applied sciences

MDPI

Article

Generation of Penetrometric Profile of the Soil Applying Machine Learning to Measure While Drilling Data from Deep Foundation Machinery

Eduardo Martínez García [1,2], Marcos García Alberti [1,*] and Antonio Alfonso Arcos Álvarez [3]

[1] Departamento de Ingeniería Civil: Construcción, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, 28040 Madrid, Spain; emartinez@menard.es
[2] Menard España, 28001 Madrid, Spain
[3] Departamento de Ingeniería y Morfología del Terreno, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, 28040 Madrid, Spain; antonio.arcos@upm.es
* Correspondence: marcos.garcia@upm.es; Tel.: +34-91-0674121

Abstract: The study performed in this article aimed to reproduce the penetrometric profile of the soil from the perforation parameters of deep foundation and ground improvement. This could allow for more easily interpretable information on the soil strength during execution as well as validate the design hypotheses. To achieve this goal, a series of Machine Learning algorithms have been used and compared with traditionally applied analytical formulas. Dynamic time warping is used to measure the likeness of the results with the expected shape. The results show that the algorithms are capable of better fitting the penetrometric profiles of the soil. Tree ensemble methods stand out with the best results.

Keywords: machine learning; rigid inclusion; penetrometer; measurement while drilling; dynamic time warping

Academic Editor: Arcady Dyskin

Received: 16 December 2024
Revised: 11 January 2025
Accepted: 21 January 2025
Published: 27 January 2025

Citation: Martínez García, E.; Alberti, M.G.; Arcos Álvarez, A.A. Generation of Penetrometric Profile of the Soil Applying Machine Learning to Measure While Drilling Data from Deep Foundation Machinery. *Appl. Sci.* **2025**, *15*, 1331. https://doi.org/10.3390/app15031331

1. Introduction

Currently, drilling rigs that work on deep foundations and ground improvements have measurement while drilling (MWD) systems that record parameters such as rotation torque or penetration rate. These parameters are related to the characteristics of the soil drilled. In the context of soil drilling, compound parameters are analytical formulas that aim to provide a more accurate assessment of soil strength based on MWD. Most of these indices tend to smooth the profile of raw values, having a greater physical meaning and easier interpretation [1]. However, they assume a form in the relationship between parameters and strength that may not necessarily fit reality [2]. These formulas provide a qualitative idea of the characteristics of the traversed soil but are not precise. For example, the results of compound parameters do not correlate well with usual soil investigation tests such as dynamic penetrometers, which limits their use.

The dynamic penetrometer is a type of test that consists of driving a cone of certain characteristics into the ground. The result of the test is the number of blows necessary to drive the cone a certain length (usually every 20 cm) along the depth reached.

Machine Learning (ML) is a branch of artificial intelligence that consists of applying a series of algorithms through computer programs to obtain information from large data sets.

In recent years, the application of ML to the fields of civil engineering and geotechnics has become an alternative to empirical methods. So much so that in 2018, the International Society of Soil Mechanics and Geotechnical Engineering (ISSMGE) created a technical

committee (TC) for Machine Learning and Big Data, TC309. Through the website of this TC, dozens of examples of the application of ML in the field of geotechnics [3] can be consulted.

The growing interest in using ML in geotechnics is logical because soil and rock properties vary extensively, and measurements are often influenced by the instruments and data processing techniques [4]. This increase has been reflected in the Soil Mechanics and Geotechnical Engineering Conference, both in the invited lectures [5,6] and in the proceedings [7–21]. Furthermore, the International Symposium on Machine Learning and Big Data in Geoscience (ISMLG) had a successful last edition in Cork [22]. Early steps of this paper were presented at said symposium [23].

The first intuitive applications for ML in geotechnics are predicting soil properties [24,25], soil types [26], or lithology [27–29]. Liang et al. even used a compound parameter, mechanical specific energy, with ML for lithology identification [30].

An early example of the use of ML to deep foundation execution data is the work of Goh [31], who used Neural Network (NN) to estimate the bearing capacity of driven piles from real data collected by Flaate [32]. The results showed that the NN achieved more reliable predictions than the driving formulas. Lee and Lee also used execution and ground parameters to estimate the bearing capacity of driven piles using NN [33]. The predictions obtained were better than with the Meyerhof formula [34]. Pal and Deswal [35] employed a Gaussian regression process (GP) to predict the bearing capacity of the pile. Part of the input data were the same as those used by Goh. The performance of the proposed model was compared with support vector machines (SVM) and empirical relationships, obtaining a better result. Alkroosh and Nikraz used gene expression programming (GEP) to correlate cone penetration test (CPT) data and pile capacity in cohesive soils [36]. On the other hand, Millán et al. used NN to predict the bearing capacity of the tip of a pile embedded in a rock mass using an extensive set of results obtained from numerical calculations [37].

Another field of combined application of MWD and ML is mining. Beattie ([38], p. 20) used NN for rock classification in an open-pit mine. Zhou et al. [39] also performed rock classification work using a Gaussian classification process and MWD data. The same authors used a clustering approach that did not require previous labelling [40]. Kadkhodaie-Ilkhchi et al. [41] conducted a comparative study of three ML techniques using MWD data from drilling for explosives in an iron mine. Khushaba et al. [42] also used MWD and ML in mining for chemical composition prediction. Although not using ML, Manzoor et al. [43] combined MWD data and close-range terrestrial digital photogrammetry (CRTDP) to establish certain relationships between drilling parameters and rock mass structures.

Galende-Hernández et al. [44] conducted a study estimating the RMR value from excavation front characterisation using MWD and expert knowledge in tunnelling with explosives. The results obtained present a good correlation. Hansen et al. [45] used NN and random forests (RF) to characterise the rock mass by its Q-class from MWD data.

Díaz et al. [46] used NN and real-time data to predict the drilling rate in a geothermal well, showing good prediction capability. Klyuchnikov et al. [47] used ML algorithms to classify rock types in directional oil well drilling. A review of this and other articles can be found in the work of Silversides [48].

In the specific case of ML application to treatment with rigid inclusions, Zhang et al. [49] used NN to analyse the behaviour of an isolated footing supported on soft soil reinforced by rigid inclusions.

For an exhaustive review of ML applications in geotechnics, readers can consult [6,10,12,50–58].

In this work, both analytical formulas and ML approaches have been compared in terms of fitting the penetrometric profile of the soil drilled during the execution of rigid inclusions. Rigid inclusions are a type of soil treatment that normally consists of

concrete columns drilled with soil displacement. Penetrometric profiles are more useful than compound parameters because of the already existing correlations with other soil parameters and because they are more intuitively understood by geotechnical engineers. Compared to previous works this article presents a more robust methodology, considers a greater number of algorithms, addresses the issue of imbalanced data, and uses a newly proposed method to compare lineal data in geotechnics.

2. Methodology

In this article, the ability to predict the penetrometric profile of the soil using different compound parameters and ML algorithms has been analysed. The starting data were the drilling records for rigid inclusions in three sites.

2.1. Compound Parameters

Compound parameters in geotechnics are analytical formulas that provide an idea of the traversed soil resistance based on drilling data.

These parameters have a similar structure where the strength of the soil would be directly proportional to the torque and applied force and inversely proportional to the drilling area and velocity.

For this article, the most common compound parameters have been used as shown in Table 1, with:

- $(t)_{dZ=0.2m}$: time to perforate 0.2 m of material;
- P: thrust applied to the drilling tool;
- V_R: rotational speed;
- V_A: perforation rate;
- S_0: area of the drilling tool;
- C_R: rotational torque.

Table 1. Compound parameters.

Name	Formula	Reference
Penetration resistance	$R_p = (t)_{dZ=0.2m}$	Möller et al., 2004 [59]
Somerton index	$S_d = P\sqrt{\frac{V_R}{V_A}}$	Somerton 1959 [60]
Drilling specific energy	$SDE = \frac{P}{S_0} + \frac{2\pi}{S_0} \cdot \frac{V_R \cdot C_R}{V_A}$	Teale 1965 [61]
Specific energy	$E_S = C_R \cdot \frac{V_R}{V_A}$	Pfister 1985 [62]

2.2. Machine Learning Algorithms

There are numerous algorithms, and the most appropriate one depends on the problem being addressed. It is not something evident since each method uses a different approach. The algorithms used for this study are listed in Table 2.

Table 2. Algorithms used in this study.

Algorithm	Abbreviation
Linear regression	-
Logistic regression	-
Support vector machines	SVM
K-nearest neighbours	KNN
Decision trees	DT
Random Forests	RF
XGBoost	XGB
Multilayer perceptron	MLP

The problem analysed can be seen as either classification or regression because the target variable is composed by numbers, but they always take discrete values. Both approaches have been tested to determine which one provides better results.

2.3. Available Data

Data from three sites with rigid inclusions were used. These inclusions were constructed using an auger that displaces the soil laterally. In all three sites, the Enteco E6050 type drill was used. An onboard Menard Emparex data acquisition system recorded depth, penetration rate, rotational speed, torque, and thrust roughly every 8 cm. All the inclusions had a diameter of 360 mm.

Site 1 is formed by an anthropic fill (silty sands, 1.60–2.20 m thick) that lies above a thin organic soil layer (silty sands, 0.4–1.0 m). Below that, there is an alluvial layer of fine sand with silt (very loose, 1.60–2.20 m), followed by altered granite, which becomes more compact with depth. Overall, the soil is homogeneous, predominantly sandy, and the water table lies near the surface. In Site 2 the top layer is an anthropic fill (sands and clayey sands) about 6 m thick. Below it, alluvial materials form interlayered clays to sands extending to about 30 m. The deepest detected layer is a dense, highly compact sand (Pliocene), which the soil treatment does not reach. Water levels have been observed between 2 and 8 m deep at different times. Overall, this site is more heterogeneous than Site 1, with varied materials and a fluctuating water table. Lastly, the soil of Site 3 is essentially a dredged fill (sandy with some fines and gravel) 4–8 m thick, placed for sea reclamation. Below it lies quaternary sands that become denser with depth. The water table is shallow, influenced by sea level. Overall, the terrain is predominantly sandy and fairly homogeneous.

In these worksites, a series of DPSH-B type penetrometers were executed according to the UNE-EN-ISO22476-2 standard [63]. The result of these tests is the number of blows required for penetrations in 20 cm sections and can be plotted against depth (Figure 1).

The features and target variable for the algorithms are included in Table 3.

Table 3. Features and target variable for the algorithm.

	Algorithm					**Target Variable**
	Depth of Perforation	**Penetration Rate**	**Rotational Speed**	**Rotational Torque**	**Thrust**	**Penetrometer Blows**
Unit	(m)	(m/h)	(rpm)	(tm)	(t)	(-)
Range	(0–13)	(12–1300)	(6.25–25.0)	(0.6–19.0)	(7.2–21.3)	(0–30)

Further information about the available data in these works can be found in the following reference [64].

To maintain a certain consistency in the study, the drilling and penetrometer data have been converted to the AGS4 format [65]. This format is intended to collect geotechnical data and can be easily converted to a Pandas data frame [66] through a dedicated library [67].

Once the data was adapted, a blow value was associated based on the closest penetrometer.

2.4. Error Induced by the Distance Between the Inclusions and the Penetrometer

In ML, one of the first and perhaps most important steps is to choose the dataset with which to train and test the model.

For this problem, one major source of error is the changes in the terrain characteristics between the position of the penetrometer and the inclusion. Larger distances typically increase the mismatch in soil properties, especially in heterogeneous terrains. In Figure 2, which represents a geological cross section, inclusions 1 and 2 would be associated with

penetrometer P1, as it is closer than penetrometer P2. Then, inclusion 1 would have a small error when interpreting the contact between levels 1 and 2. However, the error of inclusion 2 is greater due to its distance to P1.

