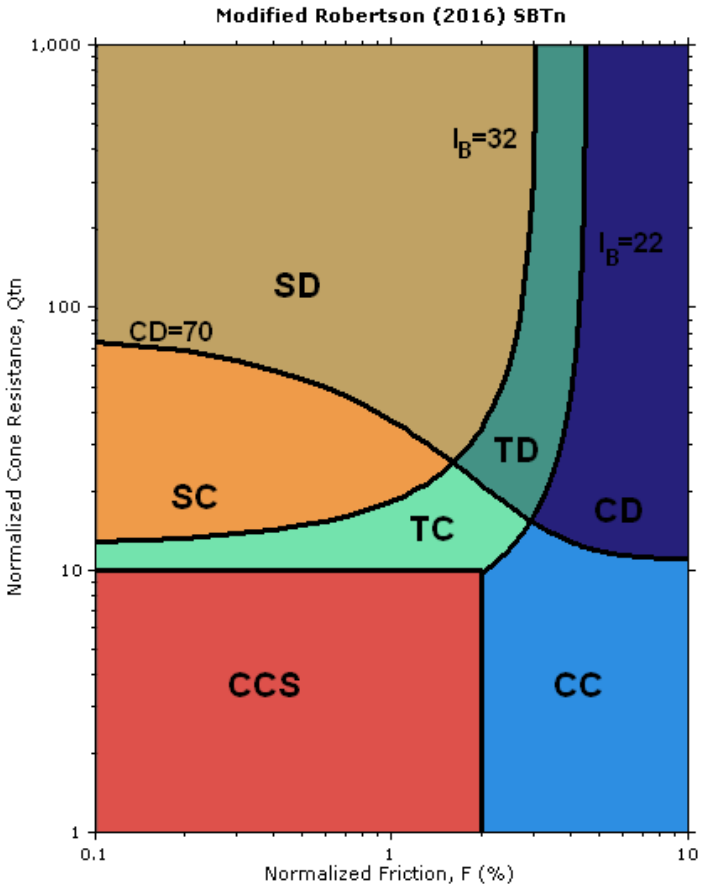


FIGURE 2.4 – Roberton's (2016) $Q_{tn} \times F_r$ chart.

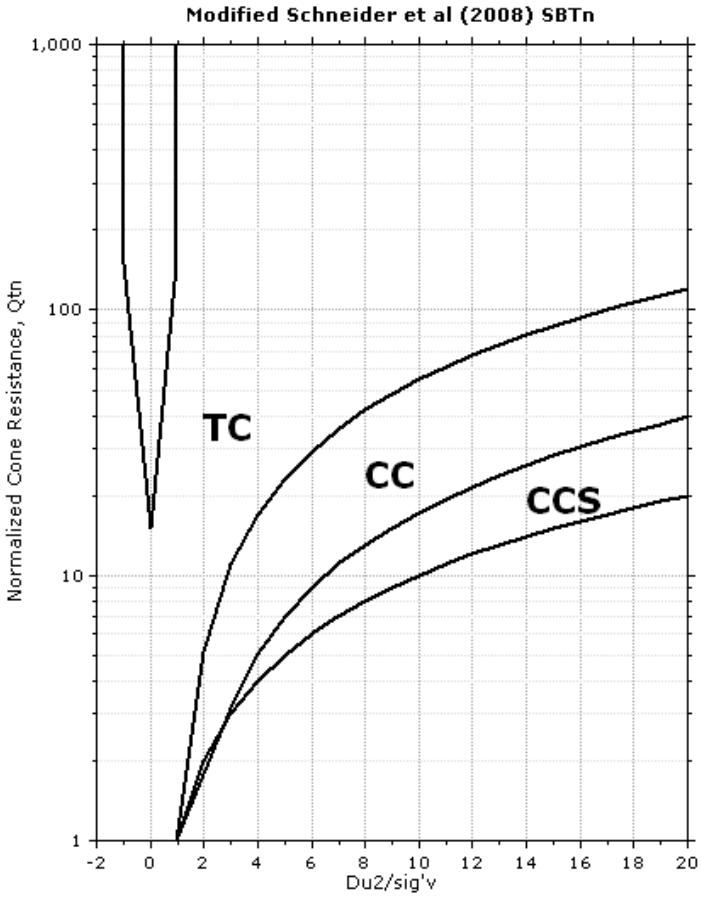


FIGURE 2.5 – Roberton's (2016) $Q_{tn} \times U_2$ chart.

3 Machine learning techniques

3.1 DWNN

The distance-based techniques make use of the hypothesis that close objects in the input features space probably belong to the same class. In other words, it means that objects from the same class tend to be concentrated in the same region of the feature space. Although, there can be overlaps between classes regions, this can be considered a valid rule. To compute distances it has to be defined a specific distance metric and one of the most common distance metrics used is the Minkowski's distance metric. It provides, for a pair $\mathbf{x}_i$ and $\mathbf{x}_j$ of objects in a feature space with dimensionality d, the following:

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[p]{\sum_{l=1}^{d} \left| x_i^l - x_j^l \right|^p} \tag{3.1}$$

In this expression, $1 \leq p < +\infty$ is a free parameter. This metric is sensitive to scale and the lower the value of p the higher the sensitivity to outliers too. The Euclidian distance corresponds to $p = 2$, which is considered in this work.

The nearest neighbors-based algorithms are a lazy learning method because it tests each object from the dataset, memorizes the distances and delivers a direct conclusion from it without creating a decision model that can guide future decisions. The simplest form of this kind of algorithm is the 1-NN (COVER; HART, 1967), which considers just the nearest neighbor for the decision. This algorithm can be described by the following pseudo-code:

Nearest neighbor pseudo-code

Inputs: training set $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1, ..., n\}$,
test object to be classified $(\mathbf{x}_t, y_t)$ and a distance function
Output: y_t

$d_{min} \leftarrow +\infty$
for each $i \in 1, ..., n$ **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

The predicted output of the tested object is the same of its nearest neighbor and an upgrade of this algorithm is the KNN. This method takes into account the k nearest neighbors to predict the output. For a classification problem as the present, the predicted output is the mode of its k nearest neighbors classifications, what corresponds to make a voting, electing the most voted class. There has to be some care with the value of k. A high value can consider too much different neighbors in the prediction and rises the computational cost as well as too different objects from the tested one. On the other hand, a low value of k may not give enough information for the prediction.

A way to improve the k value is through a calibration or training procedure in which it is increased while the accuracy rises. Another improvement that can be used is to apply weights that are functions of the distance to the neighbors votes, getting a weighted mode as the prediction. This is the essence of the DWNN algorithm (DUDANI, 1976). For the Gaussian weighting, the weights w_i are defined by Hechenbichler and Schliep (2004):

$$w(d) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d^2} \tag{3.2}$$

where d is the Minkowiski's distance.

3.2 Symbolic algorithms

The symbolic methods are knowledge-based techniques. It means that these ML techniques construct symbolic models, which are easier theoretically interpretable, differently from the black-box techniques such as ANN and SVM. The main examples of symbolic methods are the decision trees (DT) and the random forests (RF), which are explored

in this work. The DT use the divide-and-conquer strategy. This approach consists in dividing a complex problem into simpler pieces to which the same technique is applied. Then the subproblems solutions are combined. The RF is a generalization of the DT.

The main examples of symbolic methods are the decision trees (DT) and the random forests (RF), which are explored in this work. The decision tree is formally a directed acyclic graph (DAG) wherein each node is either a leaf or a division. The division nodes have two or more descendant nodes and a condition applied to a certain input attribute so that depending on the value assumed by this attribute the ascendant node conducts to a different successor. The leaf nodes are in turn those that assign a value to the target feature as a function of the output values of the training instances which fall into the node. For the classification problem, the mode is adopted as the classification function. Therefore, the decision tree can be defined recursively by the composition of decision nodes, which apply a condition to a given feature, and leaf nodes, which contain an output prediction.

In a classification problem, the DT algorithm works partitioning the feature space in order to delimit the corresponding classes regions. Each leaf node regards a specific class. A DT property is that it covers the whole feature space, allowing extrapolations. The hypothesis space defined by the DT fits in the Disjunctive Normal Form (DNF). It means that the algorithm is composed by a conditional part, which partitions the input features space following conditional operators, and a conclusion part, which establishes a prediction for the output attribute.

There are several algorithms by which a DT can be implemented. It includes the Hunt algorithm, which was one of the firsts to be conceived and is a reference for other more recent ones as the CART (BREIMAN *et al.*, 1984), the ID3 (QUINLAN, 1986), the C4.5 or J48 (QUINLAN, 1993) and others. If $\mathbf{D}_t$ is the training instances set which achieves the leaf node t, then the Hunter algorithm can be described by the following pseudo-code:

Hunt algorithm

if $x_t \in y_t, \forall x_t \in \mathbf{D}_t$
then t is a leaf node labeled with y_t
else if $\mathbf{D}_t = \emptyset$
then t is a leaf node with a *default* label y_d
else if objects in D_t belong to different classes
then divide the instances set in subsets based on some attribute
  apply the procedure to the generated subset

Furthermore, a basic algorithm for constructing a DT can be described by the following pseudo-code:

Training a DT: function CreateTree(D)

Input: Training set $D = \{(x_i, y_i), i = 1, ..., n\}$
Output: DT
if (stop criterion) **then**
  return leaf node with the label that maximizes the cost function
choose the input attribute that maximize the division criterion in D
for each partition D_i based on the attribute values
  induct a subtree $Tree_i = CreateTree(D_i)$
return tree with the decision node of the chosen attribute and descendants $Tree_i$

Relevant decisions in the DT structuring are the choice of the division criterion in the division nodes, of the feature to be used for the separation and of the stop criterion. The division method depends on the feature type. For discrete or symbolic features, a number of branches that equals the number of possible values of the feature can be adopted. For continuous inputs, it can be discretized. The choice of the feature attached to the node is based on the quality of the division made by it looking one step forward, which can be considered a lazy heuristic.

