Cristian Godoy Leiva

Site characterization using Kriging and Machine Learning approaches

Master's thesis in Geotechnics and Geohazards
June 2019

Master's thesis

NTNU
Norwegian University of
Science and Technology

Cristian Andrés Godoy Leiva

Site characterization using Kriging and Machine Learning approaches

NTNU – Trondheim
Norwegian University of
Science and Technology

Preface

This Master thesis is submitted as partial fulfillment of the MSc in Geotechnics and Geohazards at the Department of Civil and Environmental Engineering at the Norwegian University of Science and Technology (NTNU).

The aim of this study is to investigate the application of statistical methods in geotechnical engineering, in particular, the application of kriging and machine learning approaches to site characterization.

The study was performed during the spring semester and was developed from the joint interest of the supervisors, Vikas Thakur and Ivan Depina, and the student on the subject.

Trondheim, June 2019

Cristian Godoy Leiva

Acknowledgment

First of all, I express my deep gratitude to my family for their love and support. All my achievements are yours as well.

Special thanks to my supervisors Vikas and Ivan for their constant feedback and support.

Thanks to my friends in Norway who made these two years pass faster than I would have thought.

Thanks to Anders Lindgård from NGI and Samson Degago from Statens Vegvesen for sharing the information which I used for the Machine Learning chapter.

Summary and Conclusions

Site characterization has a great number of uncertainties associated with the nature of soil formation and deposition, but also due to technical and economic constraints. In geotechnical practice, these uncertainties have to be addressed in order to perform the most optimum design possible. In this scenario, statistical approaches can be useful to help engineers to gain a better insight into the soil properties and the uncertainties associated.

This master thesis focuses on using two statistical approaches to solve a major challenge faced in geotechnical site characterization: data availability. The first approach is the geostatistical interpolation of soil properties, in which a certain soil property is estimated based on measurements in different locations and the spatial relationship between data. The second approach is machine learning classification and is used to determine the soil types present on the field based on indirect measurements.

Geostatistical interpolation was performed on a slope located in Rissa, which has been studied deeply in previous works, five different CPTu test measurements were used to estimate the properties values in unsampled locations within the slope. The results show a good agreement between the properties distributions and the layering proposed for the site.

Machine learning classification was used to determine the soil classes present on two different sites based on CPTu tests measurements. The sites studied are the NGTS investigation site at Tiller, and the Skaudal bridge foundations site part of a Statens Vegvesen project. The results show a very good capacity of prediction using these algorithms compared with chart classifications commonly used in engineering practise.

Finally, it is always better for design purposes to have additional information whether it be as more field tests to cover the entire area or more laboratory tests to certainly know the properties of the soil beneath the surface. Since resources are rarely enough to fulfill all the engineering necessities, the use of the statistical approaches studied within the framework of this thesis can help engineers to improve the interpretation of field tests without increasing exploration costs.

Contents

Chapter 1

Introduction

1.1 Background

Geotechnical site characterization is a process that is always limited by several constrains whether technical or economical, making this a difficult task for geotechnical engineers which requires a great deal of engineering judgement, experience and complementary information. In the site characterization workflow, the field exploration and the following laboratory campaign are key tasks that can determine the success or failure of an engineering project. Since it is economically and technically impossible to get rid of all the uncertainties involved in the site characterization process, geotechnical engineers have to deal with them in an everyday basis. Is in this context where the use of field tests, which in general measure indirect properties, and statistical methods, which allow to make inference given incomplete information, have an important role. In this work, the use advanced statistical methods applied to field tests is going to be studied. In particular, this thesis focuses on the use of CPTu tests measurements complemented with kriging for interpolation and machine learning classification, and their application to geotechnical site characterization.

The use of these approaches in geotechnical engineering is starting to increase as new research support their success. Examples of geostatiscs in geotechnics are several: Olea (1999) applied universal kriging to water level measurements in High Plains aquifer in Kansas (2D horizontal

plane interpolation), Nadim (1988) used kriging to interpolate the depth to a weak clay layer under an offshore platform, Uzielli et al. (2010) used bayesian kriging to interpolate V_s values from seismic piezocone tests (SCPT) in clay (pseudo 3D interpolation), Firouzianbandpey et al. (2015) used universal kriging to estimate CPTu cone resistance at different depths considering soil's anisotropy (pseudo 3D interpolation).

Machine Learning given its flexibility can been applied in different ways, Rabarijoely et al. (2007) used a clustering algorithm to perform a zoning of earthquake damage, Farrokhzad et al. (2010), Samui and Sitharam (2011) and García et al. (2012) used machine learning to assess liquefaction potential in soil deposits, Krogstad et al. (2018) and Vezhapparambu et al. (2018) used Hidden Markov Models on field tests to classify soil and rock, respectively.

In the case of Kriging interpolation, a big challenge is faced when it is performed on a 2D cross section due to the difficulty associated to the determination of horizontal spatial correlation, since all the data points are located aligned in vertical lines. This is interesting since cross-sections are a main output in geotechnical engineering reports and constitute the basis for slope stability and stress-strain analyses.

On the other hand, the use of CPTu for soil classification is a widespread practice, specially using the well-known classification charts found in literature (Lunne et al., 2002). But the major challenge comes to light when the soil deposits are non-textbook soils, as in the case of quick or highly sensitive clays, very common in Norway. In this cases, alternatives should be found in order to keep the convenience of using indirect field measurements without expending a big amount of resources.

Problem Formulation

This thesis is a research study on site characterization using CPTu tests measurements complemented through the use of statistical methods, in particular geostatistics and machine learning approaches. The main focus is to gain knowledge on how both approaches can be implemented to help design engineers to achieve better characterization without an increased amount of testing. The thesis is divided in two main problems, the first is applying kriging to a well studied pro-

file in Rissa (Kornbrekke, 2012; Wolebo, 2016) and assess the performance of the interpolation. The other is to perform machine learning classification using CPTu tests from two databases to identify the presence of highly sensitive and quick clays.

1.2 Objectives

The main objectives of this project are

1. Study and gain insight in the theoretical framework of geostatistical interpolation and machine learning classification

2. Program in Python the necessary geostatistical methods to perform kriging and, if necessary customize them in order to fit the necessities of the study.

3. Learn and use the Python libraries available for machine learning (scikit-learn and hmm-learn).

4. Perform spatial interpolation of CPTu tests in a cross section, particularly profile 3-3 in Rissa slope (Kornbrekke, 2012), and assess the correctness of the interpolation results.

5. Perform the classification of the soils present at the NGTS and Vegvesen sites, putting special emphasis on the identification of quick and highly sensitive clays. Assess the quality of the prediction, the necessary number of training tests, and evaluate different classification methodologies.

1.3 Limitations

Kriging

Due to time constraints this thesis will focus on univariate kriging, despite cokriging (interpolation with the help of one or many covariates) is interesting because of the potential synergy between geotechnical and geophysical sounding. Anisotropy effects are going to be studied but limited to geometrical anisotropy, however zonal anisotropy can have a great effect in natural

soil deposits. Due to data availability, it will not be possible to assess the quality of the interpolation with the help of additional soundings, since all the data available will be used to make the predictions, however the results from Kornbrekke (2012) will be used to compare.

Machine Learning

Even though there are several classification algorithms available, this work will only focus on three of them: Logistic Regression, Naive Bayes and Hidden Markov Models. In the last two of them, Gaussian distribution is assumed for the distribution of CPTu measurements, despite the libraries allow for different distributions.

1.4   Approach

The general scientific approach to the project starts with a literature study of the theoretical background behind soil variability, geostatistics and machine learning classification.

To gain a better insight into the subject, kriging and all the associated calculation are going to be programmed in Python (Van Rossum, 1995) which will give the opportunity to customize them according to the needs of the project. The method will be applied at Rissa slope, using the available CPTu data to interpolate a continuous profile of the different measured properties. Since there is no additional sampling and testing, the assessment on the performance will be done comparing with the results of a previous study of the area. Cross-validation will be performed in order to evaluate the influence of the amount of information on the prediction accuracy.

The classification part will also be addressed using Python, but in this case all the algorithms are provided by two machine learning libraries: Scikit-Learn (Pedregosa et al., 2011) and hmmlearn (hmmlearn developers, 2010). The methodologies will be tested on two datasets representing different sites with presence of quick and highly sensitive clays. The different classification models will be trained by sequentially adding information, and the accuracy will be tested by comparing the classification results with the actual layering on the sites.

1.5 Structure of the Report

The rest of the thesis is structured as follows. Chapter 2 gives an introduction to basic geostatistical concepts. Chapter 3 gives a theoretical framework on machine learning and classification algorithms. Chapter 4 presents the results of the application of ordinary kriging to the CPTu tests performed in profile 3-3 in Rissa. Chapter 5 shows the results of machine learning classification on the datasets provided by NGTS and Statens Vegvesen. And, finally, Chapter 6 presents a summary and the conclusions of the project, and the discussion of the results.

Chapter 2

Geostatistics

This chapter is dedicated to state the theoretical background behind geostatistics. First a review of random functions and quantification of spatial correlation, continuing with the different kriging methods, a summary of their computation algorithms and technical aspects considered for their implementation.

The main bibliography used in this chapter is Journel (1978), Deutsch (2002) and Olea (1999).

2.1 Quantification of Spatial Correlation

2.1.1 Regionalization and Random Function Concepts

A variable that is distributed in space it is said to be *"regionalized"*. Regionalized variables usually describe natural properties, as ore body's metal grade, soil's friction angle, reservoir's piezometric elevation, population density, rainfall measurements, among others, such phenomena are called "regionalization". From a mathematical point of view a regionalized variable is a function $f(x)$ which takes a value at every point x within a region or spatial domain. A regionalized variable have two characteristics:

  i. A local, random behavior depending on some probability distribution (random variable)

  ii. An average aspect which requires a certain functional representation

The concept of *random functions* takes into account this dual behavior of regionalized variables.

The uncertainty of an unsampled value z is modeled through the probability distribution of random variable, $Z(x)$, which depends on the location x. A random function is a set of random variables defined over the field of interest ($\{Z(x), z \in$ study area$\}$). While a random variable is characterized by its cumulative distribution function (cdf), a random function is characterized by the set of all of its cdfs.

$$F(x_1, ..., x_N; z_1, ..., z_N) = P(Z(x_1) \leq z_1, ..., Z(x_N) \leq z_N) \tag{2.1}$$

2.1.2 Statistical Moments

Expectation or first order moment: If $f(x)$ is the probability density function (pdf) of $Z(x)$, then the expectation is a function of x defined by.

$$E(Z(x)) = \int x f(x) dx = \mu(x) \tag{2.2}$$

Second-order moments: In geostatistics, usually three second order moments are considered:

  i. The variance:
$$Var(Z(x)) = E((Z(x) - \mu(x))^2) \tag{2.3}$$

 ii. The covariance:
$$C(x_1, x_2) = E((Z(x_1) - \mu(x_1))(Z(x_2) - \mu(x_2))) \tag{2.4}$$

iii. The variogram (also called semivariogram):

$$\gamma(x_1, x_2) = \frac{1}{2} Var(Z(x_1) - Z(x_2)) \tag{2.5}$$

In simple words covariance measures the similarity between the spatial data, while variogram measures the average dissimilarity between the data.

2.1.3   Stationarity

Stationarity is an assumption that has to be made in order to allow sampling in different locations (or at different times) be used to infer the random variable cdf over the entire domain. Depending on the properties of the random function, different types of stationarity can be defined.

Strict stationarity: A random function is said to be stationary, in the strict sense, if its multivariate cdf is invariant under any traslation, that is:

$$F(x_1, ..., x_N) = F(x_1 + h, ..., x_N + h) \tag{2.6}$$

for any traslation vector h.

Second order stationarity: A random function is said to be second order stationary when:

i. The mathematical expectation, $E(x)$, exists and does not depend on spatial location, x. In other words,

$$E(Z(x)) = \mu \tag{2.7}$$

ii. for any pair of random variables $Z(x), Z(x + h)$ the covariance exists and depends on the separation distance h.

$$C(h) = E(Z(x + h) \cdot Z(x)) - \mu^2 \tag{2.8}$$

At $h = 0$ the stationary covariance $C(0)$ equals the stationary variance σ^2, that is,

$$C(0) = Var(Z(x)) = \sigma^2 \tag{2.9}$$

Intrinsic stationarity: A random function $Z(x)$ is intrinsic when:

i. The mathematical expectation, $E(x)$, exists and does not depend on spatial location, x.

ii. For all vectors h the increment $Z(x + h) - Z(x)$ has a finite variance which does not depend

on x. In other words, the variogram only depends on the increment vector h.

$$\gamma(x, x + h) = \gamma(h) \tag{2.10}$$

A second order stationary random function is intrinsic but the converse is not true.

The variogram and the covariance can be linked through the following equation:

$$\gamma(h) = C(0) - C(h) \tag{2.11}$$

This allows to estimate the variogram (whose advantages will be discussed later in this chapter) to study the spatial correlation of the data and then convert to covariance for the interpolation over a second order random field.

Quasi stationarity: If the random variable $Z(x)$ is only homogeneous in certain area defined by the limit b such as $|h| < b$. In those cases the covariance function, $C(x, x + h)$, is defined for distances $|h| < b$ being locally stationary. In practical terms, neighborhoods should be defined inside of which the expectation and covariance can be considered as stationary and the data is sufficient in order to make possible a statistical inference.

Stationarity is a property of the random function needed for inference and it is not a characteristic of the phenomenon under study. It is a decision that has to be made by the user and can not be assessed from data. In this work it is assumed that the random function is second order stationary and the verification of this assumption is not part of the scope of this thesis.

2.1.4 Empical Variogram

In geostatistics, the variogram is preferred for calculations, even though it might be converted to covariance afterwards. Among the benefits of using the variogram Olea (1999) mentions the following:

i. In a probabilistic framework, the variogram does not require knowledge of the mean of the random function for its estimation.

ii. The variogram require less assumptions for its existence compared with the covariance.

iii. The estimation of the variogram is insensitive to the addition of a constant to the random function, whereas the covariance is.

iv. Estimation of the variogram is less sensitive than the covariance estimation to failure to comply with the requirement that the trend must be constant (simple and ordinary kriging).

The most common estimator of the variogram from experimental data is defined as the average squared difference of values separated approximately by h.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} [z(x) - z(x+h)]^2 \tag{2.12}$$

Where $N(h)$ is the number of pairs for lag (h).

The algorithm to obtain the experimental variogram of a sampling is described below:

1. Identify all the possible combination of pairs within the data. Compute the distance and azimuth angle of each pair.

2. The variogram can be omnidirectional (when the search of pairs is made disregarding the direction) or directional, if this is the case, only the pairs oriented with a defined azimuth angle, α, within an angular and linear tolerance should be considered. The linear tolerance around the search direction is called bandwidth.

3. Set the lag increment, the maximum recommended lag is half of the maximum dimension of the sampling area. The pairs whose distance to each other is equal to the lag distance (plus or less a lag tolerance) are grouped in that lag.

4. Use the estimator defined in Equation 2.12 to compute the experimental variogram for each lag. It is recommended to discard lags where $N(h) < 30$. Basically, the variogram of a lag, is the mean value of the variogram of each pair of points in that lag.

Figure 2.1 shows how to define a lag for a given point, considering the distances and tolerances.
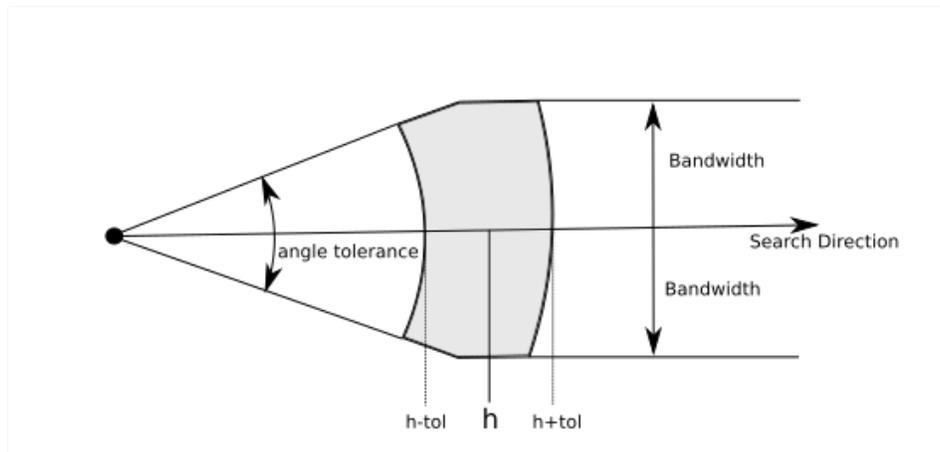
Figure 2.1: Illustration for lag definition for a given point, it is defined by the lag distance (h) and tolerance (tol), direction, angle tolerance and bandwidth. The shaded area is the area of allowable points for the given lag. (modified from Deutsch (2002))

The following figure shows the variogram cloud, which is the variogram of each pair of data, the lags discretization and, finally, the variogram of each lag.
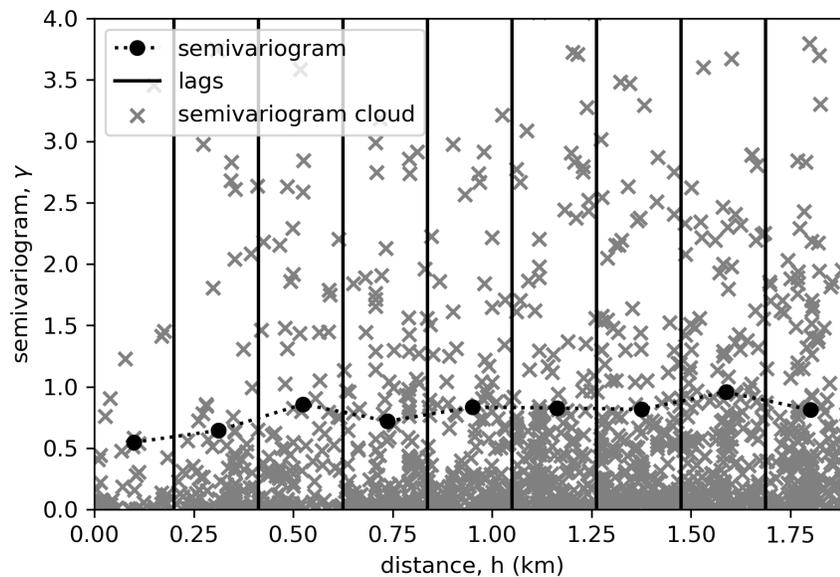
Figure 2.2: The experimental variogram computation involves computing each pair of data's variogram, the discretization of distances into lags and the averaging of lag's variograms.

2.1.5 Variogram Models

Kriging procedure (and other geostatistical procedures as simulation) involves inverting matrices whose terms are calculated from the variogram. If the discrete values are used, this can lead to singular matrices, multiple solutions or negative mean square errors in the kriging system. The solution is to use functions, called permissible or licit variogram functions, some of the most commonly used ones are presented below.

Nugget Effect:

$$\gamma(h) = \begin{cases} 0, & h = 0. \\ C_0, & h > 0. \end{cases} \tag{2.13}$$

Spherical:

$$\gamma(h) = \begin{cases} C(\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^2), & 0 \le h < a. \\ C, & a \le h. \end{cases} \tag{2.14}$$

Exponential:

$$\gamma(h) = C(1 - e^{-\frac{3h}{a}}) \tag{2.15}$$

Gaussian:

$$\gamma(h) = C(1 - e^{-3(\frac{h}{a})^2}) \tag{2.16}$$

In the definitions presented above, C is called the sill, C_0 is the nugget and a is the range of the

model. Figure 2.3 shown below, present the four variogram models mentioned before, however, there are many more who have not been mentioned for the sake of simplicity, the reader is referred to Olea (1999) for more information about variogram models.
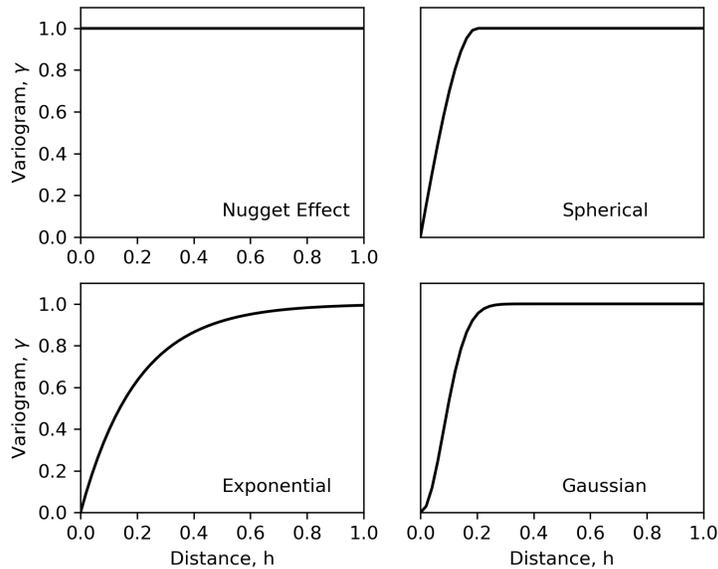
Figure 2.3: Four typical variogram models: nugget effect, spherical, exponential and Gaussian (parameters used: $C_0 = 1.0$, $a = 0.2$).

An important property is that a variogram model can be constructed as a possitive sum of licit variograms functions:

$$\gamma(h) = \sum_{i=1}^{n} \gamma_i(h) \tag{2.17}$$

Figure 2.4 shows the result of combining a nugget effect and a Gaussian models, it shows the parameters that usually define a variogram: nugget, sill and range. The difference between the sill and the nugget is often called partial sill. The combination of a nugget effect plus one of the remaining licit functions is commonly used since it reflects the uncertainty at small (close to zero) distance, which is usually associated with measurement errors. The range is the distance at which the data relate to each other, the practical range is defined as the distance at which the variogram value is 95% of the sill. The sill value, reflects the maximum dissimilarity between data for a given distance.
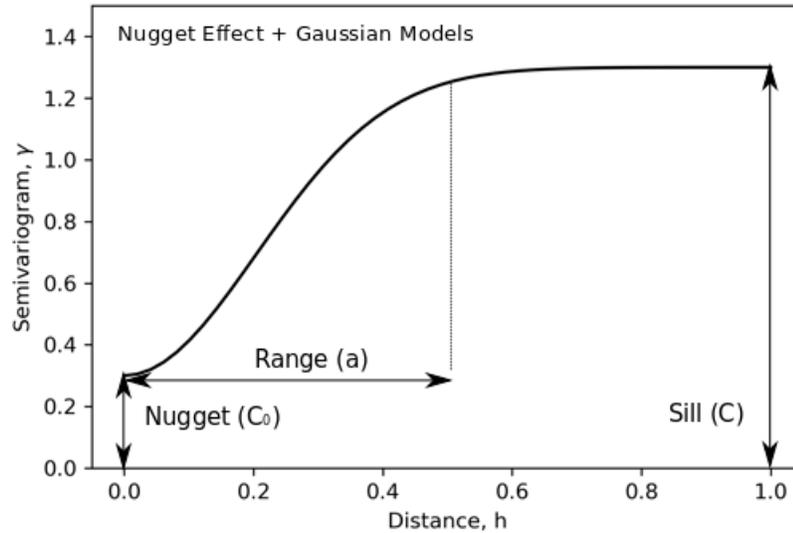
Figure 2.4: Variogram model made by summing a nugget effect and a gaussian models, along-side with the model parameters C_0, C and a, the value $C - C_0$ is often called *partial sill*.

In a soil deposit, layered or with vertical variation of the properties, it is expected that the rate of change of the properties would be greater in the vertical direction compared with the horizontal. This leads to the range to be shorter in the vertical than the one in the horizontal direction.

2.2 Kriging

Kriging is a generic name adopted by geostaticians for a set of generalized least-squares regression alogorithms accounting for data spatially related. Kriging estimators are defined (with some variations) as:

$$\hat{Z}(x) = \mu(x) + \sum_{i=1}^{n} \lambda_i [Z(x_i) - \mu(x_i)] \tag{2.18}$$

Where: $\hat{Z}(x)$ is the kriging estimator of $Z(x)$, λ_i is the weight associated to x_i, $\mu(x)$ and $\mu(x_i)$ are the expected values of $Z(x)$ and $Z(x_i)$, respectively; n is a subset of the sampled space, called neighborhood.

The error, thus, can be defined as the random variable $\hat{Z}(x) - Z(x)$. The goal of all kriging

variants is to minimize the error variance σ_E^2:

$$\sigma_E^2(x) = Var(\hat{Z}(x) - Z(x)) \tag{2.19}$$

under the constraint of unbiasedness, this is:

$$E(\hat{Z}(x) - Z(x)) = 0 \tag{2.20}$$

Depending on how the trend is modeled a different kriging variant is defined (Goovaerts, 1997):

1. Simple Kriging: considers the mean $\mu(x)$ to be known and constant throughout the study area.

2. Ordinary Kriging: considers an unknown and constant mean but takes into account its fluctuations by using a local neighborhood in the computations.

3. Universal Kriging: considers that the mean varies smoothly in the study area, and is modeled as a linear combination of functions of the coordinates.

The three variants of Kriging are summarized in the following sections, for a more detailed explanation of the mathematical foundations of them the reader is encouraged to check Olea (1999).

2.2.1   Simple Kriging

The Simple Kriging (SK) estimator of Z, a second order stationary random function with mean μ, at point x_0 is defined by the following equation,

$$\hat{Z}(x_0) = \mu + \sum_{i=1}^{n} \lambda_i (Z(x_i) - \mu) \tag{2.21}$$

The minimum mean square error for simple kriging is defined by:

$$\sigma_{SK}^2(x_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(x_0, x_i) \tag{2.22}$$

The solution with minimum variance will be the one in which the derivative of the variance with respect to the weights is zero:

$$\frac{\partial(\sigma_{SK}^2)}{\partial(\lambda_i)} = 0 \tag{2.23}$$

The solution of the system defined by the previous equations can be expressed in terms of matrices for simplicity.

The covariance matrix is defined as:

$$\mathbf{C} = \begin{pmatrix} C(x_1, x_1) & C(x_1, x_2) & ... & C(x_1, x_n) \\ C(x_2, x_1) & C(x_2, x_2) & ... & C(x_2, x_n) \\ ... & ... & ... & ... \\ C(x_n, x_1) & C(x_n, x_2) & ... & C(x_n, x_n) \end{pmatrix} \tag{2.24}$$

The weights vector (λ) and the covariance vector (c) are defined as following,

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \end{pmatrix} \tag{2.25} \qquad \mathbf{c} = \begin{pmatrix} C(x_0, x_1) \\ C(x_0, x_2) \\ ... \\ C(x_0, x_n) \end{pmatrix} \tag{2.26}$$

And, finally, the vector with the random variables is,

$$\mathbf{Y} = \begin{pmatrix} Z(x_1) - \mu \\ Z(x_2) - \mu \\ ... \\ Z(x_n) - \mu \end{pmatrix} \tag{2.27}$$

Then the optimum weights are found by solving the equation:

$$\mathbf{C} \cdot \lambda = \mathbf{c} \tag{2.28}$$

The simple kriging estimator of Z at the point x_0 is defined by:

$$\hat{Z}_{SK}(x_0) = \mu + \mathbf{Y}^T \lambda = \mu + \mathbf{Y}^T \mathbf{C}^{-1} \mathbf{c} \tag{2.29}$$

While the variance of the estimation is computed as:

$$\sigma^2_{SK}(x_0) = C(0) - \mathbf{c}^T \lambda = C(0) - \mathbf{c}^T \mathbf{C}^{-1} \mathbf{c} \tag{2.30}$$

2.2.2 Ordinary Kriging

Ordinary kriging is a improvement of simple kriging by discarding the requirement of previous knowledge of the random function's mean. In the formulation this entails a constrained optimization that is solved through Lagrange method of multipliers.

The ordinary kriging estimator is given by a linear combination of the sampled random variables at sites x_i:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{2.31}$$

with the constrain that $\sum_{i=1}^{n} \lambda_i = 1$ (this ensures unbiasedness of the estimator).

The estimation variance is defined by:

$$\sigma^2(x_0) = C(0) + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j) - 2 \sum_{i=1}^{n} \lambda_i C(x_i, x_0) \tag{2.32}$$

The optimal weight of the constrained optimization problem can be calculated using the

Lagrange method of multipliers through the use of a new objective function called Lagrangian function, defined by:

$$L = \sigma^2(x_0) + 2k(\sum_{i=1}^{n} \lambda_i - 1) \tag{2.33}$$

where k is called the Lagrange multiplier.

With the optimal weights, the minimum mean square error for ordinary kriging is:

$$\sigma_{OK}^2(x_0) = \sum_{i=1}^{n} \lambda_i C(x_0, x_i) - k \tag{2.34}$$

As in simple kriging, the algorithm is simpler using matrices. The covariance matrix is given by:

$$\mathbf{C} = \begin{pmatrix} C(x_1, x_1) & C(x_1, x_2) & ... & C(x_1, x_n) & 1 \\ C(x_2, x_1) & C(x_2, x_2) & ... & C(x_2, x_n) & 1 \\ ... & ... & ... & ... & ... \\ C(x_n, x_1) & C(x_n, x_2) & ... & C(x_n, x_n) & 1 \\ 1 & 1 & ... & 1 & 0 \end{pmatrix} \tag{2.35}$$

The weights vector (λ) and the covariance vector (c) are:

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \\ k \end{pmatrix} \tag{2.36} \qquad\qquad \mathbf{c} = \begin{pmatrix} C(x_0, x_1) \\ C(x_0, x_2) \\ ... \\ C(x_0, x_n) \\ 1 \end{pmatrix} \tag{2.37}$$

The system to be solved is:

$$\mathbf{C} \cdot \lambda = \mathbf{c} \tag{2.38}$$

The ordinary kriging estimate is defined by:

$$\hat{Z}_{OK}(x_0) = \mathbf{Z}^T \cdot \lambda = \mathbf{Z}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c} \tag{2.39}$$

The estimation variance is then defined by:

$$\sigma^2_{OK}(x_0) = C(0) - \mathbf{c} \cdot \lambda = C(0) - \mathbf{c}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c} \tag{2.40}$$

2.2.3 Universal Kriging

This kind of kriging is based on the hypothesis that the random function is not stationary and its expected value (called drift) varies through the space by a systematic trend, which in this case is assumed a polynomial function of the coordinates (Emery, 2016).

The estimator is the same as in ordinary kriging, shown in equation 2.31: $\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$.

The residual of the random function is defined as the difference between the random function and its expected value.

$$Y(x) = Z(x) - E(Z(x)) \tag{2.41}$$

The mean, also called the drift, is defined as a linear combination of k functions of the coordinates:

$$\mu(x) = E(Z(x)) = \sum_{i=0}^{k} a_i f_i(x), \text{ with } f_0(x) = 1 \tag{2.42}$$

Usually, the mean is modeled with a with a first- or second-order polynomial function as $\mu(x, y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 x y + a_5 y^2$.

A particular case of universal kriging when the mean is $\mu(x) = a_0 = \mu$ is ordinary kriging. Thus, universal kriging follows a methodology similar to the one shown for ordinary kriging.

The unbiasedness requirement is

$$\sum_{i=1}^{n} \lambda_i f_i(x_i) = f_j, \text{ for } j = 0,1,...,k \tag{2.43}$$

The minimum mean square error for universal kriging is given by,

$$\sigma_{UK}^2(x_0) = C_Y(0) - \sum_{i=1}^{n} \lambda_i C_Y(x_0, x_i) - \mu_0 - \sum_{j=1}^{k} \mu_j f_j(x_0) \tag{2.44}$$

where C_Y is the covariance function of the residuals $Y(x)$ instead of the random function $Z(x)$.