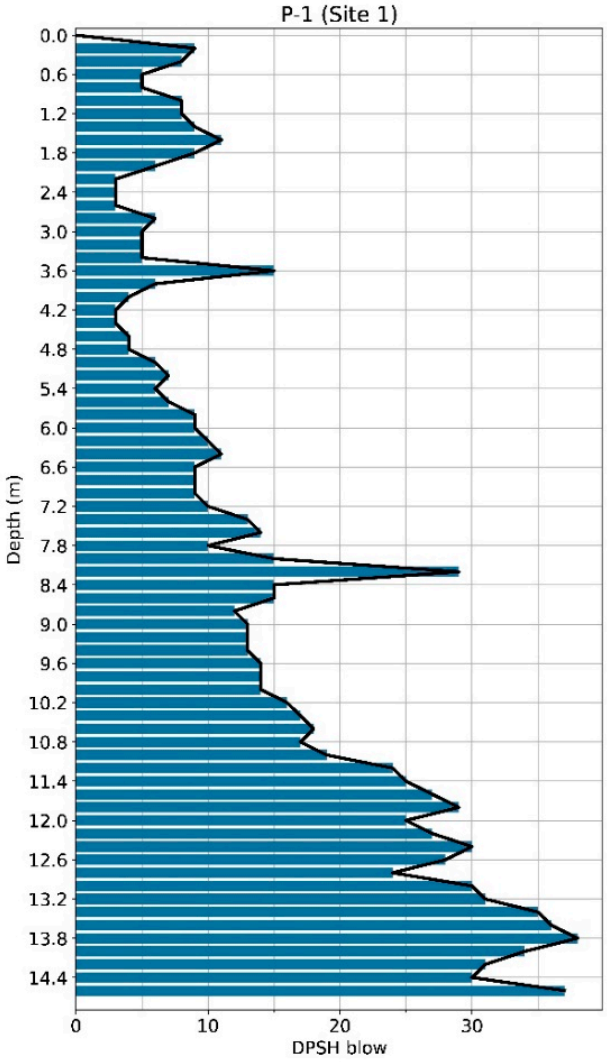

Figure 1. Representation of penetrometer P-1 of Site 1.

While selecting only inclusions very close to penetrometers minimises this error, it may also reduce the dataset size significantly. Bunieski [9] addressed this problem, proposing methods to evaluate the error introduced by considering increasing distances to the penetrometers. In this article, this effect was evaluated for the studied sites.

2.5. Distribution of the Penetrometer Blows

Another source of error arises from the uneven distribution of blow count values.

Figure 3 shows the number of samples for each blow value, considering a distance between the inclusion and its nearest penetrometer of, at most, 2 m.

As can be seen, there was a significant difference in the distribution of blows. This heterogeneity could bias the algorithms.

There are various techniques to tackle the problem of imbalanced data. In this article, SMOTE (Synthetic Minority Over-sampling) [68] was used.

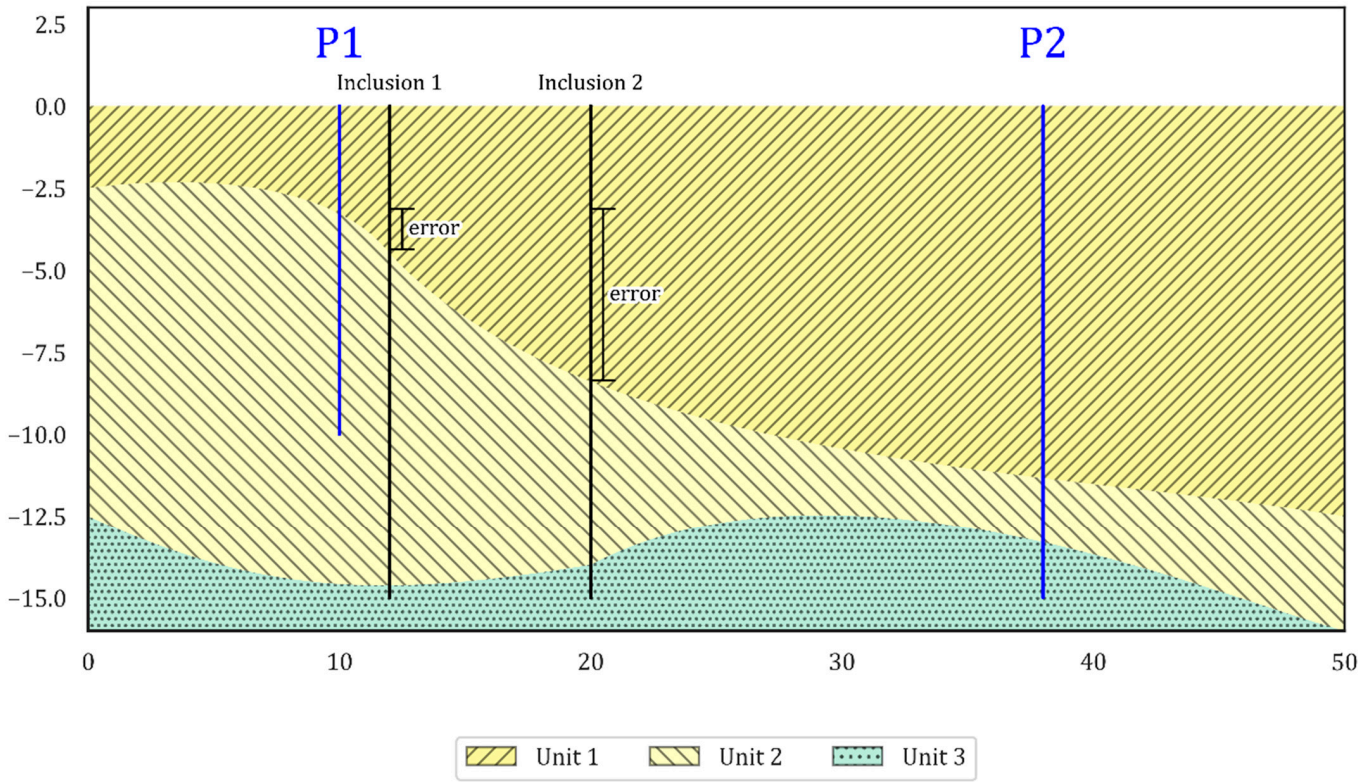

Figure 2. Error introduced by distance between inclusion and penetrometer.

2.6. Hyperparameters

A search was made for the most appropriate hyperparameters using GridSearchCV [69], running 500 different combinations of the datasets. The scores function for choosing the best combination of hyperparameters were accuracy for classification and the coefficient of determination for regression. The algorithms and the set of verified hyperparameters can be found in Table 4.

Table 4. Evaluated hyperparameters.

Algorithm	Evaluated Hyperparameters
Decision trees (classification)	**'criterion'**: 'gini', 'entropy'; **'splitter'**: 'best', 'random'; **'max_depth'**: 5, 10, 15, None; **'min_samples_split'**: 2, 4, 6, 8; **'min_samples_leaf'**: 2, 3, 4, 5; **'max_leaf_nodes'**: None, 100, 50, 10; **'ccp_alpha'**: 0, 0.005, 0.015, 0.03
Decision trees (regression)	**'criterion'**: 'squared_error', 'friedman_mse', 'absolute_error'; **'splitter'**: 'best', 'random'; **'max_depth'**: 5, 10, 15, None; **'min_samples_split'**: 2, 4, 6, 8; **'min_samples_leaf'**: 2, 3, 4, 5; **'max_leaf_nodes'**: None, 100, 50, 10; **'ccp_alpha'**: 0, 0.005, 0.015, 0.03
Random forests	**'n_estimators'**: 100, 300, 500; **'max_depth'**: 5, 10, 15, 20, None; **'min_samples_split'**: 2, 5, 10; **'min_samples_leaf'**: 1, 2, 5, 10; **'max_features'**: 'log2', 'sqrt', None
K-nearest neighbours	**'n_neighbors'**: 2, 4, 8, 16; **'p'**: 2, 3
Linear regression	
Logistic regression	**'C'**: 0.001, 0.01, 0.1, 1; **'penalty'**: 'l1', 'l2'
Support vector machines	**'C'**: 0.1, 1, 10, 100, 1000; **'gamma'**: 1, 0.1, 0.01, 0.001, 0.0001

Table 4. *Cont.*

Algorithm	Evaluated Hyperparameters
XGBoost	**'eta'**: 0.01, 0.025, 0.1; **'gamma'**: 0.05, 0.5, 1.0; **'max_depth'**: 5, 12, 25; **'min_child_weight'**: 1, 3, 7; **'subsample'**: 0.6, 0.8, 1.0; **'colsample_bytree'**: 0.6, 0.8, 1.0; **'lambda'**: 0.01, 0.1, 1.0; **'alpha'**: 0, 0.5, 1.0
Multilayer perceptron	**'activation'**: 'relu', 'identity', 'logistic', 'tanh'; **'solver'**: 'adam', 'lbfgs', 'sgd'; **'alpha'**: 0.0001, 0.01, 0.1; **'learning_rate'**: 'constant', 'invscaling', 'adaptive'; **'learning_rate_init'**: 0.001, 0.1, 1.0

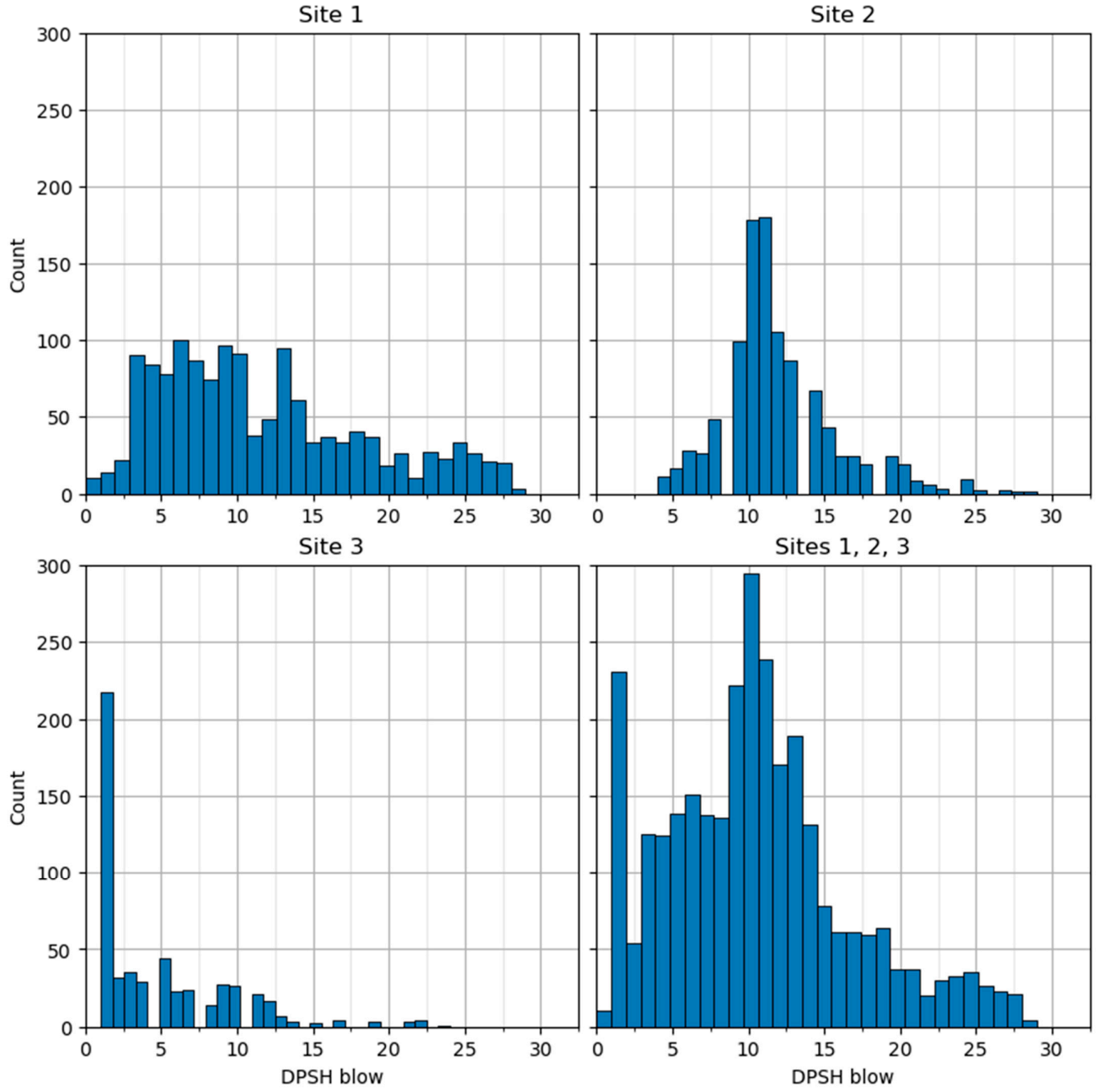

Figure 3. Number of samples for each blow count value for maximum distances of 2 m.

This computationally intensive process used the Magerit-3 supercomputer belonging to the Universidad Politécnica de Madrid.

Once the best hyperparameters have been obtained, the behaviour of the resulting models is analysed. The algorithms using the best hyperparameters are trained with 70%

of the samples and tested on the rest 30%. The values shown are the mean values from 500 different train-test splits.