For classification, this quality can be given by impurity functions, which evaluate the classes distributions in the node. If t is a division node and p_i is the probability of observing an example belonging to class c_i in t, then the impurity $i(t)$ of t is a function applied over the classes proportions in t. Mathematically:

$$i(t) = \phi(p_1, p_2, ..., p_k) \tag{3.3}$$

where ϕ is the impurity function and k is the number of classes.

If S is a conditional test that divides the objects into two subsets **L** and **R**, then the impurity decrease by S can be measured by:

$$\delta(S) = \phi(p_1, ..., p_k) - P_L \cdot \phi(p_{1,L}, ..., p_{k,L}) - P_R \cdot \phi(p_{1,R}, ..., p_{k,R}) \tag{3.4}$$

where P_L and P_R are the probabilities of the example going to the **L** or the **R** subsets, respectively.

A common impurity measure is the Gini index. For a given node t with a proportion

p_i for each class i among k classes, the Gini index is given by:

$$Gini(t) = 1 - \sum_{i=1}^{k} p_i^2 \tag{3.5}$$

Another way to measure the impurity decrease is through the information gain IG, which is function of the entropy $H(\mathbf{D})$ of the division node instances set and the expected entropy $E(A, \mathbf{D})$ of the v subsets generated with the division by the attribute A. The entropy and the expected entropy are respectively defined as:

$$H(\mathbf{D}) = -\sum_{i=1}^{k} p_i \cdot log_2(p_i) \tag{3.6}$$

$$E(A, \mathbf{D}) = \sum_{j=1}^{v} \sum_{i=1}^{k} p_{i,j} H(\mathbf{D}_i) \tag{3.7}$$

The information gain $IG(A, \mathbf{D})$ achieved with the choice of the attribute A for the splitting of the dataset $\mathbf{D}$ is then given by:

$$IG(A, \mathbf{D}) = H(\mathbf{D}) - E(A, \mathbf{D}) \tag{3.8}$$

Pruning is usually employed to avoid overfitting and to reduct noise. This procedure consists in replacing some division nodes by leaf nodes. A bias is introduced to simplify the model following the principle of the Occam's razor. It states that if a simpler hypothesis explains data then it is enough. The pruning can be made while DT is trained, called pre-pruning, or after DT is trained, called post-pruning. This procedure looks for the balance between the complexity of the model and the training error. The CART algorithm uses post-pruning and the Gini index as impurity measure. The C4.5 and the C5.0, which is a modification of the previous, uses the post-pruning and the IG as impurity measure. The ID3 algorithm is applicable only for nominal features, using post-pruning and the IG.

The advantages of DT include the fact that it does not require data normalization, the possibility of training data intervals extrapolation and the approximation of Bayes error on the limit. It is also a robust models because they are invariant to monotonous transformations on inputs and less sensitive to outliers. Moreover, there is a feature selection inside its construction, it produces a more interpretable model than other ML techniques and it can be more efficient than other algorithms depending on the strategy used for training. Some drawbacks are the computational cost for dealing with continuous inputs and the instability to training set variations.

A way to overcome this instability and to avoid overfitting is using random forests

(RF). This technique consists in generating different decision trees by bootstrap aggregating or bagging. In other words, trees are generated from training set samples taken with replacement and also by getting features samples (BREIMAN, 2001). The final predictions for a classification problem are given by voting. Another technique is the adaptive boosting, which is made by giving in each new training set sampling a higher probability of choice for the classes more confused by the previous trees created. The resulting model is called boosted decision trees, referred just as DT in this work.

4 Dataset analysis

The full dataset employed in this work is composed by 111 soundings of which 38 were provided by Professor P. K. Robertson and the other 73 were made available online by Professor P. W. Mayne [1]. The dataset provided by Professor P. K. Robertson is described in Table 4.1 while the one provided by Professor Paul Mayne is described in Table 4.2. The soil unit weight wasn't available for all soundings and it was automatically estimated inside the CPeT-IT environment as well as all normalizations. The ground water table was in turn available for all soundings and did not had to be estimated.

However, the soil geological age was not available for all soundings but just for those provided by Professor P. K. Robertson. It was independently determined based on the information given in the work of (ROBERTSON, 2016). The soundings for which the soil age could be determined are shown in Table 4.3. This partial dataset is called here as geological dataset. As it is an initial study around the introduction of this feature as an input, the geological age was treated as an ordered discrete input called here classification of geology (CG). It is supported by the fact that, as the interval between geological periods and epochs decreases, the soil changing with time increases because the soil becomes more degraded and vulnerable to the weathering agents.

The presented datasets are briefly analyzed to justify some methodological procedures. The analysis of the objects distribution among classes are one useful analysis because it allows the identification of the data real diversity and the balancing need. Thus, histograms were plotted considering each one of the classification methods, the ISG and the FSB, and each one of the datasets, the full and the geological one. The results are shown in Figure 4.1.

It can be seen that, for the FSB method, all classes have enough members to make a general analysis possible for the full as well as for the geological datasets. On the other hand, for the ISG method, the geological dataset is compromised. For that reason, the geological age introduction analysis is made just for the FSB method. The fact that this classification method identifies the geological age as an important factor make this limitation less harming.

[1] http://geosystems.ce.gatech.edu/Faculty/Mayne/Research/index.html

TABLE 4.1 – 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

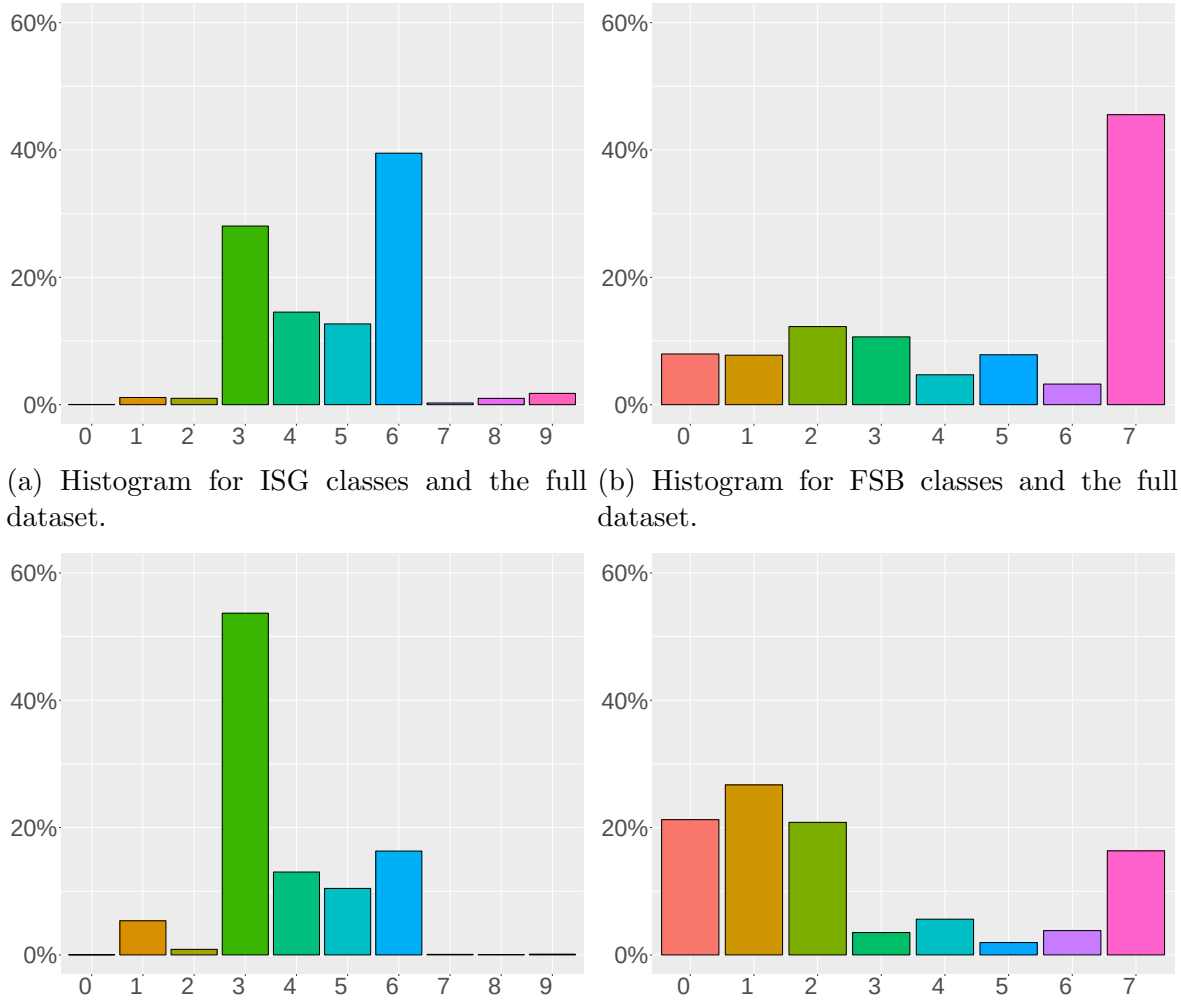
TABLE 4.2 – Dataset from P. W. Mayne.

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 4.3 – Geological dataset (Classification of Geology – CG).

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

The class 0 observed for both methods regards to those objects that could not be classified. For the ISG method, it occurs when the object falls outside the $Q_{tn} \times F_r$ chart. For the FSB method, it occurs when the charts predictions do not match or when the soil is structured, meaning cemented or aged. As this class is relatively wide for the FSB method, it is more deeply investigated later.



(a) Histogram for ISG classes and the full dataset.

(b) Histogram for FSB classes and the full dataset.

(c) Histogram for ISG classes and the geological dataset.

(d) Histogram for FSB classes and the geological dataset.

FIGURE 4.1 – Histograms.