Defining the covariance matrix as:

$$\mathbf{C} = \begin{pmatrix}
C_Y(x_1, x_1) & \dots & C_Y(x_1, x_n) & 1 & f_1(x_1) & f_2(x_1) & \dots & f_k(x_1) \\
C_Y(x_2, x_1) & \dots & C_Y(x_2, x_n) & 1 & f_1(x_2) & f_2(x_2) & \dots & f_k(x_2) \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
C_Y(x_n, x_1) & \dots & C_Y(x_n, x_n) & 1 & f_1(x_n) & f_2(x_n) & \dots & f_k(x_n) \\
1 & \dots & 1 & 0 & 0 & 0 & \dots & 0 \\
f_1(x_1) & \dots & f_1(x_n) & 0 & 0 & 0 & \dots & 0 \\
f_2(x_1) & \dots & f_2(x_n) & 0 & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
f_k(x_1) & \dots & f_k(x_n) & 0 & 0 & 0 & \dots & 0
\end{pmatrix} \tag{2.45}$$

The weights vector (λ) and the covariance vector (c) are:

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ -k_0 \\ -k_1 \\ -k_2 \\ \dots \\ -k_k \end{pmatrix} \tag{2.46} \qquad\qquad \mathbf{c} = \begin{pmatrix} C_Y(x_0, x_1) \\ C_Y(x_0, x_2) \\ \dots \\ C_Y(x_0, x_n) \\ 1 \\ f_1(x_0) \\ f_2(x_0) \\ \dots \\ f_k(x_0) \end{pmatrix} \tag{2.47}$$

Additionally, the matrix Z whose elements are the random variables (n elements) and k+1 zeros.

$$
\mathbf{Z} = \begin{pmatrix} Z(x_1) \\ Z(x_2) \\ \ldots \\ Z(x_n) \\ 0 \\ 0 \\ \ldots \\ 0 \end{pmatrix}
\tag{2.48}
$$

As before, the system to be solved is:

$$
\mathbf{C} \cdot \lambda = \mathbf{c}
\tag{2.49}
$$

The universal kriging estimation is:

$$
\hat{Z}_{UK}(x_0) = \mathbf{Z}^T \cdot \lambda = \mathbf{Z}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c}
\tag{2.50}
$$

And the estimation variance is,

$$
\sigma^2_{UK}(x_0) = C_Y(0) - \mathbf{c}^T \cdot \lambda = C_Y(0) - \mathbf{c}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{c}
\tag{2.51}
$$

Universal and ordinary kriging are both minimum mean square error, unbiassed, exact interpolators that automatically corrects for clustering in the sampling (Olea, 1999).

Two drawbacks for this method would be that is necessary to specify a trend model for the random function, and besides, a variogram for the residuals (which is not the same as the variogram of the data minus the trend as stated in Goovaerts (1997)) and not for the random function as before. According to Emery (2019) it is not advisable to use Universal Kriging when it is not sure that the drift can be modeled as linear combination of the coordinates, in that case Ordinary kriging and constant in the neighborhood mean would be better choice.

2.2.4 Search Neighborhood

When performing kriging, if all the measured data is used, the covariance matrix (**C**) is not changed by changing the location of the estimation point. On the other hand, the larger the sampled space, the larger the matrix to invert, and in that way, the longer the solution will take. As the distance between the sampled and unsampled point increases, the weight associated tends to be zero, since covariance tend to zero as well. The latter can imply instability of the system to solve (Davis and Morris, 1997).

The reduction of the sampling space can be performed in different ways, for example defining maximum radius and select a certain amount of neighbors by quadrant. If there is a preferential search direction, an elliptical search would be a clever choice. In this project, the nearest sampling points are chosen given a tolerance distance and a circular or elliptical search is made depending on the anisotropy of the site in study. Also the search area is divided in four quadrants with fixed number of points to be considered in each of them, in order to distribute in a better way the information around the point to be estimated (Emery, 2016), this can be seen in Figure 2.5.
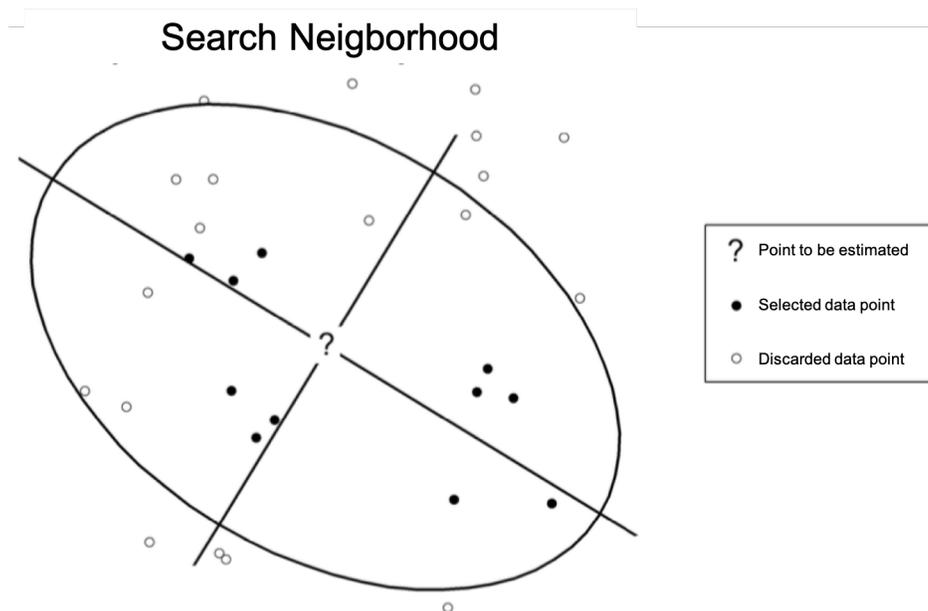
Figure 2.5: Search neighborhood, with elliptical shape and search by quadrant (modified from Emery (2016)).

2.2.5 Anisotropy

Anisotropy can be assessed by inspecting the experimental variogram in different directions. The variogram is considered isotropic if changing the direction does not produce significant changes on it. A common case is the geometric anisotropy, shown in Figure 2.6, where the range varies with the direction of the variogram. This kind of anisotropy can be addressed by rescaling the sampling space in order for the range to become constant.
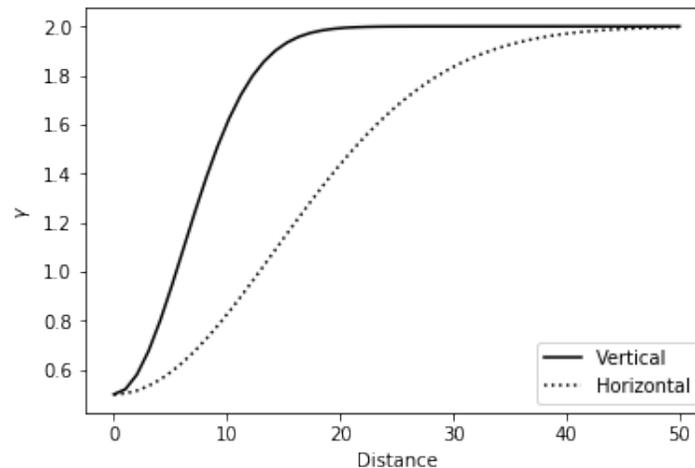
Figure 2.6: Geometric anisotropy, the ranges in these variograms are 15 m in the vertical direction and 35 m in the horizontal.

Another kind of anisotropy is zonal anisotropy which is not covered in this thesis but Goovaerts (1997) offers a comprehensive review of this and other interesting variogram features.

2.2.6 Cross-validation

Cross-validation is a tool for testing interpolation methods without requiring additional data. The methodology for cross-validation consists in leave out a measured value or a subset of measured values, and use the rest of them to predict it with the same kriging model that the one used to interpolate the unsampled data. The error is computed by comparing the measured and estimated values.

Emery (2016) recommends some graphical cues for a better understanding of the cross-validation results, the first being the correlation cloud between the actual and the

estimated values. Additionally, he recommends to plot a histogram of the standardized errors, a good estimation is considered if it is centered around 0, with the mean of the standardized error being close to 0. The standardized error is given by the following equation:

$$SE(x_i) = \frac{\hat{Z}(x_i) - Z(x_i)}{\hat{\sigma}(x_i)} \tag{2.52}$$

According to Olea (1999) cross-validation is the only way to indirectly test kriging, despite it does not indicate if the interpolation is correct, it can help draw conclusions by comparison. In this thesis, the cross-validation methodology to be used will be the "leave one group out" (Scikit-learn developers, 2019) in which an entire group (subset) of data is left out in order to make the prediction, in this work a group is represented by a different CPTu test.

Chapter 3

Machine Learning

Machine learning is a series of statistical and computational techniques to allow a computer to learn from experience (data) without relying on a predetermined equation as a model. The algorithms improve their performance as the number of samples available for learning increases. Machine learning algorithms find patterns in data that generate insight and help the user to make better decisions and predictions. They are used every day to make critical decisions in medicine, business, engineering and many other areas of knowledge.

3.1   Types of Machine Learning

Machine Learning methods can be classified in different types depending on the application. One of the most used criteria to differentiate them is if it is supervised or unsupervised. In supervised machine learning, the training data which feed the algorithm include the desired solutions, called labels or class variable. Two typical supervised tasks are regression and classification. Regression predicts continuous responses, for example, the correlation between SPT blow count and soil density as studied by Puri et al. (2018); classification on the other hand, predicts discrete responses, for example, if a soil is susceptible to liquefaction under cyclic loading based on the SPT test, as the work done by Samui and Sitharam (2011).

In unsupervised machine learning, on the other hand, the system do not require the target

solution, so it has to be able of recognize patterns in the data itself. Clustering is a widely used unsupervised machine learning method. For example, an application of clustering can be the zoning of earthquake damage given the peak ground acceleration, lithology and topographic zone as done by Rabarijoely et al. (2007).

The focus of the application of machine learning techniques to geotechnical engineering issued in this work will be the classification of soils using CPTu test. Thereby the methods to be studied are basically classification methods such as Logistic Regression and Naive Bayes classificators, and Hidden Markov Models applied to supervised classification. Nevertheless, towards the end of this document a small clustering exercise is performed given the ability of Hidden Markov Model to perform such kind of tasks.

3.2 The Classification Problem

According to Asiri (2018): *"Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y)".* In general, classificators build a model based on training data before getting the data to be classified.

Classification in a probabilistic framework consists in quantify the probability that a sample to belong to a class (Y) given a set of observations (X), in mathematical terms this will be $P(Y|X)$ (Flach, 2012), where Y is the variable defining the classes (targets) and X is a set of observations. In geotechnical terms, this can be the probability that a soil sample classifies as "sand" given that its angle of internal friction is 28°, $P(Y = "sand"|X = \phi = 28°)$.

3.2.1 Classification metric: The accuracy score

There are many metrics to evaluate the classification results (Scikit-learn developers, 2019), one of the simplest is the one called Accuracy Score. In a binary classification context (being the class labels 1 and 0), there are four possible outcomes: True Positive (TP), True Negative

(TN), False Positive (FP) and False Negative (FN), representing a observation correctly put into class 1, correctly put into class 0, incorrectly put into class 1 and incorrectly put into class 0, respectively. The accuracy score will be in this case (Marsland, 2011):

$$Accuracy\ Score = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

In a multiclass classification, the essence of the score is the same but counting all the samples where the classification matches the target value over the total number of samples classified. Despite there are many improved metrics for multiclass classification (Mosley, 2013), the accuracy score will be used in this work to evaluate the classification results due to its simplicity.

3.3 Logistic Regression Classification

Logistic regression is a linear model for classification rather than a regression, it is commonly used to estimate the probability that an instance belongs to a particular class (for example, the probability that soil is quick clay or not). If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled "1"), or else it predicts that it does not (it belongs to the negative class, labeled "0"). This makes it a binary classifier.

In a binary classification, the probability that a set of observations X belongs to a class Y (=1) is defined by:

$$h_\theta(X) = \frac{1}{1 + exp(-(\theta^T X))} = P(Y = 1 | X; \theta) \tag{3.2}$$

Assuming that the observations are independently Bernoulli distributed, the likelihood function is given by:

$$L(\theta|x) = P(Y|X; \theta) = \prod P(y_i | x_i; \theta) = \prod h_\theta(x_i)^{y_i} (1 - h_\theta(x_i))^{(1 - y_i)} \tag{3.3}$$

The likelihood, or frequently the logarithm of it, is maximized using an optimization technique such as gradient descent (Wikipedia contributors, 2019a).

When the data is multinomial, as in geotechnical classification problems where a soil can be part of one of many soil classes, like clay, silt, sand, gravel, and so on; the formulation changes as described below.

With a given set of input parameters X, the multinomial logistic regression, first compute a score $s_k(X)$ for each class k, then estimates the probability of each class by applying a normalized exponential function (also called Softmax function) to the score. The equation to compute the scores is just a equation for linear regression prediction: $s_k(X) = \theta_k^T \cdot X$. The probability $P(Y = k|X;\theta)$ that the input belongs to a class k is given by:

$$P(Y = k|X;\theta) = \frac{exp(s_k(X))}{\sum_{j=1}^{K} exp(s_j(X))} \tag{3.4}$$

Being K the total number of classes.

The optimal parameter vector θ is determined by a maximum a posteriori estimation (MAP), which is an extension of maximum likelihood method mentioned above. The problem to solve is shown in the equation below, and is solved by using an iterative optimization procedure such as gradient descent (Géron, 2017).

$$\hat{y} = \arg\max_k P(Y = k|X;\theta) = \arg\max_k s_k(x) = \arg\max_k (\theta_k^T \cdot x) \tag{3.5}$$

3.4 Naive Bayes Classification

Naive Bayes is a supervised learning algorithm based on applying Bayes theorem with the assumption of conditional independence between every pair of features given the value of the class variable. Bayes theorem states the following relationship, given class variable Y and dependent feature vectors X_i:

$$P(Y|X_1,...,X_n) = \frac{P(Y)P(X_1,...,X_n|Y)}{P(X_1,...,X_n)} \qquad (3.6)$$

The conditional independence assumes that $P(X_i|Y, X_1,...,X_{i-1}, X_{i+1},...,X_n) = P(X_i|Y)$, so the relationship is simplified to:

$$P(Y|X_1,...,X_n) = \frac{P(Y)\prod_{i=1}^{n}P(X_i|Y)}{P(X_1,...,X_n)} \qquad (3.7)$$

The equation above represents the "naive" part of the classifier, since the conditional independence assumption is rarely true in most real-world applications, however, despite this simplification, its performance is still satisfactory compared with other classifiers (Zhang, 2004).

The *a priori* probability of the classes, $P(Y)$, is estimated from the training set, while an assumption is required regarding the distribution of the likelihood function, $P(X_i|Y)$. In a Gaussian Naive Bayes Classifier, the likelihood is assumed to be Gaussian.

$$P(X_i|Y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}}exp\left(-\frac{(X_i-\mu_Y)^2}{2\sigma_Y^2}\right) \qquad (3.8)$$

The parameters of the likelihood function (σ_Y and μ_Y) are estimated by maximum likelihood. The value $P(X_1,...,X_n)$ is constant given the input. The classification problem is solved using a maximum a posteriori (MAP) estimation (Scikit-learn developers, 2019), which means:

$$\hat{Y} = \underset{Y}{\arg\max}\, P(Y)\prod_{i=1}^{n}P(X_i|Y) \qquad (3.9)$$

3.5   Hidden Markov Models

Consider a system that can be described at any depth (or time) as being in one of N different states S_i. At regularly space depth intervals the system can change the state (possibly back to the same state) according to a set of probabilities associated with each state. The depths (or

times) associated with the state changes are denoted as t, the actual state at certain depth is denoted as q_t. Considering a first order Markov chain, the system can be described probabilistically as only depending on just the current and the previous state, according to the next expression:

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j | q_{t-1} = S_i) = p_{ij} \tag{3.10}$$

p_{ij} are called the transition probabilities and define the transition matrix P, defining the probability of being in a state given the state in the previous depth. In other words, the transition matrix is defined by: $P = p_{ij}$ with the rows (i) being the previous state and columns (j) the current state. Both the columns and the rows must sum 1. The size of P is defined by the total number of different states in the problem to model.

The initial probabilities $\pi_i = P(q_1 = S_i)$ are the probabilities that the first value of the sequence is defined by the state i.

If the states are not observable (hidden) and can only be observed through another set of stochastic processes that produce a sequence of observations (O), it is said that the model is a Hidden Markov Model. In this thesis context, the hidden states are the soil classes (if a soil is clayey or sandy, etc.), the observations are any combination of the measurements made by the CPTu test (tip resistance, side friction, pore pressure, or the normalization of them).

Summing up a Hidden Markov Model is completely characterized by: the number of states (N), the observations (O), the transition matrix (P), the starting probabilities (π_i) and the probability distribution of the observations. In the case of a Normal distribution, a mean matrix is required μ with the expected value of each observation variable for each state, and a variance matrix Σ that can be the same for each state, in which case it is said that the model has tied covariance, or a different matrix for each state defining a full covariance model, this influences the relative orientation of the likelihood functions for the different classes, as can be seen in Figure 3.1 below.
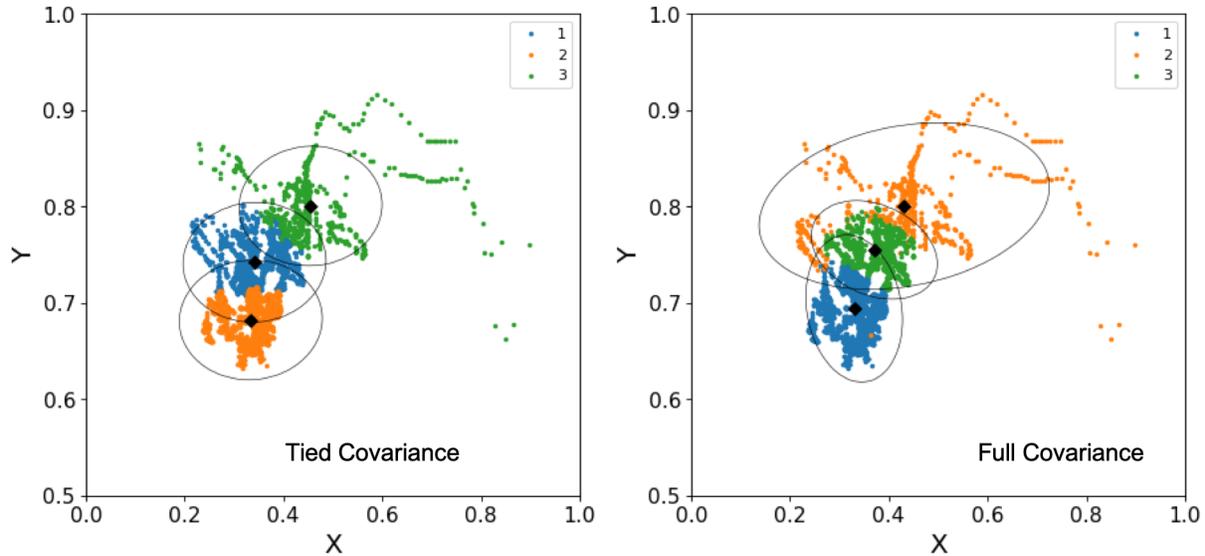
Figure 3.1: Likelihood contour of two Gaussian distributions with tied (left) and full (right) covariance model.

Concerning HMM there are three kind of problems that must be solved in order for the model to be applicable:

1. Given the observation sequence and a model, compute the probability of observations given the model $P(O|Model)$. This is solved by the Forward-Backward algorithm (Baum and Eagon, 1967).

2. Given the observation sequence and a model, choose the corresponding state sequence which is optimal. This is solved with the help of the Viterbi algorithm (Forney, 1973).

3. Given the observations, estimate the model parameters (train the model). This problem is solved iteratively by the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

All these algorithms are part of the Python library hmmlearn (hmmlearn developers, 2010) which was used in this work to solve the problems modeled through a Hidden Markov Model. For a better understanding of the algorithms and the theory behind HMM, it is recommended to read the work of Rabiner (1989) on which was based this section and that presents a well explained summary of applied Hidden Markov Models.

Chapter 4

Site Characterization using Ordinary Kriging

Rissa is a town located in the Indre Fosen municipality in the Trøndelag county. This site became famous after a huge quick clay landslide in 1978 (Gregersen, 1981). After this event efforts have been made in order to characterize the site and assess the stability of its slopes. In that framework a comprehensive study was made by Kornbrekke (2012) in which field and laboratory tests were gathered and processed to characterize the properties of a site near Rein church, in particular the profile called 3-3. This chapter presents the results of applying Ordinary Kriging to the data from CPTu testing for the continuous interpolation of the three parameters measured by CPTu test: tip resistance (q_t), sleeve friction (f_s) and excess pore pressure (Δu_2) in the slope.

All the computations performed within this chapter where done using Python (Van Rossum, 1995) as programming environment and through custom made scripts and functions for data handling, variogram determination and fitting, kriging interpolation, cross-validation and, visualization and plotting.

4.1 Study Profile and Test Data

For the present investigation, a set of five CPTu performed in Rein church's slope were used whose names are: C2, C3, C4, C5 and C6. The locations of these CPTu tests are shown in Figure 4.1 as well as a segment of the profile 3-3 studied in Kornbrekke (2012).

For a cross-section construction, a projection of each test was performed into the profile 3-3, the criteria used to do this was to project them to the same elevation. The resulting cross-section is presented in Figure 4.2 shown below.
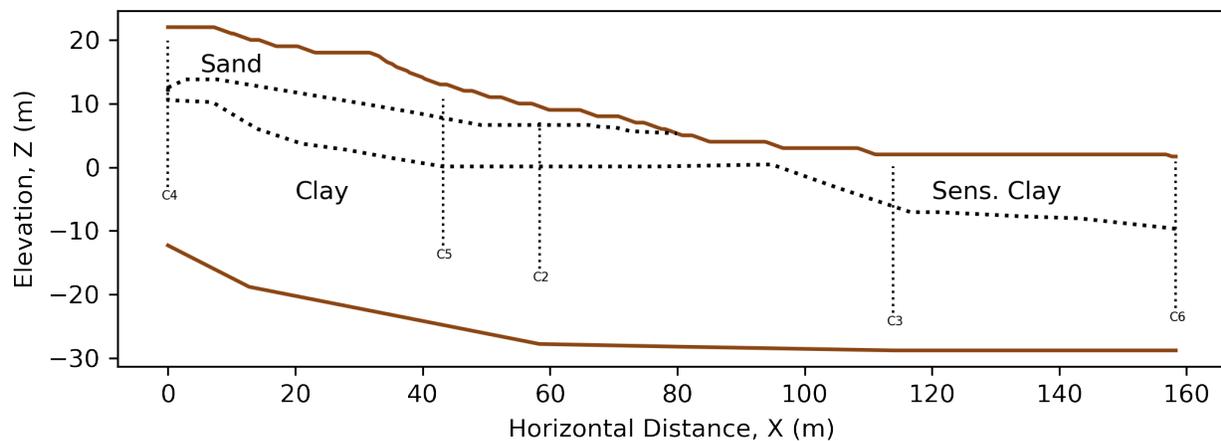
Figure 4.2: Cross-section of profile 3-3 with the projected CPTu tests and the layering proposed by Kornbrekke (2012)

The CPTu data was filtered in order to reduce the total amount of data points, and smoothed with a rolling mean (Wikipedia contributors, 2019b) to get rid of unrepresentative peaks. Figure 4.3 shows the results from the CPTu test projected on the study profile.

Figure 4.1: Plan view of the CPTu testing setup and the projection on profile 3-3.
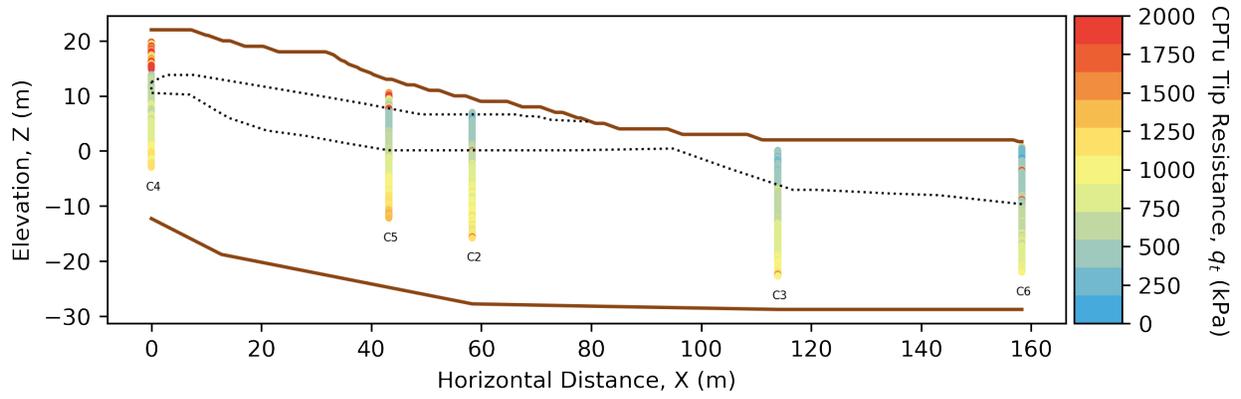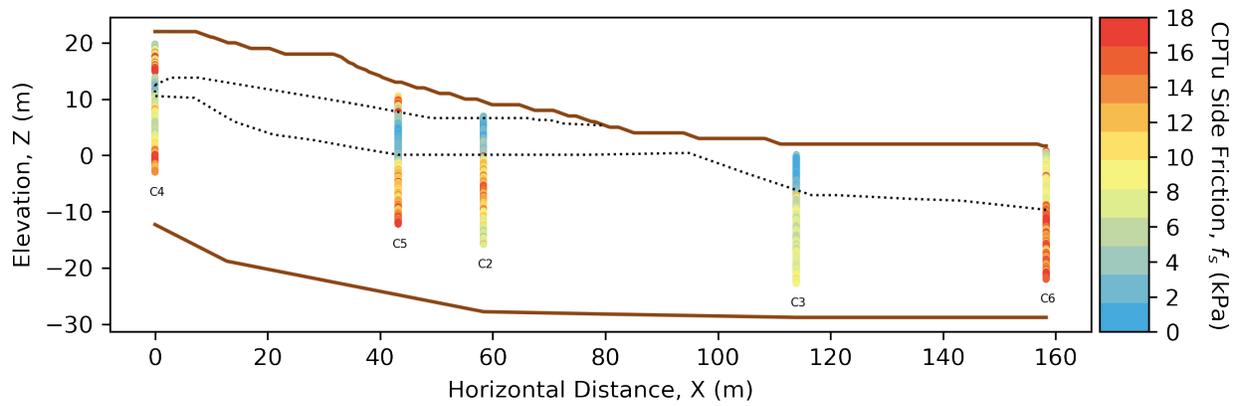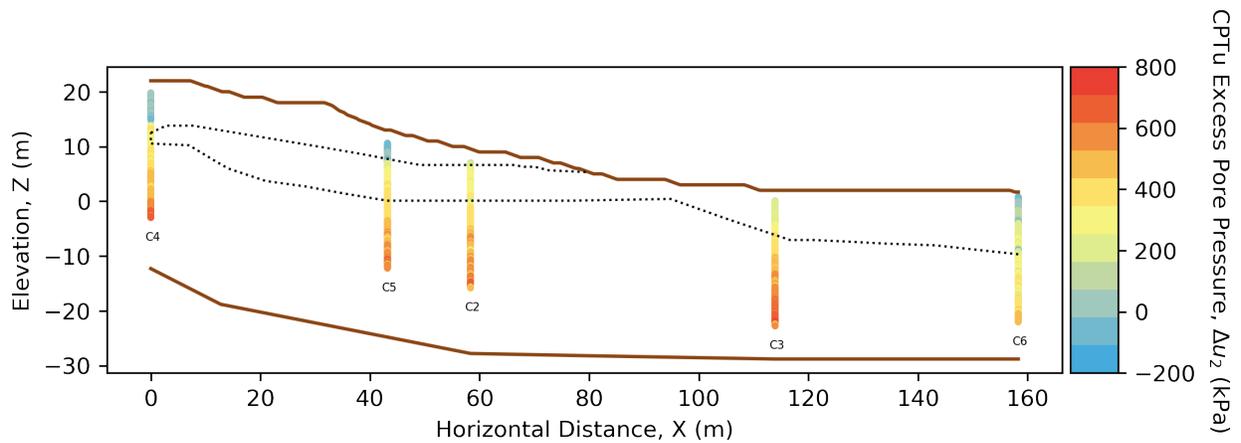
Figure 4.3: CPTu measurements projected on the study profile for: Tip Resistance (a), Sleeve Friction (b) and Excess Pore Pressure (c).

4.2   Variogram Models

For the variogram computation, since the data is not aligned in the X, Z plane shown in Figure 4.2, a three dimensional approach should be used to compute the distance in order to account for the right variogram's distances between each sampling point.

The experimental variograms were fitted using a nested Nugget plus Spherical model by recommendation of Emery (2019). The maximum lag distance for the computation was one large enough to express the total variance of the data.

To fit the horizontal variogram, since there are less data points per lag in that direction compared to the vertical, the result is more scattered. If only geometric anisotropy is assumed, then the only different parameter between vertical and horizontal variograms would be the range. Thus, the horizontal variogram was fitted keeping constant the sill and nugget from the vertical.

4.2.1   Tip resistance variograms

The experimental variograms for tip resistance are shown in Figure 4.4. The fitting parameters are:

Nugget (C_0)	$0\ kPa^2$
Partial Sill $(C - C_0)$	$251038.77\ kPa^2$
Vertical Range (a_V)	$2\ m$
Horizontal Range (a_H)	$170\ m$
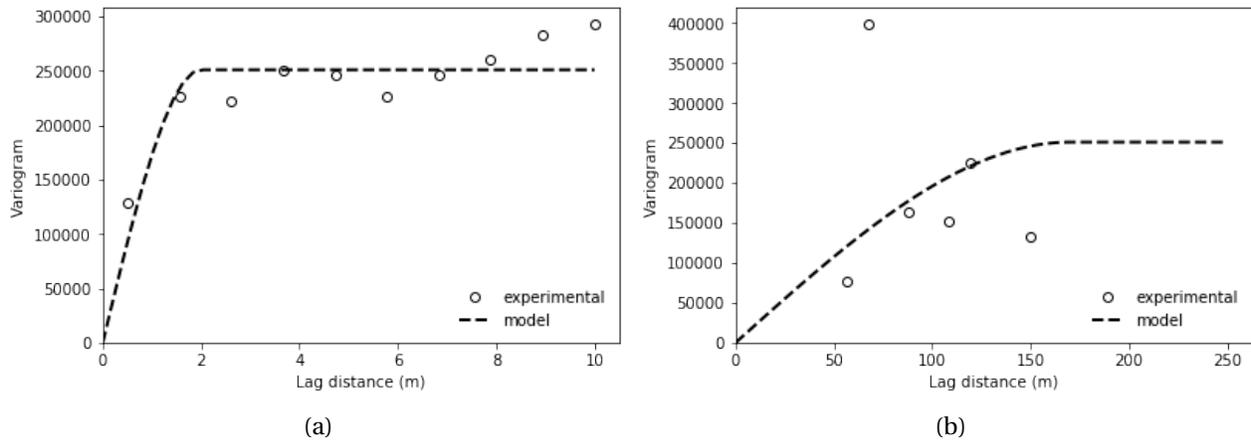
Figure 4.4: Vertical (a) and Horizontal (b) variograms for tip resistance q_t.