2.7. Measuring Performance and Dynamic Time Warping

To measure the performance of the algorithms and compound parameters, different metrics can be used. However, it is convenient to represent the results graphically to avoid being deceived [70].

In the situation studied, the goal is to match the overall "shape" of the penetrometric profile. At the 4th ISMLG, Charles et al. proposed the use of dynamic time warping (DTW) to determine similarity between pairs of cone penetration tests (CPT) [71].

DTW [72] is an algorithm designed for the comparison of time series. This algorithm tries all possible combinations to shift one time series until it matches another (Figure 4). The minimum displacement distance gives us an idea of the similarity between both series.

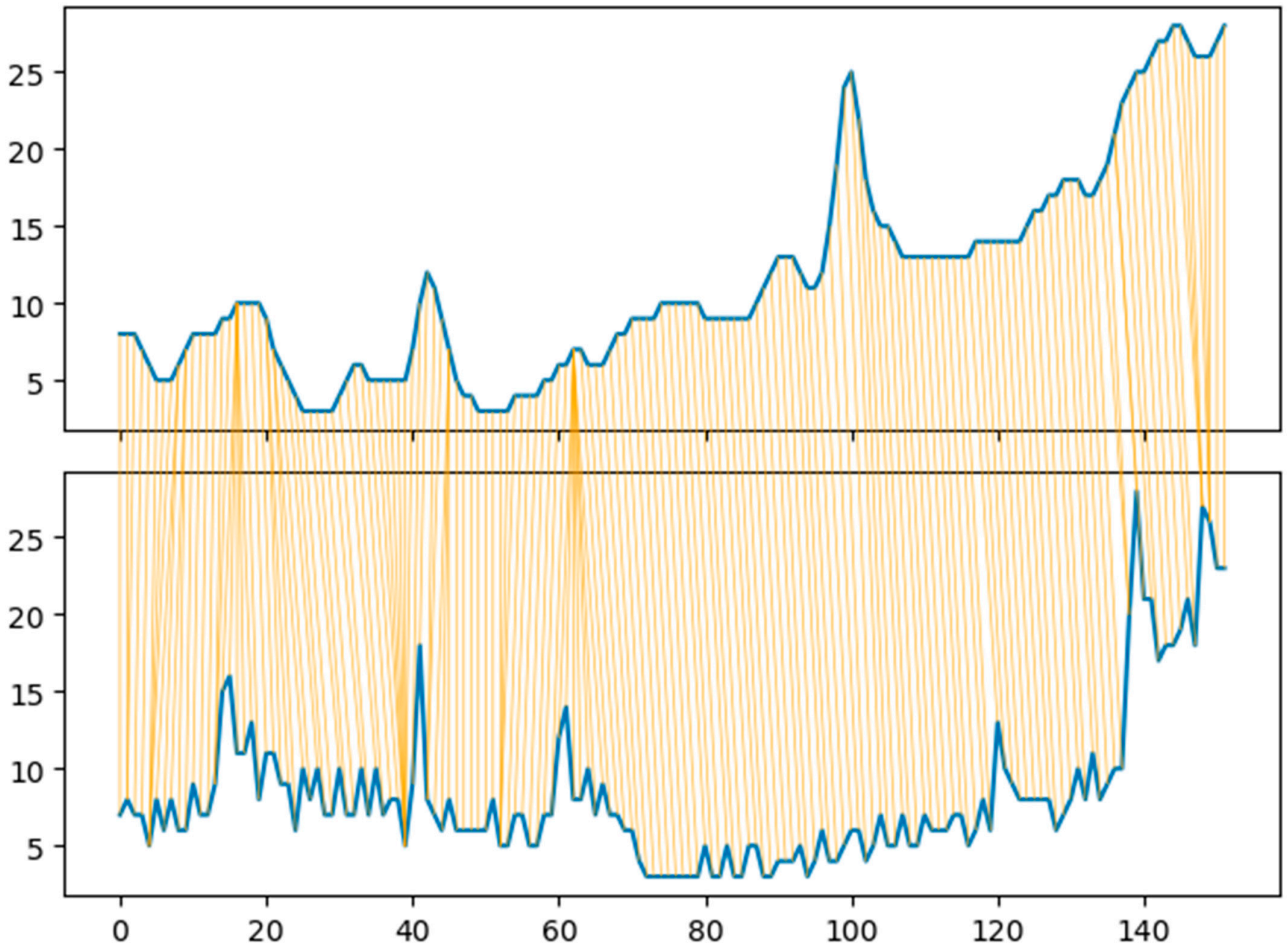

Figure 4. Comparison between two series (in blue) using DTW with a window of 5. The yellow lines represent the path to convert one series into the other.

In this paper, DTW is used to compare the expected and predicted penetrometric profiles for each drilled inclusion. To limit the possibilities that DTW considers, a window of 5 is used. Otherwise, the algorithm could try paths with no geotechnical meaning (Figure 5).

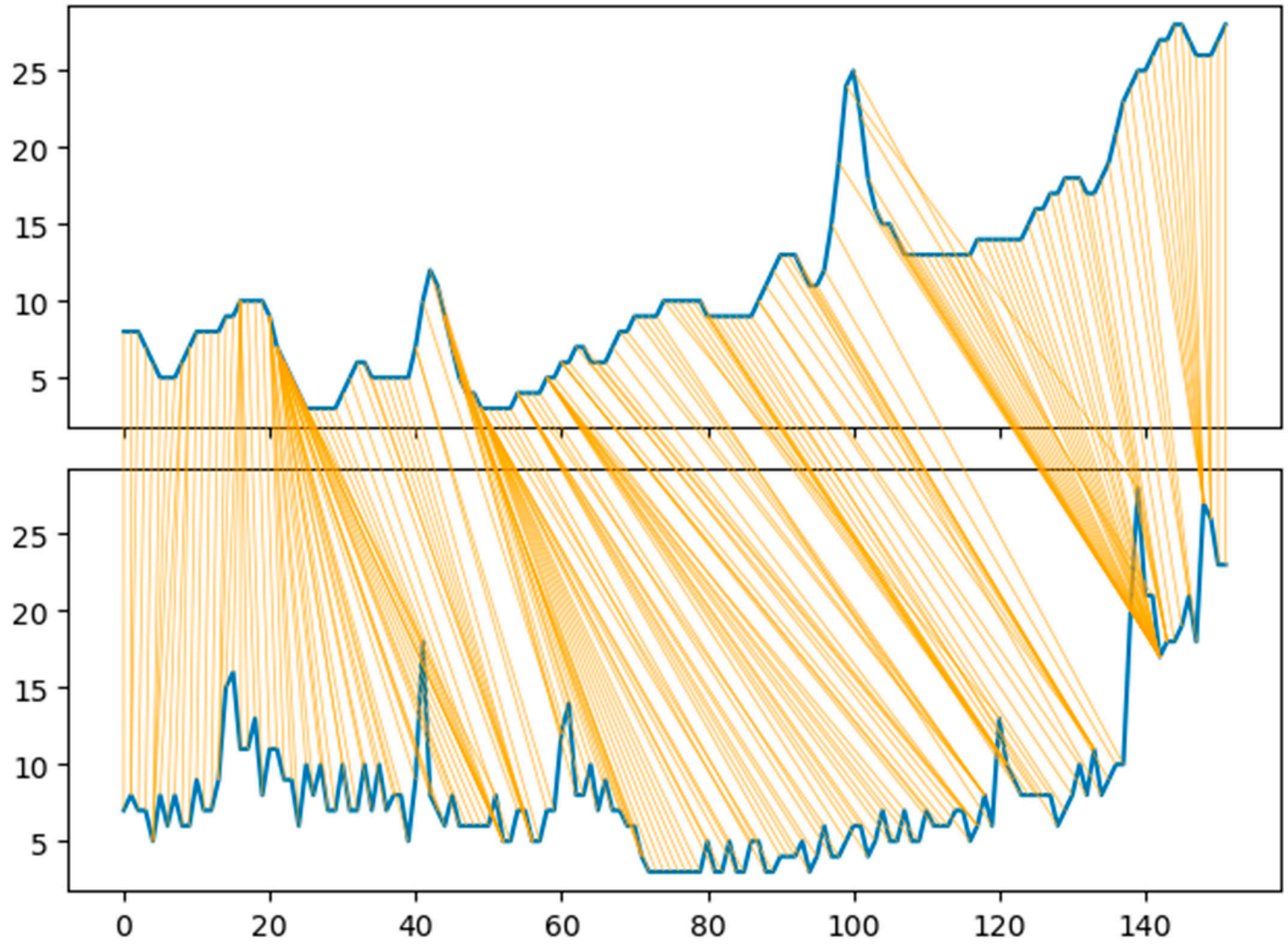

Figure 5. Comparison between two series using DTW with no window applied.

Long series can have longer distances than short ones, even in cases where the fit between series is better. To consider this, the obtained distance is divided by the length of the series. This gives a metric, d, with a similar meaning to MAE but considering the resemblance in shape of the two series.

The methodology followed is summarised in Figure 6.

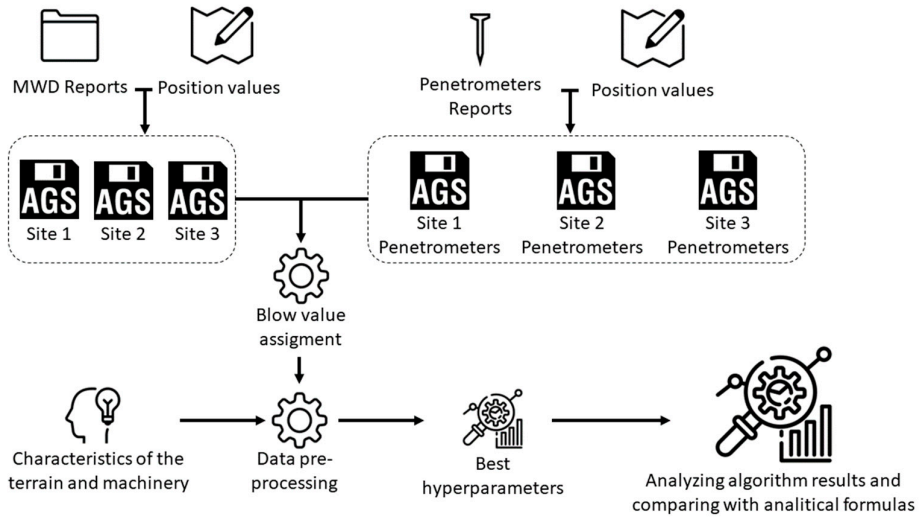

Figure 6. Scheme of the methodology followed.

3. Results

3.1. Attained Metrics

Table 5 shows the different metrics obtained when comparing the penetrometer blows and the compound parameters. The data corresponds to distances of 2 m between the inclusion and the penetrometer. To obtain the errors and distances for the compound parameters, they are scaled to have the same maximum and minimum values as the penetrometer blows for each site.

Table 6 shows the metrics values for the different algorithms studied. As in the case of Table 5, the values shown are for distances of 2 m between the inclusion and the penetrometer.

In every site the best algorithms outperform the compound parameters in terms of R, MAE, RMSE, and d.

Figure 7 shows the average profiles obtained for the different compound parameters, as well as the real average penetrometric profile. The parameter values are scaled between 0 and 30 for easier comparison.