The box-plots for each input feature were plotted by class to verify the need of data cleaning. The results are presented in Figure 4.2 for the ISG classes and in Figure 4.3 for the FSB classes. The box-plots highlight the first, second and third quartiles and indicate as outliers the black points out the upper and lower limits defined by the dotted lines. These points are supposed to be outliers but it depends on the features values statistical distribution. Nonetheless, the normalizations reduce these points, suggesting that they tend to reduce outliers, creating a more concentrated distribution. In spite of this reduction, it can be seen that there are still candidates to be outliers and some data

cleaning procedure is convenient.

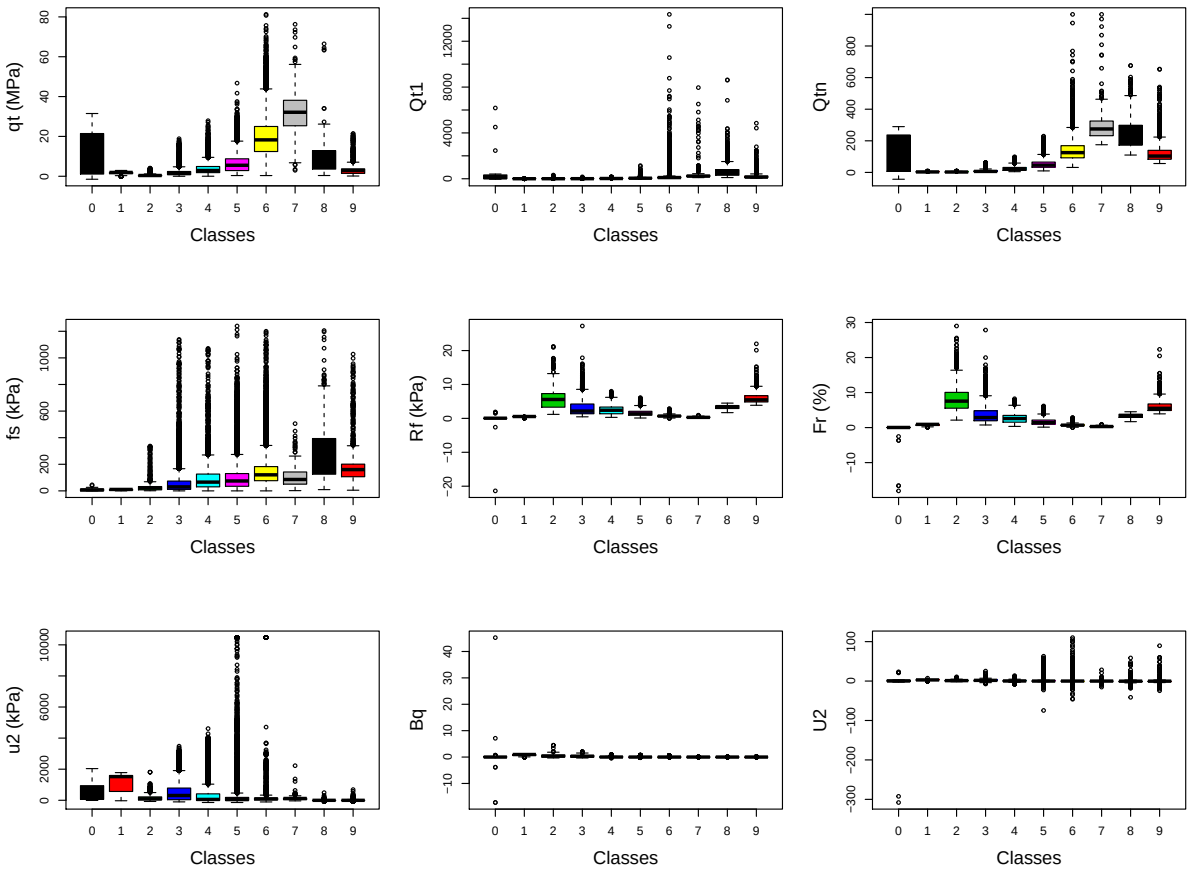


FIGURE 4.2 – Box-plots of features values for ISG classes.

One more possible analysis is the evaluation of the correlation between features. It is important to verify if some pair of features tends to be interchangeable and if the analysis of different normalizations is valid. To identify consistent correlations, all classes have to be analyzed together. Thus, correlation matrices considering the full and the geological datasets were plotted. These matrices are presented in Figure 4.4. Matrices without the geological age correspond to the full dataset.

It can be noticed that the raw total cone resistance q_t is in fact strongly positively correlated with the lateral friction f_s and the friction ratio R_f reduces this correlation, suggesting that this attribute could contribute to distinguish more clearly the objects than the raw lateral friction. On the other hand, just the raw pore pressure seems to be correlated with depth, what is reduced by the normalizations.

The normalized excess pore pressure B_q of (ROBERTSON, 1990) is strongly negatively correlated with the normalized cone resistances and the normalization U_2 reduces it indeed, confirming the statements found in literature (SCHNEIDER *et al.*, 2008). The cone resistance normalization Q_{t1} is strongly positively correlated to the normalization Q_{tn},

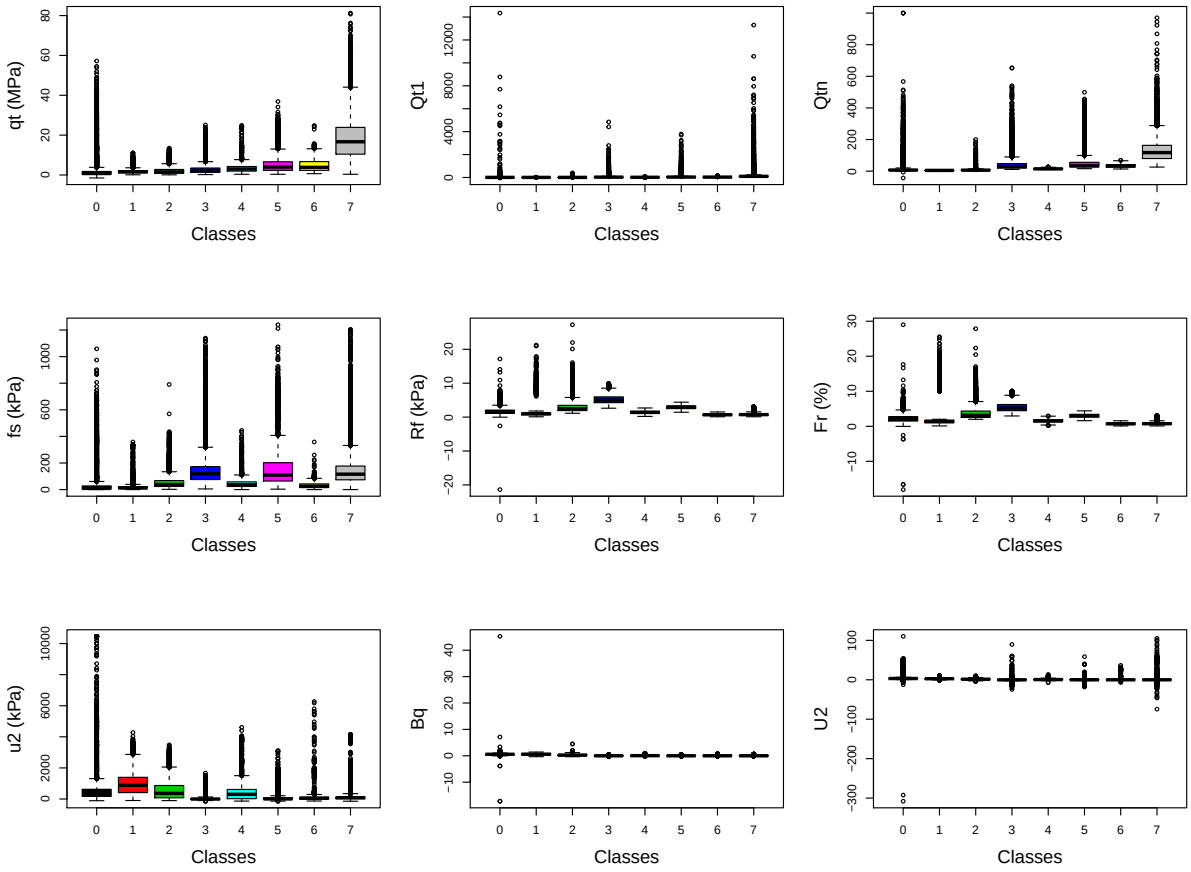


FIGURE 4.3 – Box-plots of features values for FSB classes.

suggesting that they are interchangeable. The geological age does not show significant correlation to any other parameter, what can be justified by the fact that all classes were put together.

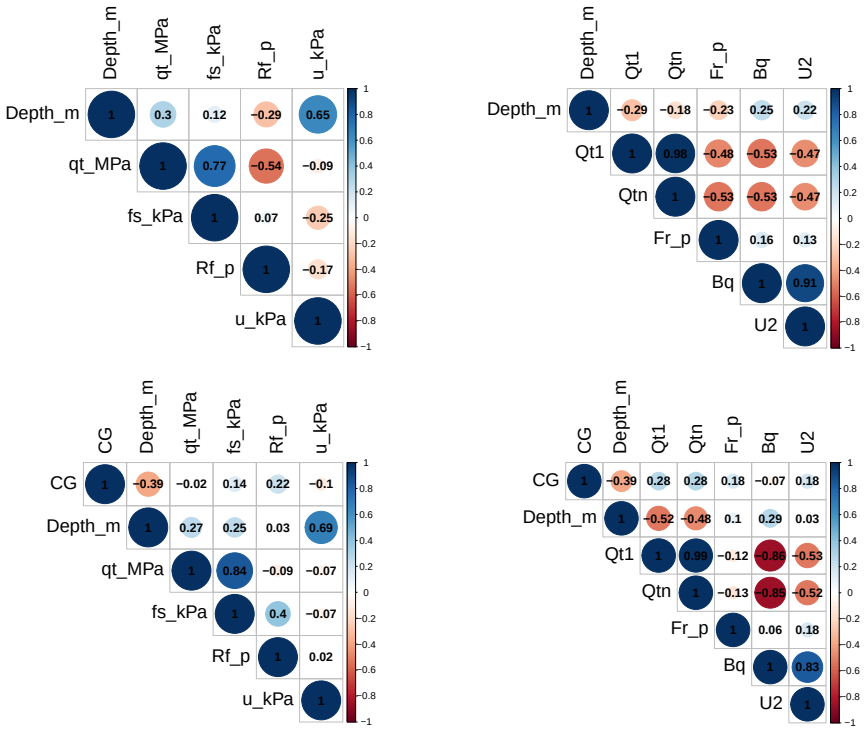


FIGURE 4.4 – Correlation matrices for all input features.