4.2.2 Sleeve friction variograms

The experimental variograms for sleeve friction are shown in Figure 4.5. The fitting parameters are:

Nugget (C_0)	14.91 kPa^2
Partial Sill $(C - C_0)$	23.09 kPa^2
Vertical Range (a_V)	9.52 m
Horizontal Range (a_H)	128.5 m
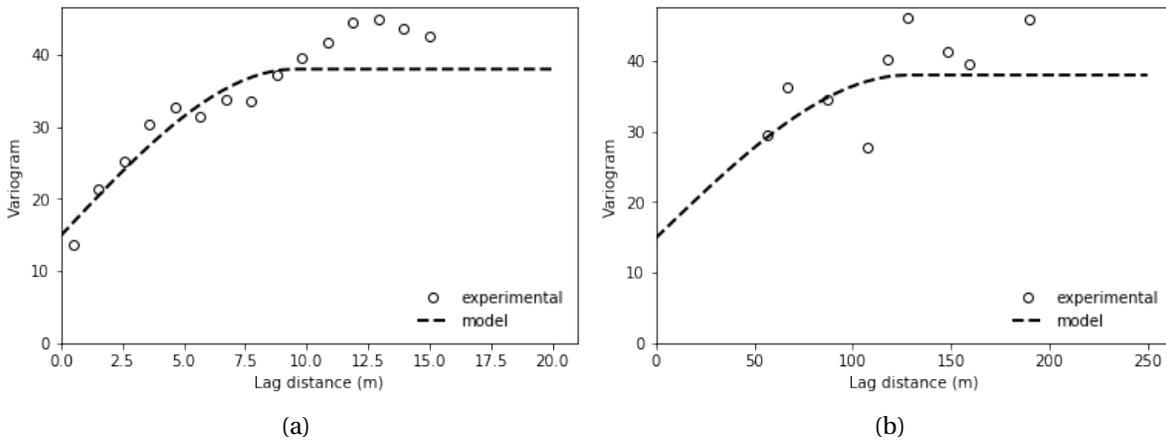
Figure 4.5: Vertical (a) and Horizontal (b) variograms for sleeve friction f_s.

4.2.3 Excess pore pressure variograms

In the case of excess pore pressure behind the cone tip Δu_2, the variogram showed a quadratic trend which indicates a linear trend in the modeled variable (because variograms are calculated with squared values), this can be seen in Figure 4.6.
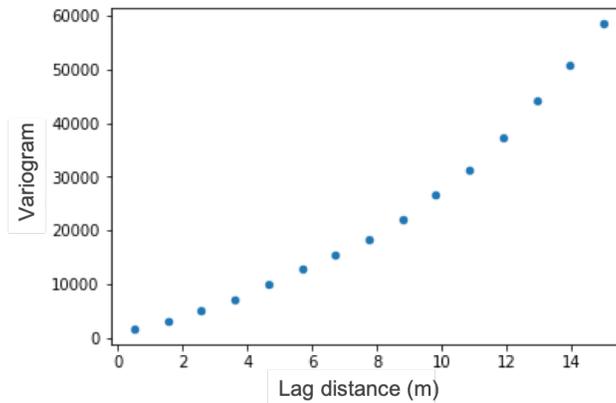
Figure 4.6: Linear trend on Δu_2 expressed as a quadratic trend in the variogram.

For a better calculation of the variogram, and according to the reccomendations of Goovaerts (1997), the variogram should be computed using the residuals (the values minus the trend) then the kriging interpolation is performed for them. After the interpolated values are computed, the trend is added to get the values of Δu_2. The experimental variograms for the residuals of excess pore pressure are shown in Figure 4.7. The fitting parameters are:

Nugget (C_0)	1119.9 kPa^2
Partial Sill $(C - C_0)$	3460.1 kPa^2
Vertical Range (a_V)	5.9 m
Horizontal Range (a_H)	80 m
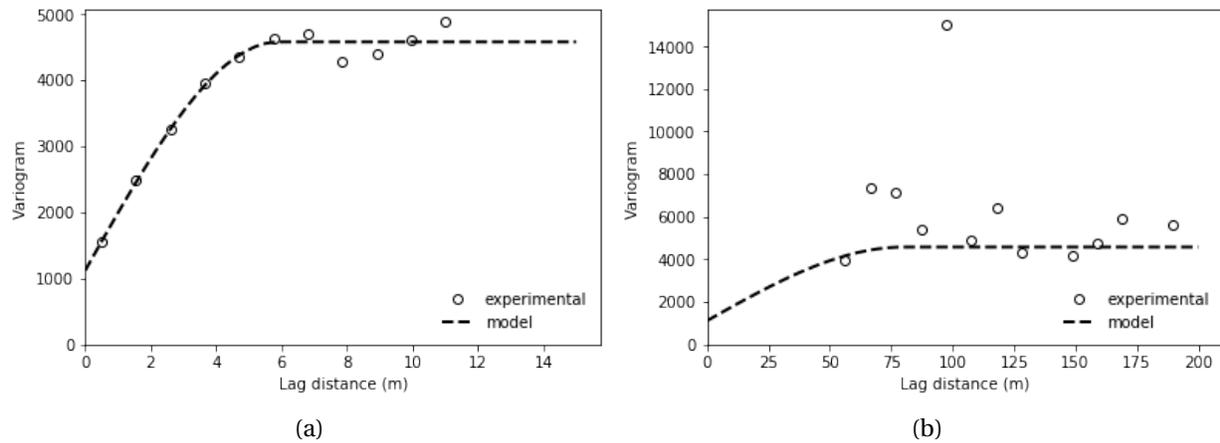
Figure 4.7: Vertical (a) and Horizontal (b) variograms for the residuals of excess pore pressure behind the cone tip Δu_2.

4.2.4 Conclusions about the variogram models

Constraining the variograms with the total variance allowed to get a good fit. Having the vertical variogram fitted allowed to fit the horizontal variogram's range. The results presented in Figures 4.4 to 4.7 show that all the variogram were successfully fitted considering the great variability on the data, and allow to perform kriging interpolation of the parameters over the entire area as it will be shown in the next section.

The values found in literature for vertical and horizontal range (expressed as correlation length) are smaller than the ones proposed here, for example, Firouzianbandpey et al. (2014) estimated values of around 0.5 m for the vertical and 2.0 m for the horizontal; Stuedlein et al. (2012) estimated values of less than 1.2 m and less than 10 m for vertical and horizontal respectively; Uzielli et al. (2005) found values ranging between 0.16 m to 1.11 m for the vertical range. Fenton (1999) explains that the different values on the correlation length can be due to different domain of the model, this means that larger models might have larger correlation length compared to smaller ones.

4.3   Ordinary Kriging Interpolation

The interpolation mesh was spaced 0.6 m and 0.7 m in the vertical and horizontal direction, respectively, giving a mesh of 11645 points. The search neighborhood in all cases is elliptical, with anisotropies (horizontal range over vertical range, $\frac{a_H}{a_V}$) of 85 for q_t and 14 for f_s and Δu_2, and a maximum search distance of 10 m and 10 neighbours per quadrant.

The results of the ordinary kriging of the CPTu tests at Rissa are shown in Figure 4.8. As a reference, the average runtime of the interpolation of the 11645 points comprising the mesh is around 4.3 minutes (260 seconds). The results show the ability of the kriging method to smoothly interpolate between the data points and to extrapolate outside the area defined by the soundings. It is also evident that the anisotropy is reflected in the shape of the contours of the estimated parameters.

The variance plots in Figure 4.9 show, as it would be to expect, a higher variance on the extrapolation zones (where there is no actual data) and is minimal (equals to the nugget of the variogram) in the CPTu lines, and in between them the variance increases with the distance.

To assess the goodness of the estimation, since all the measured data was used for the interpolation, it is interesting to compare the results of the interpolation with the layering proposed by Kornbrekke (2012), which was defined using field and laboratory testing. The results are shown in Figure 4.10. It is possible to see a good match between the layering and the changes in tip resistance predicted by the interpolation, and specially in the sleeve friction map (which can be related with the remoulded resistance of the soil) where lower values are reached within the sensitive clay area. The excess pore pressure agrees with what is to be expected, especially for the sand layer where the lowest (even negative) pressure values are reached.
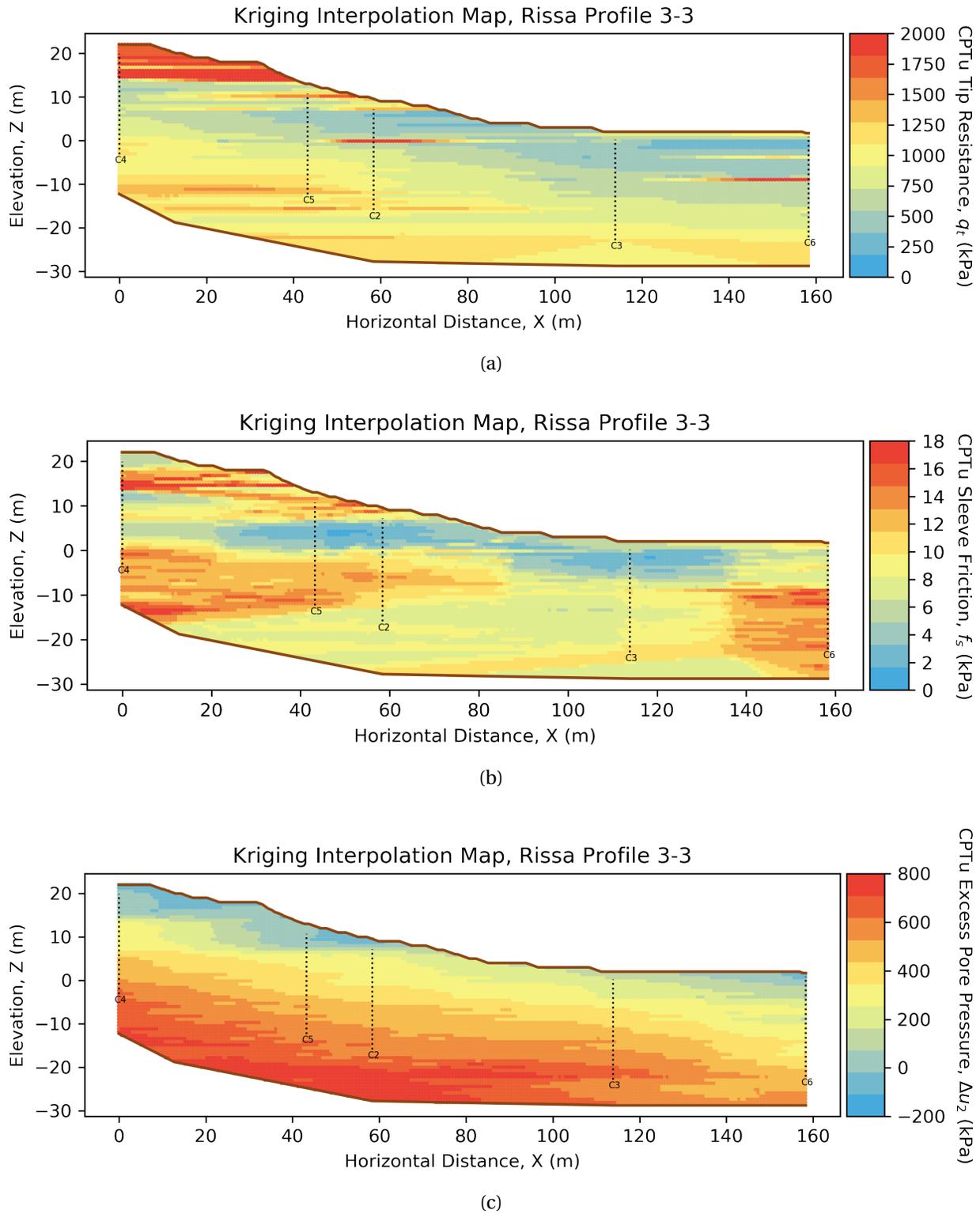
Figure 4.8: Kriging interpolation maps for: Tip Resistance (a), Sleeve Friction (b) and Excess Pore Pressure (c).
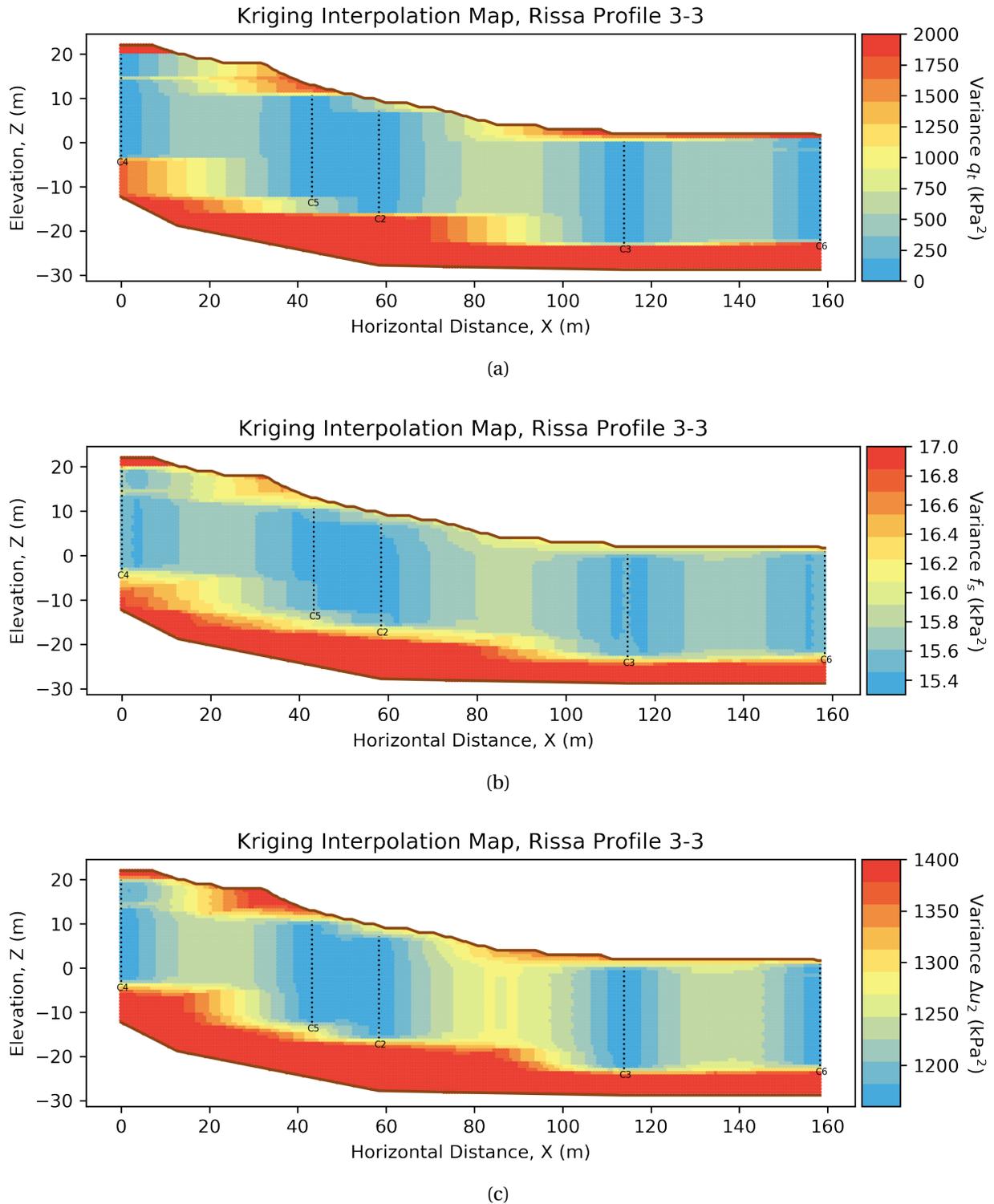
Figure 4.9: Kriging variance maps for: Tip Resistance (a), Sleeve Friction (b) and Excess Pore Pressure (c).
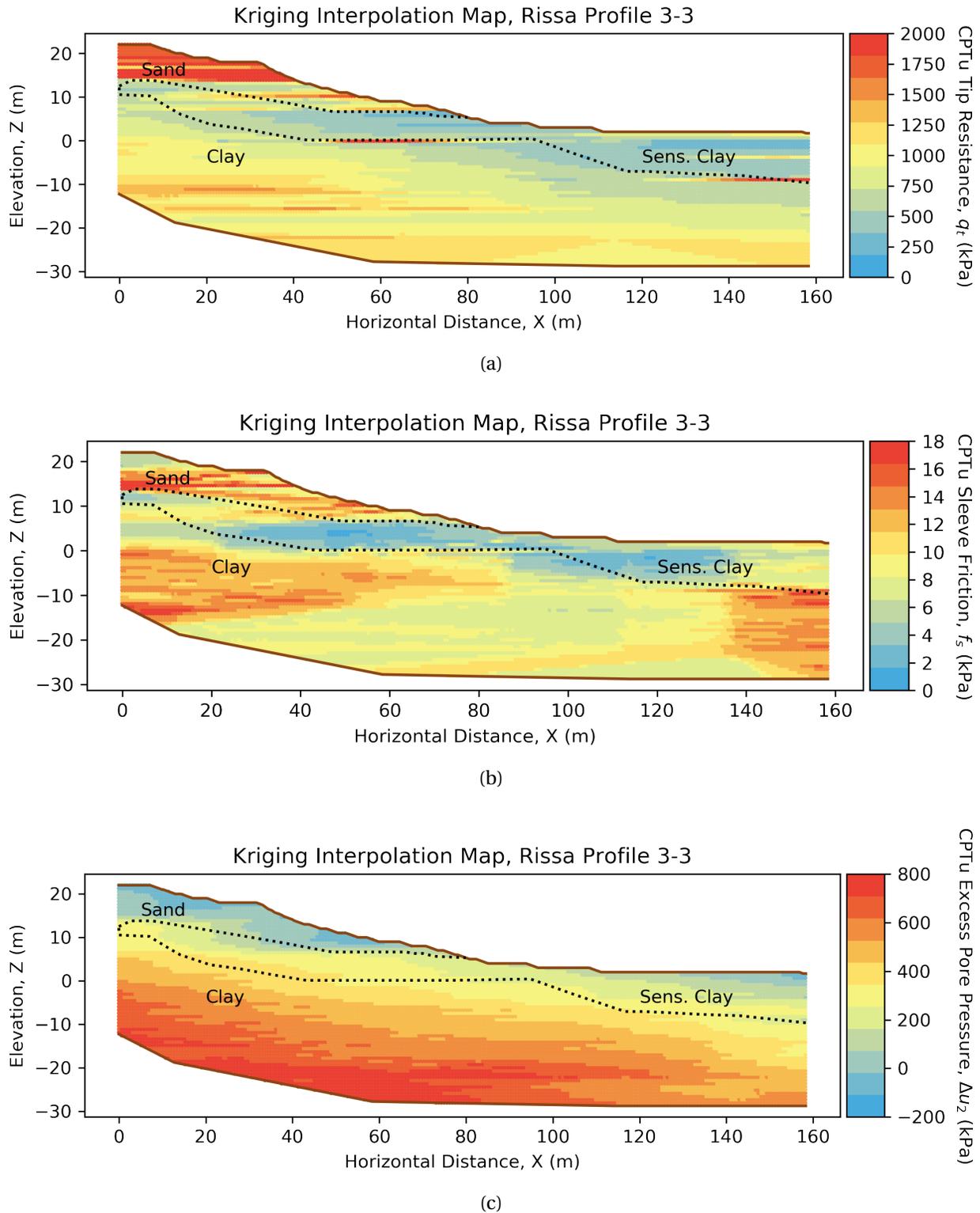
Figure 4.10: Kriging interpolation maps for: Tip Resistance (a), Sleeve Friction (b) and Excess Pore Pressure (c), alongside with the layering proposed by Kornbrekke (2012).

As mentioned before, it is difficult to quantify the goodness of fit of a geostatistical interpolation without performing new sampling, but with the cross-validation technique a sense of the performance of the estimation model can be obtained, as well as a grasp of the influence of each test in the overall estimation.

The cross-validation was performed by removing a complete CPTu test from the data available for the interpolation, and estimating the values at its location using the remaining data. The results are measured in terms of the standardized error explained in Chapter 2. In this case, the CPTu test at the profile borders are always kept since its influence constraining the solution is vital, and also because the main focus of this methodology is to interpolate rather than extrapolate values.

The results are shown in Figures 4.11 and 4.12, the associated standardized errors means of the estimations are -0.47, 0.45 and -0.18 for q_t, f_s and Δu_2 (residuals), respectively, the histograms of the errors show that they are centered around 0. It can be noticed a good fit for the majority of the values, however, the outliers present in the tests, are not accounted by the kriging interpolation. Nevertheless, these outliers representing the peak values which are typical of CPTu tests, are seldom used with design purposes because they do not represent the overall behavior of the soil but minor discontinuities or testing flaws.

Regarding the influence of leaving a CPTu test out, it is possible to see that C3 is the less influential one, and it is thought that this is because of the lack of outliers on it. While, leaving C2 and C5 produce a bigger dispersion between the measured and the estimated values, mainly at the outliers.
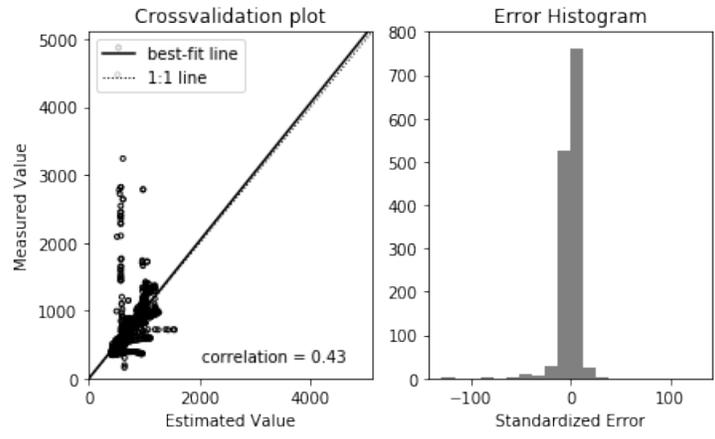
(a)
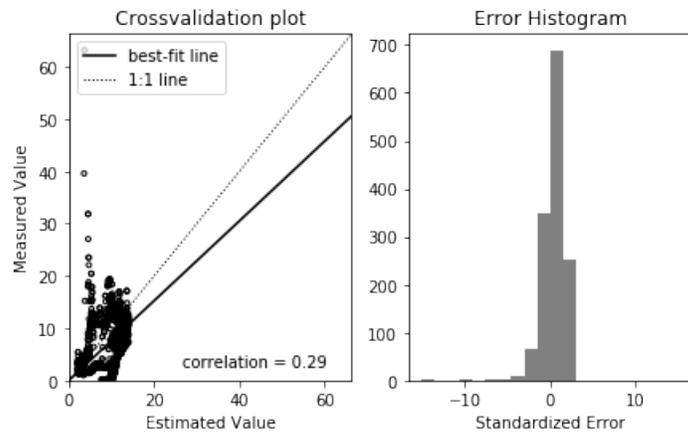
(b)
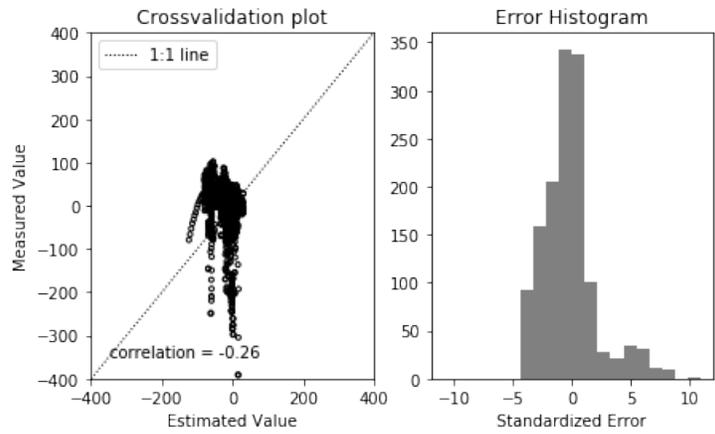
(c)

Figure 4.11: Cross-validation plots, measured vs estimated scatter plot (left) and standardized error histogram (right), for tip resistance (a), sleeve friction (b) and excess pore pressure (residuals) (c).
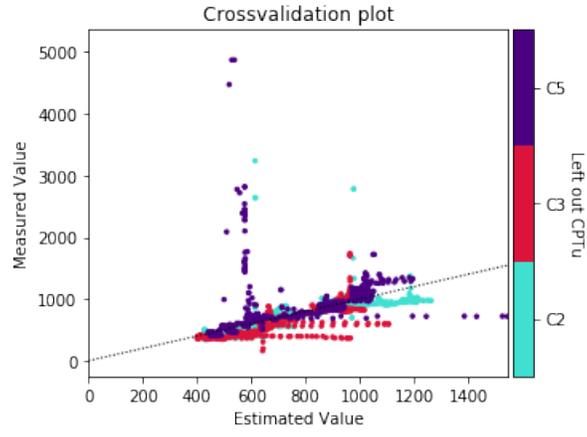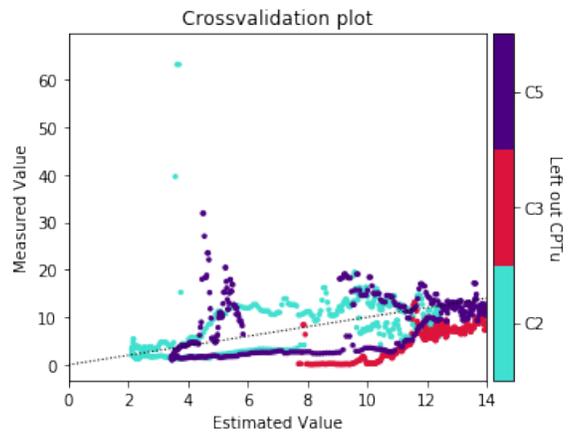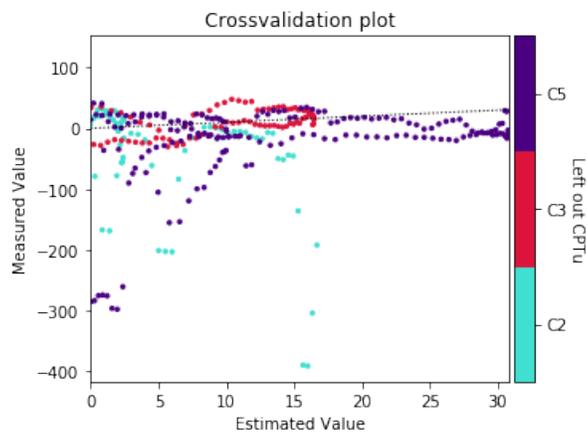
Figure 4.12: Cross-validation plots, measured vs estimated scatter plot grouped by the left out CPTu test, for tip resistance (a), sleeve friction (b) and excess pore pressure (residuals) (c).

Chapter 5

Site Characterization using Machine Learning

In Norway, a big concern in any construction project is the presence of highly sensitive or quick clays, since they can easily determine the feasibility of them. Since CPTu tests are widespread and are present in almost every geotechnical exploration program in Norway, it would be convenient to have a way to determine whether a soil profile contains quick clays or not based on the CPTu test results. Classification charts based on CPTu tests are meant to be global, indicating general behavior rather than specific soil characteristics.

In this context, the use of machine learning approaches is ideal since local data can be used to help a model to "learn" which measured data characterizes a certain kind of soil, highly sensitive or quick clays in this case. With few information the results will not be satisfactory, but as the exploration advances, the model will learn from the newly generated data and adjust to give better results.

The following chapter studies how machine learning techniques can be used to improve the identification of highly sensitive and quick clay soils from CPTu tests.

All the computations performed within this chapter were done using Python (Van Rossum, 1995) as programming environment, complemented by Scikit-Learn machine learning library (Pedregosa et al., 2011) and hmmlearn Hidden Markov Model library (hmmlearn developers,

2010), alongside with custom made scripts and functions for data handling, visualization and plotting.

5.1 Datasets

This chapter describe the datasets used to test the machine learning approach with real data from one investigation project (NTGS) and one engineering challenge faced by Statens Vegvesen, both involving the presence of quick and highly sensitive clays.

5.1.1 NGTS dataset

NGTS stands for Norwegian Geo-Test Sites and is a research consortium led by the Norwegian Geotechnical Institute (NGI) and with the participation of NTNU and other organizations. Its main focus is to develop field laboratories for testing, verification and control of new methods and equipment for site investigations and foundation engineering (NGI, 2019). Within the NGTS framework, one of the important study subjects are quick clays, being the site at Tiller chosen for that matter which is located near Trondheim. Figure 5.1 presents the location of the CPTu tests, while Figure 5.2 shows the summary of the tests alongside with the layering of the site. For the present work, 31 CPTu test were used (CPTu test C18 was discarded because of high sleeve friction which was not representative of the site). Detailed plots of each test can be found in Appendix A

The layering of the site consists of 2.5 m of dry crust, followed by a clay layer up to a depth of 7.5 m on top of a quick clay layer 12.5 m thick. The water table is at 1.5 m from the surface. The terrain is flat so the features described above are expected to have few variations over the study area.
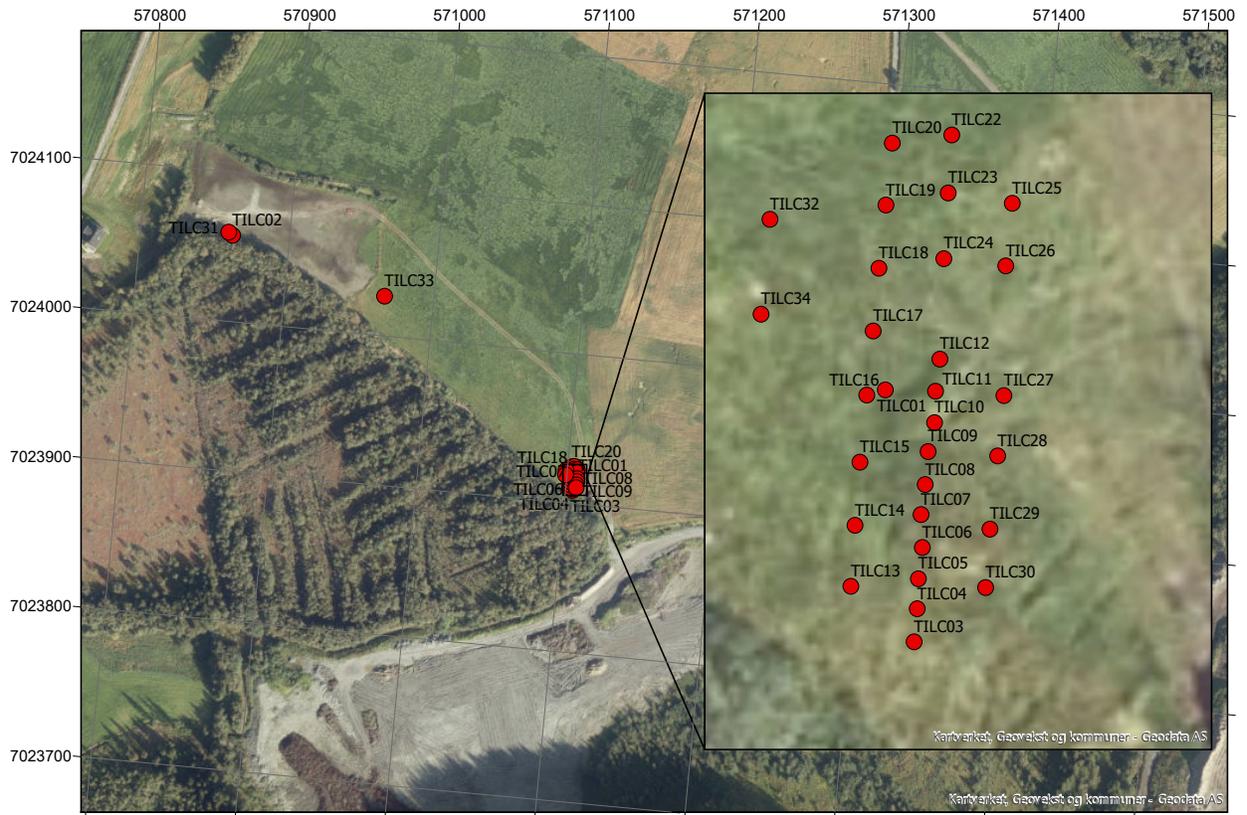
Figure 5.1: CPTu test layout at NGTS site.
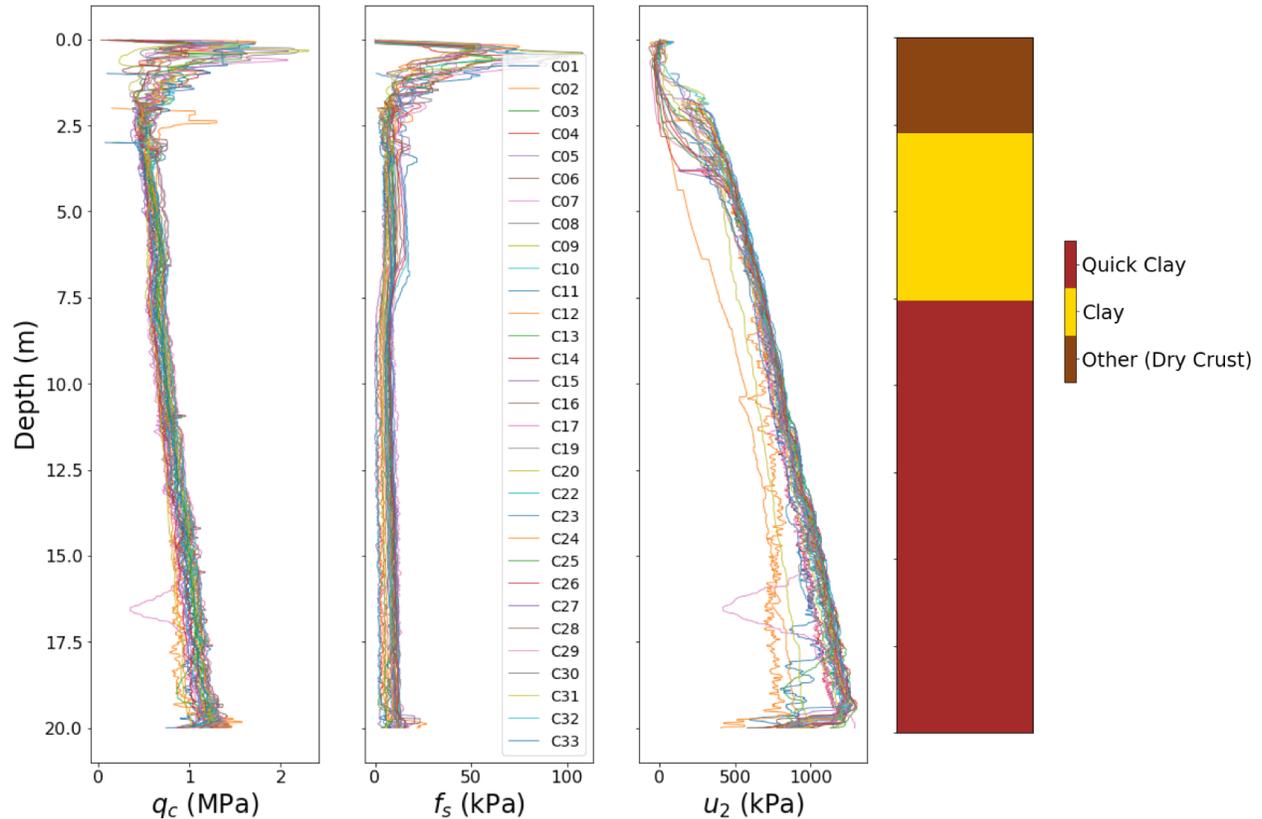
Figure 5.2: CPTu test summary and layering at NGTS site, the plots show the tip resistance, sleeve friction and pore pressure over the depth.