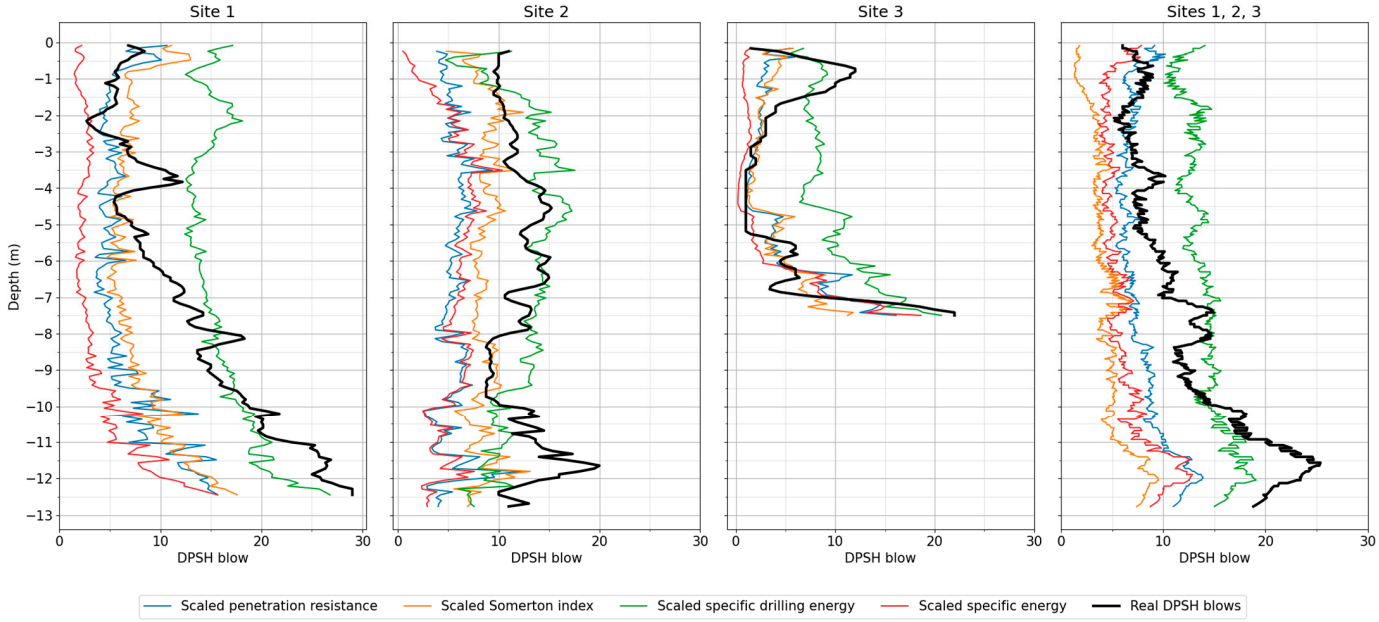

Figure 7. Mean profiles of compound parameters and real mean penetrometric profile.

Figure 8 shows the predicted penetrometric profiles on the test set by the different algorithms along with the actual mean penetrometric profile.

The profile predicted by the algorithms generally fits quite well with the actual penetrometric profile. With traditional analytical formulas, the fit is lower.

Additionally, another data point obtained during the drilling is the depth reached. If we add depth as a variable to the algorithms, we obtain the correlation and determination coefficients values of Table 7.

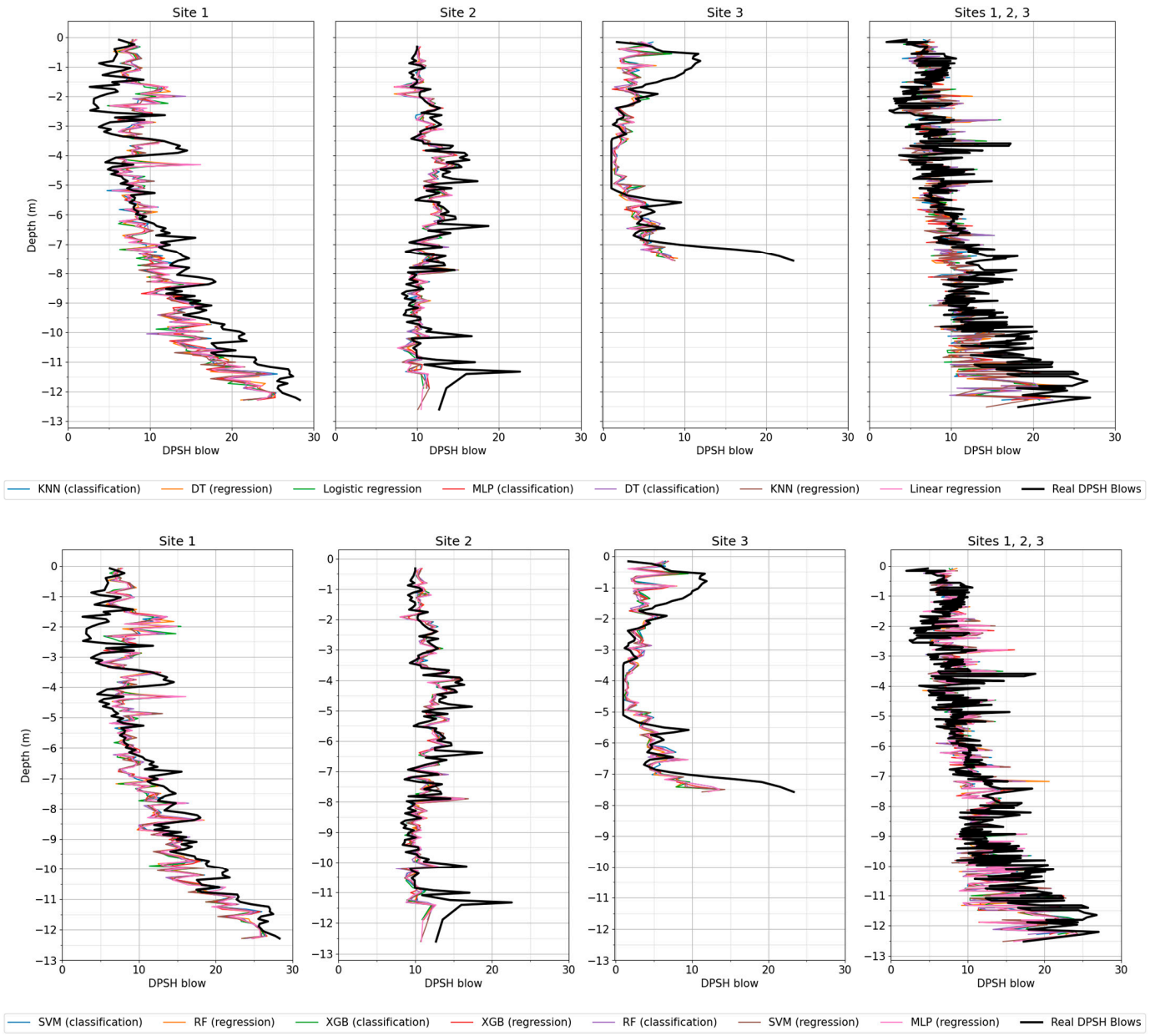

Figure 8. Mean penetrometric profiles predicted by the algorithms and real mean penetrometric profile.

When taking depth into account, the metrics significantly improve for all sites. In addition to the checks with and without depth, oversampling has been considered, obtaining the results in Tables 8 and 9.

Table 5. Metrics obtained when comparing the penetrometer blows and the relevant compound parameters. Bold marks the best results for each metric and compound parameter.

	Site 1				**Site 2**				**Site 3**				**Site 1, 2, 3**			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
Penetration resistance	0.45 ± 0.04	7.02 ± 0.23	8.95 ± 0.25	6.78 ± 1.91	−0.06 ± 0.03	5.13 ± 0.15	6.46 ± 0.17	2.26 ± 0.97	**0.44 ± 0.01**	**3.00 ± 0.03**	**4.16 ± 0.03**	**2.45 ± 0.29**	0.36 ± 0.04	6.12 ± 0.28	7.80 ± 0.32	3.92 ± 2.51
Somerton index	0.42 ± 0.04	6.28 ± 0.30	7.81 ± 0.34	5.81 ± 1.20	**−0.10 ± 0.02**	5.03 ± 0.27	6.36 ± 0.26	2.66 ± 1.12	0.41 ± 0.01	3.31 ± 0.04	4.28 ± 0.04	2.61 ± 0.39	**0.40 ± 0.04**	**4.96 ± 0.26**	**6.47 ± 0.29**	3.77 ± 1.84
Drilling specific energy	0.45 ± 0.03	**5.03 ± 0.36**	**6.60 ± 0.39**	**4.99 ± 1.90**	0.09 ± 0.03	4.93 ± 0.26	6.18 ± 0.33	2.92 ± 2.10	0.24 ± 0.01	5.65 ± 0.12	6.78 ± 0.13	3.62 ± 1.91	0.37 ± 0.04	5.50 ± 0.27	6.79 ± 0.33	3.81 ± 2.20
Specific energy	**0.58 ± 0.02**	7.29 ± 0.40	8.99 ± 0.40	6.12 ± 1.79	0.01 ± 0.03	**4.81 ± 0.15**	**6.11 ± 0.17**	**2.13 ± 0.90**	**0.44 ± 0.01**	3.05 ± 0.04	4.63 ± 0.04	2.49 ± 0.22	0.37 ± 0.04	6.72 ± 0.43	8.39 ± 0.43	**3.63 ± 2.23**

Table 6. Metrics obtained when comparing the penetrometer blows and values given by the algorithms. Bold marks the best results for each metric and algorithm.

	Site 1				**Site 2**				**Site 3**				**Sites 1, 2, 3**			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
Linear regression	0.56 ± 0.03	4.64 ± 0.14	5.88 ± 0.16	4.65 ± 1.25	0.39 ± 0.04	2.45 ± 0.11	3.47 ± 0.19	3.69 ± 2.52	0.49 ± 0.05	3.12 ± 0.15	3.98 ± 0.21	2.46 ± 0.29	0.51 ± 0.02	4.07 ± 0.10	5.35 ± 0.12	4.44 ± 2.10
DT (Classification)	0.67 ± 0.04	3.87 ± 0.13	5.63 ± 0.19	2.72 ± 0.10	0.39 ± 0.04	2.52 ± 0.15	4.04 ± 0.22	2.89 ± 0.54	0.26 ± 0.1	3.28 ± 0.28	5.30 ± 0.46	2.73 ± 0.13	0.62 ± 0.02	3.53 ± 0.10	5.35 ± 0.16	2.03 ± 0.15
DT (Regression)	0.70 ± 0.02	3.81 ± 0.13	5.10 ± 0.18	2.96 ± 0.08	0.43 ± 0.05	2.40 ± 0.14	3.46 ± 0.23	2.91 ± 0.3	0.47 ± 0.06	2.98 ± 0.15	4.14 ± 0.25	2.07 ± 0.15	0.65 ± 0.03	3.52 ± 0.18	4.80 ± 0.17	3.15 ± 0.18
KNN (Classification)	0.69 ± 0.02	3.83 ± 0.11	5.64 ± 0.18	5.10 ± 0.09	0.41 ± 0.04	2.44 ± 0.17	3.96 ± 0.19	3.87 ± 0.18	0.30 ± 0.07	3.00 ± 0.23	5.00 ± 0.21	2.60 ± 0.08	0.66 ± 0.02	3.36 ± 0.11	5.21 ± 0.17	5.68 ± 0.12
KNN (Regression)	0.77 ± 0.02	3.38 ± 0.14	4.52 ± 0.20	3.80 ± 0.06	0.49 ± 0.02	2.25 ± 0.09	3.32 ± 0.14	3.69 ± 0.09	0.49 ± 0.04	2.94 ± 0.19	3.87 ± 0.27	2.41 ± 0.07	0.72 ± 0.02	3.09 ± 0.04	4.30 ± 0.06	5.12 ± 0.09
Logistic regression	0.42 ± 0.05	5.25 ± 0.2	7.36 ± 0.32	4.87 ± 0.28	0.31 ± 0.02	2.47 ± 0.10	3.68 ± 0.19	3.43 ± 0.05	0.35 ± 0.07	3.13 ± 0.18	5.25 ± 0.27	2.76 ± 0.06	0.38 ± 0.01	4.66 ± 0.12	6.54 ± 0.17	5.11 ± 0.10
MLP (Classification)	0.73 ± 0.03	3.47 ± 0.11	5.09 ± 0.22	6.03 ± 0.28	0.35 ± 0.06	2.46 ± 0.16	3.88 ± 0.26	5.23 ± 0.53	0.33 ± 0.12	2.86 ± 0.28	4.89 ± 0.38	2.28 ± 0.21	0.62 ± 0.03	3.51 ± 0.12	5.31 ± 0.16	6.06 ± 0.36
MLP (Regression)	**0.79 ± 0.02**	3.23 ± 0.17	4.31 ± 0.21	5.94 ± 0.42	0.46 ± 0.03	2.30 ± 0.11	3.23 ± 0.14	3.99 ± 0.34	0.58 ± 0.05	2.88 ± 0.17	3.76 ± 0.23	2.49 ± 0.13	0.71 ± 0.02	3.24 ± 0.11	4.40 ± 0.14	6.41 ± 0.75
RF (Classification)	0.73 ± 0.03	3.38 ± 0.16	5.08 ± 0.27	**1.10 ± 0.05**	0.52 ± 0.03	2.18 ± 0.07	3.56 ± 0.17	**0.82 ± 0.27**	0.38 ± 0.07	2.91 ± 0.26	5.02 ± 0.42	1.93 ± 0.15	0.72 ± 0.02	**2.87 ± 0.11**	4.60 ± 0.17	**0.78 ± 0.10**
RF (Regression)	**0.79 ± 0.02**	**3.16 ± 0.12**	4.36 ± 0.16	1.36 ± 0.03	0.54 ± 0.07	**2.13 ± 0.12**	3.25 ± 0.24	1.31 ± 0.17	0.63 ± 0.05	2.46 ± 0.16	**3.34 ± 0.24**	1.08 ± 0.03	**0.75 ± 0.02**	2.89 ± 0.05	**4.10 ± 0.12**	1.51 ± 0.10
SVM (Classification)	0.72 ± 0.02	3.53 ± 0.15	5.23 ± 0.21	6.00 ± 0.24	0.49 ± 0.04	2.22 ± 0.14	3.52 ± 0.24	4.43 ± 0.37	0.40 ± 0.13	2.78 ± 0.22	4.81 ± 0.32	2.36 ± 0.11	0.68 ± 0.02	3.28 ± 0.11	4.91 ± 0.16	6.42 ± 0.29