5 Methodology

5.1 Description

The methodology presented here is general and valid to any ML technique with few adaptations. It is applied to the datasets with the outputs given by CPeT-IT software v2.0.2.5 using symbolic algorithms, namely boosted decision trees (DT) and random forests (RF), and the distance-weighted k-nearest neighbor (DWNN) algorithm for comparison. These algorithms are tested with 10 inputs and outputs combinations:

- Basic combinations – for ISG and FSB outputs (full 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 – just 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 – for ISG and FSB outputs (full dataset)
  - ISG inputs: Q_{tn} and F_r (%)
  - FSB inputs: Q_{tn}, F_r (%) and U_2

The 10-fold cross-validation procedure is employed (STONE, 1974). It means that each one of the 10 combinations above are tested 10 times with different training and testing sets, resulting in 100 testing cases for each one of the techniques, totalizing 300. For each combination, the original dataset is divided into 10 partitions called folds maintaining the classes proportions. Each one of these partitions are tested one by one. From the 9 remaining folds, one is taken as the validation set and the other 8 compose the training set, used to train the technique for calibration and for testing. The folds are previously

generated and fixed for all combinations that use the same original dataset in order to make the comparisons more reliable.

The calibration is done by varying the adjusting parameters of the technique and measuring its performance on the validation set. For the DWNN, the number of nearest neighbors considered for voting k is increased by 2 from $k = 1$ so that the number of electors is odd, reducing the possibiliy of ties. For DT, the number of trees t constructed with the adaptive boosting method is increased by the number of classes c from $t = c$. For the RF, the calibration procedure is not necessary, because a high number of trees can be generated by bagging without a compromising computational cost. The number of trees for the RF was fixed in 500 by default.

The calibrated technique is then tested with the testing set, giving an accuracy value. For a complete 10-fold testing procedure, 10 accuracy values are obtained. A mean value and a standard deviation are calculated to represent the performance of the technique for the features combination. These statistically representative values are used to compare the techniques with the statistical test of Friedman (FRIEDMAN, 1937), the post-hoc Nemenyi statistics (NEMENYI, 1963) and the statistical test of Wilcoxon (WILCOXON, 1945).

The symbolic algorithms have an inner feature selection procedure, allowing features importance evaluation based on some criterion. For DT, the importance of a feature can be measured by usage or by split according to the function used for implementing it. The usage importance is given by the percentage of training samples that passes through division nodes associated with that feature. The split importance is the fraction of division nodes associated with the feature. For the RF, on the other hand, these kind of easily interpretable measure can not be made due to the incompatibility between the features of the trees that compose the forest, leading to the adoption of the mean decrease in Gini, which is the importance measure considered by the function used for implementing it. This measure is defined as the mean of the decrease in Gini index for the feature divided by the variance of this decrement.

5.2 Preprocessing

The proposed methodology for preprocessing data is summarized in Figure 5.1. It is applied for each 10-fold testing case as the training, validation and testing sets vary and have to be preprocessed from the beginning. As the inputs and outputs also vary for different tested combinations, for each one of the 300 testing cases the preprocessing procedure have to be applied.

A first and fundamental point that has to be defined is the performance measure to be adopted. As can be seen in Section 3.1, the data is strongly imbalanced. Considering that

the folds have the same classes proportions of the original dataset, they are imbalanced too. Thus, the performance measure has to take this unbalancing into account in order to reflect the reality because the testing and the validation sets are not balanced. The general performance measure for multi-class problems is the accuracy, which is calculated by the number of right predictions divided by the total number of objects. Given the number of right predictions of each class, if it can be converted to an equivalent number of right predictions that would be achieved if the classes were balanced, then the equivalent accuracy for this case would be the mean recall. This is the reason why the mean recall was adopted as the performance measure.

The normalizations are made just limiting each feature values in the training set to the interval between 0 and 1. It means that the lowest feature value in the training set becomes 0 and the highest becomes 1. These initial lowest and highest values for each feature are used to transform with the same rule the values of the testing and validation sets. Mathematically, if v is a value of the feature A, then the normalized value v' is given by:

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)} \tag{5.1}$$

where $\min(A)$ and $\max(A)$ are respectively the minimum and maximum values of A in the training set.

The data cleaning procedure proposed consists in a double filtering. The first step is a univariate statistical pre-filtering identifying possible outliers. For each feature of each class the values that are outside the interval $[Q_1 - 1.5IQ, Q_3 + 1.5IQ]$ are selected as potential outliers, where Q_1 is the first quartile, Q_3 is the third quartile and $IQ = Q_3 - Q_1$ is the interquartile range. This need for a second filtering procedure is due to the fact that depending on the statistical distribution of the feature values the instances identified by the univariate statistics may not be outliers or may not disturb the classification. The pre-filtered objects are then subject to an Edit Nearest Neighbor (ENN) filtering (WILSON, 1972). It is based on the idea that just those objects that disturb the classification have to be considered harming. The ENN applied for filtering purposes could be applied to the whole training set, removing noise too. Nevertheless, beyond the high computational cost, it could remove relevant information and noise are already partially dealt by the calibration procedure. Therefore, the combination of the procedures seems to be a better option and it is reinforced by empirical evidence.

In the balancing step, the final number of elements adopted for the classes was the minimum value between 1000 and twice the number of elements in the minority class. Thus, the classes with a number of objects greater than this final number have to be reduced and the other ones have to be increased. As the CPT produces a lot of similar

data due to the several measurements obtained within each soil layer, a random choice of the objects is made for the majority classes reduction. For the minority classes increment, each new artificial object is estimated from a certain number of real objects. This is the SMOTE procedure (CHAWLA *et al.*, 2002). In the present case, the number of real objects considered is equal to $d + 1$, where d is the feature space dimension. It follows the idea that with $d+1$ objects a d-simplex geometry can be defined in a d-dimensional space and the new object stays at its centroid, filling the space and staying outside the hyperplanes defined by d or less points. The Figure 5.2 shows an example of this procedure applied to the $Q_{tn} \times F_r$ space.

Nonetheless, the balancing tends to remove information from the majority classes. Thus, the gain with the better filling of classes regions can be lower than the loss of information. This is the reason why a pre-test of the balanced and imbalanced training sets is performed before training and testing the technique. This pre-testing is done using the validation dataset.

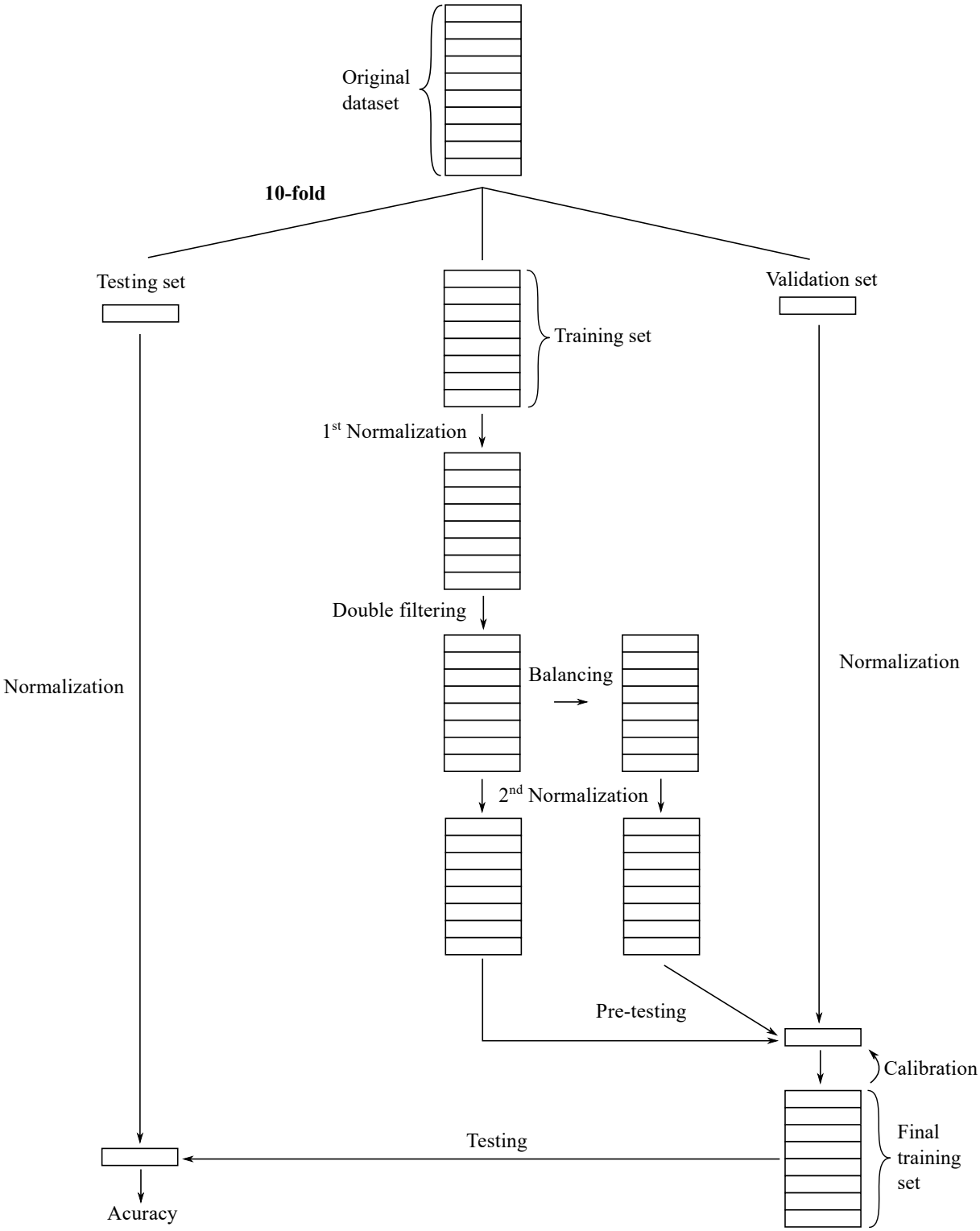


FIGURE 5.1 – Summarized methodology for data preprocessing.