5.1.2 Vegvesen dataset

This dataset is made by seven CPTu tests which are part of the studies for the construction of the county road 715 connecting Keiserås and Olsøy which passes through an area of high risk of quick clay slides where the foundations of Skaudal bridge were placed. The soil layering of the site is not as constant as it is at the NGTS site, however, the common sequence is a stiff upper layer followed by clay on top of a thick quick clay layer, in deep it can appear clay or stiffer layers. The locations of the tests are shown in Figure 5.3, while the layering is shown in Figure 5.4. It is important to note that the layering, in this case, is a proposed layering made by the student based on the information of the data report of the site made by Statens Vegvesen (2013). The summary of the tests in terms of the parameters measured is shown in Figure 5.5, detailed plots of the tests considered for this dataset can be found in Appendix A.
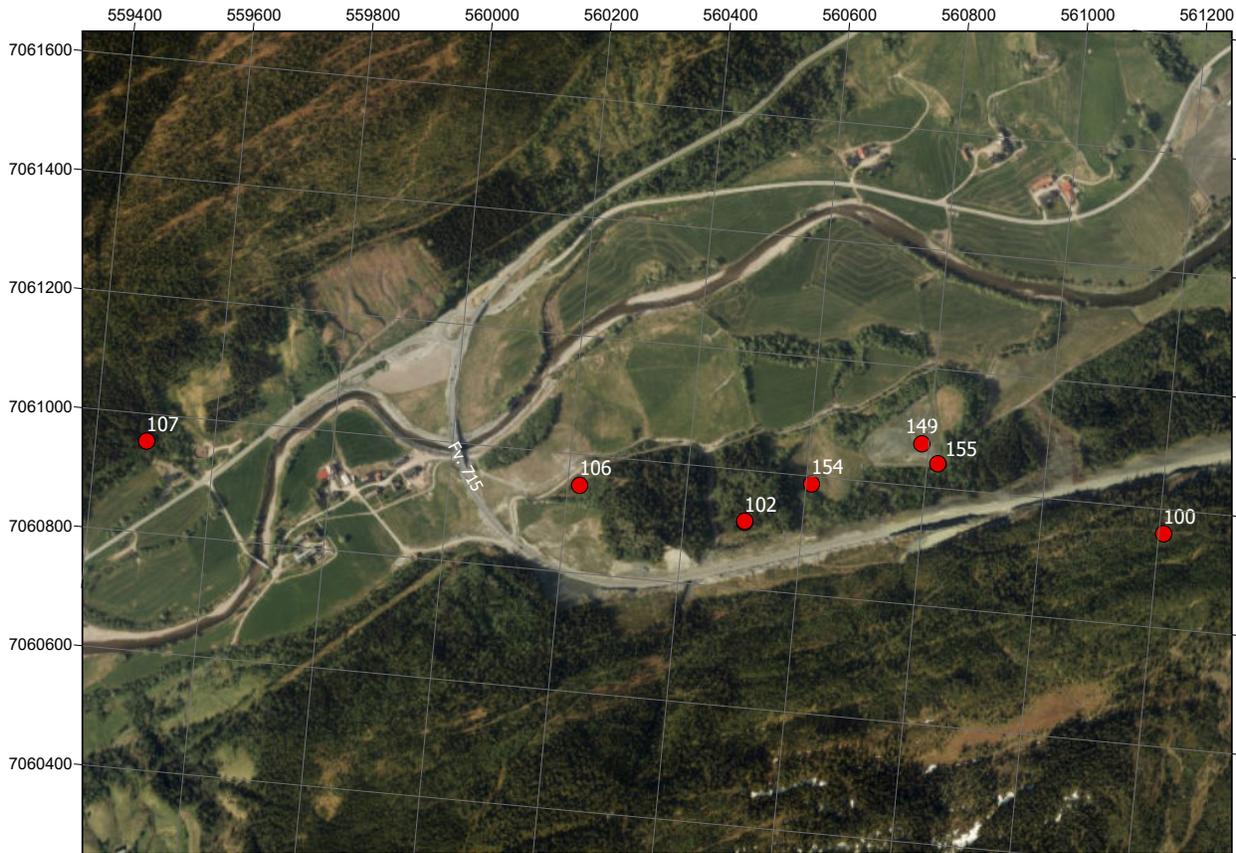
Figure 5.3: CPTu test layout at Vegvesen site, Fv. 715 Keiresås-Olsøy.
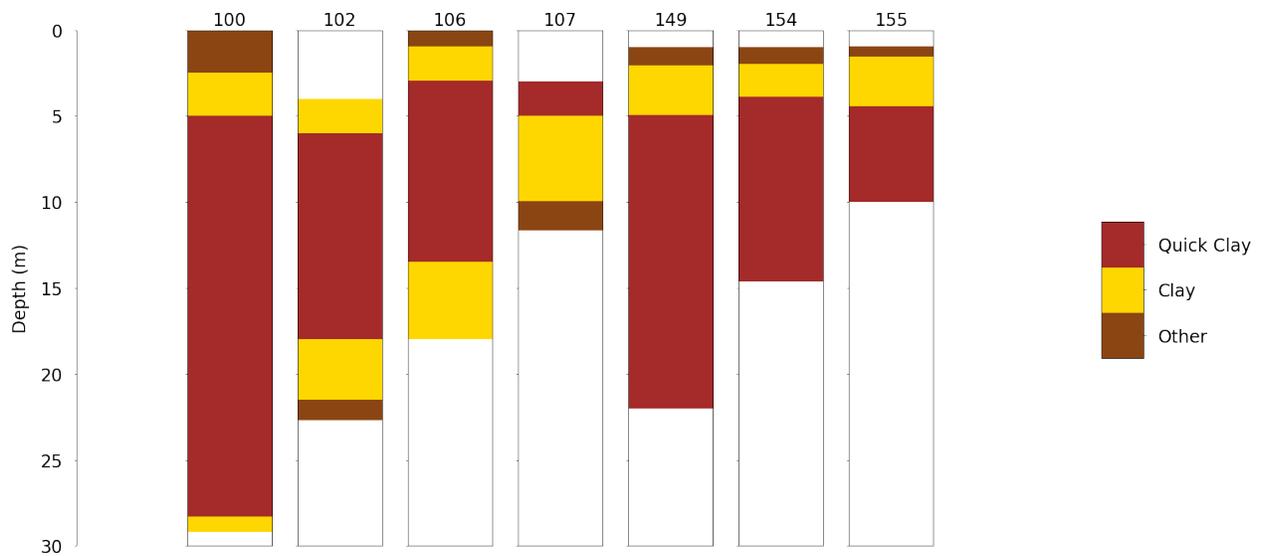
Figure 5.4: Layering of Vegvesen site.
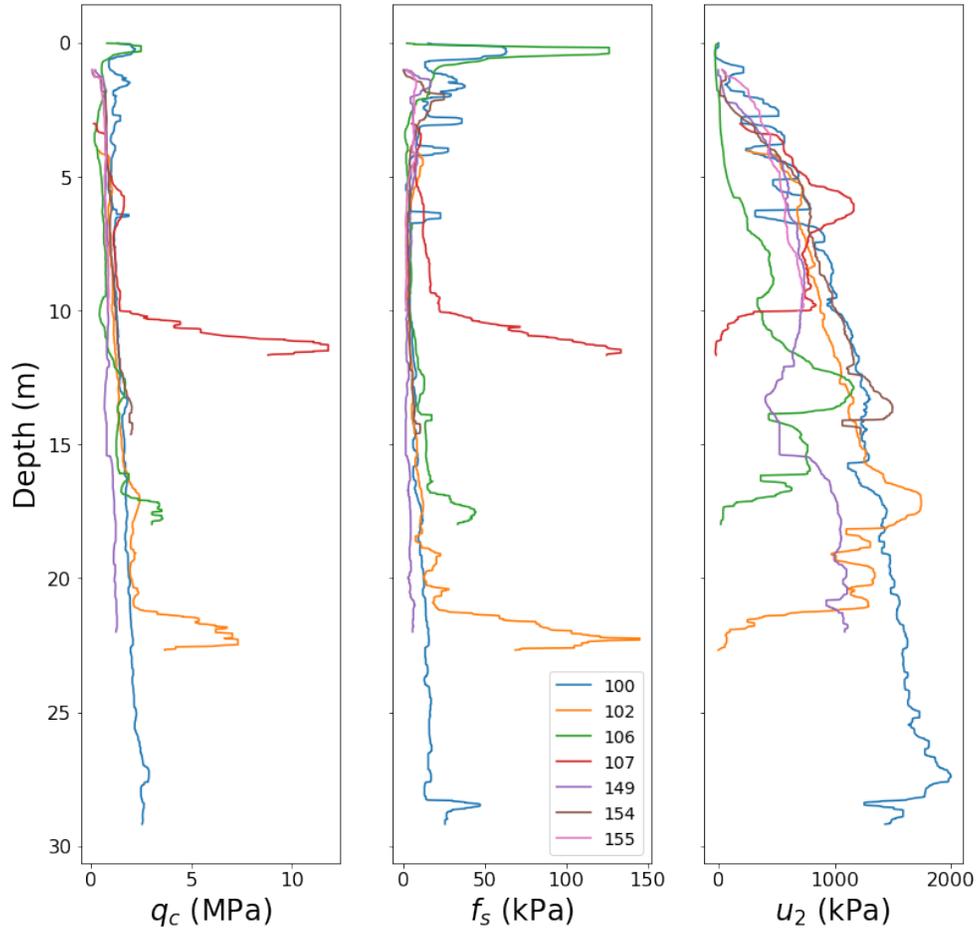
Figure 5.5: Summary of CPTu test at Vegvesen site.

5.2 Data processing

The CPTu test data was received as raw files in ".cpt" format with measurement of depth, tip resistance (q_c), sleeve friction (f_s) and pore pressure behind the cone (u_2). The tip resistance value was corrected from the effects of pore pressure acting at the conical tip, using the formula: $q_t = q_c + (1 + a) \cdot u_2$, where a is the net area ratio dependent on probe design and q_t is the corrected tip resistance. The normalized parameters were computed according to the following equations:

Normalized cone resistance (Q_t):

$$Q_t = \frac{q_t - \sigma_{v0}}{\sigma'_{v0}} \tag{5.1}$$

Normalized friction ratio (F_r):

$$F_r = \frac{f_s}{q_t - \sigma_{v0}} \tag{5.2}$$

Pore pressure ratio (B_q):

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}} = \frac{\Delta u_2}{q_t - \sigma_{v0}} \tag{5.3}$$

Normalized excess pore pressure (U_2):

$$U_2 = \frac{\Delta u_2}{\sigma'_{v0}} \tag{5.4}$$

Where σ_{v0} and σ'_{v0} are the in situ total and effective stresses, respectively, and u_0 is the in situ pore pressure.

The data were inspected to check for abnormal measurements (like negative sleeve friction) and was smoothed using a median statistical filter, as recommended by Wickremesinghe (1989), in order to remove unwanted spikes. The results of such smoothing can be seen in Figure 5.6 below.
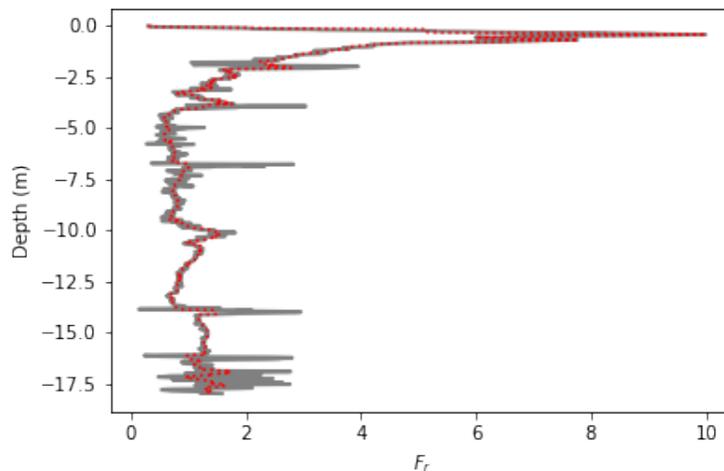
Figure 5.6: Vegvesen dataset CPTu 106 Friction ratio, raw data (solid gray line) compared with the smoothed through statistical filter (dashed red line).

In the present study, for the Machine Learning classification, a logarithmic transformation is done in order to normalize the data in the range of (0,1). The transformations done are presented below. The normalized excess pore pressure (U_2) is preferred over the pore pressure

ratio (B_q).

$$Q_t^{norm} = \frac{2 \cdot log_{10}(Q_t) + 1}{9} \tag{5.5}$$

$$F_r^{norm} = \frac{log_{10}(F_r) + 4.5}{6} \tag{5.6}$$

$$U_2^{norm} = \frac{log_{10}(U_2 + 40)}{2.2} \tag{5.7}$$

5.3 Classification using charts

To have a point to compare the Machine Learning approach, the classification first was performed using well-known charts which consider in their classification schemes sensitive soils. The charts used those recommended by Robertson (1990), Eslami and Fellenius (1997), Schneider et al. (2008), Robertson (2016) and Gylland et al. (2017). The metric used to evaluate the accuracy is the Accuracy Score. Since this part of the work is focused on predicting the appearance of highly sensitive and quick clays from the measurements of CPTu tests, only three soil classes will be taken into account: Sensitive, Clayey and other (more competent) soils. The classification results will be adjusted consequently in order to measure the classification accuracy.

5.3.1 Robertson (1990)

This classification chart is based on an extensive database of CPTu tests, using the three normalized parameters introduced above to define soil behavior types. The chart is presented in Figure 5.7 and the respective soil classes are summarized in Table 5.1.
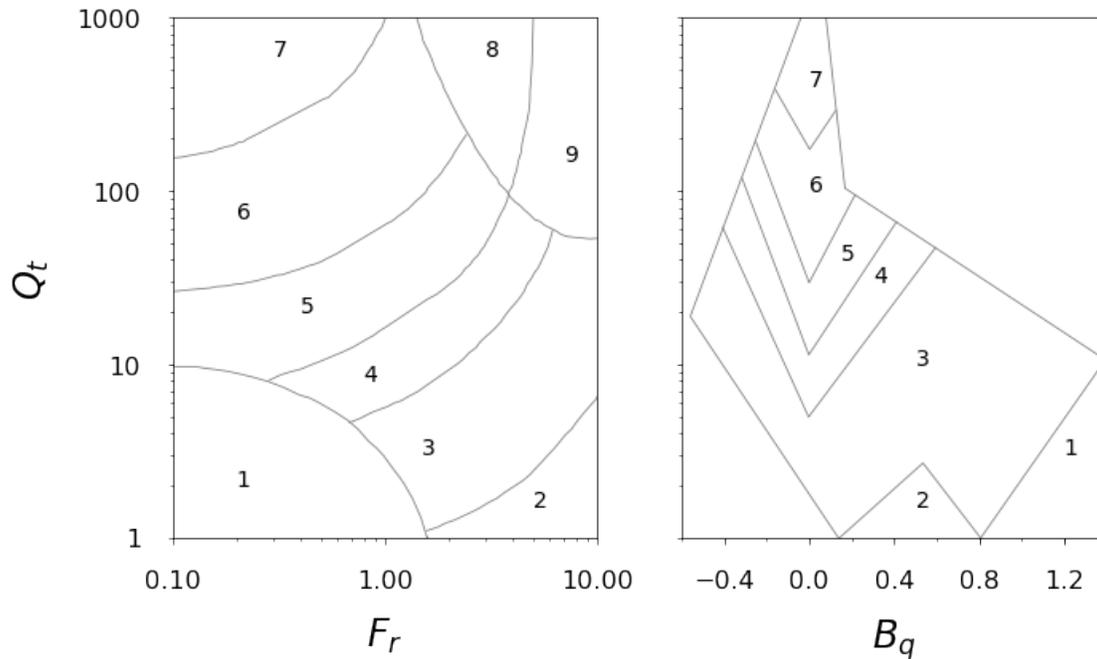
Figure 5.7: Soil behavior type classification chart based on normalized CPTu (Robertson, 1990).

Table 5.1: Soil behavior type classification chart based on normalized CPTu (Robertson, 1990).

Zone	Soil Behavior Type	Zone	Soil Behavior Type	Zone	Soil Behavior Type
1	Sensitive, fine grained	4	Silt mixtures clayey silt to silty clay	7	Gravelly sand to sand
2	Organic soils - peats	5	Sand mixtures; silty sand to sand silty	8	Very stiff sand to clayey sand
3	Clays - clay to silty clay	6	Sands; clean sands to silty sands	9	Very stiff fine grained

The results of using this classification chart are shown in Figure 5.8. For comparison, soil classes 3 and 4 are considered as clayey, while 5, 6, 7, 8 and 9 are considered as other. The accuracy score of this classification using the study datasets are shown below:

NGTS: $Q_t - F_r$ Chart: 70%

              $Q_t - B_q$ Chart: 53%

Vegvesen: $Q_t - F_r$ Chart: 28%

              $Q_t - B_q$ Chart: 27%

By looking at the chart is evident that the classification results for Vegvesen dataset will have

low accuracy score because of the high Q_t of the site's sensitive clays compared with the zone defined by Robertson. On the other hand, the results for NGTS site show a better agreement with the chart, specially the $Q_T - F_r$ plot.
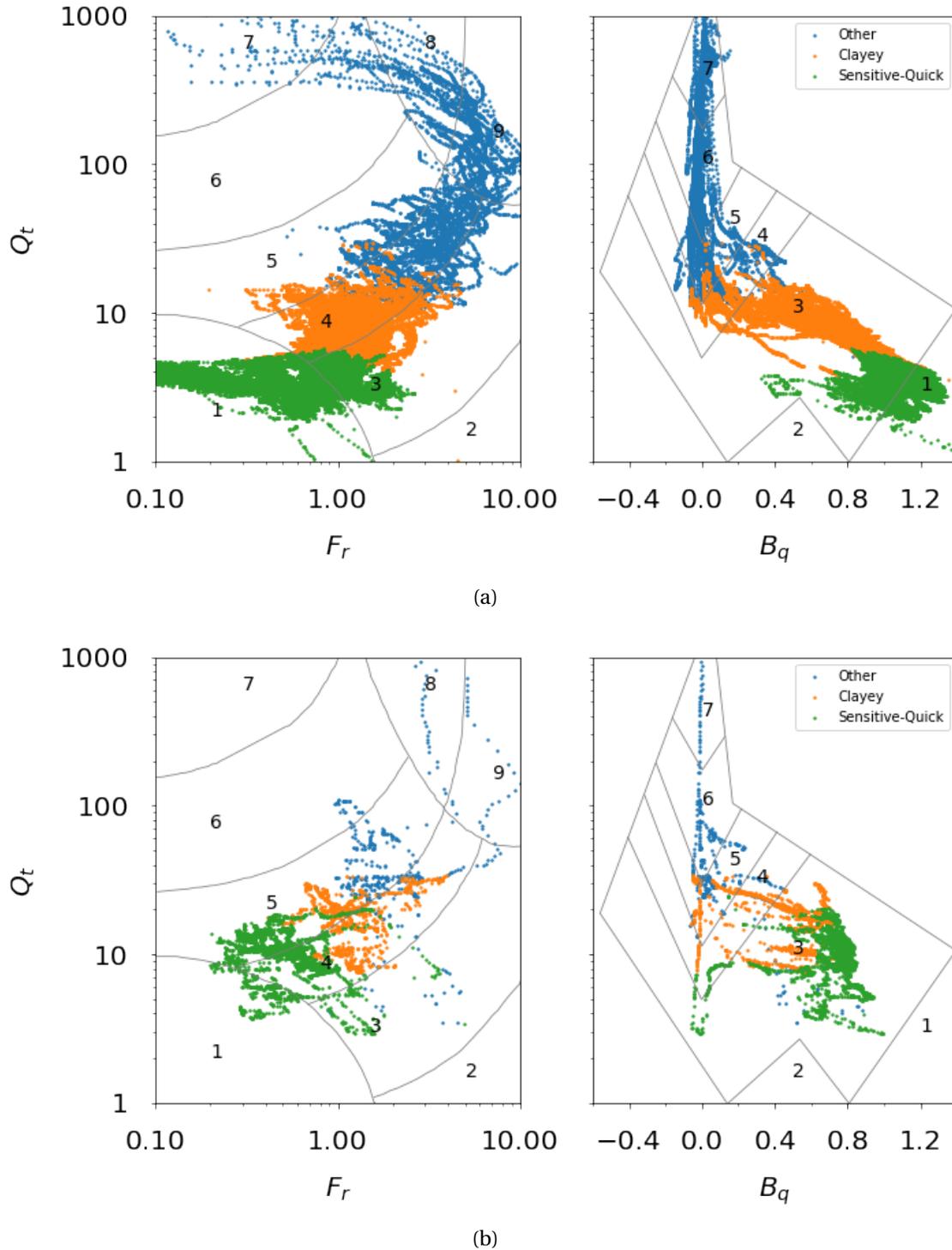
Figure 5.8: Robertson (1990) Classification showing the NGTS (a) and Vegvesen (b) datasets.

5.3.2 Eslami and Fellenius (1997)

This classification chart was developed when investigating the use of CPT in pile design and was made using data from 20 sites in 5 countries. In this case, an "effective" cone resistance and the sleeve friction values are used instead of the normalized ones. The effective cone resistance is defined as $q_E = q_t - u_2$. The chart defines five classes: 1. Sensitive and Collapsible Clay and/or Silt, 2. Clay and/or Silt, 3. Silty Clay and/or Clayey Silt, 4. Sandy Silt and/or Silty Sand, and 5. Sand and/or Sandy Gravel. Figure 5.9 shows the zones where each soil class is defined.
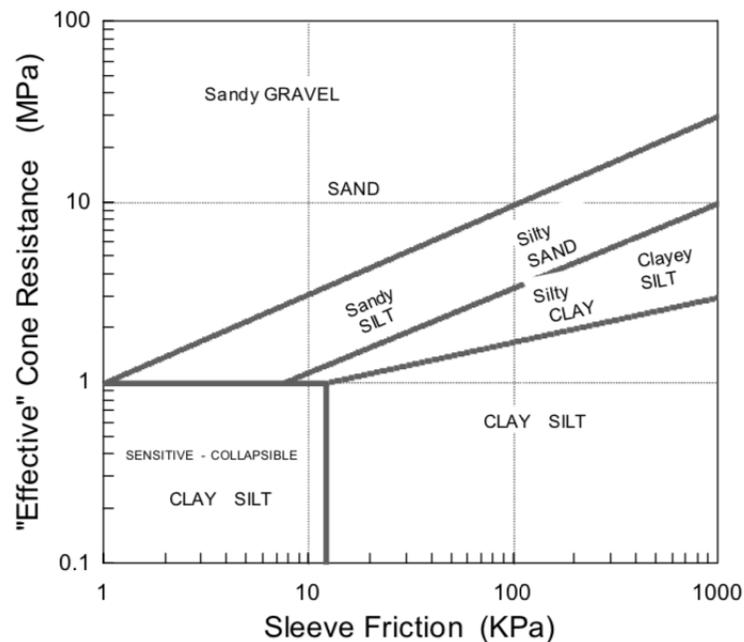
Figure 5.9: Soil type categories classification chart (Eslami and Fellenius, 1997).

The results of using the Eslami and Fellenius chart with the datasets in study are shown in Figure 5.10. It is possible to see a better agreement in the identification of sensitive soils for the Vegvesen dataset compared with Robertson's 1990 classification. In both datasets, but clearer in NGTS, is possible to see a major overlap between Clayey and Quick Clay soils. The accuracy scores are 63% for NGTS and 74% for Vegvesen.
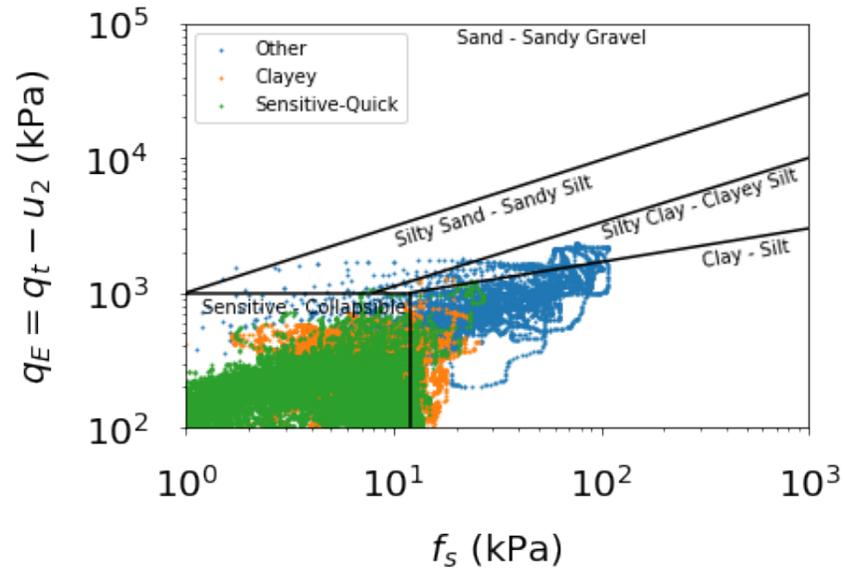
(a)
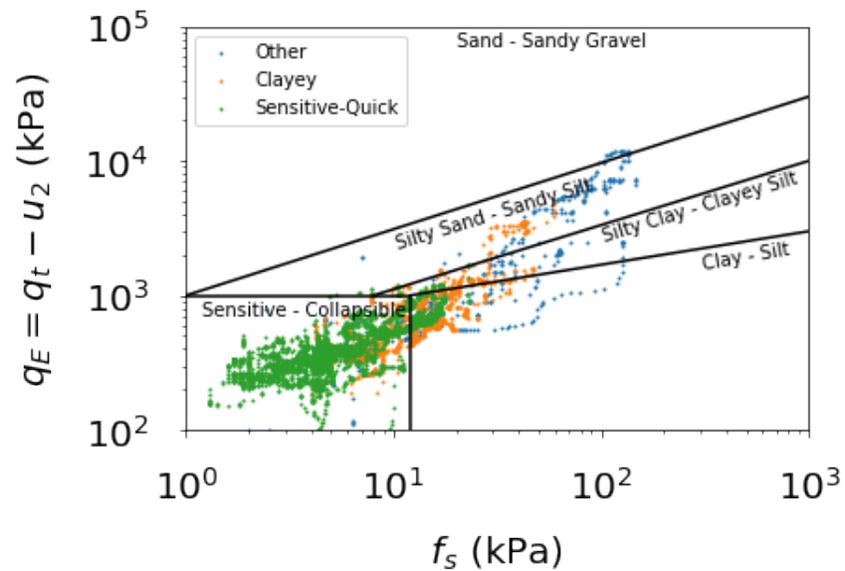
(b)

Figure 5.10: Eslami and Fellenius (1997) Classification showing the NGTS (a) and Vegvesen (b) datasets.

5.3.3 Schneider et al. (2008)

The work done by Schneider et al. (2008) focuses on improving the simple classification charts available at that time in order to take into account the undrained penetration effects on

penetration resistance. The chart is plotted on Q_t - U_2 space which gives improved results for a range of soil types. According to the authors: *"Trends in normalized piezocone response observed in this study may influence design decisions, particularly in non-textbook geomaterials, such as silts, heavily overconsolidated clays, loams, sensitive clays, and mixed soil types".* The database used in this study includes sensitive soils from Norway and Canada. The classification chart, which can be seen in Figure 5.11, is divided into five different zones: 1a. Silts and low rigidity index (I_r) clays, 1b. Clays, 1c. Sensitive clays, 2. Essentially drained sands, and 3. Transitional soils.
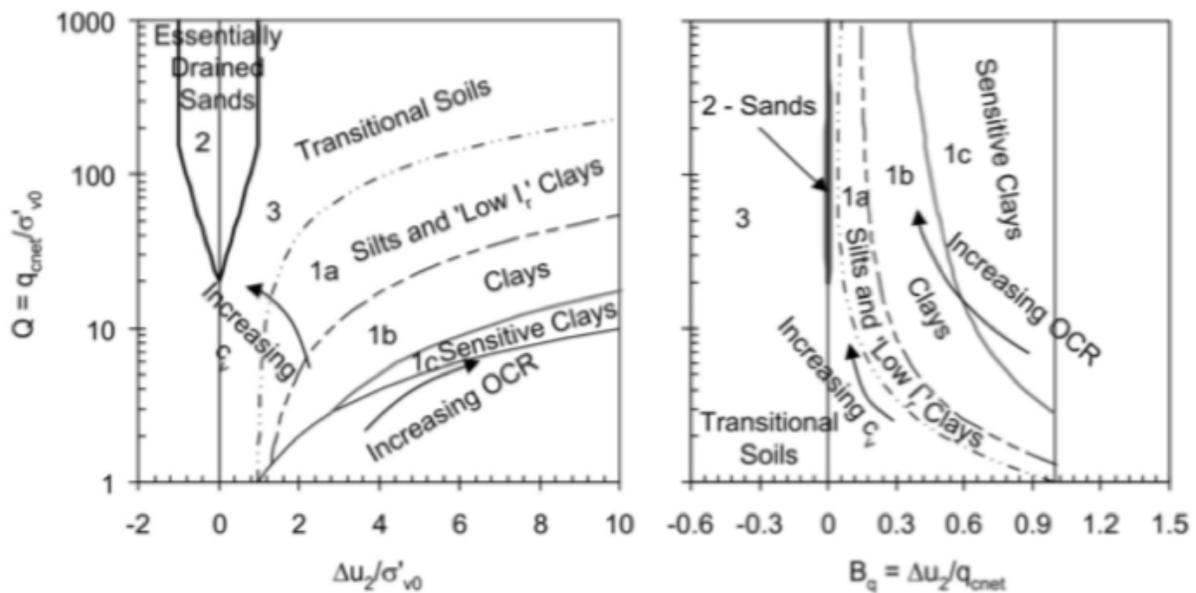
Figure 5.11: Schneider et al. (2008) proposed classification, on the left the chart is plotted on Q_t-$\frac{\Delta u_2}{\sigma'_{v0}}$, while the right one is on the common Q_t-B_q space.