Table 6. *Cont.*

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
SVM (Regression)	0.76 ± 0.02	3.32 ± 0.09	4.66 ± 0.17	8.68 ± 0.32	0.45 ± 0.04	2.32 ± 0.04	3.52 ± 0.10	4.69 ± 0.22	0.51 ± 0.04	2.74 ± 0.18	4.14 ± 0.3	2.53 ± 0.2	0.69 ± 0.01	3.20 ± 0.08	4.51 ± 0.12	9.68 ± 0.65
XGB (Classification)	0.71 ± 0.02	3.55 ± 0.08	5.23 ± 0.15	1.55 ± 0.04	0.42 ± 0.05	2.22 ± 0.09	3.58 ± 0.13	3.06 ± 0.08	0.29 ± 0.11	3.18 ± 0.35	5.28 ± 0.51	2.86 ± 0.17	0.66 ± 0.02	3.18 ± 0.10	4.99 ± 0.20	1.01 ± 0.14
XGB (Regression)	**0.79 ± 0.03**	3.21 ± 0.15	**4.34 ± 0.23**	1.79 ± 0.04	**0.56 ± 0.05**	2.20 ± 0.14	**3.18 ± 0.27**	1.20 ± 0.14	**0.62 ± 0.06**	**2.44 ± 0.21**	3.62 ± 0.36	**1.07 ± 0.05**	0.74 ± 0.01	3.01 ± 0.08	4.19 ± 0.11	2.01 ± 0.06

Table 7. Metrics obtained when comparing the penetrometer blows and the values given by the algorithms considering depth. Bold marks the best results for each metric and algorithm.

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
Linear regression	0.86 ± 0.01	2.86 ± 0.09	3.60 ± 0.14	2.28 ± 0.71	0.39 ± 0.04	2.46 ± 0.11	3.47 ± 0.18	3.70 ± 2.96	0.54 ± 0.04	2.91 ± 0.15	3.85 ± 0.25	2.17 ± 0.26	0.67 ± 0.02	3.69 ± 0.08	4.65 ± 0.10	4.45 ± 2.37
DT (Classification)	0.89 ± 0.01	1.91 ± 0.09	3.29 ± 0.13	0.82 ± 0.03	0.62 ± 0.11	1.81 ± 0.22	3.22 ± 0.48	1.71 ± 0.75	0.95 ± 0.01	0.62 ± 0.07	1.44 ± 0.16	0.22 ± 0.02	0.85 ± 0.01	2.00 ± 0.08	3.45 ± 0.11	1.05 ± 0.22
DT (Regression)	0.90 ± 0.01	2.10 ± 0.13	3.08 ± 0.20	1.41 ± 0.05	0.74 ± 0.04	1.50 ± 0.12	2.65 ± 0.22	1.07 ± 0.43	0.96 ± 0.01	0.62 ± 0.06	1.31 ± 0.07	0.26 ± 0.02	0.87 ± 0.01	1.94 ± 0.08	3.08 ± 0.14	1.39 ± 0.17
KNN (Classification)	0.90 ± 0.01	2.26 ± 0.12	3.41 ± 0.22	1.99 ± 0.03	0.63 ± 0.06	1.85 ± 0.07	3.28 ± 0.22	4.09 ± 0.12	0.66 ± 0.07	1.81 ± 0.35	3.47 ± 0.56	1.18 ± 0.06	0.86 ± 0.01	2.02 ± 0.09	3.42 ± 0.17	5.34 ± 0.08
KNN (Regression)	0.90 ± 0.01	2.13 ± 0.06	2.99 ± 0.13	1.88 ± 0.05	0.69 ± 0.06	1.65 ± 0.10	2.7 ± 0.25	3.86 ± 0.08	0.78 ± 0.04	1.65 ± 0.18	2.91 ± 0.31	1.20 ± 0.05	0.88 ± 0.01	1.88 ± 0.05	2.93 ± 0.10	4.97 ± 0.08
Logistic regression	0.85 ± 0.01	2.77 ± 0.15	3.97 ± 0.17	2.38 ± 0.07	0.26 ± 0.08	2.60 ± 0.15	3.85 ± 0.24	3.73 ± 0.06	0.31 ± 0.06	2.84 ± 0.19	4.99 ± 0.24	2.53 ± 0.14	0.62 ± 0.03	3.80 ± 0.21	5.35 ± 0.23	5.05 ± 0.11
MLP (Classification)	0.88 ± 0.01	2.17 ± 0.12	3.43 ± 0.18	3.48 ± 0.58	0.44 ± 0.04	2.33 ± 0.16	3.66 ± 0.25	5.01 ± 0.49	0.84 ± 0.04	1.41 ± 0.11	2.62 ± 0.23	1.14 ± 0.11	0.82 ± 0.01	2.34 ± 0.07	3.68 ± 0.07	5.46 ± 0.21
MLP (Regression)	0.91 ± 0.01	2.10 ± 0.08	2.99 ± 0.13	2.78 ± 0.18	0.64 ± 0.04	2.09 ± 0.14	2.93 ± 0.02	4.96 ± 0.51	0.90 ± 0.02	1.34 ± 0.15	2.00 ± 0.21	1.47 ± 0.18	0.85 ± 0.02	2.36 ± 0.09	3.25 ± 0.18	5.54 ± 0.40
RF (Classification)	0.91 ± 0.02	1.67 ± 0.13	2.98 ± 0.26	**0.40 ± 0.04**	0.78 ± 0.04	1.35 ± 0.10	2.46 ± 0.20	0.70 ± 0.49	0.91 ± 0.03	0.89 ± 0.15	1.87 ± 0.34	0.21 ± 0.03	0.90 ± 0.01	1.49 ± 0.09	2.73 ± 0.18	**0.47 ± 0.19**
RF (Regression)	**0.94 ± 0.01**	**1.56 ± 0.05**	**2.41 ± 0.15**	0.67 ± 0.02	0.83 ± 0.03	1.34 ± 0.08	2.17 ± 0.14	1.08 ± 0.26	0.97 ± 0.00	0.52 ± 0.05	1.03 ± 0.09	0.26 ± 0.01	0.92 ± 0.01	1.53 ± 0.04	2.40 ± 0.07	0.77 ± 0.07
SVM (Classification)	0.90 ± 0.02	2.01 ± 0.16	3.21 ± 0.32	3.08 ± 0.28	0.62 ± 0.05	1.86 ± 0.09	3.17 ± 0.23	4.19 ± 0.22	0.83 ± 0.02	1.38 ± 0.09	2.58 ± 0.18	1.24 ± 0.07	0.85 ± 0.01	2.06 ± 0.05	3.38 ± 0.11	5.87 ± 0.31

Table 7. *Cont.*

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
SVM (Regression)	0.90 ± 0.01	2.06 ± 0.07	3.07 ± 0.18	2.15 ± 0.12	0.62 ± 0.04	1.94 ± 0.11	2.93 ± 0.20	6.61 ± 0.75	0.80 ± 0.04	1.82 ± 0.17	2.82 ± 0.27	1.82 ± 0.14	0.85 ± 0.01	2.27 ± 0.06	3.30 ± 0.10	9.74 ± 0.44
XGB (Classification)	0.90 ± 0.01	1.78 ± 0.12	3.14 ± 0.20	0.45 ± 0.04	0.52 ± 0.06	1.98 ± 0.14	3.44 ± 0.27	2.47 ± 0.61	0.85 ± 0.04	1.23 ± 0.20	2.51 ± 0.38	1.17 ± 0.09	0.88 ± 0.02	1.70 ± 0.12	3.06 ± 0.22	0.55 ± 0.09
XGB (Regression)	0.93 ± 0.01	1.59 ± 0.09	2.50 ± 0.18	0.48 ± 0.02	**0.86 ± 0.02**	**1.18 ± 0.05**	**1.92 ± 0.09**	**0.59 ± 0.22**	**0.98 ± 0.00**	**0.39 ± 0.06**	**0.88 ± 0.13**	**0.16 ± 0.01**	**0.93 ± 0.01**	**1.49 ± 0.03**	**2.33 ± 0.06**	0.48 ± 0.05

Table 8. Metrics obtained when comparing the penetrometer blows and the values given by the algorithms considering oversampling. Bold marks the best results for each metric and algorithm.

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
Linear regression	0.56 +0.03	4.95 +0.16	6.19 +0.20	5.50 +1.34	0.37 ± 0.04	3.27 ± 0.16	4.16 ± 0.21	3.02 ± 5.82	0.26 ± 0.14	4.17 ± 0.20	4.91 ± 0.26	3.51 ± 0.91	0.50 ± 0.02	5.00 ± 0.14	6.29 ± 0.17	3.40 ± 6.54
DT (Classification)	0.65 ± 0.03	4.21 ± 0.20	6.04 ± 0.28	1.88 ± 0.07	0.04 ± 0.05	2.81 ± 0.14	4.28 ± 0.22	1.76 ± 0.53	0.37 ± 0.09	3.11 ± 0.23	4.78 ± 0.30	1.26 ± 0.13	0.60 ± 0.04	3.80 ± 0.15	5.74 ± 0.25	1.73 ± 0.16
DT (Regression)	0.67 ± 0.03	4.04 ± 0.18	5.67 ± 0.30	2.13 ± 0.05	0.41 ± 0.07	2.85 ± 0.17	4.20 ± 0.28	1.93 ± 0.30	0.34 ± 0.01	3.26 ± 0.31	4.71 ± 0.44	1.53 ± 0.17	0.63 ± 0.04	3.75 ± 0.20	5.40 ± 0.32	2.23 ± 0.16
KNN (Classification)	0.71 ± 0.03	3.72 ± 0.18	5.49 ± 0.30	5.33 ± 0.11	0.43 ± 0.06	2.55 ± 0.18	4.13 ± 0.24	4.74 ± 0.12	0.33 ± 0.07	3.13 ± 0.24	5.02 ± 0.27	2.68 ± 0.11	0.67 ± 0.03	3.34 ± 0.13	5.19 ± 0.21	6.38 ± 0.15
KNN (Regression)	0.72 ± 0.01	3.66 ± 0.06	5.20 ± 0.08	5.00 ± 0.16	0.46 ± 0.07	2.64 ± 0.14	4.05 ± 0.29	4.84 ± 0.17	0.31 ± 0.10	3.36 ± 0.27	4.98 ± 0.41	2.56 ± 0.18	0.69 ± 0.02	3.29 ± 0.08	4.98 ± 0.12	6.31 ± 0.22
Logistic regression	0.51 ± 0.04	6.31 ± 0.35	8.41 ± 0.38	8.74 ± 0.35	0.30 ± 0.07	5.69 ± 0.45	7.23 ± 0.43	8.24 ± 0.04	0.27 ± 0.07	3.82 ± 0.22	5.76 ± 0.33	3.76 ± 0.49	0.52 ± 0.03	5.80 ± 0.34	8.08 ± 0.42	5.85 ± 0.30
MLP (Classification)	0.65 ± 0.02	4.58 ± 0.16	6.44 ± 0.27	6.86 ± 0.67	0.35 ± 0.05	4.00 ± 0.27	5.76 ± 0.35	6.35 ± 0.31	0.29 ± 0.11	3.46 ± 0.34	5.26 ± 0.35	3.09 ± 0.39	0.59 ± 0.03	4.57 ± 0.24	6.62 ± 0.32	7.20 ± 0.51
MLP (Regression)	0.76 ± 0.01	3.59 ± 0.09	4.76 ± 0.10	6.48 ± 1.07	0.37 ± 0.02	3.22 ± 0.14	4.35 ± 0.16	6.60 ± 1.15	0.35 ± 0.13	3.83 ± 0.30	4.81 ± 0.39	3.53 ± 0.58	0.65 ± 0.03	3.93 ± 0.14	5.16 ± 0.21	11.0 ± 1.46
RF (Classification)	0.70 ± 0.02	3.81 ± 0.15	5.59 ± 0.22	**1.03 ± 0.05**	0.48 ± 0.04	2.57 ± 0.16	4.07 ± 0.21	**1.12 ± 0.40**	0.46 ± 0.06	2.81 ± 0.21	4.53 ± 0.22	**1.00 ± 0.08**	0.67 ± 0.03	3.34 ± 0.13	5.33 ± 0.18	0.89 ± 0.11
RF (Regression)	**0.77 ± 0.03**	**3.32 ± 0.16**	**4.56 ± 0.20**	1.36 ± 0.04	**0.55 ± 0.04**	**2.26 ± 0.13**	**3.27 ± 0.19**	1.36 ± 0.22	0.49 ± 0.09	2.97 ± 0.17	4.16 ± 0.28	1.33 ± 0.09	**0.73 ± 0.01**	**3.02 ± 0.08**	4.35 ± 0.12	1.27 ± 0.09
SVM (Classification)	0.67 ± 0.03	4.05 ± 0.24	5.85 ± 0.28	5.88 ± 0.17	0.41 ± 0.04	3.25 ± 0.14	4.74 ± 0.16	4.63 ± 0.25	0.33 ± 0.08	3.14 ± 0.23	5.01 ± 0.33	2.75 ± 0.21	0.64 ± 0.02	4.02 ± 0.18	5.96 ± 0.24	6.36 ± 0.29
SVM (Regression)	0.74 ± 0.03	3.65 ± 0.16	4.96 ± 0.28	9.46 ± 1.42	0.40 ± 0.08	3.14 ± 0.21	4.67 ± 0.81	19.44 ± 3.33	0.18 ± 0.17	4.16 ± 0.25	5.36 ± 0.43	4.75 ± 0.67	0.67 ± 0.01	3.80 ± 0.10	5.13 ± 0.11	20.56 ± 2.0