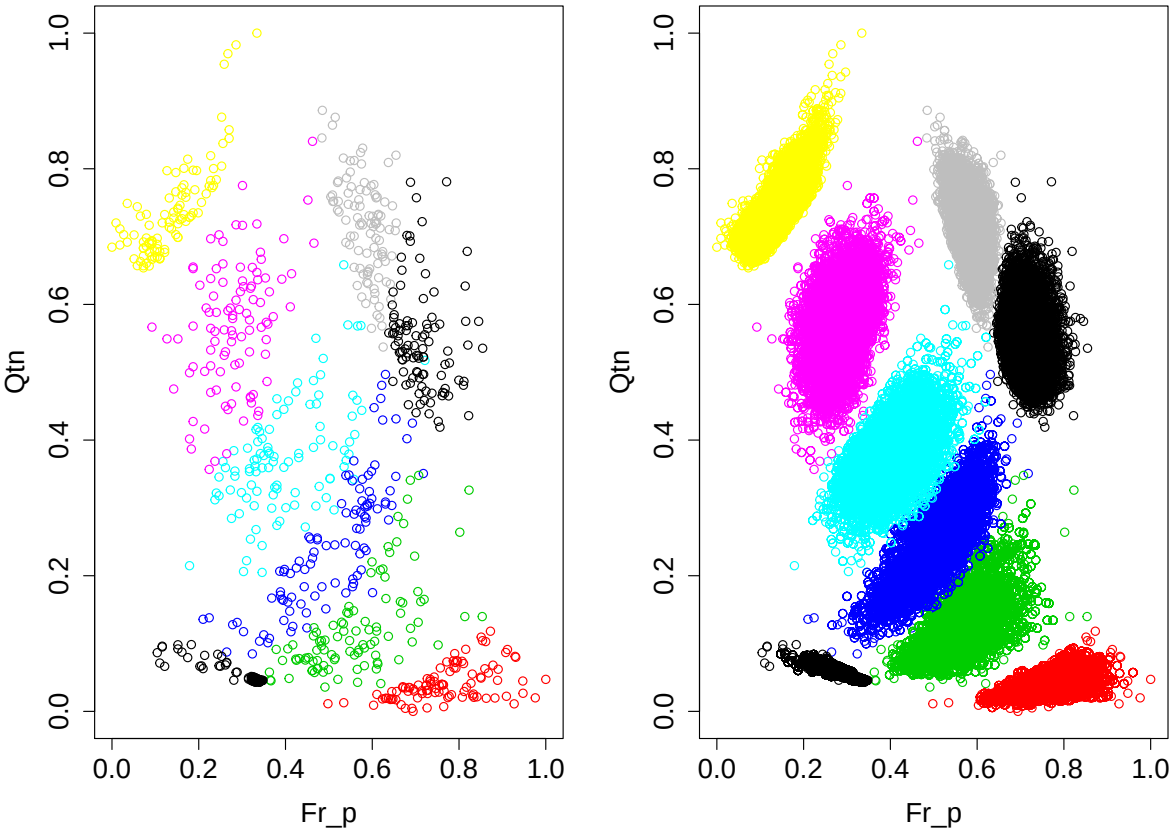


FIGURE 5.2 – Balancing example.

6 Results and Discussions

6.1 DWNN

The DWNN technique was applied to the problem through the command *kknn()* from the R *kknn* package. The results obtained for the main features combinations with the DWNN, represented by the mean accuracy (MA) of the 10-fold cross-validation procedure and the corresponding standard deviation (SD), are shown in Tables 6.1, without the geological age, and 6.2, with the geological age as an input. An additional test was made in order to evaluate the technique capacity to reproduce the classification methods using the biased inputs, in other words, the inputs used by the CPeT-IT software to generate the reference outputs. The results for this test are found in Table 6.3. As expected, the biased inputs for the ISG method give the best accuracy. However, for the FSB method, the inclusion of depth increases slightly the performance. It can be also noticed that the non normalized inputs could replace the normalized ones maintaining a reasonable accuracy and that the normalized inputs of Robertson (1990) and of Robertson (2016) are not interchangeable.

In order to evaluate some eventual specific technique weakness, it was calculated the mean confusion matrices for the biased features combinations. These matrices were obtained calculating the mean of the 10-fold confusion matrices. Their lines correspond to the predicted values while their columns correspond to the reference values. The recall of

TABLE 6.1 – Results for the DWNN with the full dataset and without the geological age as input.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
ISG	z, q_t, f_s, u_2	1	90.23	0.66
	z, Q_{t1}, F_r, B_q	1	89.40	0.90
	z, Q_{tn}, F_r, U_2	1	93.13	0.70
FSB	z, q_t, f_s, u_2	1	90.28	0.48
	z, Q_{t1}, F_r, B_q	1	88.82	1.62
	z, Q_{tn}, F_r, U_2	1	93.77	0.57

TABLE 6.2 – Results for the DWNN with the geological dataset and the geological age included as input.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	1	91.03	1.04
	CG, z, Q_{tn}, F_r, U_2	1	94.73	0.73

TABLE 6.3 – Results for the DWNN with the full dataset and the inputs used by the CPeT-IT to generate the reference outputs.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
ISG	Q_{tn}, F_r	1	96.58	0.59
FSB	Q_{tn}, F_r, U_2	3	93.06	0.34

each class is calculated dividing the elements in the principal diagonal by the sum of the elements in the corresponding column. Therefore, the classes with lower recall are those for which the technique has inferior predictive performance. The matrices for the ISG and FSB methods are, respectively:

$$\mathbf{C}_{ISG} = \begin{bmatrix} 96.5 & 0 & 10.8 & 4.5 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 84.1 & 87.1 & 0 & 0 & 0 & 0 & 0 & 0.2 \\ 0.1 & 1.3 & 2203.4 & 50.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 87.3 & 1139.6 & 47.9 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 44.1 & 1014.4 & 29.8 & 0 & 0.2 & 0.2 \\ 0 & 0 & 0 & 0 & 4.3 & 3295.4 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 33.9 & 22.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10.6 & 19.8 & 0.1 & 84.4 & 0.8 \\ 0 & 0 & 11.5 & 5.6 & 7.7 & 0 & 0 & 0.5 & 150.9 \end{bmatrix} \tag{6.1}$$

$$\mathbf{C}_{FSB} = \begin{bmatrix} 487 & 53 & 43.3 & 2.4 & 4.5 & 2.9 & 0.5 & 3.6 \\ 47.9 & 605.7 & 4.7 & 0 & 2.3 & 0 & 0 & 0 \\ 57 & 3.1 & 988.4 & 11.2 & 3.9 & 0.2 & 0 & 0 \\ 15.6 & 0 & 9.9 & 890.1 & 0 & 7 & 0 & 0 \\ 10.3 & 3.3 & 2.7 & 0 & 385.8 & 3 & 2.4 & 0.1 \\ 23.4 & 0 & 0.3 & 7.5 & 3 & 652.3 & 0 & 11.3 \\ 4.8 & 0 & 0 & 0 & 2.4 & 0 & 270.2 & 6.1 \\ 35.4 & 0 & 0 & 0.2 & 0.1 & 5.8 & 4.2 & 3875.1 \end{bmatrix} \tag{6.2}$$

For the ISG method, the classes go from 1 to 9, while for the FSB method they go

TABLE 6.4 – Results for the DWNN and the full dataset without the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	z, q_t, f_s, u_2	1	90.51	0.48
	z, Q_{t1}, F_r, B_q	1	89.23	0.71
	z, Q_{tn}, F_r, U_2	1	95.22	0.40

TABLE 6.5 – Results for the DWNN and the geological dataset with the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	1	88.94	1.39
	CG, z, Q_{tn}, F_r, U_2	1	94.47	1.48

from 0 to 7, because the class 0 is included in this case. It can be seen that, for the ISG method, the more hardly identified classes are 3, 4, 5 and 6. On the other hand, for the FSB method, the class 0 is the main source of confusion, with a recall of 71.47%, followed by class 1, which is mainly confused with class 0. Thus, removing the class 0 from the training, validation and testing sets, it is obtained a mean accuracy of 97.61% with a standard deviation of 0.30% for the 10-fold with the biased inputs and the following confusion matrix:

$$\mathbf{C}_{FSB} = \begin{bmatrix} 658.1 & 5.3 & 0 & 5.8 & 0 & 0.1 & 0 \\ 2.3 & 1025.7 & 15.7 & 5.3 & 0.4 & 0 & 0 \\ 0 & 13.1 & 888.1 & 0 & 6.9 & 0 & 0.1 \\ 4.7 & 4.3 & 0 & 384.5 & 4.9 & 4 & 0.3 \\ 0 & 0.9 & 7.5 & 3.6 & 654.4 & 0 & 19.7 \\ 0 & 0 & 0 & 2.5 & 0 & 268.6 & 11.2 \\ 0 & 0 & 0.1 & 0.3 & 4.6 & 4.6 & 3864.9 \end{bmatrix} \tag{6.3}$$

With the class 0 removal, the results in Tables 6.4 and 6.5 are also obtained for the remaining features combinations. It can be seen that the performance gain, when it occurs, is modest and that there is even loss of performance for some combinations, what is observed for those with the geological age. It is justified by the existence of many structured soil instances in class 0 and the geological age aids their identification. Therefore, the elimation of the class 0 is more beneficial for the biased combinations but it can be maintained with the DWNN without severe harm.