Figure 5.12 shows both dataset plotted on the $Q_t - \frac{\Delta u_2}{\sigma'_{v0}}$ proposed by Schneider et al. (2008). It is possible to see that the sensitive clays from the Vegvesen dataset show a behavior closer to the one predicted by the scheme, giving an accuracy score of 75%, while in the NGTS dataset, there are more cases of "False Positives" meaning that a big fraction of the clay layer would be classified as Sensitive when they are not, however, the accuracy score is still 75% as well.
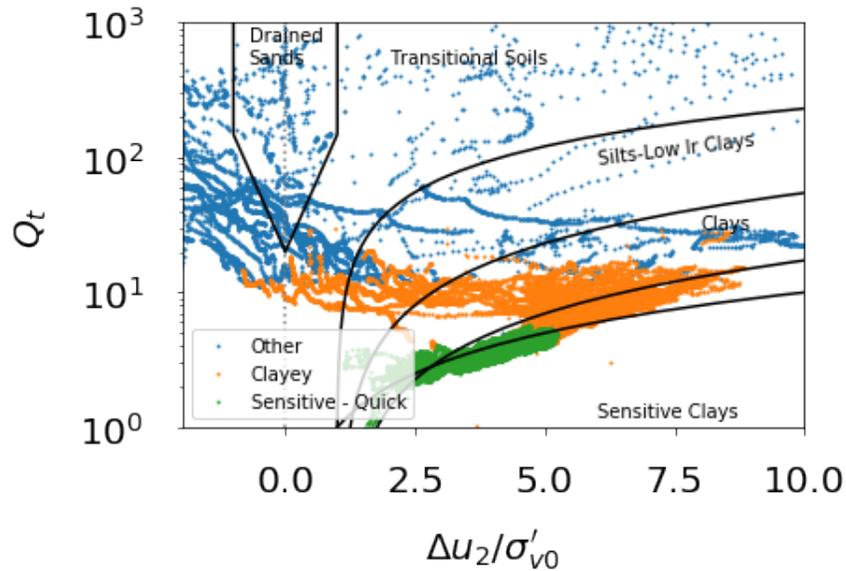
(a)
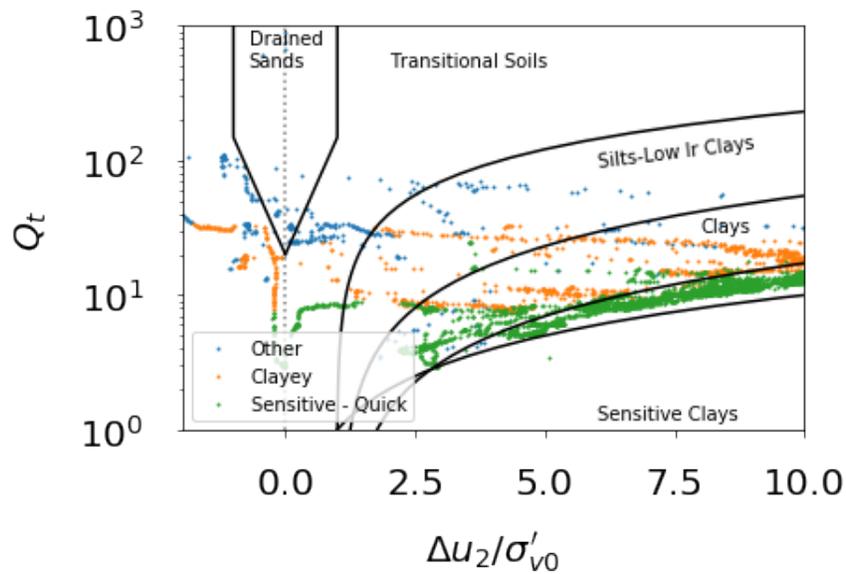
(b)

Figure 5.12: Schneider (2008) Classification showing the NGTS (a) and Vegvesen (b) datasets.

5.3.4 Robertson (2016)

This work presents new correlations in the form of contours of key parameters on the Robertson's 1990 Soil Behavior Type chart with normalized parameters (SBTn) updating also the classification system to use behaviour-based descriptions. Robertson (2009) updated his

original classification by updating the normalization of tip resistance including a variable stress exponent (n) which depends on a soil behavior type index (I_c). The 2016 classification introduces contractive-dilative boundaries (CD) representing the behavior of the soil and a modified soil behaviour type index, I_B. Figure 5.13 shows the new classification system overlaying the original one.
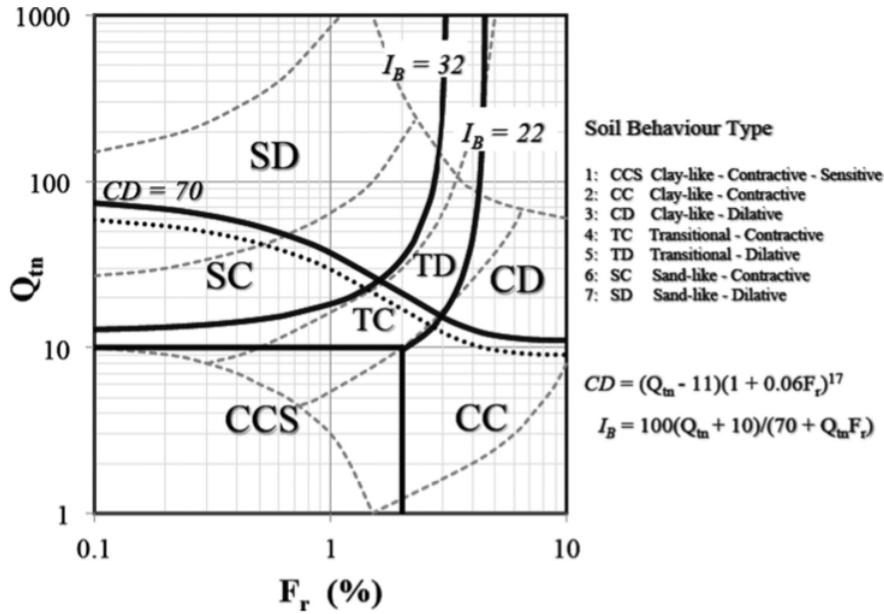
Figure 5.13: Robertson (2016) SBTn chart based on Q_{tn}–F_r (solid lines show proposed new soil behaviour type boundaries, and dashed lines show boundaries suggested by Robertson (1990)).

The normalized tip resistance is computed iteratively according to the following expressions:

$$Q_{tn} = \frac{(Q_t - \sigma_{v0})}{p_a} \left(\frac{p_a}{\sigma'_{v0}}\right)^n \tag{5.8}$$

$$n = 0.381 \cdot I_c + 0.05 \cdot \frac{\sigma'_{v0}}{p_a} - 0.15 \tag{5.9}$$

$$I_c = \left((3.47 - \log(Q_{tn}))^2 + (\log(F_r) + 1.22)^2\right)^{0.5} \tag{5.10}$$

While the CD boundary and the modified soil behavior type index are computed as follow:

$$CD = (Q_{tn} - 11)(1 + 0.06 \cdot F_r)^{17} \tag{5.11}$$

$$I_B = \frac{100 \cdot (Q_{tn} + 10)}{70 + Q_{tn} \cdot F_r} \tag{5.12}$$

Figure 5.14 show both dataset plotted on the $Q_{tn} - F_r$ space, again the NGTS dataset present a big fraction of the Clayey soil classifying as Sensitive, but the other two soil classes seem to fall well within the correct groups. The Vegvesen dataset shows a big fraction of Sensitive soils classifying as Transitional or Clay Contractive. The accuracy score for NGTS is 75% while for Vegvesen is 52%.
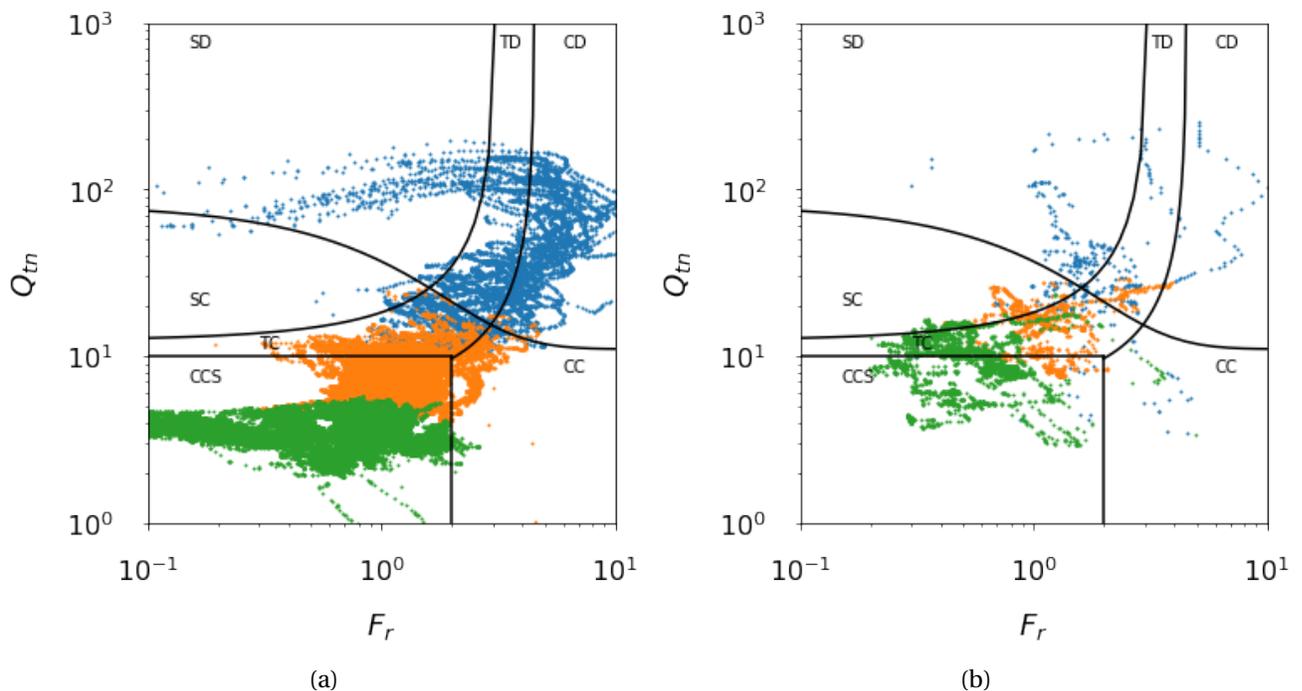
Figure 5.14: Robertson (2016) Classification showing the NGTS (a) and Vegvesen (b) datasets.

5.3.5   Gylland et al. (2017)

This work proposes a classification chart specially focused on the identification of sensitive clays and is based on tests performed in Norway. This is done using parameters following the

same philosophy as Robertson (1990) but with different normalization, as shown below.

$$N_{mc} = \frac{q_t - \sigma_{v0}}{\sigma'_A + a} \tag{5.13}$$

$$B_{q1} = \frac{\Delta u_1}{q_t - \sigma_{v0}} \tag{5.14}$$

$$R_{fu} = \frac{f_s}{\Delta u_1} \tag{5.15}$$

Where the reference stress $\sigma'_A = \sigma'^m_c + \sigma'^{(1-m)}_{v0}$, σ'_c is the effective pre-consolidation stress, a is the attraction, m is the SHANSEP-framework exponent (typically between 0.7 - 0.8 for Norwegian clays), and Δu_1 is the excess pore pressure at the tip of the cone. The classification chart is shown in Figure 5.15 below.
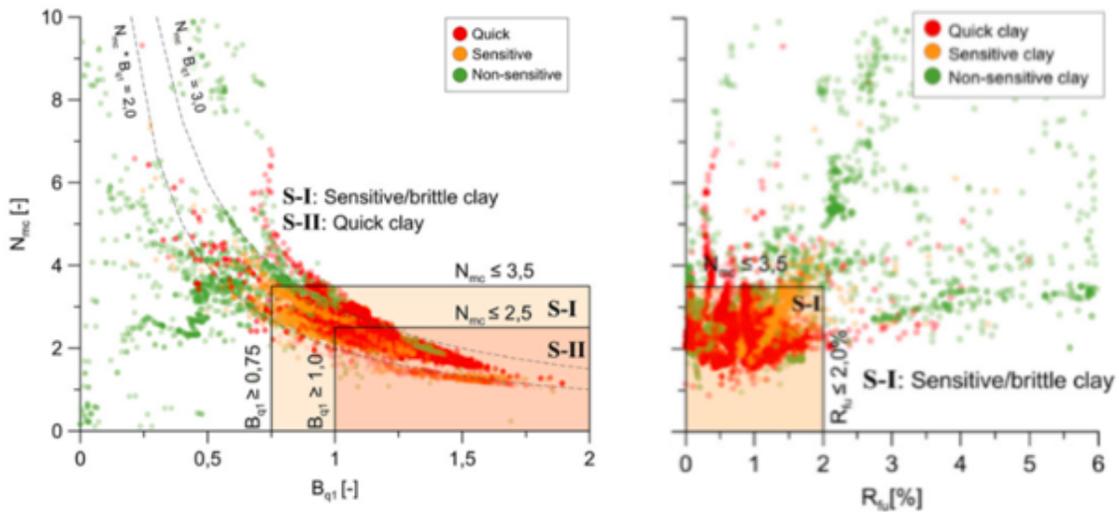
Figure 5.15: Gylland classification charts.

The main drawback of this scheme is that is necessary to know parameters that do not come necessarily associated with the CPT test itself, as the attraction and pre-consolidation stress, and requires also the knowledge of the pore pressure at the tip (u_1) which is not usually measured in spite of the one behind the cone (u_2). In that case, is necessary to use correlation involving the measured parameters to estimate the ones from the model.

Figure 5.16 shows both datasets plotted on the chart proposed by Gylland et al. (2017), in this

case, it is evident that the NGTS dataset shows better agreement compared with the Vegvesen one. The accuracy score, in this case, is only a binary classification score due to the nature of the classification proposed by the authors, the results are summarized below:

NGTS: $N_{mc} - B_{q1}$ Chart: 86%

                $N_{mc} - R_{fu}$ Chart: 87%

Vegvesen: $N_{mc} - B_{q1}$ Chart: 39%

                $N_{mc} - R_{fu}$ Chart: 40%

The Vegvesen dataset is plotted almost completely out of the red shaded area defining sensitive clays, showing a different behavior of the sensitive clays present in the area compared with the ones that were part of the dataset used by the authors, which included the NGTS site at Tiller.
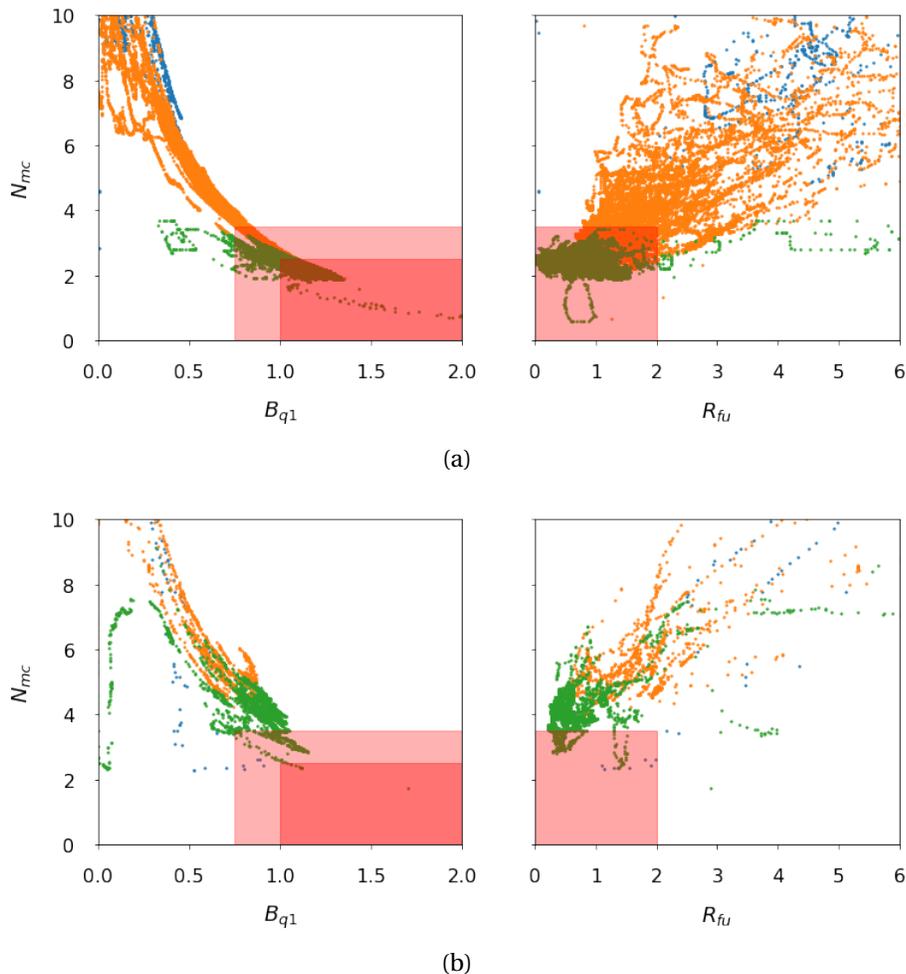
Figure 5.16: Gylland (2017) Classification showing the NGTS (a) and Vegvesen (b) datasets.

5.3.6 Summary of classification charts

A summary of the accuracies of each classification is presented on Table 5.2. It is possible to see that the results are quite good for general classification charts, but they are not very consistent, showing big differences between each dataset, being the exception the Schneider chart which scores the same accuracy for both. The accuracy and consistency problem from classification charts arises because of the datasets that were used to define each classification set are whether very global or very local, which makes them difficult to classify accurately specific cases as the quick clays in Norway, which have particular properties that make them unique. In that sense, it would be better to develop local approaches that allow to get feedback from the field and predict with higher accuracy the appearance of quick or highly sensitive clays.

In the next section, a Machine Learning approach will be studied where data from the field will be used to train a model which is able to learn from the data and improve as new knowledge is gathered for making better predictions.

Table 5.2: Summary of accuracy scores for the different classification charts studied.

Authors	Parameters	Accuracy Score (%)	
		NGTS	Vegvesen
Robertson (1990)	$Q_t - F_r$	70	28
	$Q_t - B_q$	53	27
Eslami & Fellenius (1997)	$q_E - f_s$	63	74
Schneider (2008)	$Q_t - U_2$	75	75
Robertson (2016)	$Q_{tn} - F_r$	75	52
Gylland (2017)	$N_{mc} - B_{q1}$	86	39
	$N_{mc} - R_{fu}$	87	40

5.4 Machine Learning Classification Results

This section presents the results of applying machine learning to the datasets from NGTS site and Vegvesen described before.

5.4.1 Train and test datasets

Machine learning is about learning some properties from a dataset and then testing those properties against another dataset. A common practice in machine learning is to evaluate an algorithm by splitting a dataset into two. One of those sets is called the training set, on which the model learns some properties; the other set is called the testing set, on which the learned properties are tested. Geotechnical site exploration, generally because of resources constrains, is done sequentially, with a couple of exploration teams working at the same time. To take this into account, the splitting of the dataset will be done sequentially, using the data from one CPTu to train the model, then predict the next one and use it to re-train the model and predict the next one, and so on. The CPTu data to be used will be selected randomly as well as the testing one. Since the draws are made randomly, the process will be repeated at least 30 times to account for different combinations of CPTu tests in the training and testing datasets.

5.4.2 Results on NGTS Dataset

Logistic Regression Classifier

A logistic regression classificator was sequentially trained and the results of the predictions are shown in Figure 5.17, using $Q_t^{norm} - F_r^{norm}$, $Q_t^{norm} - U_2^{norm}$ and the three of them as predictors.

The results show a high accuracy score even for the first estimation (using only one CPTu test to train) and a sharp increase in it afterwards. It can be noticed that using just four tests to train the classificator, accuracies of at least 80% are reached. It is also possible to see that the $Q_t^{norm} - U_2^{norm}$ scheme shows higher accuracies with fewer data. The use of the three parameters to train the classificator shows improved accuracies as well as less variability.

It is important to highlight that NGTS site is a highly homogeneous site with a regular layering sequence and low dispersion of the parameters measured. These results are not expected to happen on non-homogeneous sites. And because of its homogeneity would be difficult to use a model trained in this site to predict the classification of the soils at a different site.
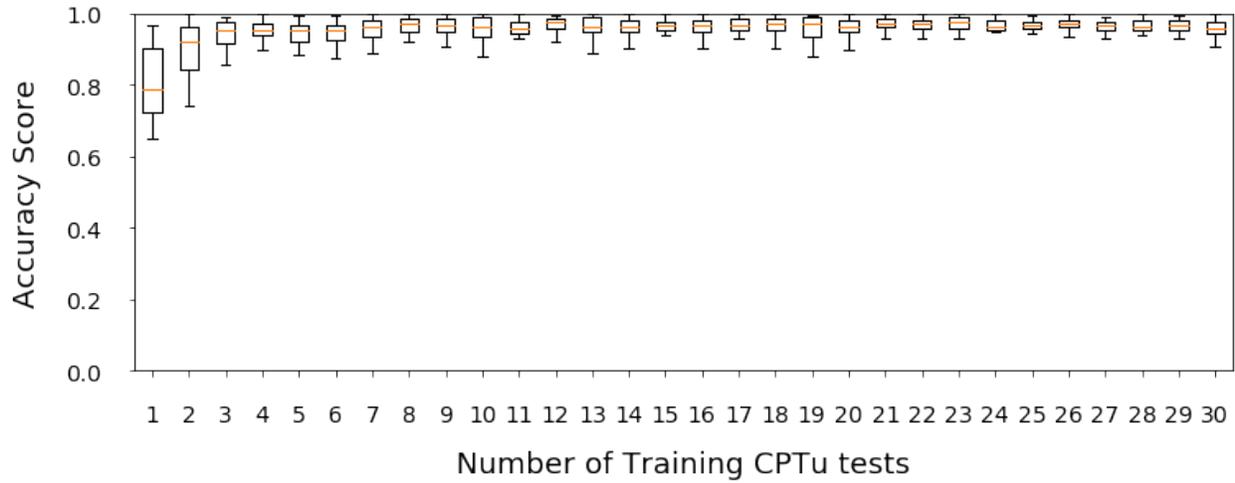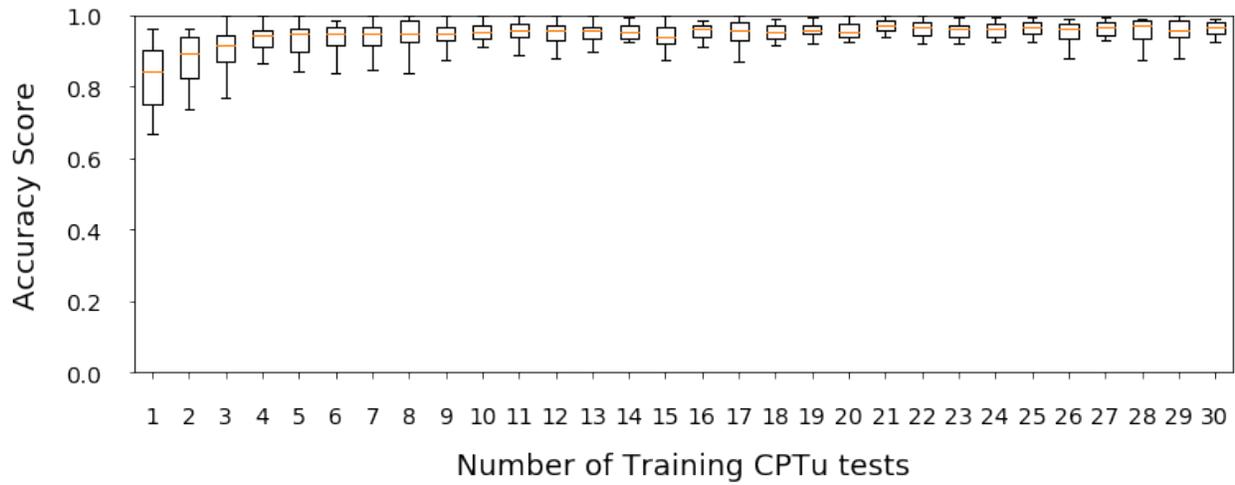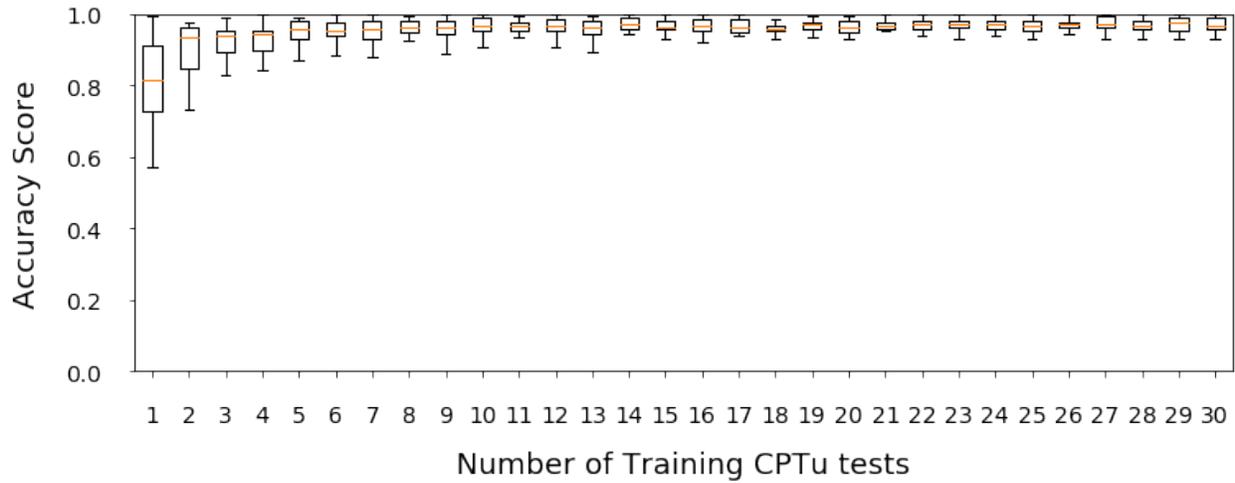
Figure 5.17: Results of a sequentially trained Logistic Regression classificator on the NGTS dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

Using all the data to train the model it is possible to get an equivalent of the classification charts shown before. This is done by meshing the domain defined by any set of two parameters (Q_t-F_r, Q_t-U_2) and use the classificator to predict the class that each point of the mesh belongs. The results for the typical Q_t-F_r and Q_t-U_2 are shown below in Figure 5.18. Since the Logistic Regression classificator is a linear model for classification, the boundaries defining each class will be straight lines, outside the area defined by the data it is not expected that the charts will be representative even for data in the same sites.

(a)

(b)
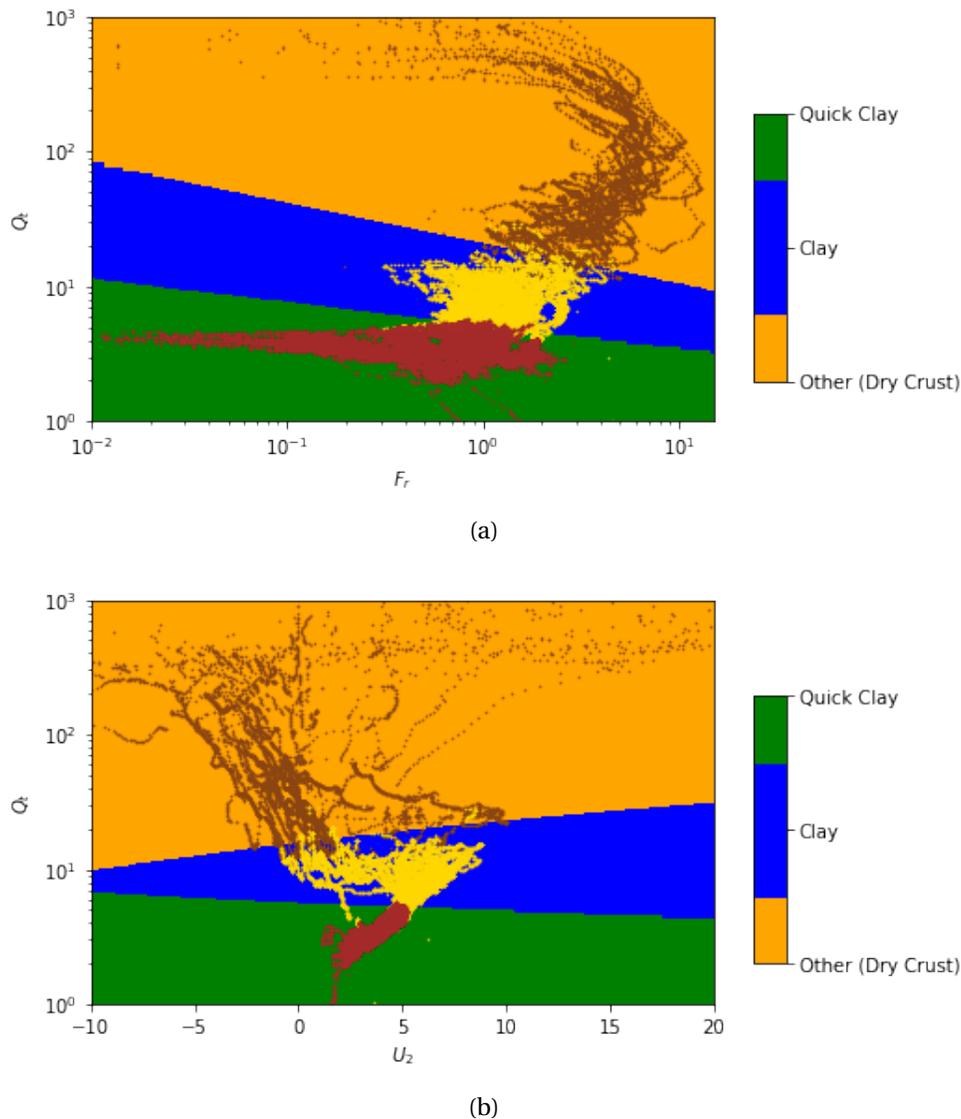
Figure 5.18: Classification charts for NGTS site obtained by Logistic Regression.