Table 8. *Cont.*

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
XGB (Classification)	0.68 ± 0.02	3.83 ± 0.20	5.72 ± 0.30	1.05 ± 0.06	0.43 ± 0.08	2.68 ± 0.21	4.25 ± 0.29	1.21 ± 0.32	0.43 ± 0.06	2.79 ± 0.21	4.59 ± 0.30	1.63 ± 0.07	0.67 ± 0.02	3.39 ± 0.11	5.31 ± 0.17	**1.00 ± 0.15**
XGB (Regression)	0.76 ± 0.02	3.39 ± 0.14	4.68 ± 0.18	1.13 ± 0.04	0.49 ± 0.04	2.40 ± 0.08	3.54 ± 0.16	1.16 ± 0.19	**0.51 ± 0.06**	**2.71 ± 0.14**	**3.95 ± 0.30**	1.14 ± 0.11	**0.73 ± 0.01**	3.08 ± 0.04	**4.40 ± 0.08**	1.05 ± 0.06

Table 9. Metrics obtained when comparing the penetrometer blows and the values given by the algorithms considering depth and oversampling. Bold marks the best results for each metric and algorithm.

	Site 1				Site 2				Site 3				Sites 1, 2, 3			
	R	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**	**R**	**MAE**	**RMSE**	**d**
Linear regression	0.86 ± 0.01	2.94 ± 0.09	3.74 ± 0.14	2.46 +0.79	0.37 ± 0.04	3.41 ± 0.18	4.26 ± 0.21	7.25 ± 4.80	0.39 ± 0.08	3.75 ± 0.19	4.55 ± 0.24	3.18 ± 0.46	0.66 ± 0.02	4.23 ± 0.09	5.29 ± 0.12	5.50 ± 3.44
DT (Classification)	0.87 ± 0.01	2.17 ± 0.13	3.63 ± 0.22	0.83 ± 0.06	0.57 ± 0.06	2.03 ± 0.14	3.52 ± 0.24	1.45 ± 0.35	0.75 ± 0.08	1.60 ± 0.29	2.98 ± 0.43	0.63 ± 0.15	0.80 ± 0.01	2.35 ± 0.10	3.96 ± 0.15	1.04 ± 0.14
DT (Regression)	0.88 ± 0.01	2.28 ± 0.10	3.47 ± 0.14	1.15 ± 0.03	0.68 ± 0.05	1.90 ± 0.16	3.12 ± 0.29	0.99 ± 0.17	0.73 ± 0.07	1.77 ± 0.20	3.15 ± 0.41	0.78 ± 0.12	0.83 ± 0.01	2.33 ± 0.09	3.72 ± 0.13	1.07 ± 0.07
KNN (Classification)	0.89 ± 0.02	2.21 ± 0.12	3.37 ± 0.24	2.33 ± 0.15	0.68 ± 0.06	1.68 ± 0.12	2.99 ± 0.25	4.53 ± 0.10	0.64 ± 0.07	1.83 ± 0.26	3.66 ± 0.52	1.32 ± 0.11	0.86 ± 0.01	1.99 ± 0.07	3.42 ± 0.17	5.81 ± 0.15
KNN (Regression)	0.89 ± 0.01	2.26 ± 0.07	3.34 ± 0.13	2.40 ± 0.11	0.69 ± 0.04	1.79 ± 0.14	2.99 ± 0.21	4.62 ± 0.17	0.66 ± 0.07	1.90 ± 0.22	3.44 ± 0.41	1.27 ± 0.08	0.86 ± 0.01	2.03 ± 0.06	3.33 ± 0.16	5.57 ± 0.13
Logistic regression	0.84 ± 0.01	3.49 ± 0.11	4.67 ± 0.12	4.70 ± 0.47	0.38 ± 0.04	4.51 ± 0.21	6.17 ± 0.25	7.45 ± 0.40	0.57 ± 0.13	2.58 ± 0.29	4.24 ± 0.51	2.18 ± 0.15	0.68 ± 0.02	4.66 ± 0.15	6.41 ± 0.23	6.25 ± 0.17
MLP (Classification)	0.84 ± 0.03	2.7 ± 0.18	4.11 ± 0.38	3.78 ± 0.57	0.52 ± 0.06	2.80 ± 0.19	4.42 ± 0.29	5.51 ± 0.47	0.66 ± 0.09	1.98 ± 0.27	3.52 ± 0.49	1.53 ± 0.18	0.77 ± 0.02	3.03 ± 0.12	4.66 ± 0.17	6.65 ± 0.47
MLP (Regression)	0.90 ± 0.01	2.34 ± 0.10	3.16 ± 0.13	3.31 ± 0.37	0.60 ± 0.03	2.65 ± 0.11	3.56 ± 0.14	7.64 ± 1.20	0.77 ± 0.09	1.94 ± 0.30	2.88 ± 0.55	2.07 ± 0.52	0.82 ± 0.01	2.75 ± 0.05	3.71 ± 0.13	6.50 ± 0.33
RF (Classification)	0.90 ± 0.01	1.88 ± 0.12	3.26 ± 0.20	0.47 ± 0.03	0.68 ± 0.07	1.70 ± 0.16	3.09 ± 0.36	0.82 ± 0.04	0.79 ± 0.03	1.43 ± 0.14	2.91 ± 0.29	0.54 ± 0.06	0.88 ± 0.01	1.75 ± 0.09	3.15 ± 0.13	**0.54 ± 0.13**
RF (Regression)	**0.93 ± 0.01**	**1.82 ± 0.10**	**2.70 ± 0.18**	0.73 ± 0.02	0.74 ± 0.05	**1.61 ± 0.09**	**2.52 ± 0.20**	1.12 ± 0.31	0.87 ± 0.05	1.37 ± 0.25	2.30 ± 0.47	0.58 ± 0.13	**0.90 ± 0.01**	1.76 ± 0.05	**2.69 ± 0.11**	0.82 ± 0.12
SVM (Classification)	0.88 ± 0.01	2.24 ± 0.10	3.52 ± 0.17	2.97 ± 0.30	0.61 ± 0.05	2.11 ± 0.11	3.49 ± 0.17	3.74 ± 0.16	0.66 ± 0.07	1.86 ± 0.25	3.55 ± 0.49	1.36 ± 0.12	0.81 ± 0.02	2.40 ± 0.09	3.89 ± 0.12	5.78 ± 0.18
SVM (Regression)	0.89 ± 0.01	2.22 ± 0.09	3.26 ± 0.15	5.97 ± 1.16	0.55 ± 0.06	2.66 ± 0.23	3.90 ± 0.31	13.39 ± 3.37	0.74 ± 0.08	2.04 ± 0.17	3.11 ± 0.35	2.79 ± 0.37	0.83 ± 0.02	2.62 ± 0.11	3.72 ± 0.21	14.05 ± 1.23
XGB (Classification)	0.90 ± 0.02	1.88 ± 0.13	3.27 ± 0.25	**0.46 ± 0.04**	0.65 ± 0.04	1.82 ± 0.14	3.22 ± 0.24	**0.67 ± 0.11**	0.82 ± 0.03	**1.21 ± 0.12**	2.55 ± 0.17	1.20 ± 0.07	0.85 ± 0.01	1.97 ± 0.08	3.50 ± 0.11	0.60 ± 0.11
XGB (Regression)	0.92 ± 0.01	1.83 ± 0.08	2.75 ± 0.15	0.60 ± 0.02	**0.75 ± 0.04**	1.64 ± 0.10	2.54 ± 0.19	0.88 ± 0.19	**0.89 ± 0.04**	1.28 ± 0.22	**2.17 ± 0.52**	**0.48 ± 0.11**	**0.90 ± 0.01**	**1.74 ± 0.05**	2.71 ± 0.10	0.68 ± 0.09

In most cases oversampling leads to a worse performance. The best option would be to take depth into account without oversampling in this case (Figure 9).

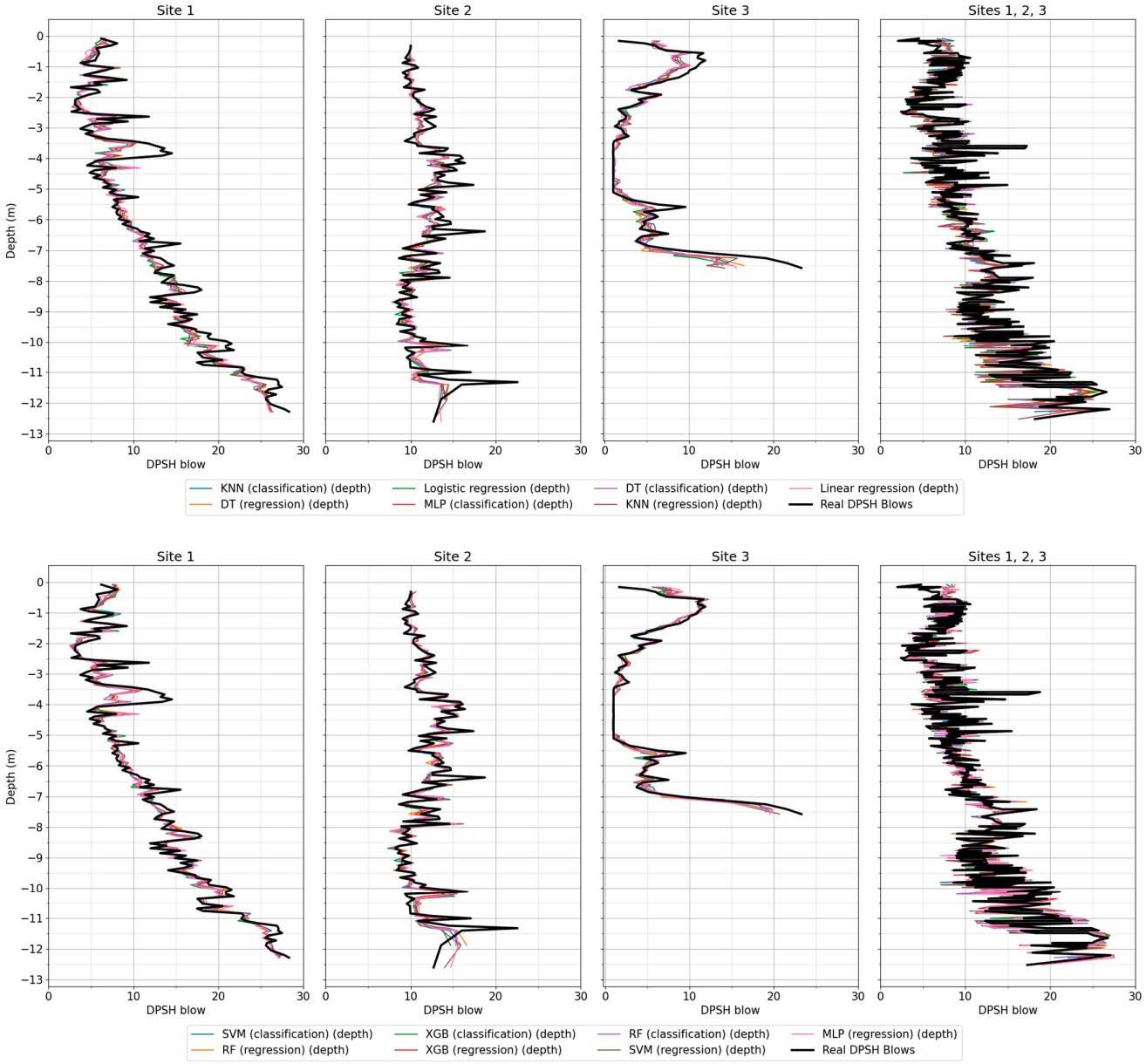

Figure 9. Mean penetrometric profiles predicted by the algorithms and real mean penetrometric profile considering depth.