TABLE 6.6 – Results for the DT and the full dataset without the geological age.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
ISG	z, q_t, f_s, u_2	27	91.71	0.76
	z, Q_{t1}, F_r, B_q	18	95.81	0.60
	z, Q_{tn}, F_r, U_2	9	97.60	0.43
FSB	z, q_t, f_s, u_2	24	91.32	0.47
	z, Q_{t1}, F_r, B_q	24	96.40	0.22
	z, Q_{tn}, F_r, U_2	40	97.31	0.22

TABLE 6.7 – Results for the DT and the geological dataset with the geological age included as input.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	8	91.66	0.90
	CG, z, Q_{tn}, F_r, U_2	8	97.01	0.84

6.2 Decision Trees

The DT with adaptive boosting was applied to the problem with the modified C4.5 algorithm of Quinlan (1993) through the command *C5.0()* from the R *C50* package. The results are shown in Tables 6.6, without the geological age for the full dataset, and 6.7, with the geological age for the geological dataset. The technique was also applied to the biased features combination, producing the results present in Table 6.8. The greater mean accuracy and lower standard deviation achieved by the DT in comparison with the DWNN points that the DT has a higher capacity to reproduce the classification methods charts. Comparing the results of the biased features combinations with the others, it is seen that the addition of depth raises their performances both for the ISG and the FSB methods.

The DWNN with Gaussian weighting and the boosted DT were compared based on their mean accuracies for the 10 tested combinations. The test of Wilcoxon points out the statistical prevalence of the boosted DT over the DWNN with a p-value of 0.20%, assuming that a p-value lower than 5% ensures that the null hypothesis of performance equivalence can be rejected.

TABLE 6.8 – Results for the DT with the inputs used by the CPeT-IT to generate the reference outputs.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
ISG	Q_{tn}, F_r	18	96.97	0.57
FSB	Q_{tn}, F_r, U_2	32	94.69	0.25

TABLE 6.9 – Importance of the non normalized inputs.

	By usage (%)		**By split (%)**	
	ISG	FSB	ISG	FSB
z	100	100	25.43	28.87
q_t	100	100	29.65	25.60
f_s	100	100	29.48	23.86
u_2	99.11	99.99	15.44	21.67

With the DT, the best inputs were also those of Robertson (2016). However, those of Robertson (1991) had a close performance, mainly for the FSB outputs. In this case, the strong correlation between Q_{t1} and Q_{tn} is reflected on the performance, unlike what is observed for the DWNN. It is probably due to the trees capacity to be invariable to monotonous transformations. It can be also noticed that, for DT, the classification methods had similar performances for the best-performing and the raw inputs. Nonetheless, the FSB method have a slightly better performance for the inputs of Robertson (1991).

The raw inputs performance was higher than 90%, showing that these parameters can be used for the soil classification task, even with the biasing produced by the way the outputs were generated. For the ISG method, Q_{tn} and F_r were the parameters considered to generate them, while for the FSB method, they were Q_{tn}, F_r and U_2.

The DT shows the advantage of allowing an internal input features importance analysis and an automatic feature selection during the model construction. For the DT created with the R *C50* package, the inputs importance can be evaluated under two perspectives. One regards to the relative amount of objects from the training set that pass through a node with the feature to be classified. This kind of importance is called by usage. The other is related to the percentage of decision nodes corresponding to the feature, which is called importance by split.

The importances of the input features without the geological age for each one of these perspectives are shown in Tables 6.9, 6.10 and 6.11. With the geological age included, the resulting importances are in Tables 6.12 and 6.13. These importances represent the mean of the importances obtained for the 10-fold testing cases.

With respect to the usage, all features show themselves with almost the same relevance, meaning that practically all of the training set objects use all input features to be classified. Therefore, it is convenient to analyze the features importance with respect to the split. It is observed that, for the non normalized inputs, depth has high relevance, mainly for the FSB method, for which the pore pressure has a comparable value too, what does not happen to the ISG method. It occurs because the pore pressure is not used to generate the ISG outputs and also, for this classification method, the agreement between the charts is lower, making the pore pressure less considered.

TABLE 6.10 – Importance of the normalized inputs of Robertson (1991).

	By usage (%)		**By split (%)**	
	ISG	FSB	ISG	FSB
z	99.12	100	21.08	29.33
Q_{t1}	100	100	33.81	28.20
F_r	100	100	31.77	25.16
B_q	96.90	99.87	13.34	17.31

TABLE 6.11 – Importance of the normalized inputs of Robertson (2016).

	By usage (%)		**By split (%)**	
	ISG	FSB	ISG	FSB
z	91.76	99.94	11.71	25.74
Q_{tn}	100	100	37.00	29.20
F_r	100	100	37.03	26.48
U_2	89.53	99.92	14.27	18.58

Even with the normalization of Robertson (1991), the depth maintains a relatively high importance for both classification methods, mainly for the FSB. It does not occur for the ISG method, for which the normalizations reduce its importance. It is expected, since the biasing increases and part of the depth information is aggregated by the normalizations.

The non normalized inputs contain just a portion of the information of the normalized inputs, which are complemented by depth. This relevance analysis allows to observe that not all the information coming from depth is aggregated into the normalizations, because their combination with depth does not exclude its relevance. In other words, it is seen that, even for ISG, the depth still contributes significantly. For the normalizations of Robertson (2016), the importance of depth is even higher. Regarding the geological age, its the feature with the lowest number of associated decision nodes.

The confusion matrices corresponding to the testing combinations of Table 6.8 are,

TABLE 6.12 – Importance of the non normalized inputs with the geological age.

	By usage (%)	**By split (%)**
CG	100	7.73
z	99.98	29.58
q_t	99.99	26.32
f_s	99.98	20.81
u_2	100	15.56

TABLE 6.13 – Importance of the normalized inputs of Robertson (2016) with the geological age.

	By usage (%)	**By split (%)**
CG	96.56	11.95
z	99.81	22.83
Q_{tn}	100	24.79
F_r	100	24.61
U_2	99.065	15.82

respectively:

$$\mathbf{C}_{ISG} = \begin{bmatrix} 96 & 0 & 7.1 & 2.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 84.4 & 18.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 1 & 2297.5 & 41.5 & 0 & 0 & 0 & 0 & 0.2 \\ 0.5 & 0 & 73.4 & 1148.1 & 46.1 & 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 46.3 & 1015.7 & 33.4 & 0 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0 & 7.5 & 3317.5 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 17.8 & 22.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9.2 & 10.2 & 0 & 84 & 0.8 \\ 0 & 0 & 3.4 & 5.2 & 7 & 0 & 0 & 0.5 & 150.6 \end{bmatrix} \tag{6.4}$$

$$\mathbf{C}_{FSB} = \begin{bmatrix} 501.6 & 37.2 & 33.8 & 1.4 & 3.7 & 2.2 & 0 & 7.6 \\ 39.5 & 627.8 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 51.2 & 0.1 & 1010.9 & 1.5 & 1.6 & 0 & 0 & 0 \\ 16 & 0 & 2.2 & 902.3 & 0.1 & 3.4 & 0 & 0 \\ 11.9 & 0 & 2 & 0 & 393.3 & 1.3 & 1.7 & 0 \\ 22.1 & 0 & 0 & 6.2 & 1.9 & 662.5 & 0 & 12.1 \\ 5.6 & 0 & 0 & 0 & 1.2 & 0 & 272.9 & 11.9 \\ 33.5 & 0 & 0 & 0 & 0.2 & 1.8 & 2.7 & 3864.6 \end{bmatrix} \tag{6.5}$$

It is noticed that the class 0 also confuses this technique. With the addition of depth

TABLE 6.14 – Results for the DT without the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	z, q_t, f_s, u_2	14	91.84	0.53
	z, Q_{t1}, F_r, B_q	21	98.33	0.27
	z, Q_{tn}, F_r, U_2	7	99.10	0.20

as an input, it is obtained:

$$\mathbf{C}_{FSB} = \begin{bmatrix} 591.4 & 8.7 & 10.1 & 0.8 & 1.1 & 0.8 & 0 & 5.9 \\ 7.2 & 656.2 & 0.4 & 0 & 0.1 & 0 & 0 & 0 \\ 14.9 & 0.2 & 1035.8 & 2.1 & 1.5 & 0 & 0 & 0 \\ 10.8 & 0 & 2 & 902.4 & 0.2 & 3.1 & 0 & 0 \\ 7.3 & 0 & 1 & 0.1 & 396.5 & 1.2 & 2.1 & 0.1 \\ 17.9 & 0 & 0 & 6 & 1.6 & 664.1 & 0.1 & 8.4 \\ 3.8 & 0 & 0 & 0 & 1 & 0 & 272.6 & 8.7 \\ 28.1 & 0 & 0 & 0 & 0 & 2 & 2.5 & 3873.1 \end{bmatrix} \tag{6.6}$$

It is realized that there is an improvement of the technique capacity to distinguish FSB class 0. Removing this class, it is obtained a mean accuracy of 99.20% with a standard deviation of 0.18% and the following mean confusion matrix:

$$\mathbf{C}_{FSB} = \begin{bmatrix} 664.2 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 1045.1 & 1.9 & 1.4 & 0 & 0 & 0 \\ 0 & 2.1 & 904.9 & 0 & 3 & 0 & 0 \\ 0 & 1.4 & 0 & 398 & 1.7 & 1.6 & 0 \\ 0 & 0 & 4.6 & 1.8 & 664.5 & 0 & 11.2 \\ 0 & 0 & 0 & 0.8 & 0 & 272.4 & 11 \\ 0 & 0 & 0 & 0 & 2 & 3.3 & 3874 \end{bmatrix} \tag{6.7}$$

Removing the FSB class 0, the results obtained for the other combinations are shown in Tables 6.14 and 6.15. It is observed that there is, in general, a performance gain, except for the geological combinations.