Naive Bayes Classifier

The results of using a Naive Bayes classificator on the NGTS dataset are shown in Figure 5.19. It is possible to see that it shows more scatter in the accuracy of the first estimation, specially using F_r^{norm} but it increases quickly afterwards. The major advantage of using a Naive Bayes approach compared with the Logistic Regression is that the time it takes to run can be 10 times less, so in big datasets it can make a major difference.
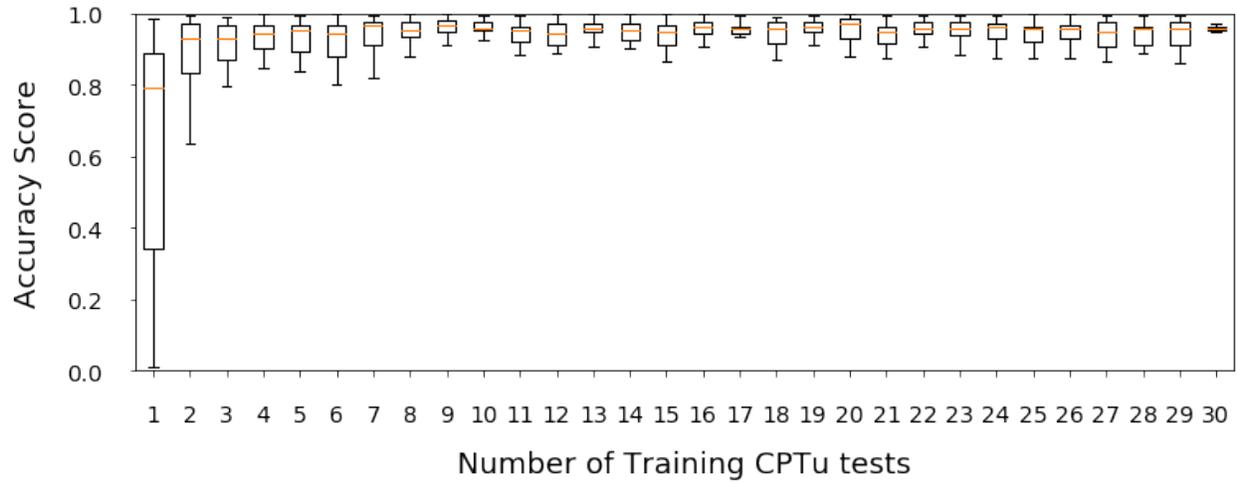
(a)
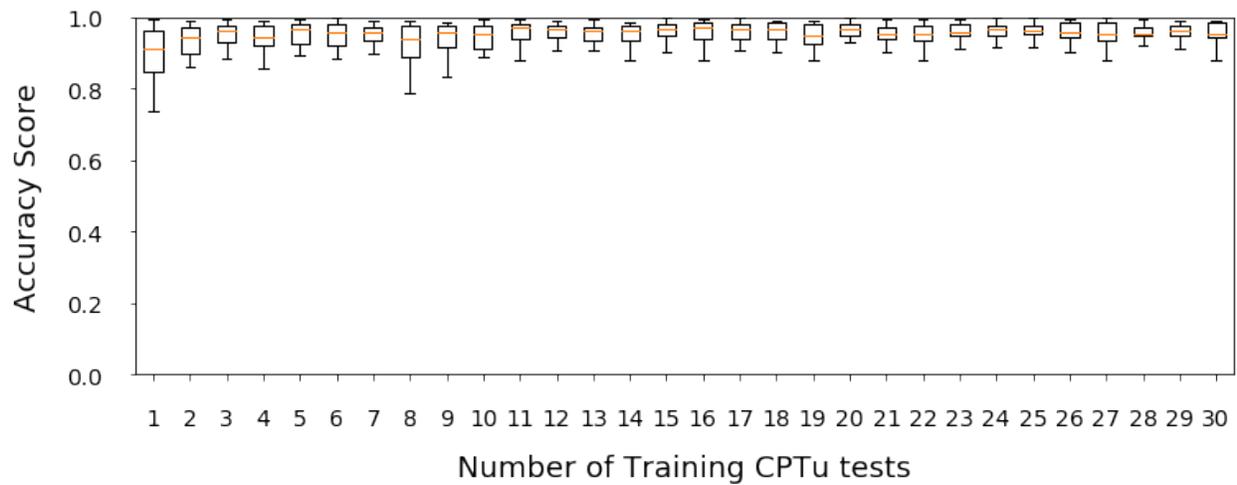
(b)
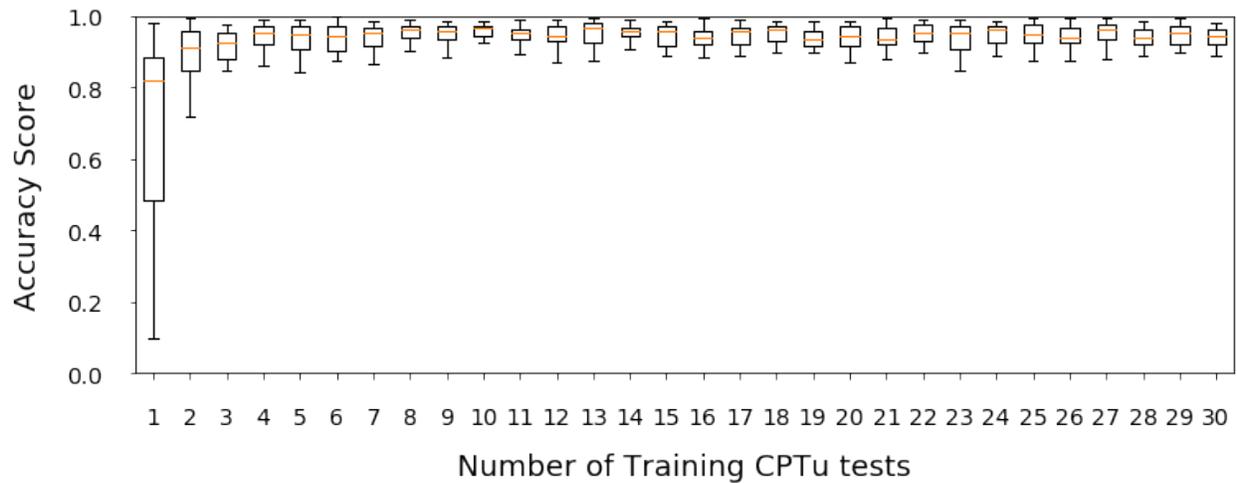
(c)

Figure 5.19: Results of a sequentially trained Naive Bayes classificator on the NGTS dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

In the same way as done with Logistic Regression, classification charts can be obtained by using the Naive Bayes classificator. In this case it is possible to see that the decision boundaries (boundaries separating classes) are elliptical containing better the data that define them, however this does not make sense in a general classification context specially at low Q_t values where a soil can be classified as a dry crust material with very low tip resistance.

(a)

(b)
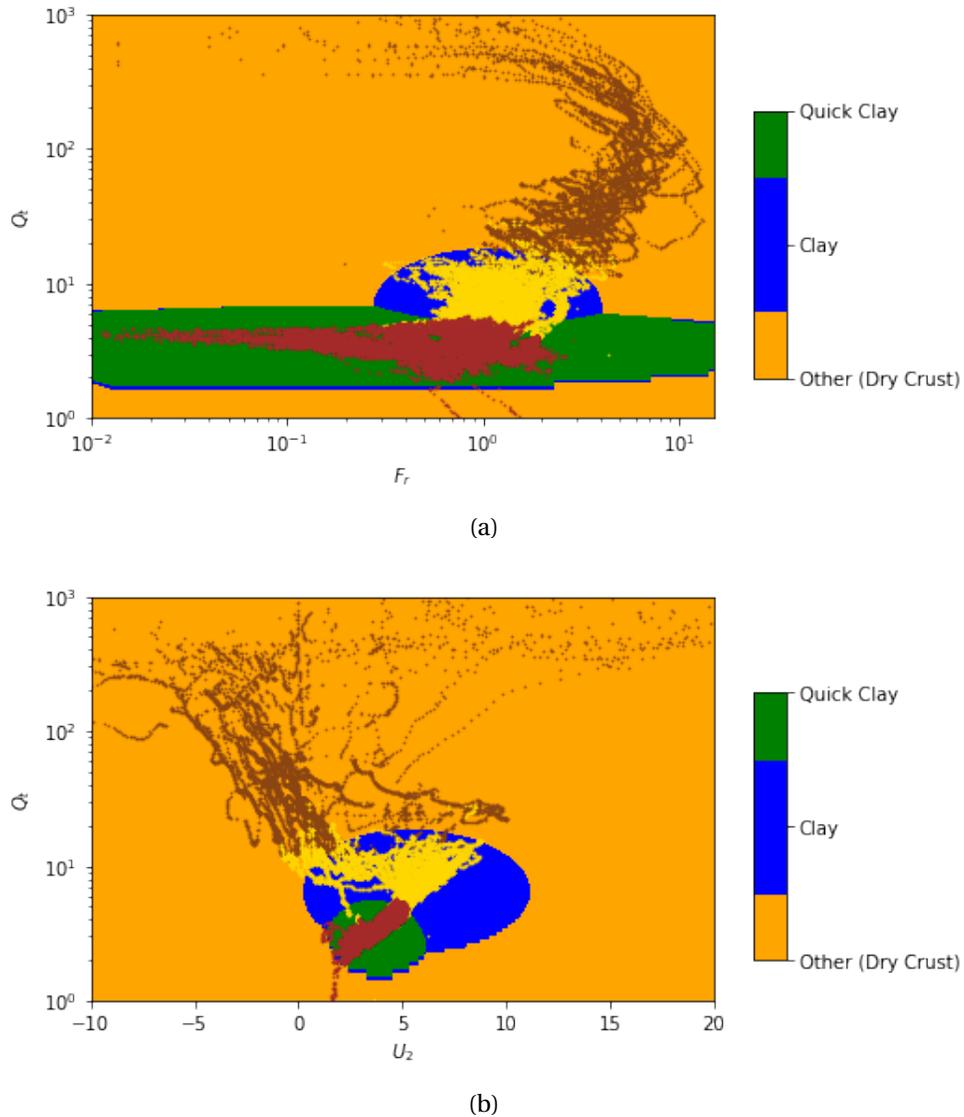
Figure 5.20: Classification charts for NGTS site obtained by Naive Bayes.

Hidden Markov Model

For this part a Hidden Markov Model (HMM) was trained in a semi-supervised way. The model parameters (transition matrix, means, covariances and starting probabilities) were estimated from the training data and then the model was allowed to update (optimize) the values of the transition matrix and covariances in the Expectation-Maximization stage, the rest remained fixed. The decoding algorithm to determine the most likely sequence of states (soil classes) was the Viterbi algorithm. The sequential training was perfomed, but due to restriction in the programmed code, it was only possible to use the CPTu test which contained the three soil classes defined, this reduced the number of combinations available but it is thought that they are enough to draw conclusions. The results of the sequential training and classification are shown in Figure 5.21.
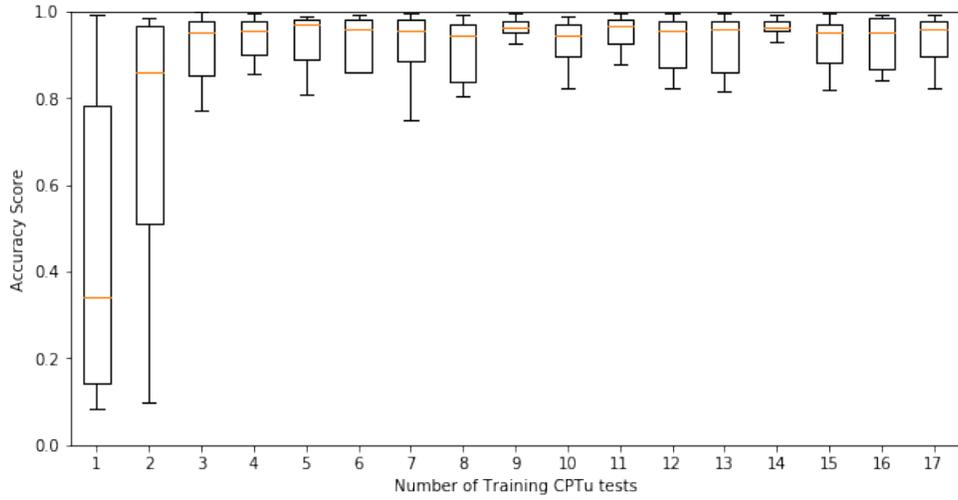
(a)
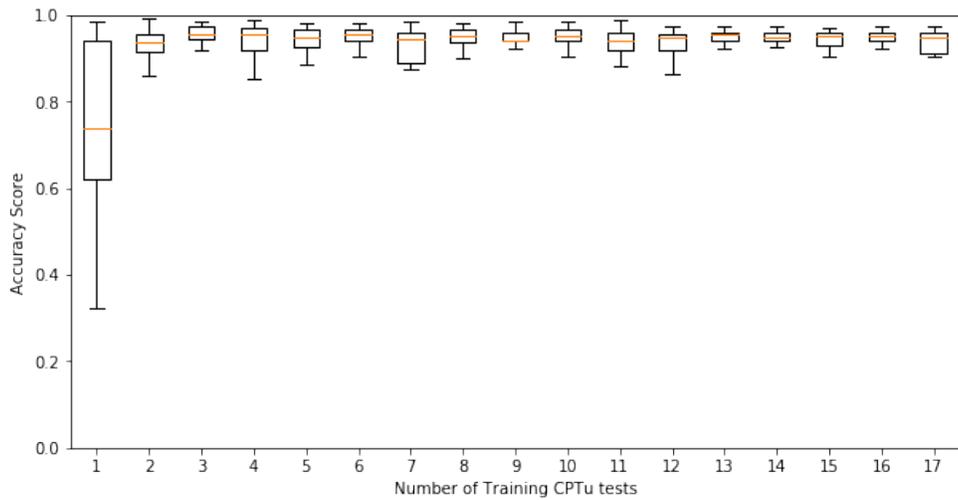
(b)
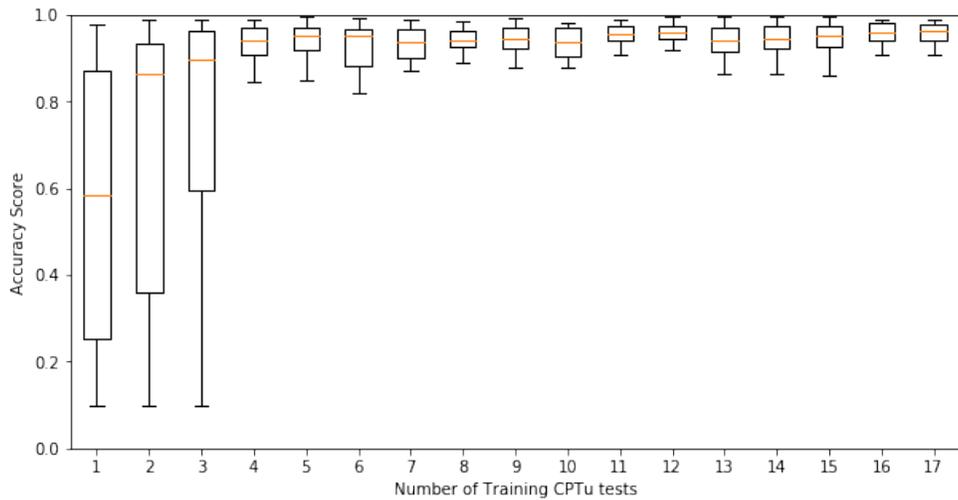
(c)

Figure 5.21: Results of a sequentially trained HMM on the NGTS dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

It is possible to see that despite the prediction is less accurate with few data, it increases quickly yielding results similar to the ones obtained using the other two methods. The advantage of using the HMM is that, since it takes into account how likely is to change from one class (hidden state) to another, the predicted profiles do not have unrepresentative thin layers within another. This can be observed in Figure 5.22 for CPTu C04 and C10.

Site profiles

In order to compare visually the different classification methods, Figures 5.22 and 5.23 shows some selected CPTu tests with the actual site layering alongside the machine learning classification. The remaining profiles can be seen in Appendix B, while the accuracies for each profile estimation is presented in Table 5.3

For the profile estimation, only seven CPTu test were used for training for each test to be classified. The criteria used to select the training dataset was to sort the test by name and use the seven closest to the one to be classified.

The mean accuracies are 95% for logistic regression, 96% for Naive Bayes and 91% for HMM, these values are quite close to each other, but a visual comparison shows despite the high mean accuracy of HMM it can present interpolations which are not close to the reality, as in the case of C05, this can be due to the training set chosen and can be improved with more data in the training phase.
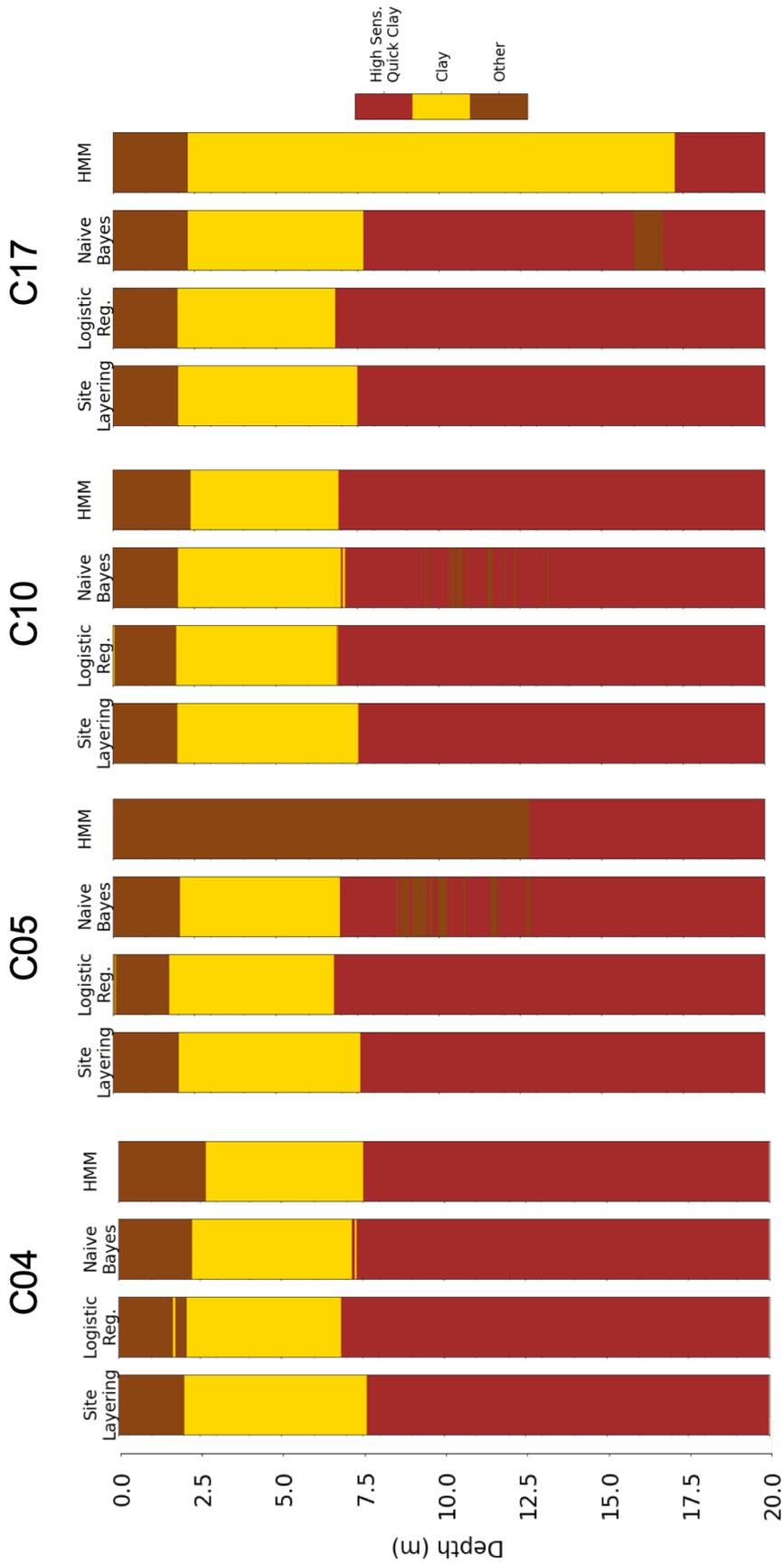
Figure 5.22: Comparison of machine learning classification for the NGTS dataset, profile view of the CPTu test C04, C05, C10 and C17.
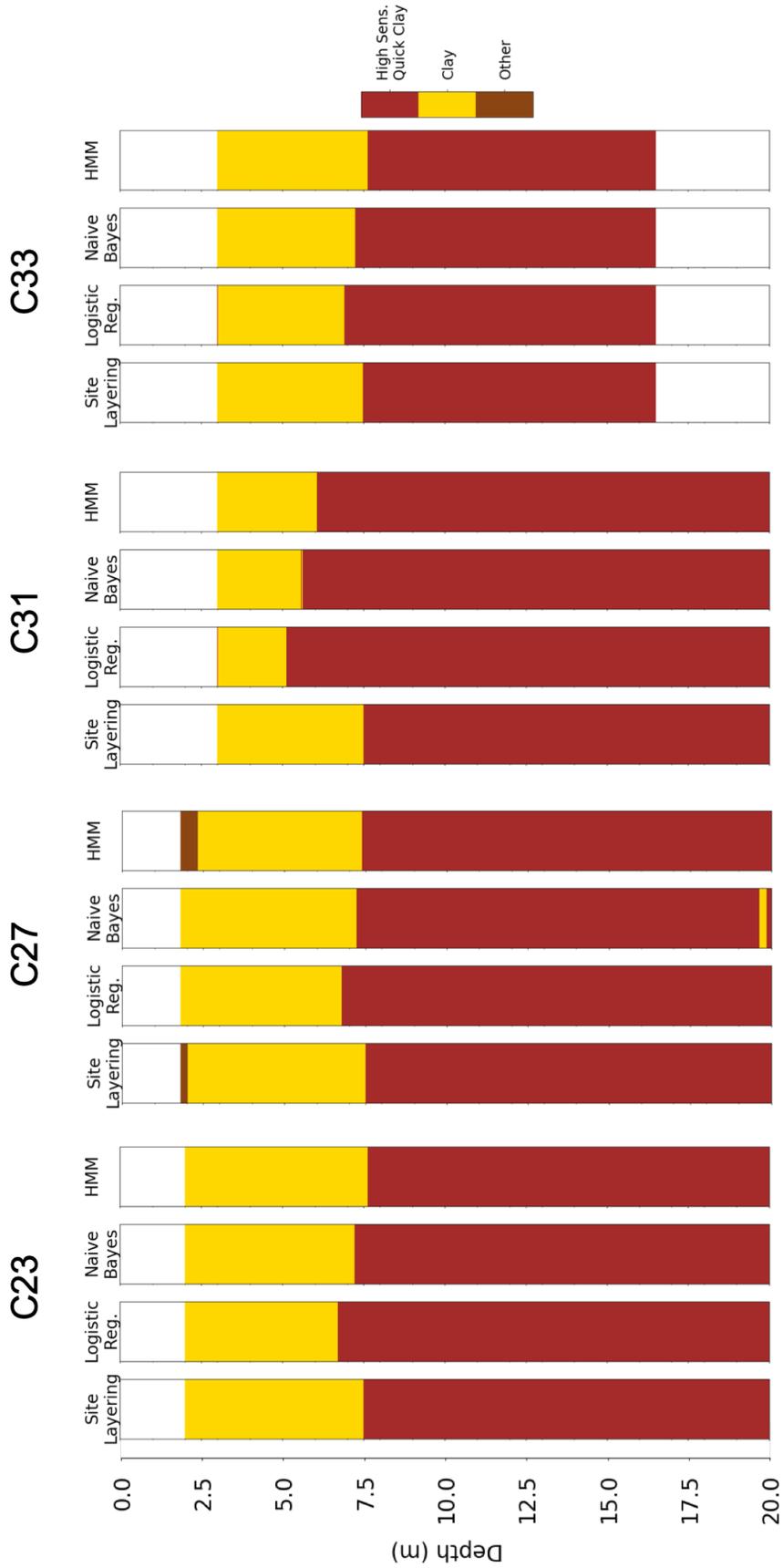
Figure 5.23: Comparison of machine learning classification for the Vegvesen dataset, profile view of the CPTu test C23, C27, C31 and C33.

Table 5.3: Comparison of machine learning classification for the NGTS dataset in terms of accuracy score.

CPTu	Accuracy Score		
	Logistic Reg.	Naive Bayes	HMM
C01	96%	90%	76%
C02	89%	89%	91%
C03	98%	96%	90%
C04	95%	97%	96%
C05	94%	88%	46%
C06	97%	98%	97%
C07	96%	95%	95%
C08	99%	97%	88%
C09	99%	98%	96%
C10	96%	94%	95%
C11	98%	98%	96%
C12	95%	98%	99%
C13	96%	95%	90%
C14	95%	96%	98%
C15	93%	95%	95%
C16	94%	96%	99%
C17	96%	93%	50%
C19	97%	100%	97%
C20	90%	93%	94%
C22	97%	99%	77%
C23	96%	98%	99%
C24	96%	98%	99%
C25	94%	94%	98%
C26	96%	97%	98%
C27	95%	96%	98%
C28	97%	97%	91%
C29	95%	97%	88%
C30	96%	98%	92%
C31	86%	89%	92%
C32	97%	100%	99%
C33	96%	98%	99%

5.4.3 Results on Vegvesen Dataset

Since the Vegvesen dataset is much smaller than NGTS, and since the layering of the site is not as homogeneous, classification using machine learning is more challenging because in the

process of splitting the dataset to train and test, a major share of the information is lost

(because it is not possible to use it to train). So in this case, besides performing the sequential

training and prediction, a cross-validation technique will be used to assess the performance of

the classificators. The cross-validation technique used will be the "Leave One Group Out"

which means for each CPTu test to be classified, the remaining six will be used to train the

model.

Logistic Regression Classifier

The results are shown in Figure 5.24 below. It is possible to see a more scattered behavior and a

less accuracy in general compared with NGTS site, however, given the fact that this dataset has

fewer CPTu test and the soil layering is more complex, the results are good, and compared with

the classification charts they improve the results. This is more evident when analyzing the

cross-validation results shown in Table 5.4.
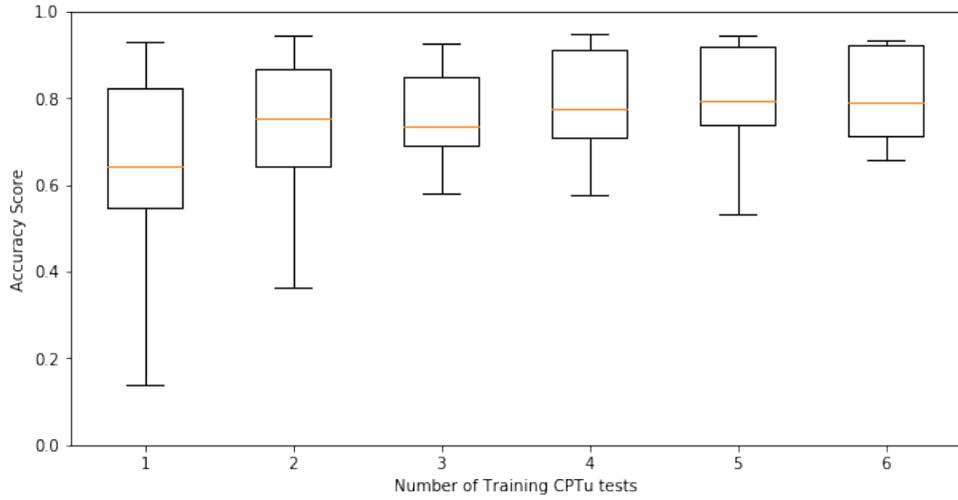
(a)
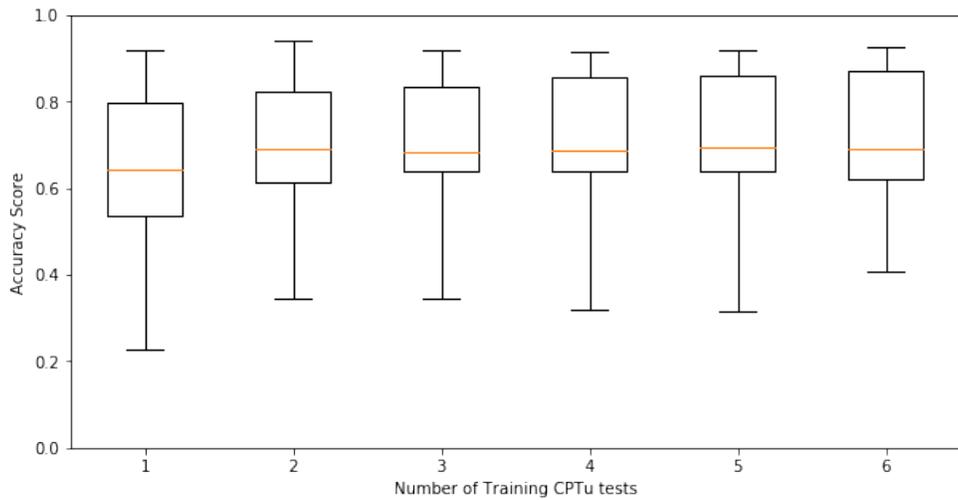
(b)
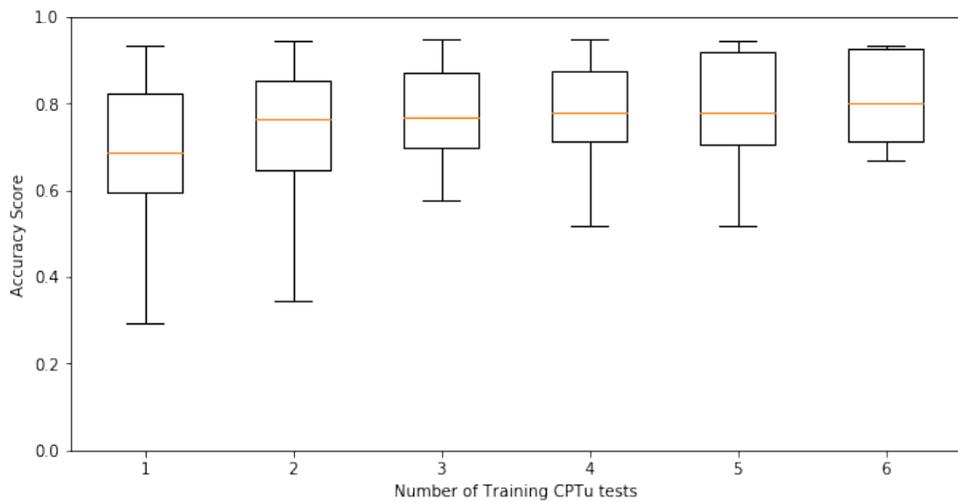
(c)

Figure 5.24: Results of a sequentially trained Logistic Regression classificator on the Vegvesen dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

Table 5.4: Results of "Leave One Group Out" Cross-validation with Logistic Regression, Vegvesen dataset.