3.2. Effect of Distance

Figure 10 shows the value of the correlation coefficient for the different algorithms based on the maximum distance between the data and its corresponding penetrometer.

In general, the behaviour was similar in the different sites. The maximum correlation coefficient was not reached at the minimum distance since the number of samples is small. The highest correlations were reached at about 2 m, from which the correlation coefficients decrease until reaching an asymptotic branch.

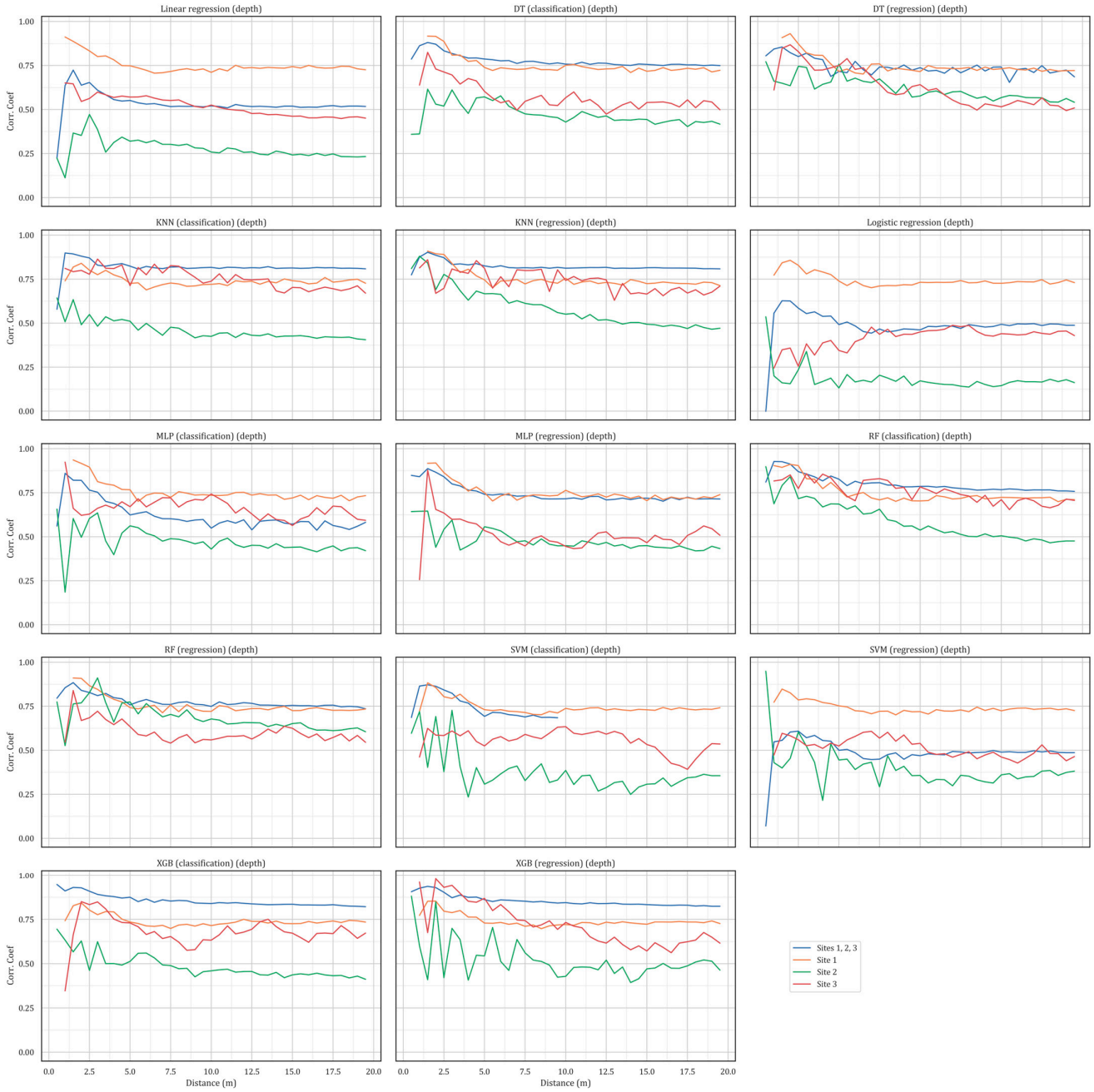

Figure 10. Values of the correlation coefficient for the different algorithms depending on the maximum distance to the penetrometer considered.

4. Discussion

In this article the correlation between the drilling data of rigid inclusions and the values of dynamic penetrometers has been analysed. The goal was for drilling data to provide the same information as penetrometers.

The results obtained show that the analytical formulas are too rigid for the range of terrains analysed. ML algorithms showed to be better adapted to the penetrometric profile of the ground.

It was also decided to introduce depth as an input variable for the algorithms. By taking the depth into account, a significant improvement in the coefficients was achieved.

It must be considered that by adding depth to the set of input variables, we are turning the data into a kind of time series, in the sense that the values can now be ordered sequentially. Time series present a very strong correlation with the variable that functions

as time, masking other possible relationships. It could cause the algorithms to be unable to predict changes in the nature of the terrain layers.

However, the nature of the terrain is related to the depth at which it is found. We are indirectly giving information to the algorithms such as the increase in friction in the tool or the difficulty in displacing the ground depending on the depth. The introduction of the depth must be done with caution, considering that it could mask the presence of layers that do not follow the prevailing trend.

On the other hand, the sequential nature of the drilling data allows the use of techniques such as DTW. This metric is easily understandable for the visual comparison of series in a plot.

Regarding oversampling, it has not proven useful in this case.

About the effect of distance when choosing the data set, it is observed that the maximum correlation coefficient was not reached at the minimum distance since the number of samples was small and the model is not behaving well in the training set. The highest correlations were reached considering a distance of about 2 m from which the correlation coefficients decreased. This validates the distance chosen to initially train the algorithms.

Regarding the algorithms used, after the analysis carried out, the regression approach seems to have better results.

The algorithms that have shown the best results for these datasets have been the tree ensemble methods (RF and XGB). This is expected, as these algorithms are top performers for tabular data. In general, XGB regressor has shown the best results.

The models perform slightly better in the train set than in the test set, which is to be expected. However, in some cases, the performance in the train set is perfect, a signal of overfitting. This could be solved with some adjustment to the hyperparameters given by grid search that are adjusting too much to the train set.

This paper is one of the few that deal with data from rigid inclusions. Compared with previous works this article presents a more robust methodology and addresses the issue of imbalanced data. The paper employs a shape-based metric—dynamic time warping (DTW)—to compare the predicted and actual penetrometric profiles, allowing for a more nuanced assessment of profile similarity than standard error metrics alone. The use of DTW for this kind of application is a newly proposed approach. In addition, the performance of the most important kind of algorithm has been analysed. The proposed methodology is a starting point for the characterisation of the bored soil based on the analysis of drilling data during the execution of rigid inclusions.

5. Conclusions

ML algorithms can reproduce with a reasonable level of accuracy the soil penetrometric profile from the drilling values of rigid inclusions. The results obtained represent an improvement compared to the traditional analytical formulas, which do not faithfully adapt to the blow values for the range and type of terrain analysed.

Attention must be paid to possible errors introduced during the selection of the data set (especially important is the distance between the inclusions and their nearest penetrometer). In practice, a distance of 2 m presented good results, but this depends on the site, number of available penetrometers, etc. Practitioners must evaluate what suit is best for their current situation.

Taking depth into account improves overall predictions, but engineers and researchers should remain alert to the possibility that abrupt changes in soil properties may be overlooked by algorithms relying heavily on depth as a predictor.

Tree ensemble methods are confirmed as the algorithms to use in this kind of problem, especially the XGB regressor.

The methodology to use DTW for drilling data proposed in this paper gives a useful metric to measure the similarity between obtained and expected results, going beyond traditional error measurements to compare overall profile shapes.

In this way, ML algorithms applied to drilling data can become a useful tool to gain more information about the terrain and validate the design hypotheses in deep foundations and ground improvement. Nonetheless, the identified limitations and the sensitivity to dataset selection should guide practitioners in responsibly adopting these methods for real-world applications.

Author Contributions: Conceptualisation, E.M.G., A.A.A.Á. and M.G.A.; methodology, E.M.G., A.A.A.Á. and M.G.A.; software, E.M.G.; validation, A.A.A.Á. and M.G.A.; investigation, E.M.G., A.A.A.Á. and M.G.A.; data curation, E.M.G.; writing—original draft preparation, E.M.G.; writing—review and editing, A.A.A.Á., E.M.G. and M.G.A.; visualisation, E.M.G., A.A.A.Á. and M.G.A.; supervision, A.A.A.Á. and M.G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Menard España and are available from the authors with the permission of Menard España.

Conflicts of Interest: Author Eduardo Martínez García was employed by the company Menard España. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Laudanski, G.; Reiffsteck, J.-L.; Benoît, J. Experimental study of drilling parameters using a test embankment. In *Geotechnical and Geophysical Site Characterization 4*; CRC Press: Boca Raton, FL, USA, 2012.
2. Burch, T.K. *Model-Based Demography*; Demographic Research Monographs; Springer International Publishing: Cham, Switzerland, 2018. [CrossRef]
3. ISSMGE. Reference List for Machine Learning and Its Applications in Geotechnical Engineering. Available online: https://www.issmge.org/news/reference-list-for-machine-learning-and-its-applications-in-geotechnical-engineering-part-i-need-to-know-knowledge0 (accessed on 25 June 2022).
4. Jaksa, M.; Liu, Z. Editorial for Special Issue "Applications of Artificial Intelligence and Machine Learning in Geotechnical Engineering". *Geosciences* **2021**, *11*, 399. [CrossRef]
5. Chen, L.; Wang, J.; Li, J. Deep Learning for Rainfall Prediction: Data Enhancement, Forecasting, and Estimation. In Proceedings of the 20th ICSMGE-State of the Art and Invited Lectures, Sydney, Australia, 1–5 May 2022; Rahman, M., Jaksa, M., Eds.; Australian Geomechanics Society: Sydney, Australia, 2022; pp. 535–540.
6. Liu, Z.; Lacasse, S. Machine learning in geotechnical engineering: Opportunities and applications. In Proceedings of the 20th ICSMGE-State of the Art and Invited Lectures, Sydney, Australia, 1–5 May 2022; Rahman, M., Jaksa, M., Eds.; Australian Geomechanics Society: Sydney, Australia, 2022; pp. 543–558.
7. Abdallah, A. Artificial neural network prediction of the water retention curve from physical soil parameters: Comparing continuous and pointwise approaches. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
8. Asr, A.A. An evolutionary modelling approach to compressed air tunneling in non-homogeneous and layered geo-materials to predict air losses. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
9. Bunieski, S. Practical methodologies to apply machine learning algorithms to ground improvement data. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
10. Congress, S.S.C.; Gajurel, A.; Chimauriya, H.; Puppala, A.J. A review of the applications of artificial intelligence techniques in geotechnical engineering disciplines. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
11. Haasnoot, J.; Madlener, P.; Tiggelman, L. Cooperation leads to a digital toolbox for future geotechnical engineers. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.