TABLE 6.15 – Results for the DT with the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**Elected k**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	8	90.72	1.62
	CG, z, Q_{tn}, F_r, U_2	8	98.05	0.87

TABLE 6.16 – Results for the RF without the geological age.

Classification Method	**Inputs**	**MA (%)**	**SD (%)**
ISG	z, q_t, f_s, u_2	91.53	0.69
	z, Q_{t1}, F_r, B_q	96.09	0.52
	z, Q_{tn}, F_r, U_2	97.44	0.40
FSB	z, q_t, f_s, u_2	91.43	0.41
	z, Q_{t1}, F_r, B_q	96.38	0.31
	z, Q_{tn}, F_r, U_2	97.27	0.24

6.3 Random Forests

The RF was implemented with the R *randomForest* package and the command *randomForest()*. This package uses the algorithm of Breiman (2001), which consists in following the bagging strategy explained in Section 3.1. As mentioned in Section 5.1, it was used 500 trees. The results are presented in Tables 6.16, 6.17 and 6.18.

Considering a confidence of 5%, the statistical test of Friedman identifies a statistical difference between the techniques with a p-value of 0.06%. The results obtained with the post-hoc statistics of Nemenyi are represented in Figure 6.1. Comparing the RF with the DWNN and the boosted DT with the statistical test of Wilcoxon, the RF is equivalent to the DT with a p-value of 43.16% and both techniques are superior to the DWNN with the same p-value of 0.20%.

The normalized inputs of Robertson (2016) outperforms the remaining and, in this case, the normalized inputs of Robertson (1991) have the second best performance for both classification methods. However, the non normalized inputs still give a satisfactory accuracy, greater than 90%. The introduction of the geological age does not produce a significant performance change.

TABLE 6.17 – Results for the RF with the geological age.

Classification Method	**Inputs**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	91.78	1.21
	CG, z, Q_{tn}, F_r, U_2	97.31	0.58

TABLE 6.18 – Results for the RF and the inputs used by the CPeT-IT to generated the reference outputs.

Classification Method	**Inputs**	**MA (%)**	**SD (%)**
ISG	Q_{tn}, F_r	97.31	0.48
FSB	Q_{tn}, F_r, U_2	94.63	0.19

TABLE 6.19 – Importance of the non normalized inputs.

	ISG	FSB
z	1680.18	6451.32
q_t	3169.23	11492.45
f_s	2235.98	8838.76
u_2	913.65	5394.92

The RF was also employed to make a feature selection and an evaluation of features importances, but with the Mean Decrease Gini as reference. This measure is used by the R *randomForest* package. The importance values obtained are given in Tables 6.19, 6.20 and 6.21 without the geological age and in Tables 6.22 and 6.23 with the inclusion of the geological age.

It is observed that, for the non normalized inputs, the depth importance is greater than for the normalized. However, the RF indicates a higher contribution of q_t and f_s than depth both for normalized and non normalized inputs. Considering normalized inputs, for the ISG method, the pore pressure is the less relevant feature. On the other hand, for the FSB method, the pore pressure has greater relevance than depth with normalized inputs and both parameters have significantly lower relevance than the cone resistance and the friction ratio.

These results show that there is not a perfect agreement between normalized and the non normalized inputs in concert with depth, but the normalized inputs can incorporate some of the depth information and this parameter represents an important piece of the normalized inputs. It is noticed also that the cone resistance and the lateral friction or the friction ratio alternates their importance order, maintaining, however, one close to each other.

With respect to the geological age, it is the feature with the lowest information gain (IG), being greater for the non normalized inputs, suggesting that if these features were considered for clustering, it would be interesting to include it as an input. However, it can be due to the biasing, making the extra information apparently less relevant. The higher IG with the non normalized inputs suggests that with a lower biasing this feature can add relevant information.

TABLE 6.20 – Importance of the normalized inputs of Robertson (1991).

	ISG	FSB
z	881.08	3124.33
Q_{t1}	2966.58	12395.40
F_r	3544.31	12203.96
B_q	606.30	3901.28

TABLE 6.21 – Importance of the normalized inputs of Robertson (2016).

	ISG	FSB
z	941.29	3065.81
Q_{tn}	5469.33	14980.16
F_r	4868.57	13900.60
U_2	848.47	4329.23

TABLE 6.22 – Importance of the non normalized inputs with the geological age.

	Mean Decrease Gini
CG	1141.45
z	2127.33
q_t	2571.23
f_s	2173.04
u_2	1942.97

TABLE 6.23 – Importance of the normalized inputs of Robertson (2016) with the geological age.

	Mean Decrease Gini
CG	690.89
z	1298.69
Q_{tn}	2754.43
F_r	2686.12
U_2	1411.57

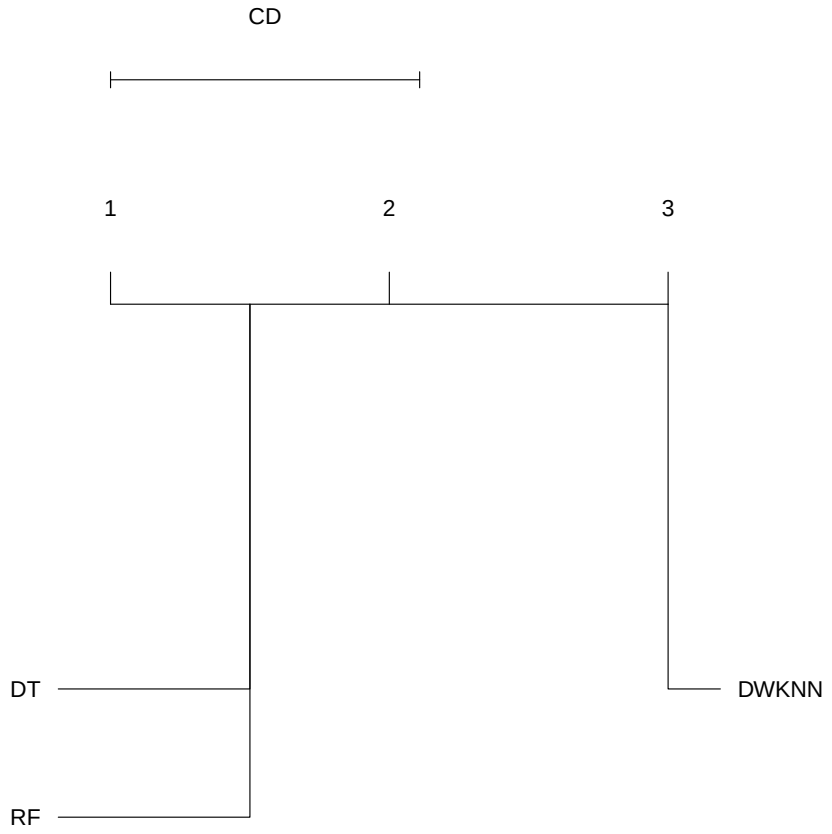


FIGURE 6.1 – Statistical test of Friedman with the post-hoc Nemenyi statistics for the RF, the DT and the DWNN.

The confusion matrices for the biased combinations of Table 6.18 are, respectively:

$$\mathbf{C}_{91} = \begin{bmatrix} 96.5 & 0 & 10 & 3.2 & 0.3 & 0 & 0 & 0 & 0 \\ 0 & 84.8 & 20.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.6 & 2299.7 & 42.2 & 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 67.1 & 1156.3 & 46.9 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 35.1 & 1017.3 & 32.1 & 0 & 0.1 & 0.2 \\ 0 & 0 & 0 & 0 & 3.7 & 3320.6 & 0.6 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 13.9 & 22.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 & 12.3 & 0 & 84.3 & 0.2 \\ 0 & 0 & 2.8 & 7.1 & 7.3 & 0 & 0 & 0.5 & 151.3 \end{bmatrix} \tag{6.8}$$

$$\mathbf{C}_{16} = \begin{bmatrix} 506.9 & 36.4 & 34.8 & 1.2 & 3.1 & 2.3 & 0.1 & 10.4 \\ 38.3 & 628.3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 48.2 & 0.4 & 1007.3 & 2.2 & 1 & 0.1 & 0 & 0 \\ 15.8 & 0 & 3.6 & 899.9 & 0 & 3.6 & 0 & 0 \\ 12.1 & 0 & 2.6 & 0.1 & 394.9 & 2.1 & 2.1 & 0.1 \\ 24.6 & 0 & 0 & 8 & 2 & 662.6 & 0.1 & 40.3 \\ 5.6 & 0 & 0 & 0 & 1 & 0 & 273.7 & 36.4 \\ 29.9 & 0 & 0 & 0 & 0 & 0.5 & 1.3 & 3809 \end{bmatrix} \tag{6.9}$$

As well as for the DWNN and for the DT, the RF gets confused with classes 3, 4, 5 and 6 for the ISG method and with the FSB class 0. With the inclusion of depth, the confusion matrix obtained for the FSB biased combination is the following:

$$\mathbf{C}_{16} = \begin{bmatrix} 591.2 & 8.4 & 12.5 & 0.4 & 0.8 & 0.3 & 0 & 8.5 \\ 7.7 & 656.5 & 0.9 & 0 & 0 & 0 & 0 & 0 \\ 15.4 & 0.2 & 1032.5 & 2 & 1.6 & 0 & 0 & 0 \\ 11.3 & 0 & 2.2 & 903.4 & 0 & 2.8 & 0 & 0 \\ 7.7 & 0 & 1.2 & 0.1 & 397.5 & 1.7 & 2.3 & 0.1 \\ 18.9 & 0 & 0 & 5.5 & 1.4 & 665.6 & 0 & 21.5 \\ 4.2 & 0 & 0 & 0 & 0.7 & 0 & 273.1 & 23 \\ 25 & 0 & 0 & 0 & 0 & 0.8 & 1.9 & 3843.1 \end{bmatrix} \tag{6.10}$$