"Left Out"	Cross-validation Score		
CPTu	$Q_t - F_r$	$Q_t - U_2$	All
100	88%	87%	88%
102	71%	65%	71%
106	78%	69%	78%
107	66%	41%	67%
149	93%	85%	93%
154	92%	93%	93%
155	79%	62%	80%
Median	79%	69%	80%

Naive Bayes Classifier

The results are shown in Figure 5.25 below. As with the Logistic Regression, the results are more scattered and less accurate, but when using the remaining six tests for training the accuracies increase considerably, as shown in Table 5.5. In the case of the $Q_t - F_r$ classification, the results show a 91% accuracy which is much more accurate than any of the classification charts.
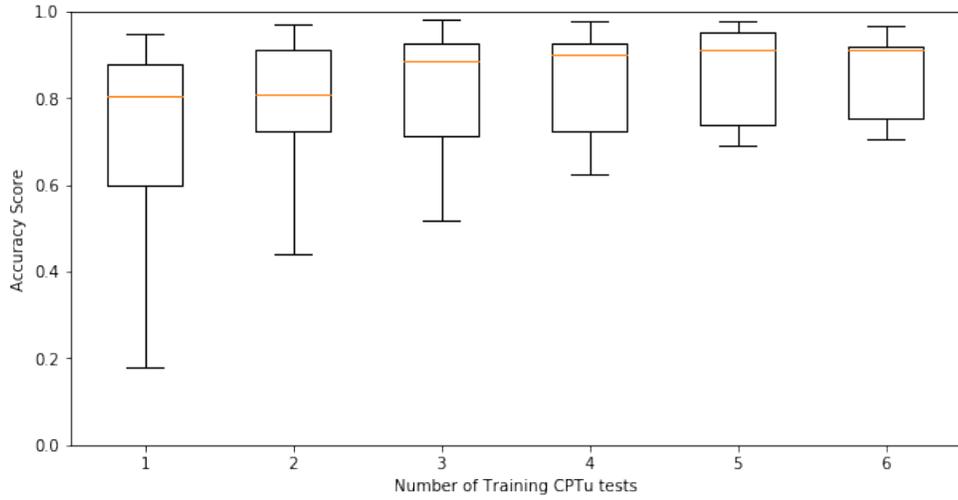
(a)
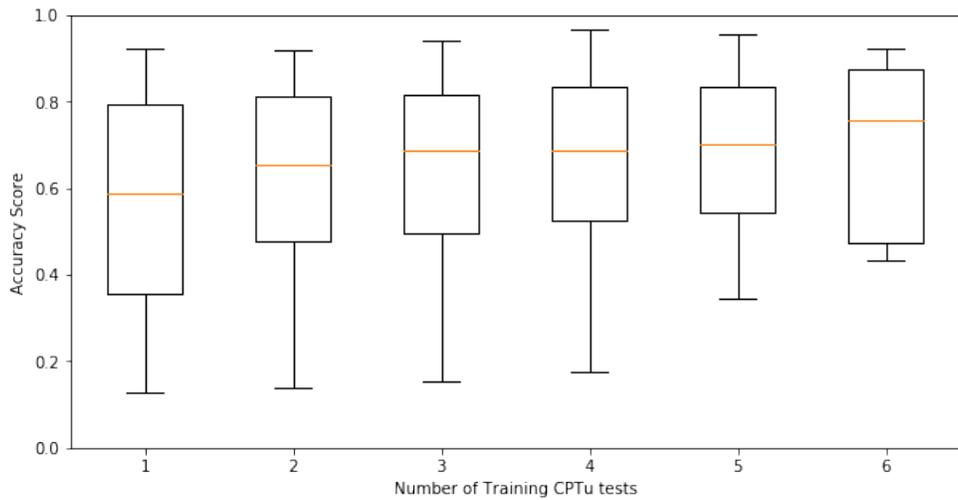
(b)
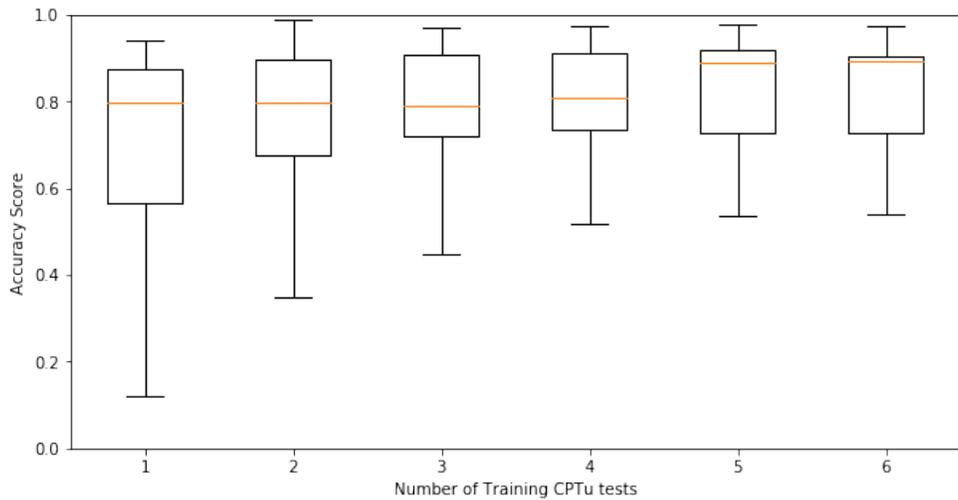
(c)

Figure 5.25: Results of a sequentially trained Naive Bayes classificator on the Vegvesen dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

Table 5.5: Results of "Leave One Group Out" Cross-validation with Naive Bayes, Vegvesen dataset.

"Left Out"	Cross-validation Score		
CPTu	$Q_t - F_r$	$Q_t - U_2$	All
100	92%	92%	90%
102	75%	79%	78%
106	71%	47%	54%
107	72%	43%	73%
149	97%	87%	98%
154	91%	67%	89%
155	93%	76%	92%
Median	91%	76%	89%

Hidden Markov Model

The results of the sequential training and classification are shown in Figure 5.26. While the results of the cross-validation are shown on Table 5.6.
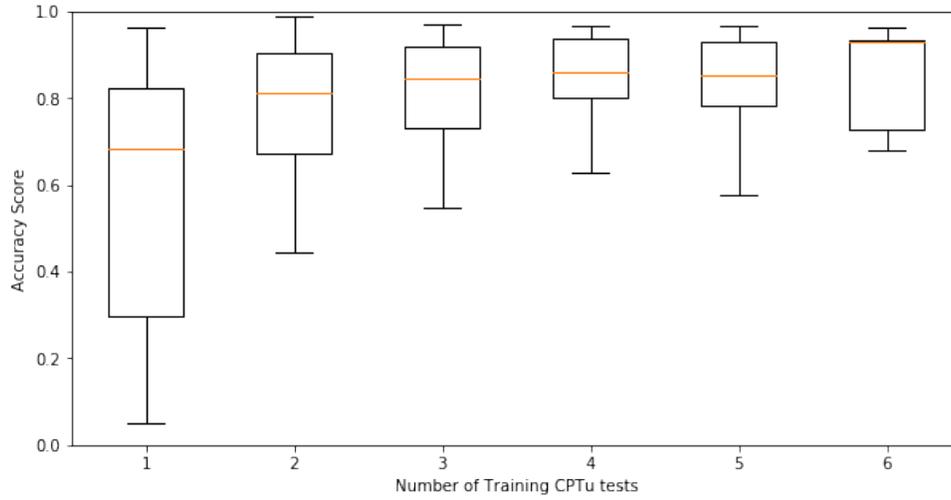
(a)
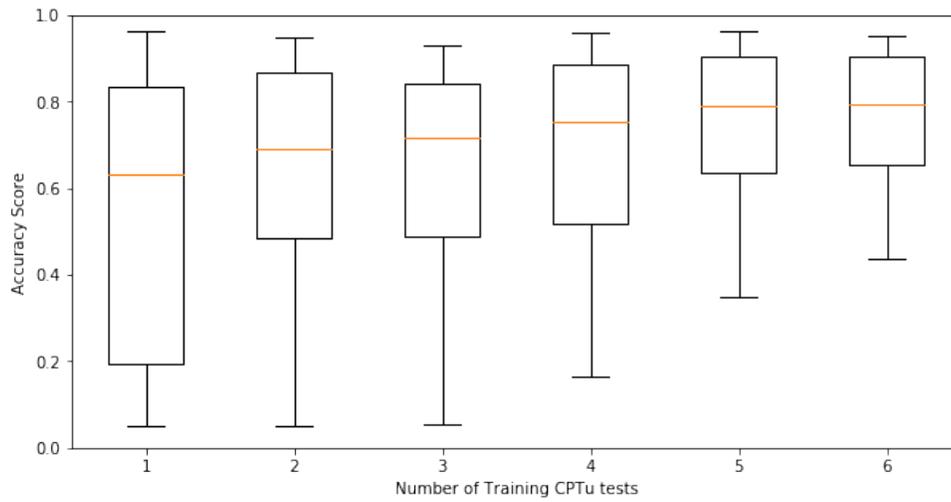
(b)
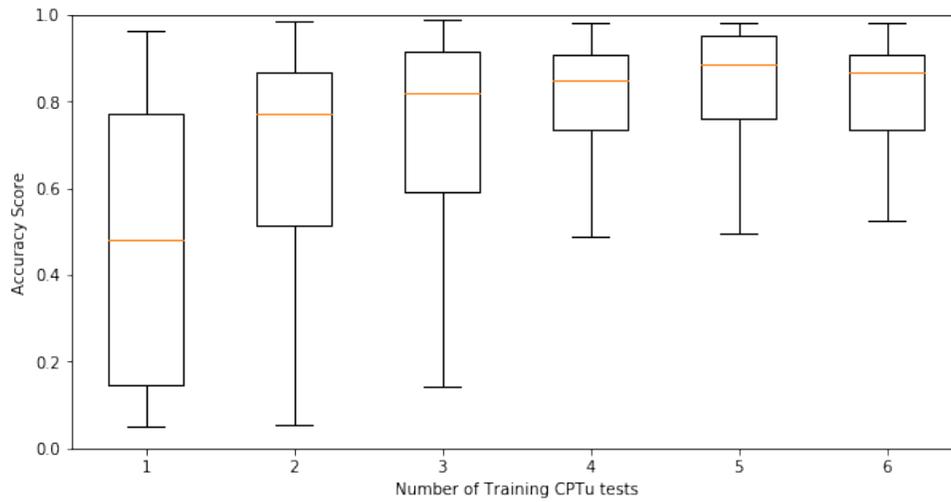
(c)

Figure 5.26: Results of a sequentially trained HMM on the Vegvesen dataset using Q_t^{norm} and F_r^{norm} (a), Q_t^{norm} and U_2^{norm} (b), and all of them (c) as predictors.

Table 5.6: Results of "Leave One Group Out" Cross-validation with the Hidden Markov Model, Vegvesen dataset.

"Left Out" CPTu	Cross-validation Score		
	$Q_t - F_r$	$Q_t - U_2$	All
100	93%	89%	91%
102	87%	85%	87%
106	70%	18%	30%
107	71%	81%	77%
149	67%	44%	48%
154	90%	68%	91%
155	95%	92%	96%
Median	87%	81%	87%

The results show high accuracies that in general are reached after using four tests to train the model.

5.4.4   Site profiles

The results of the profiles estimated by machine learning classification of the CPTu tests from Vegvesen are shown in Figures 5.27 and 5.28, while the detailed figures are attached in Appendix B. Table 5.7 shows the accuracies of the predictions, in this case six tests were used to train the machine learning model to estimate the remaining one, thus the values shown in the table are the same as in the cross-validation part. The mean accuracies are 81% for logistic regression, 84% for Naive Bayes and 82% for HMM, although these values are close to each other, as in the NGTS dataset the Naive Bayes classification shows the highest mean of the three.
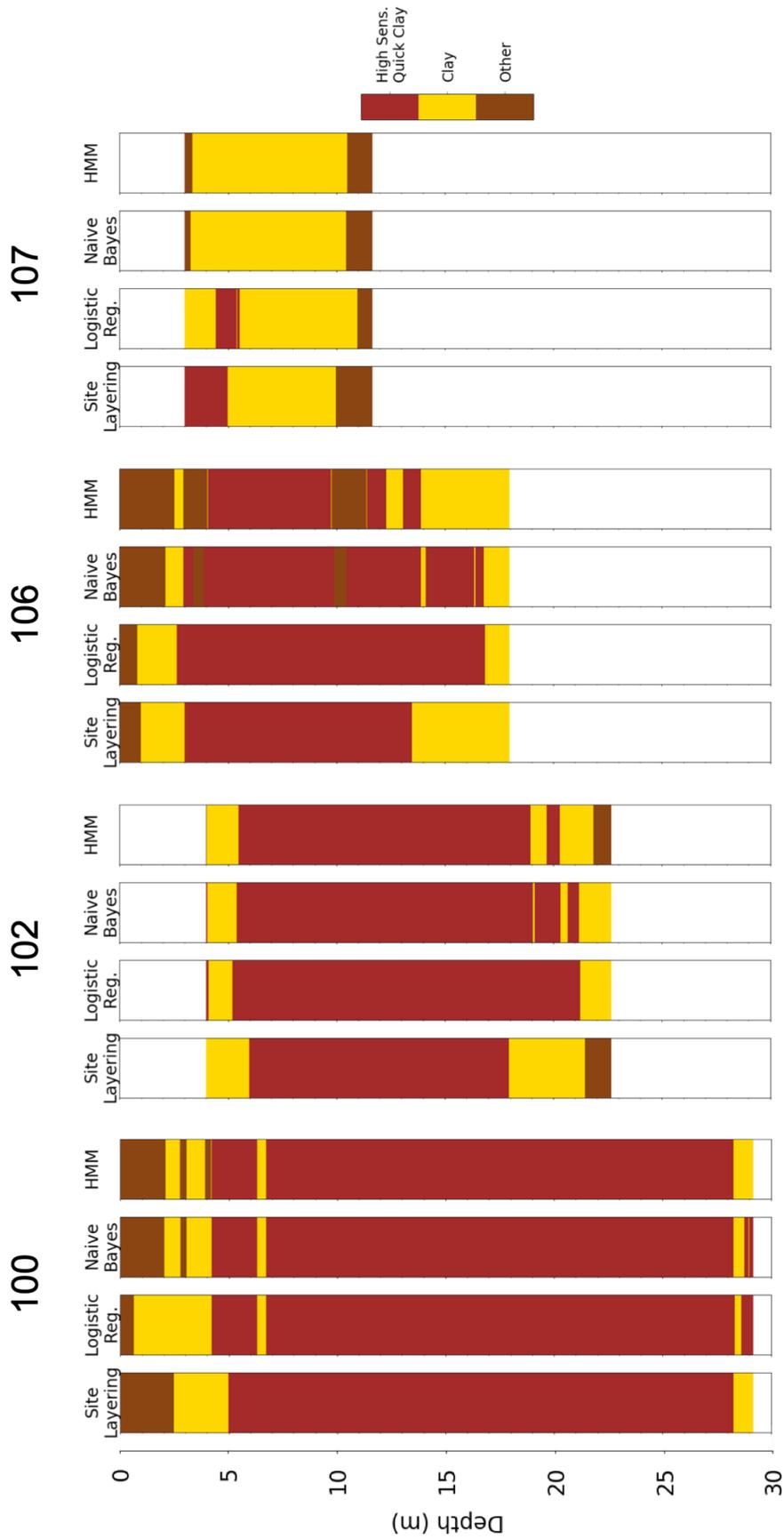
Figure 5.27: Comparison of machine learning classification for the Vegvesen dataset, profile view of the CPTu test 100, 102, 106 and 107.
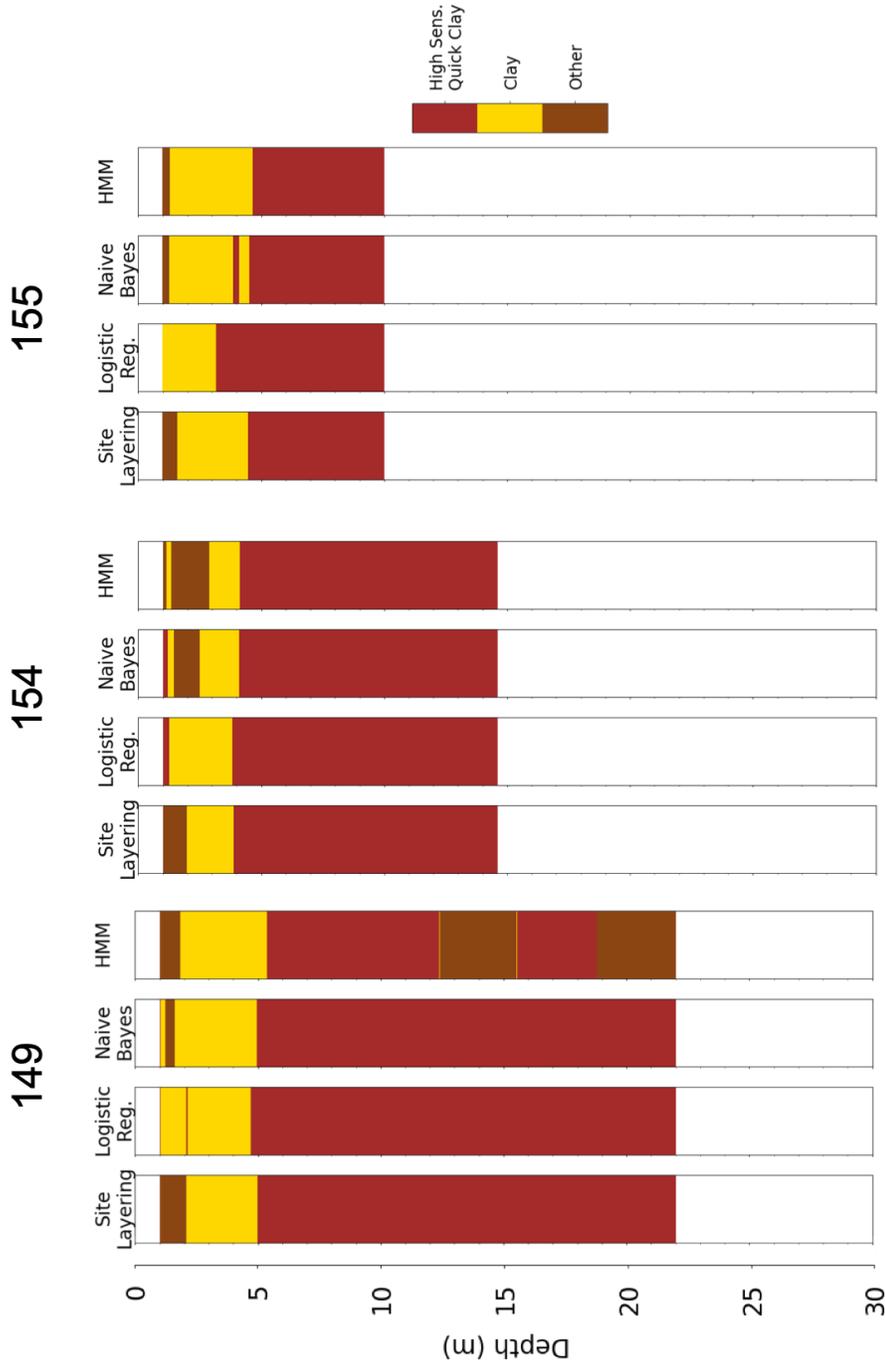
Figure 5.28: Comparison of machine learning classification for the Vegvesen dataset, profile view of the CPTu test 149, 154 and 155.

Table 5.7: Comparison of machine learning classification for the NGTS dataset in terms of accuracy score.

CPTu	Accuracy Score		
	Logistic Reg.	Naive Bayes	HMM
100	88%	92%	93%
102	71%	75%	87%
106	78%	71%	70%
107	66%	72%	71%
149	93%	97%	67%
154	92%	91%	90%
155	79%	93%	95%

5.4.5   Cross-site prediction

In this section, a Machine Learning model trained with the data of one site will be used to predict the layering of the other. Since the dataset from Vegvesen is short, the NGTS site will be used fully to train a Logistic Regression classificator and then predict the soil classes at Vegvesen site.

The results are summarized in the table below, it is possible to see a very poor fit which was expected as stated previously because of the homogeneity of NGTS site makes it hard to train a model with its data that can be used in another site with different soil conditions.

The recommendation here would be in case that there is not enough data from the site in study to properly train a machine learning model, it is better to perform a chart classification rather to use the data from a different site, unless there is evidence that the material will have similar behavior.

However, it would be interesting to keep adding information besides the one defined by these two datasets in order to generate a database which can be big enough to represent several site conditions, at least for the Norwegian sensitive clays.

Table 5.8: Results of using NGTS trained Logistic Regression Classificator on Vegvesen dataset.

CPTu	Accuracy Score		
	$Q_t - F_r$	$Q_t - U_2$	All
100	15%	16%	17%
102	32%	35%	35%
106	42%	40%	40%
107	61%	76%	77%
149	48%	37%	52%
154	8%	9%	10%
155	27%	34%	34%
Median	31%	31%	34%

5.4.6   Clustering with Hidden Markov Model

Another way to use the Hidden Markov Model concept is to perform clustering of the data, this is done in an unsupervised way with the only input of the target number of clusters (soil types). In general, when there is no information about how many clusters are there in the dataset, the elbow method is used to identify the most appropriate one. This is done by plotting the within-cluster dissimilarities (which is a measure of the distance of the cluster data to the cluster center) against the number of clusters (k), when there are too many clusters the dissimilarity is reduced since all the cluster data is close to its center so adding more clusters does not reduce it noticeably. The most appropriate number of clusters is the one that if one extra cluster is added does not significantly reduce the dissimilarity, this is graphically represented as a bend in a descending curve (or elbow). Only two parameters, Q_t^{norm} and F_r^{norm}, are going to be used in this clustering exercise. The HMM follows the same conditions as explained on 5.4.2 with the exception that two different models of covariance were used: tied (same covariance matrix for all classes) and full (different covariance matrix for each class), this was done in order to allow more flexibility in the clustering process. The elbow chart obtained by clustering Q_t^{norm} and F_r^{norm} from the Vegvesen dataset is shown in Figure 5.29. It is possible to see that after four clusters the dissimilarities start to reduce less for both covariance model used, so four clusters are chosen as the optimal number.
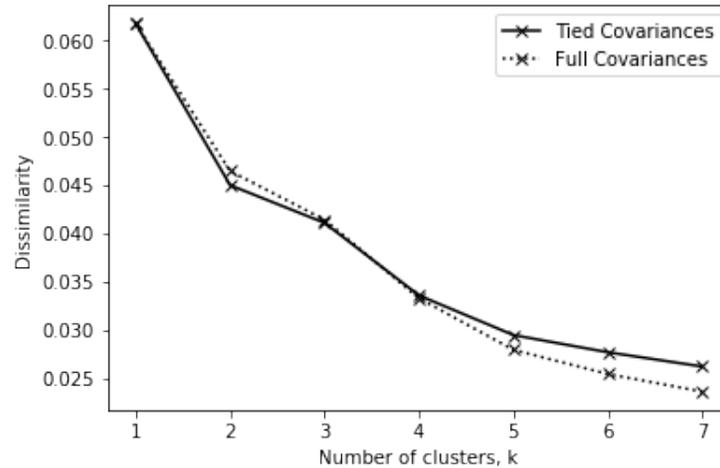
Figure 5.29: Elbow chart for clustering of Vegvesen dataset using HMM.

The data clustered alongside the likelihood function of each cluster is shown in Figure 5.30, it is possible to see that a tied covariance model reflects better what is expected from a classification of soft soils, for example, clusters 1 and 2 could represent sensitive/very soft soils, cluster 3 represents clayey soils and cluster 4 represents more competent material (dry crust or sandy clay/silt).

The clustering scheme proposed in the previous paragraph is chosen and tested against the proposed layering for the sites. The result in terms of accuracy score is 85% which is a very high value considering that the model was not trained with the actual classes and comparing with the classification chart is better than any of them, however, engineering judgement is required in order to know what represents each cluster and if two or more clusters need to be grouped together (which is different to perform a clustering analysis with fewer groups). Figure 5.30, shows a comparison between the site layering and the one obtained by the clustering analysis done.

This kind of analysis can help geotechnical engineers to interpret CPTu test in a better way by grouping data with similar characteristic before classifying it whether using only the data measured or by complementing it with laboratory tests.
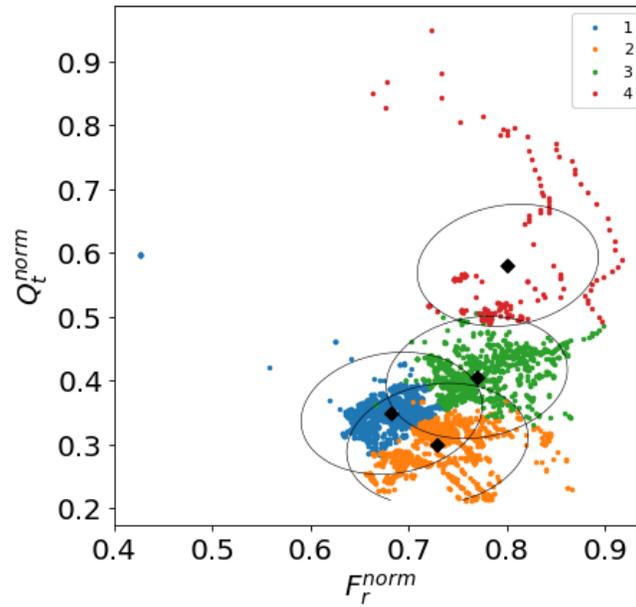
(a)
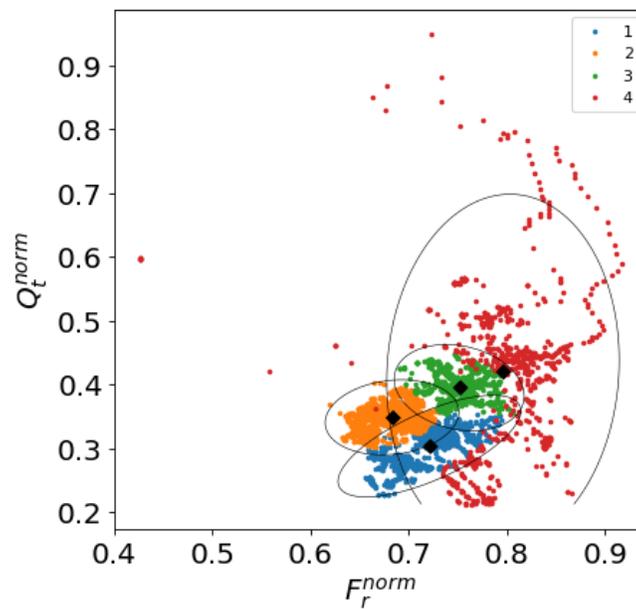
(b)

Figure 5.30: Results of the clustering of Vegvesen dataset using HMM for tied covariance (a) and full covariance (b) models. Different colors show different clusters while ellipses show a contour of the likelihood function of each cluster.
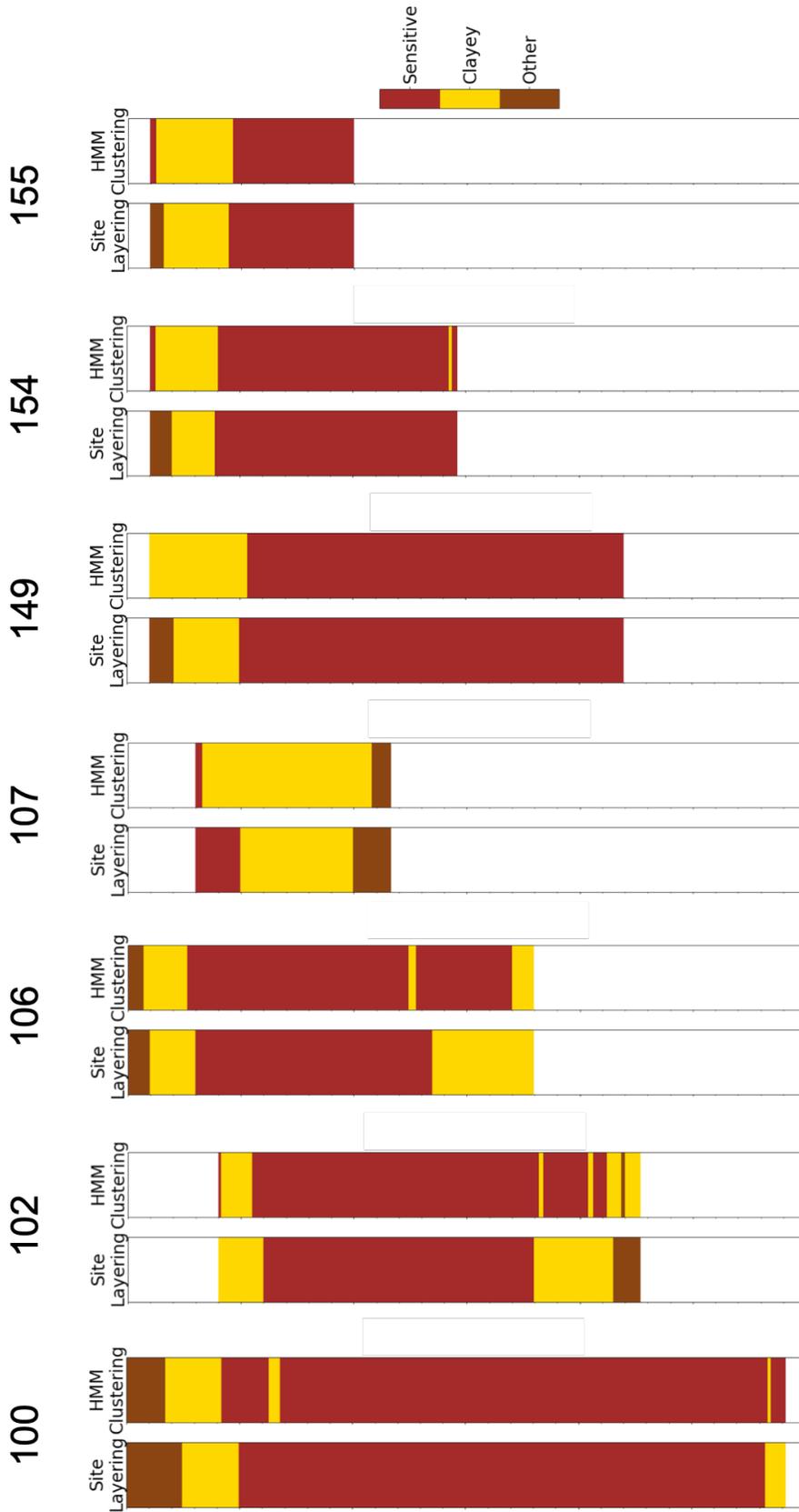
Figure 5.31: Site layering versus layering obtained from the clustering of Vegvesen dataset using HMM for tied covariance.

Chapter 6

Summary and Recommendations for Further Work

6.1 Summary and Conclusions

This thesis gave insight into the application of statistical methods on geotechnical site characterization.

The main geostatistical methods were programmed in Python and used to interpolate (and extrapolate) CPTu tests measurements performed in Rissa slope. The major challenge of this task was the determination of horizontal variograms since the data resolution in the horizontal plane is considerably less compared with the vertical one. This was solved by constraining the variograms with the total variance and assuming a geometric anisotropy. The results of the interpolation show a good agreement between the distribution of the different parameters and the layers limits defined by a previous study done in the area. Cross-validation results show low standardized errors, even though the model is unable to predict outliers in the measurements.

Regarding machine learning, a set of Python libraries were used to train three different learning algorithms: Logistic Regression, Naive Bayes and Hidden Markov Models. The results show the ability of this kind of methods to learn from the data as it is incorporated to the model, in particular, good classification accuracies are reached with just four training CPTu tests.

However, these methods are not meant to be used as general classification solutions (unless they are trained with a big dataset that includes different soil conditions). The major challenge here is to get enough data to train the models and also have enough laboratory results to test it.

Finally, it would be interesting to keep researching these methodologies and their application in the geotechnical field since they have proved to yield good results that can help engineers to optimize both field and laboratory tests.

6.2 Discussion

Geotechnical field investigations are, in general, discrete and, mainly because of economic constraints, they represent only a small fraction of the area to be studied. In that scenario, geotechnical engineers should find ways to assess how the properties are distributed within the area of interest. Statistical methods can be a really useful toolbox to help engineers to gain a better understanding of both the property distribution and the uncertainties associated, and also to make predictions based on the information already gathered.