12. Jong, S.C.; Ong, D.E.L.; Oh, E. A Bayesian approach to the prediction of strength parameters. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
13. Kim, Y.; Kim, J.; Park, J.; Yun, T.S. Deep learning based classification model of rock mass based on RMR of tunnel face. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
14. Meng, J.; Mattsson, H.; Laue, J. An artificial neural network approach for three-dimensional slope stability prediction. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
15. Shi, C.; Wang, Y. Data-driven subsurface stratigraphy from prior knowledge and sparse site-specific measurements using multiple point statistics. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
16. Smyrniou, E.; Nuttall, J.; Coelho, B.Z. Using neural network components to model soil constitutive behaviour. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
17. Tinoco, J.; Correia, A.G. When artificial neural networks and genetic algorithms work together on soil embankments stability condition identification. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
18. Wang, Z.Z.; Goh, S.H. Adaptive training of convolutional neural networks for slope reliability analysis in spatially variable soils. In Proceedings of the 7th International Young Geotechnical Engineers Conference, Sydney, Australia, 29 April–1 May 2022.
19. Xu, J.-J.; Tang, C.-S.; Cheng, Q. Soil desiccation crack recognition and quantification using deep learning. In Proceedings of the 7th International Young Geotechnical Engineers Conference, Sydney, Australia, 29 April–1 May 2022; Scott, B., Ed.; Australian Geomechanics Society: Sidney, Australia, 2022; pp. 229–234.
20. Yousefpour, N.; Pouragha, M. Evaluation of residual strength of rocks based on acoustic emission data using AI. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022; Rahman, M., Jaksa, M., Eds.; Australian Geomechanics Society: Sydney, Australia, 2022; pp. 5113–5118.
21. Zhang, L.-M.; Xiao, T. Storm-based forecasting of man-made slope failures using machine learning. In Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering, Sydney, Australia, 1–5 May 2022.
22. *ISSMGE TC309*; Abstract Book of 4th International Symposium on Machine Learning and Big Data in Geoscience. University College Cork: Cork, Ireland, 2023. Available online: https://www.ismlg2023.com/abstracts (accessed on 31 January 2024).
23. García, E.M.; Alberti, M.G.; Arcos, A.A. Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values Using Machine Learning. In *Abstract of the 4th International Symposium on Machine Learning and Big Data in Geoscience*; University College Cork: Cork, Ireland, 2023.
24. Zhao, T.; Wang, Y. Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods. *Eng. Geol.* **2020**, *265*, 105430. [CrossRef]
25. Li, Y.; Rahardjo, H.; Satyanaga, A.; Rangarajan, S.; Lee, D.T.-T. Soil database development with the application of machine learning methods in soil properties prediction. *Eng. Geol.* **2022**, *306*, 106769. [CrossRef]
26. Kim, Y.; Yun, T.S. How to classify sand types: A deep learning approach. *Eng. Geol.* **2021**, *288*, 106142. [CrossRef]
27. Ren, Q.; Zhang, H.; Zhang, D.; Zhao, X.; Yan, L.; Rui, J.; Zeng, F.; Zhu, X. A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes. *Expert Syst. Appl.* **2022**, *202*, 117278. [CrossRef]
28. Shi, H.; Ma, W.; Xu, Z.; Lin, P. A novel integrated strategy of easy pruning, parameter searching, and re-parameterization for lightweight intelligent lithology identification. *Expert Syst. Appl.* **2023**, *231*, 120657. [CrossRef]
29. Zhu, X.; Zhang, H.; Zhu, R.; Ren, Q.; Zhang, L. Classification with noisy labels through tree-based models and semi-supervised learning: A case study of lithology identification. *Expert Syst. Appl.* **2024**, *240*, 122506. [CrossRef]
30. Liang, H.; Chen, H.; Guo, J.; Bai, J.; Jiang, Y. Research on lithology identification method based on mechanical specific energy principle and machine learning theory. *Expert Syst. Appl.* **2022**, *189*, 116142. [CrossRef]
31. Goh, A.T.C. Back-propagation neural networks for modeling complex systems. *Artif. Intell. Eng.* **1995**, *9*, 143–151. [CrossRef]
32. Flaate, K. *An Investigation of the Validity of Three Pile Driving Formulae in Cohesionless Material*; Publication No. 56; Norwegian Geotechnical Institute: Oslo, Norway, 1964; pp. 11–22.
33. Lee, I.-M.; Lee, J.-H. Prediction of pile bearing capacity using artificial neural networks. *Comput. Geotech.* **1996**, *18*, 189–200. [CrossRef]
34. Meyerhof, G.G. Bearing capacity and settlement of pile foundations. *J. Geotech. Geoenviron. Eng.* **1976**, *102*, 197–228. [CrossRef]
35. Pal, M.; Deswal, S. Modelling pile capacity using Gaussian process regression. *Comput. Geotech.* **2010**, *37*, 942–947. [CrossRef]
36. Alkroosh, I.; Nikraz, H. Predicting axial capacity of driven piles in cohesive soils using intelligent computing. *Eng. Appl. Artif. Intell.* **2012**, *25*, 618–627. [CrossRef]
37. Millán Muñoz, M.Á.; Galindo Aires, R.; Santos de Alencar, A.T. Red neuronal para el cálculo de la capacidad portante de cimentaciones superficiales sobre roca. In Proceedings of the XI Simposio Nacional de Ingeniería Geotécnica, Mieres, Spain, 24–27 May 2022.

38. Beattie, N. Monitoring-While-Drilling for Open-Pit Mining in a Hard Rock Environment. Master's Thesis, Queen's University, Kingston, ON, Canada, 2009.
39. Zhou, H.; Monteiro, S.T.; Hatherly, P.; Ramos, F.; Nettleton, E.; Oppolzer, F. Spectral Feature Selection for Automated Rock Recognition using Gaussian Process Classification. In Proceedings of the Australasian Conference on Robotics and Automation (ACRA), Sydney, Australia, 2–4 December 2009.
40. Zhou, H.; Hatherly, P.; Monteiro, S.T.; Ramos, F.; Oppolzer, F.; Nettleton, E.; Scheding, S. Automatic rock recognition from drilling performance data. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, St. Paul, MN, USA, 14–18 May 2012; pp. 3407–3412. [CrossRef]
41. Kadkhodaie-Ilkhchi, A.; Monteiro, S.T.; Ramos, F.; Hatherly, P. Rock Recognition From MWD Data: A Comparative Study of Boosting, Neural Networks, and Fuzzy Logic. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 680–684. [CrossRef]
42. Khushaba, R.N.; Melkumyan, A.; Hill, A.J. A Machine Learning Approach for Material Type Logging and Chemical Assaying from Autonomous Measure-While-Drilling (MWD) Data. *Math. Geosci.* **2022**, *54*, 285–315. [CrossRef]
43. Manzoor, S.; Liaghat, S.; Gustafson, A.; Johansson, D.; Schunnesson, H. Establishing relationships between structural data from close-range terrestrial digital photogrammetry and measurement while drilling data. *Eng. Geol.* **2020**, *267*, 105480. [CrossRef]
44. Galende-Hernández, M.; Menéndez, M.; Fuente, M.J.; Sainz-Palmero, G.I. Monitor-While-Drilling-based estimation of rock mass rating with computational intelligence: The case of tunnel excavation front. *Autom. Constr.* **2018**, *93*, 325–338. [CrossRef]
45. Hansen, T.F.; Erharter, G.H.; Marcher, T.; Liu, Z.; Tørresen, J. Improving face decisions in tunnelling by machine learning-based MWD analysis. *Geomech. Tunn.* **2022**, *15*, 222–231. [CrossRef]
46. Diaz, M.B.; Kim, K.Y.; Shin, H.-S.; Zhuang, L. Predicting rate of penetration during drilling of deep geothermal well in Korea using artificial neural networks and real-time data collection. *J. Nat. Gas Sci. Eng.* **2019**, *67*, 225–232. [CrossRef]
47. Klyuchnikov, N.; Zaytsev, A.; Gruzdev, A.; Ovchinnikov, G.; Antipova, K.; Ismailova, L.; Muravleva, E.; Burnaev, E.; Semenikhin, A.; Cherepanov, A.; et al. Data-driven model for the identification of the rock type at a drilling bit. *J. Pet. Sci. Eng.* **2019**, *178*, 506–516. [CrossRef]
48. Silversides, K.L.; Melkumyan, A. Machine learning for classification of stratified geology from MWD data. *Ore Geol. Rev.* **2022**, *142*, 104737. [CrossRef]
49. Zhang, J.; An, L.; Li, C.; Dias, D.; Jenck, O. Artificial neural network response assessment of a single footing on soft soil reinforced by rigid inclusions. *Eng. Struct.* **2023**, *281*, 115753. [CrossRef]
50. Dikshit, A.; Pradhan, B.; Alamri, A.M. Pathways and challenges of the application of artificial intelligence to geohazards modelling. *Gondwana Res.* **2021**, *100*, 290–301. [CrossRef]
51. Ma, Z.; Mei, G. Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Sci. Rev.* **2021**, *223*, 103858. [CrossRef]
52. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
53. Moayedi, H.; Mosallanezhad, M.; Rashid, A.S.A.; Jusoh, W.A.W.; Muazu, M.A. A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: Theory and applications. *Neural Comput. Appl.* **2020**, *32*, 495–518. [CrossRef]
54. Tehrani, F.S.; Calvello, M.; Liu, Z.; Zhang, L.; Lacasse, S. Machine learning and landslide studies: Recent advances and applications. *Nat. Hazards* **2022**, *114*, 1197–1245. [CrossRef]
55. Xie, J.; Huang, J.; Zeng, C.; Jiang, S.-H.; Podlich, N. Systematic Literature Review on Data-Driven Models for Predictive Maintenance of Railway Track: Implications in Geotechnical Engineering. *Geosciences* **2020**, *10*, 425. [CrossRef]
56. Zhang, W.; Zhang, R.; Wu, C.; Goh, A.T.C.; Lacasse, S.; Liu, Z.; Liu, H. State-of-the-art review of soft computing applications in underground excavations. *Geosci. Front.* **2020**, *11*, 1095–1106. [CrossRef]
57. Zhang, W.; Li, H.; Li, Y.; Liu, H.; Chen, Y.; Ding, X. Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artif. Intell. Rev.* **2021**, *54*, 5633–5673. [CrossRef]
58. Zhang, P.; Yin, Z.-Y.; Jin, Y.-F. Machine Learning-Based Modelling of Soil Properties for Geotechnical Design: Review, Tool Development and Comparison. *Arch. Comput. Methods Eng.* **2022**, *29*, 1229–1245. [CrossRef]
59. Möller, B.; Bergdahl, U.; K, E. Soil-rock sounding with MWD—A modern technique to investigate hard soils and rocks. In Proceedings of the 2nd International Conference on Site Characterization, Porto, Portugal, 19–22 September 2004.
60. Somerton, W.H. A Laboratory Study of Rock Breakage by Rotary Drilling. *Trans. AIME* **1959**, *216*, 92–97. [CrossRef]
61. Teale, R. The Concept of Specific Energy in Rock Drilling. *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.* **1965**, *2*, 57–73. [CrossRef]
62. Pfister, P. Recording Drilling Parameters in Ground Engineering. *Ground Eng.* **1985**, *18*, 16–21.
63. *UNE-EN ISO 22476-2:2008*; Geotechnical Investigation and Testing—Field Testing—Part 2: Dynamic Probing. ISO: Geneva, Switzerland, 2008.
64. García, E.M.; Alberti, M.G.; Arcos Álvarez, A.A. Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values Using Decision Trees and Random Forests. *Appl. Sci.* **2022**, *12*, 4565. [CrossRef]

65. Association of Geotechnical & Geoenvironmental Specialists. AGS Data Format. Available online: https://www.ags.org.uk/data-format/ (accessed on 10 December 2023).
66. Pandas Team. Pandas. Available online: https://pandas.pydata.org/ (accessed on 10 December 2023).
67. Association of Geotechnical & Geoenvironmental Specialists. AGS Python Library. Available online: https://gitlab.com/ags-data-format-wg/ags-python-library (accessed on 10 December 2023).
68. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
69. Scikit-learn. Sklearn.Model_Selection. GridSearchCV Class. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV (accessed on 10 December 2023).
70. Matejka, J.; Fitzmaurice, G. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1290–1294.
71. Charles, D.J.; Axtell, M.D.; Gourvenec, S. Assessing the quality of synthetic CPT training data using time series similarity. In Proceedings of the 4th International Symposium on Machine Learning & Big Data in Geoscience, Cork, Ireland, 28 August–1 September 2023.
72. Morse, M.D.; Patel, J.M. An efficient and accurate method for evaluating time series similarity. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 12–14 June 2007; pp. 569–580. [CrossRef]