The result is similar to that observed for the DT. Removing the FSB class 0, the mean accuracy achieved is slightly lower than that of the DT, being of 98.82% with a standard deviation of 0.14%. The confusion matrix in this case is given by:

$$\mathbf{C}_{16} = \begin{bmatrix} 664.4 & 1.3 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 1041.2 & 2.9 & 1.8 & 0 & 0 & 0 \\ 0 & 3.8 & 900.7 & 0.1 & 4.8 & 0 & 0 \\ 0 & 2.9 & 0.1 & 397.2 & 2.2 & 2.7 & 0.1 \\ 0 & 0.1 & 7.7 & 2 & 663.4 & 0 & 32.2 \\ 0 & 0 & 0 & 0.9 & 0 & 270.8 & 25.9 \\ 0 & 0 & 0 & 0 & 0.8 & 3.8 & 3838 \end{bmatrix} \tag{6.11}$$

With the FSB class 0 removal, the results of Tables 6.24 and 6.25 are obtained for the remaining combinations. It is seen that the performance gain is, in general, significant, except for the geological combinations, for which it stands almost steady or suffers a slight

TABLE 6.24 – Results for the RF without the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**MA (%)**	**SD (%)**
FSB	z, q_t, f_s, u_2	91.97	0.40
	z, Q_{t1}, F_r, B_q	98.44	0.19
	z, Q_{tn}, F_r, U_2	99.18	0.13

TABLE 6.25 – Results for the RF with the geological age and without the FSB method class 0.

Classification Method	**Inputs**	**MA (%)**	**SD (%)**
FSB	CG, z, q_t, f_s, u_2	90.91	1.58
	CG, z, Q_{tn}, F_r, U_2	97.95	0.78

reduction. The statistical test of Wilcoxon points an equivalence between the RF and the DT with a p-value of 18.75% in this situation.

7 Conclusions

A general methodology for the application of ML techniques for soil classification with CPT data is presented and applied to 111 CPT soundings using three different techniques and two distinct approaches. It consists in firstly dividing the dataset into training, validation and testing sets by the 10-fold strategy. Then, preprocessing procedures are run trying to ensure a good performance for different available data, including data transformation, cleaning and balancing. The whole available data is divided into two analyzed sets: a full dataset and a so-called geological dataset, which includes those soundings for which the soil geological age was available. The reference outputs for each instance were generated with the CPeT-IT software with a student license. This program uses the ISG (see Section 2.1) and the FSB methods (see Section 2.2) for classification. Preliminary discussions were made with descriptive statistics to ground the methodology procedures and to show details about the data.

With just some calibration adaptations, the proposed approach could be applied to different techniques. This approach is more rigorously established and employed in comparison to other ML applications for soil classification found in literature. Furthermore, there are few or no studies exploring a dimensionally extended feature space in this problem.

The methodology was implemented with a distance-based method, the Gaussian DWNN, and two symbolic methods, random forests (RF) and decision trees (DT) with the adaptive boosting improvement. These two last techniques make an inner features importance evaluation and selection, what can be used to analyze the contribution of each feature to the task. It could be concluded that depth introduces relevant information, even with normalized or biased inputs, and that the non normalized or raw inputs can be almost as good as the others if depth is included.

For the DWNN, the ISG method could reach a greater performance than the FSB if this last method was considered with its class 0, with non identified or structured soils, but the FSB method overcomes the other if this class is removed. However, for DT and RF, this class had a minor effect and these techniques could achieve high performance, even with this noise, and almost the same for both classification methods and inputs sets. It was

also concluded that boosted DT and RF have statistically equivalent performances with boosted DT tending to overcome RF. The comparisons were based on the statistical tests of Wilcoxon and Friedman with the post-hoc statistics of Nemenyi. A complementary and initial study was conducted including the geological age as a discrete input with some feature combinations. The importance analysis shows that this information can be relevant to the classification task.

These conclusions encourage further studies applying other ML techniques with different approaches such as the optimization-based or the multiple modeling to the problem and analyzing the adaptations required or the improvements that can be done to the methodology. The conclusions also suggest that a clustering investigation with a dimensionally increased feature space could yield interesting results. In this way, data from other in situ tests like the standard penetration test can be incorporated. Similar or deeper explorations to the presented can also eventually be conducted with even more diversified data.

Bibliography

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, 2012.

BEGEMANN, H. K. S. P. The Friction Jacket Cone as an Aid in Determining the Soil Profile. **Proceedings 6th International Conference on Soil Mechanics and Foundation Engineering**, pp. 17–20, 1965.

BHATTACHARYA, B.; SOLOMTINE, D. P. Machine learning in soil classification. **Neural Networks**, vol. 19, pp. 186–195, 2006.

BREIMAN, L. Random forests. **Machine Learning**, vol. 45, no. 1, pp. 5–32, Oct 2001. ISSN 1573-0565. Available from Internet: <https://doi.org/10.1023/A:1010933404324>.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. 1st. ed. [S.l.]: Chapman & Hall, 1984.

CAI, G.; LIU, S.; PUPPALA, A. J. 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, 2011.

CAL, Y. Soil classification by neural network. **Advances in Engineering Software**, vol. 22, pp. 95–97, 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.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, vol. 16, pp. 321–357, 2002.

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, pp. 241–248, 2009.

DOUGLAS, B. J.; OLSEN, R. S. Soil Classification Using Eletric 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**, 2004.

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.

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, 2017.

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, pp. 986–996, dec 2002.

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, pp. 173–176, 1991.

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, 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.

LUNNE, T.; POWELL, J.; ROBERTSON, P. **Cone Penetration Testing in Geotechnical Practice**. Taylor & Francis, 1997. ISBN 9780419237501. Available from Internet: <https://books.google.com.br/books?id=ofbnE1xMl_kC>.

MAYNE, P. W. Interpretation of geotechnical parameters from seismic piezocone tests. **Proceedings of 3rd International Symposium on Cone Penetration Testing**, 2014.

MAYNE, P. W.; PEUCHEN, J.; BOUWMEESTER, D. Soil unit weight estimated from CPTu in offshore soils. **Frontiers in Offshore Geotechnics II**, vol. 119, pp. 371–376, 2010.

NEMENYI, P. **Distribution-free Multiple Comparisons**. Thesis (PhD) — Princeton University, 1963.

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, CA, USA: 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**, pp. 341–356, 2002.

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 Geotehnical Journal**, vol. 28, pp. 176–178, 1991.

ROBERTSON, P. K. Interpretation of cone penetration tests – a unified approach. **Canadian Geotehnical Journal**, vol. 46, no. 11, pp. 1337–1355, 2009.

ROBERTSON, P. K. Cone penetration test (CPT)-based soil behaviour type (SBT) classification system – an update. **Canadian Geotechnical Journal**, vol. 53, pp. 1910–1927, 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**, 1986.

ROBERTSON, P. K.; WRIDE, C. E. Evaluating cyclic liquefaction potential using the cone penetration test. **Canadian Geotechnical Journal**, vol. 35, 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.

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, pp. 209–215, 2012.

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, 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, 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. vol. 13, mar 2008.

WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, vol. 1, no. 6, pp. 80–83, 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, G.; ROBERTSON, P. K.; BRACHMAN, R. W. Estimating liquefaction-induced ground settlements from cpt for level ground. **Canadian Geotechnical Journal**, vol. 39, no. 5, pp. 1168–1180, 2002.

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	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	27 de setembro de 2018	DCTA/ITA/TC-002/2018	68

5. TÍTULO E SUBTÍTULO:

Application of Machine Learning techniques for soil classification from CPT data

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; Machine learning; Soil classification; K-nearest neighbor; Decision tree; Random forest.

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Aprendizagem (inteligência artificial); Árvores de decisão; Máquinas aprendizes; Máquinas de vetores- suporte; Algoritmos; Técnicas de previsão; Computação.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Prof. Dr. Dimas Betioli Ribeiro. Publicada em 2018.

11. RESUMO:

The soil classification problem with cone penetration test (CPT) data is usually treated with bidimensional solutions such as charts or, less often, machine learning (ML) approaches in a dimensionally restricted feature space. To avoid this restriction, a multi-dimensional analysis of CPT data for soil classification is here performed by using k-nearest neighbors (KNN) and machine learning symbolic algorithms. The symbolic algorithms are able to do an inner input features relevance analysis and feature selection, calculating the features importance. These algorithms are employed in order to evaluate each input feature importance by different criteria and to analyze their performance considering up to five features including raw and normalized CPT inputs as continuous inputs and soil age as a discrete one. The dataset used is composed by 111 soundings from different locations around the world. The symbolic techniques, namely boosted decision trees (DT) and random forests (RF), are applied to the problem, studied and compared using a 10-fold cross-validation procedure. Two classification methods are considered: one influenced by soil granulometry (ISG) and the other focused on soil behaviour only (FSB). A general methodology for soil classification using ML techniques is described and followed. It covers descriptive statistical procedures and other ML techniques for data preprocessing, including data transformation, cleaning and balancing. The symbolic techniques are compared with the Gaussian distance-weighted nearest neighbor technique (DWNN). The comparisons are made with statistical hypothesis tests. The results shows that RF and boosted DT have equivalent performance and that they both perform better than the DWNN. The features importance analysis indicates that depth and soil age introduce relevant information for soil classification and that the raw inputs including depth can be enough to perform the task.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**