6.2.1 Kriging

In the present work, the use of kriging was studied, in particular, its most widespread variant: ordinary kriging. Simple kriging as its name implies is a simple method, which gives a minimum variance unbiased estimator of a random function. The main drawback is that it needs the knowledge of the mean in its formulation. Ordinary kriging overcomes this problem without many additional complexities. When the data have a trend, universal kriging can handle it by default, but at the expense of a more complicated formulation and better knowledge of the physical phenomenon in study. In general, the ordinary kriging implicitly takes into account the trend through a constant in the neighborhood mean, as the neighborhood moves the trend changes accordingly.

Regarding Rissa slope, since CPTu is a vertical sounding, the variogram estimation in that direction is straightforward since there is abundant data to use. The opposite happens in the

horizontal direction by cause of the reduced data in it (each CPTu location means only one distance in the horizontal variogram). Constraining the variogram through the total variance of the data and assuming a geometric anisotropy, allowed to fit properly both variograms. It might be a good idea to use as much data as possible for a certain location, even though it is not part of the estimation profile. If the data used is not aligned in the profile, as in this case, care should be taken in compute the variogram with the appropriate distance between the sampled points, this was taken into consideration when programming the method to compute it.

With reference to the results of the interpolation of Rissa CPTu tests using ordinary kriging, they are considered adequate since they capture the layering in an appropriate way. The variance plots show that reducing the spacing between the test improves the accuracy of the estimations (reducing the estimation variance).

The cross-validation shows that only one test (C3) can be dropped without significantly influencing the estimation results, however, the cross-validation errors are centered around zero, and the biggest errors are associated with isolated outliers that are rather difficult to estimate correctly, being this a problem that will affect any interpolation method.

6.2.2   Machine Learning

In general, classification charts meant for broad use fail in capture special soils like quick or sensitive clays, especially Norwegian sensitive clays which have distinctive features compared with other soils with similar behavior (for example Canadian quick clays). Furthermore, classification charts use each data point individually without taking into account the spatial correlation between soils in the same layer.
This study used three machine learning approaches to study a methodology to improve sensitive soil determination from CPTu tests. The approaches studied were Logistic Regression, Naive Nayes and Hidden Markov Model Classifications.
Logistic Regression uses a linear model to create a decision boundary which separates different classes, Naive Bayes approach uses a probabilistic framework based on Bayes Theorem with the simplification of conditional independence of the data, finally, HMM uses Markov Chains

and a probabilistic framework to model the spatial correlation between measured data. The approaches were tested in two different datasets, comprising soils of different characteristics. In the case of the NGTS site, where the soil layering was regular and layers characteristics are homogeneous, the three methods show excellent performance measured through an accuracy score well above 90%. Even though the three methods score high accuracies, the HMM is slightly lower than the other two, particularly low accuracies can be seen in two profiles: C05 and C17, as shown in Table 5.3 and Figure 5.22. The advantage of HMM is that it takes into account the spatial connection between the data and that can be noticed in less presence of thin layers within the bigger ones. Vegvesen dataset has fewer data available to train the models, but results also are quite favourable with accuracy scores above 80%.

The sequential training of the models, shows graphically how they learn from the data as it is incorporated, both datasets show that using just four tests to train the model improves significantly the accuracy of the classification. This can be helpful when performing site investigations where the model can be trained as the data is retrieved from the soundings and laboratory tests or even from in-place visual classification. Once the model is successfully trained, the necessity of laboratory test might be reduced. This works even in not so homogeneous datasets such as the one from Vegvesen. Also, a small test of clustering through Hidden Markov Model was performed yielding good results and showing an application of this approach that can help engineers improve the definition of geotechnical layers based on field test measurements such as CPTu.

The drawbacks of the machine learning approach are related with the fact that they use data to learn making them a site-specific solution rather than a global one. It is seen clearly on the decision boundaries defined by Naive Bayes classification (which was the highest accuracy method) representing a solution that can not be used beyond the data which defined it, and also in the low accuracies of the classification of Vegvesen dataset using a model trained with the NGTS one. A solution for that would be to incorporate as many data as possible for the case in study, for example, the identification of quick clays.

6.3   Recommendations for Further Work

Despite statistical approaches in geotechnical engineering and, in particular, its application to site characterization, are being used increasingly, there is still plenty of space for it to grow inside the geoengineering field. There are many ideas and methodologies that would have contributed to enhancing the scope of this thesis but due to time and technical limitations they were not possible to implement. Here are presented a list of recommendations for those who might be interested in further research in this field.

1. Study on potential improvements of the horizontal variogram calculation through constraining it with complementary data. In Rezvandehy and Deutsch (2018) it is explained how to use the vertical variogram and geophysical testing to constrain the horizontal one and quantify the uncertainty associated with it. Solberg et al. (2012) made resistivity measurements in almost the same profile at Rissa, this data could be valuable to perform this task.

2. An interesting step forward in this research would be the study and implementation of kriging with external drift and cokriging. They represent different approaches to handle secondary information which might be better sampled than the one in study. This fits perfectly in geotechnical engineering since, usually, geotechnical exploration comes along with geophysical field tests which have a better resolution in the space of study and is cheaper than geotechnical soundings. In Goovaerts (1997) and Olea (1999) there can be found the theoretical basis and the algorithms for both methodologies.

3. Furthermore, kriging can be used as a stochastic simulation tool, which can be useful in slope stability analyses or any other geotechnical assessment that requires uncertainty handling. The theory behind this and examples of application can be found in Olea (1999) and Emery (2007).

4. The machine learning part of this thesis can be improved by testing different methods than the ones used here, the Scikit-learn library has a big set of algorithms already available for the user, and many others that can be customized to fit particular needs. A first approach would be to use Naive Bayes with non-Gaussian distributions, Support

vector machines with different kernel functions (which can improve the linear decision boundaries obtained with Logistic Regression), Gaussian Mixture Models, Hidden Markov Models with custom emission probability, and so on.

5. Regarding the machine learning classification, it would be interesting to make a big database which comprises all the tests performed on Norwegian quick and highly sensitive clay sites. In this way a predictive model can be trained which incorporates all the information available, and check if the predictions on new sites have better accuracy than the charts generally used.

6. In geotechnical practise, it would be interesting to use both approaches in a linked way. In that sense it is recommended for future work to study the alternatives of implementing the whole workflow from CPTu data read and conversion, kriging interpolation, normalized parameters calculation, machine learning pre-processing, machine learning classification and geotechnical layering.

Bibliography

Asiri, S. (2018). Machine Learning Classifiers. `https://towardsdatascience.com/`.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363.

Davis, G. J. and Morris, M. D. (1997). Six factors which affect the condition number of matrices associated with kriging. *Mathematical geology*, 29(5):669–683.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Deutsch, C. V. (2002). Geostatistical reservoir modeling.

Emery, X. (2007). Conditioning simulations of Gaussian random fields by ordinary kriging. *Mathematical Geology*, 39(6):607–623.

Emery, X. (2016). Geoestadistica. Lecture notes (in Spanish).

Emery, X. (2019). Personal communication.

Eslami, A. and Fellenius, B. H. (1997). Pile capacity by direct CPT and CPTu methods applied to 102 case histories. *Canadian Geotechnical Journal*, 34(6):886–904.

Farrokhzad, F., Choobbasti, A. J., and Barari, A. (2010). Artificial neural network model for prediction of liquefaction potential in soil deposits.

Fenton, G. A. (1999). Random field modeling of CPT data. *Journal of geotechnical and geoenvironmental engineering,* 125(6):486–498.

Firouzianbandpey, S., Griffiths, D., Ibsen, L., and Andersen, L. (2014). Spatial correlation length of normalized cone data in sand: case study in the north of Denmark. *Canadian Geotechnical Journal,* 51(8):844–857.

Firouzianbandpey, S., Ibsen, L., Griffiths, D., Vahdatirad, M., Andersen, L., and Sørensen, J. (2015). Effect of Spatial Correlation Length on the Interpretation of Normalized CPT Data Using a Kriging Approach. *Journal of Geotechnical and Geoenvironmental Engineering,* 141:4015052.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data.* Cambridge University Press.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE,* 61(3):268–278.

García, S., Ovando-Shelley, E., Gutiérrez, J., and García, J. (2012). Liquefaction Assessment through Machine Learning Approach. In *15th World Conf. Earthq. Eng.*

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation.

Gregersen, O. (1981). The quick clay landslide in Rissa, Norway. *Norwegian Geotechnical Institute Publication,* 135:1–6.

Gylland, A. S., Sandven, R., Montafia, A., Pfaffhuber, A. A., Kåsin, K., and Long, M. (2017). CPTU classification diagrams for identification of sensitive clays. In *Landslides in Sensitive Clays,* pages 57–66. Springer.

hmmlearn developers (2010). hmmlearn: Unsupervised learning and inference of Hidden Markov Models. `https://hmmlearn.readthedocs.io/en/latest/#`.

Journel, André G; Huijbregts, C. J. (1978). *Mining geostatistics.* Academic Press, London.

Kornbrekke, H. A. (2012). *Skråningsstabilitet ved Rein Kirke med utgangspunkt i resultater fra Sherbrooke blokkprøver.* PhD thesis, NTNU.

Krogstad, A., Depina, I., and Omre, H. (2018). Cone penetration data classification by Bayesian inversion with a Hidden Markov model. In *Journal of Physics: Conference Series*, volume 1104, page 12015. IOP Publishing.

Lunne, T., Powell, J. J. M., and Robertson, P. K. (2002). *Cone penetration testing in geotechnical practice.* CRC Press.

Marsland, S. (2011). *Machine learning: an algorithmic perspective.* Chapman and Hall/CRC.

Mosley, L. (2013). A balanced approach to the multi-class imbalance problem.

Nadim, F. (1988). Geotechnical site description using Stochastic Interpolation. *Norwegian Geotechnical Institute Publication*, (176).

NGI (2019). NGTS - Norwegian Geotest Sites. `https://www.ngi.no/eng/Projects/NGTS-Norwegian-Geo-Test-Sites`.

Olea, R. A. R. A. (1999). Geostatistics for engineers and earth scientists.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and others (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Puri, N., Prasad, H., and Jain, A. (2018). Prediction of Geotechnical Parameters Using Machine Learning Techniques. *Procedia Computer Science*, 125:509–517.

Rabarijoely, S., Bilski, P., and Falkowski, T. (2007). Usage of the graph clustering algorithm to the recognition of geotechnical layers. *Annals of Warsaw University of Life Sciences - Sggw. Land Reclamation*, 38:57–67.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rezvandehy, M. and Deutsch, C. V. (2018). Horizontal variogram inference in the presence of widely spaced well data. *Petroleum Geoscience*, 24(2):219–235.

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

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

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

Samui, P. and Sitharam, T. G. (2011). Machine learning modelling for predicting soil liquefaction susceptibility. *Natural Hazards and Earth System Sciences*, 11(1):1–9.

Schneider, J. A., Randolph, M. F., Mayne, P. W., and Ramsey, N. R. (2008). Analysis of Factors Influencing Soil Classification Using Normalized Piezocone Tip Resistance and Pore Pressure Parameters. *Journal of Geotechnical and Geoenvironmental Engineering*, 134(11):1569–1586.

Scikit-learn developers (2019). Scikit-learn user guide. Release 0.21.2. `https://scikit-learn.org/stable/user_guide.html`.

Solberg, I. L., Dalsegg, E., L'Heureux, J. S., and Rønning, J. S. (2012). Resistivitetsmålinger for løsmassekartlegging ved Rein kirke i Rissa, Sør-Trøndelag. Data og tolkninger. NGU Report 2012.018. Technical report.

Statens Vegvesen (2013). Fv. 715 Keiserås - Olsøy, Parsell: Leksvik grense - Olsøy Datarapport. Nr. 2012039995-009. Technical report.

Stuedlein, A. W., Kramer, S. L., Arduino, P., and Holtz, R. D. (2012). Geotechnical characterization and random field modeling of desiccated clay. *Journal of Geotechnical and Geoenvironmental Engineering*, 138(11):1301–1313.

Uzielli, M., Mayne, P., and Facciorusso, J. (2010). Spatial characterization of Vs at Amherst NGES site from SCPT using Bayesian kriging. *2nd International Symposium on Cone Penetration Testing*, 2(May):8.

Uzielli, M., Vannucchi, G., and Phoon, K. K. (2005). Random field characterisation of stress-nomalised cone penetration testing parameters. *Geotechnique*, 55(1):3–20.

Van Rossum, G. (1995). Python reference manual.

Vezhapparambu, V. S., Eidsvik, J., and Ellefmo, S. (2018). Rock Classification Using Multivariate Analysis of Measurement While Drilling Data: Towards a Better Sampling Strategy. *Minerals*, 8:384.

Wickremesinghe, D. S. (1989). *Statistical characterization of soil profiles using in situ tests.* PhD thesis, University of British Columbia.

Wikipedia contributors (2019a). Logistic regression — {Wikipedia}{,} The Free Encyclopedia.
`https:`
`//en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=898833411.`

Wikipedia contributors (2019b). Moving average — {Wikipedia}{,} The Free Encyclopedia.
`https://en.wikipedia.org/w/index.php?title=Moving_average&oldid=897012086.`

Wolebo, A. P. (2016). *Advanced Probabilistic Slope Stability Analysis on Rissa Slope.* PhD thesis, NTNU.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, 1(2):3.

Appendix A

CPTu Tests Detailed Plots

NGTS CPTu C01
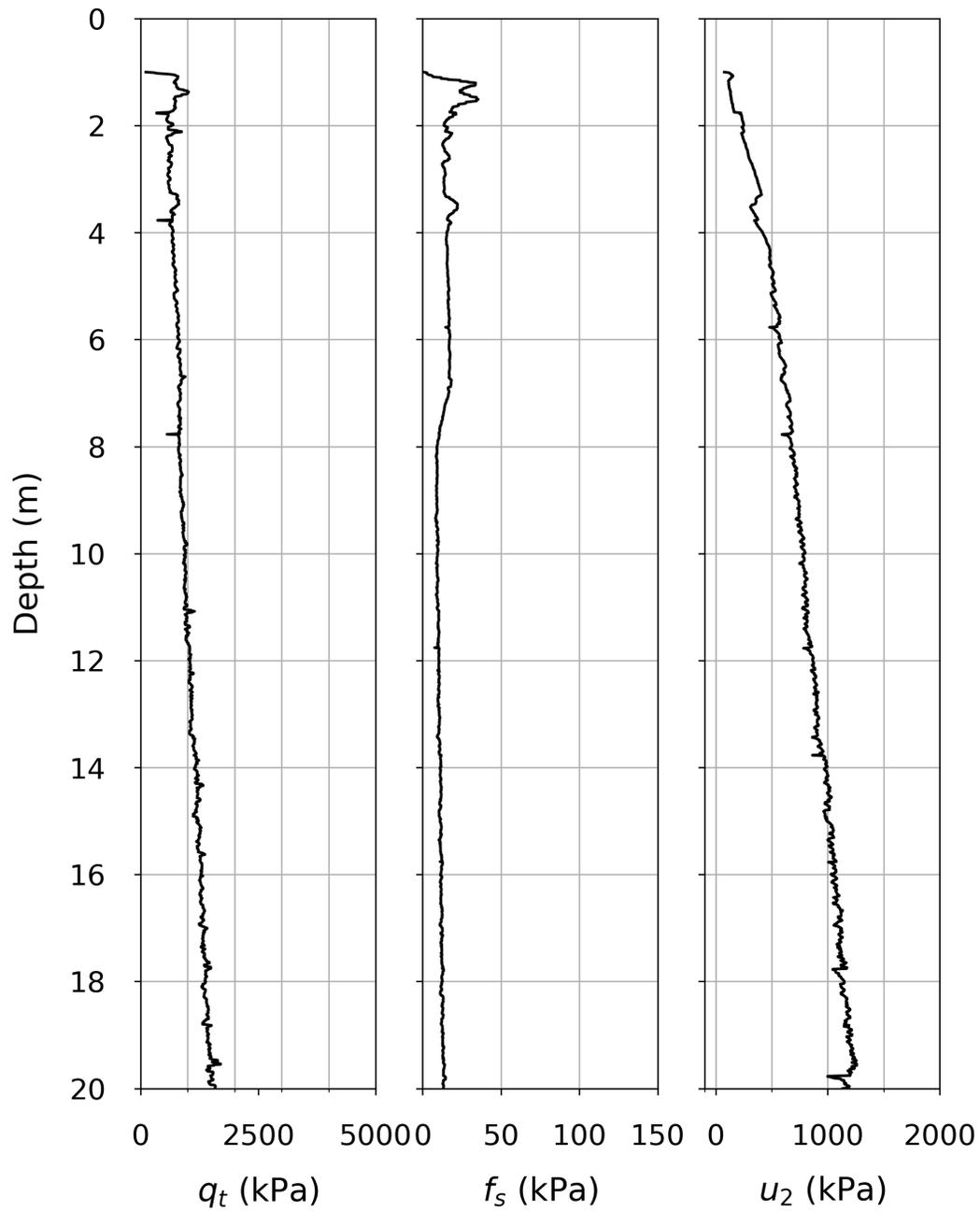
NGTS CPTu C02
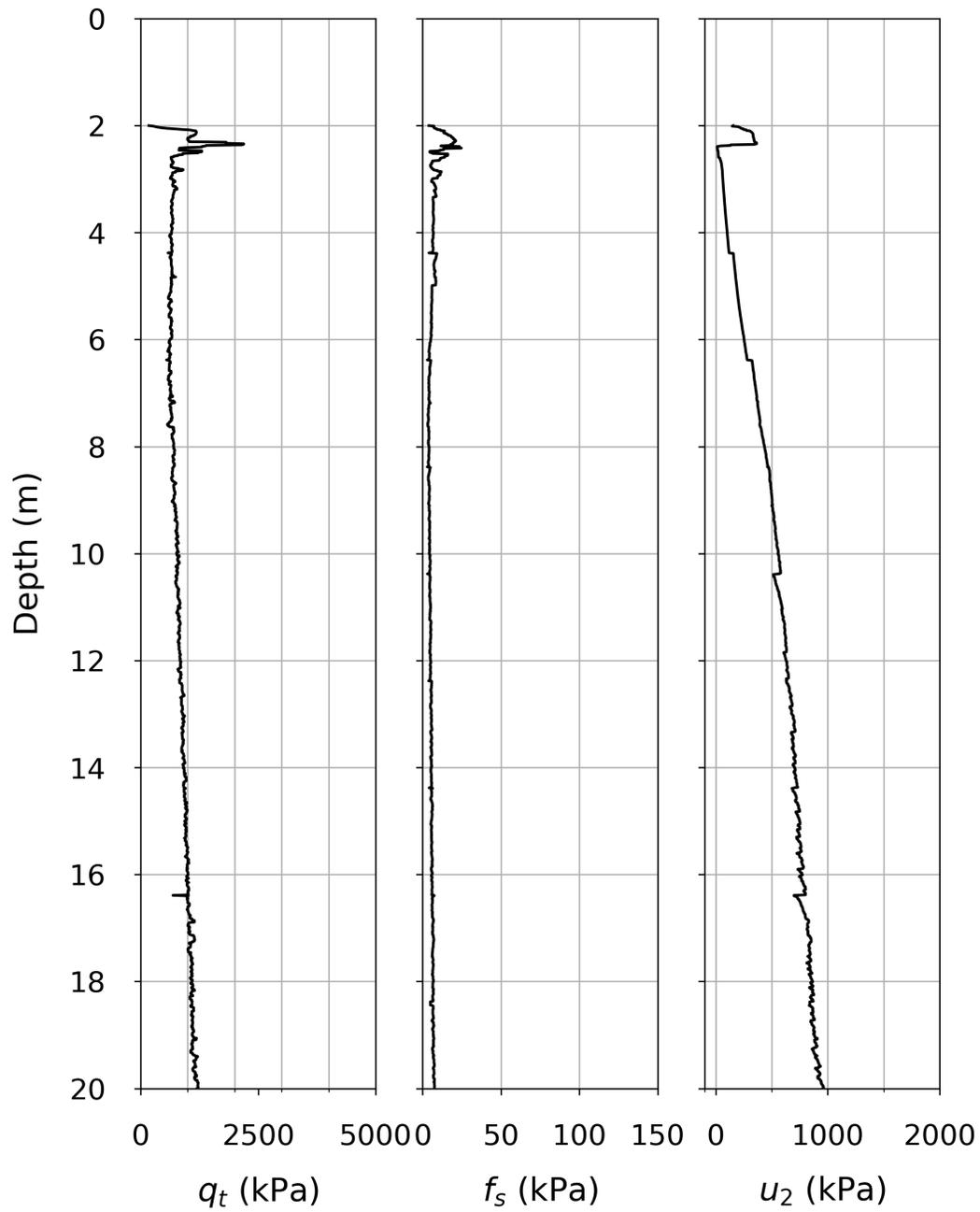
NGTS CPTu C03
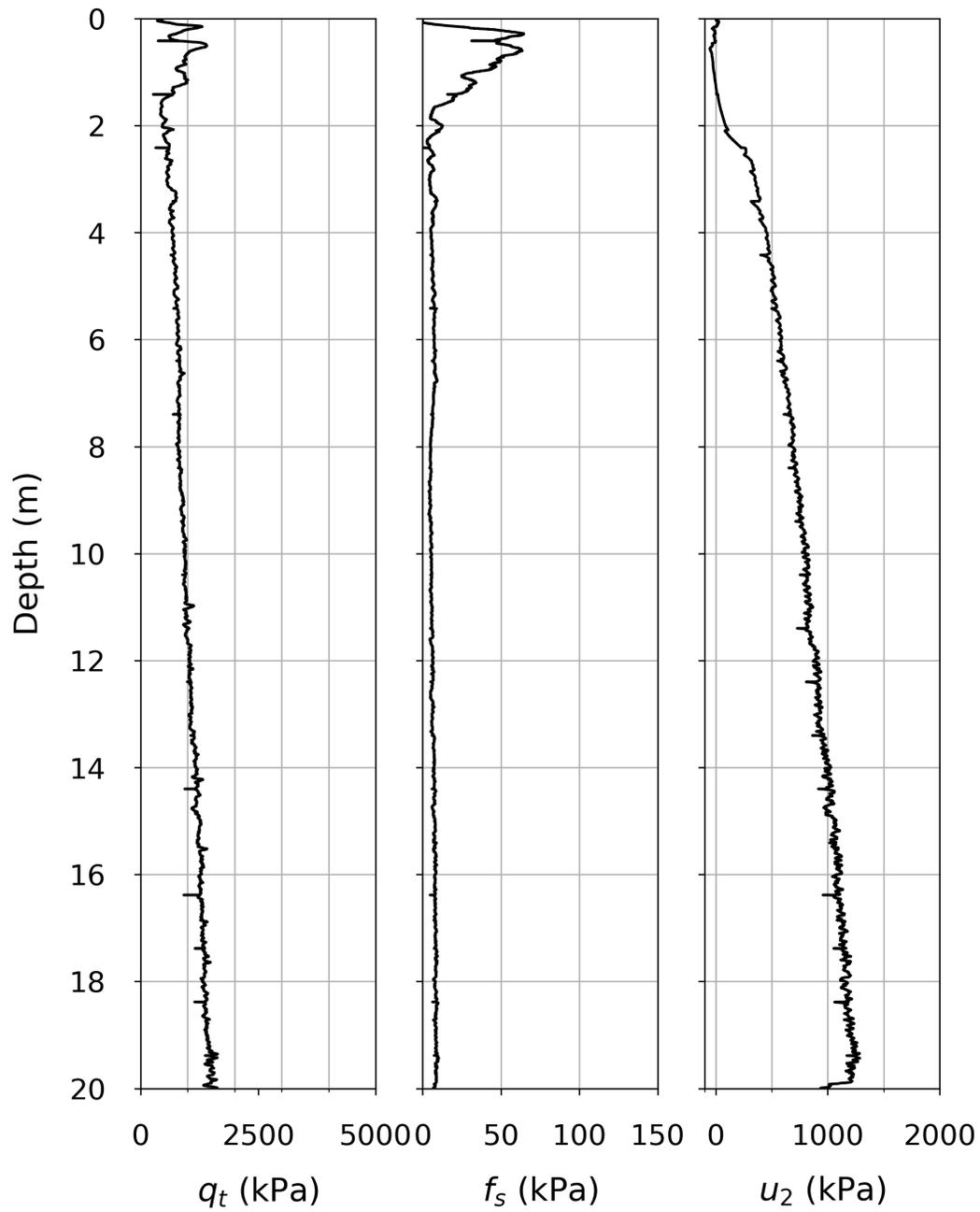
NGTS CPTu C04
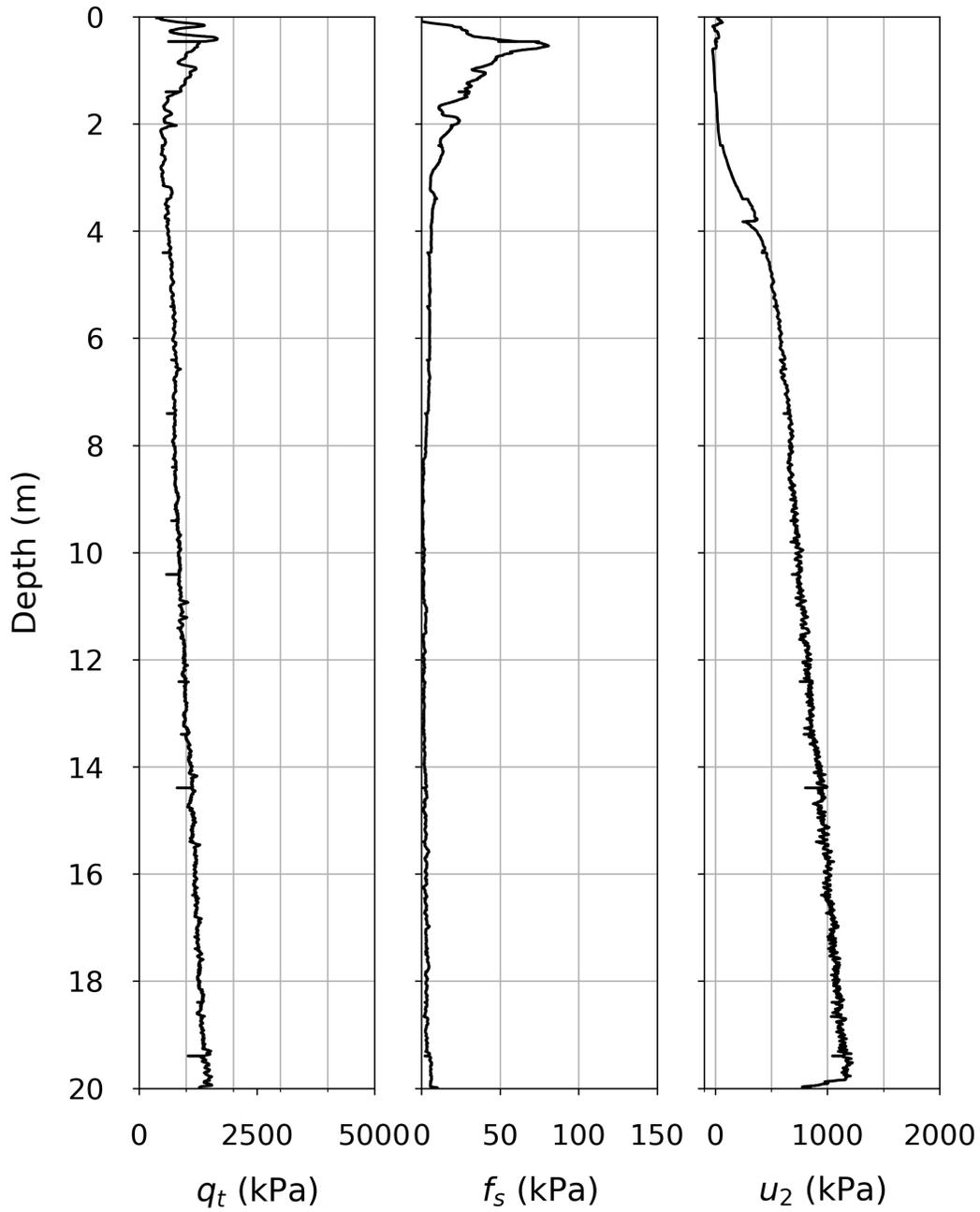
NGTS CPTu C05
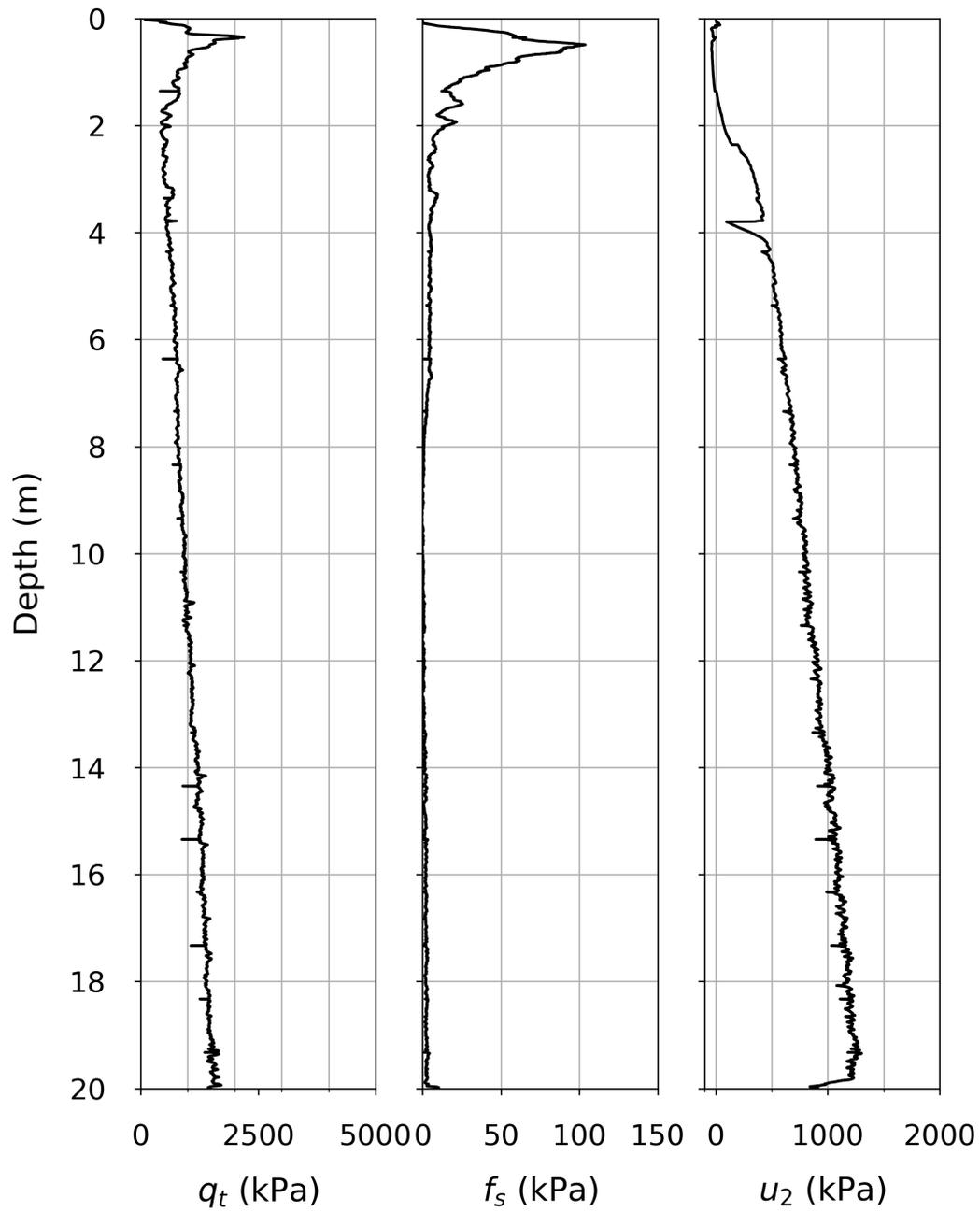
NGTS CPTu C06
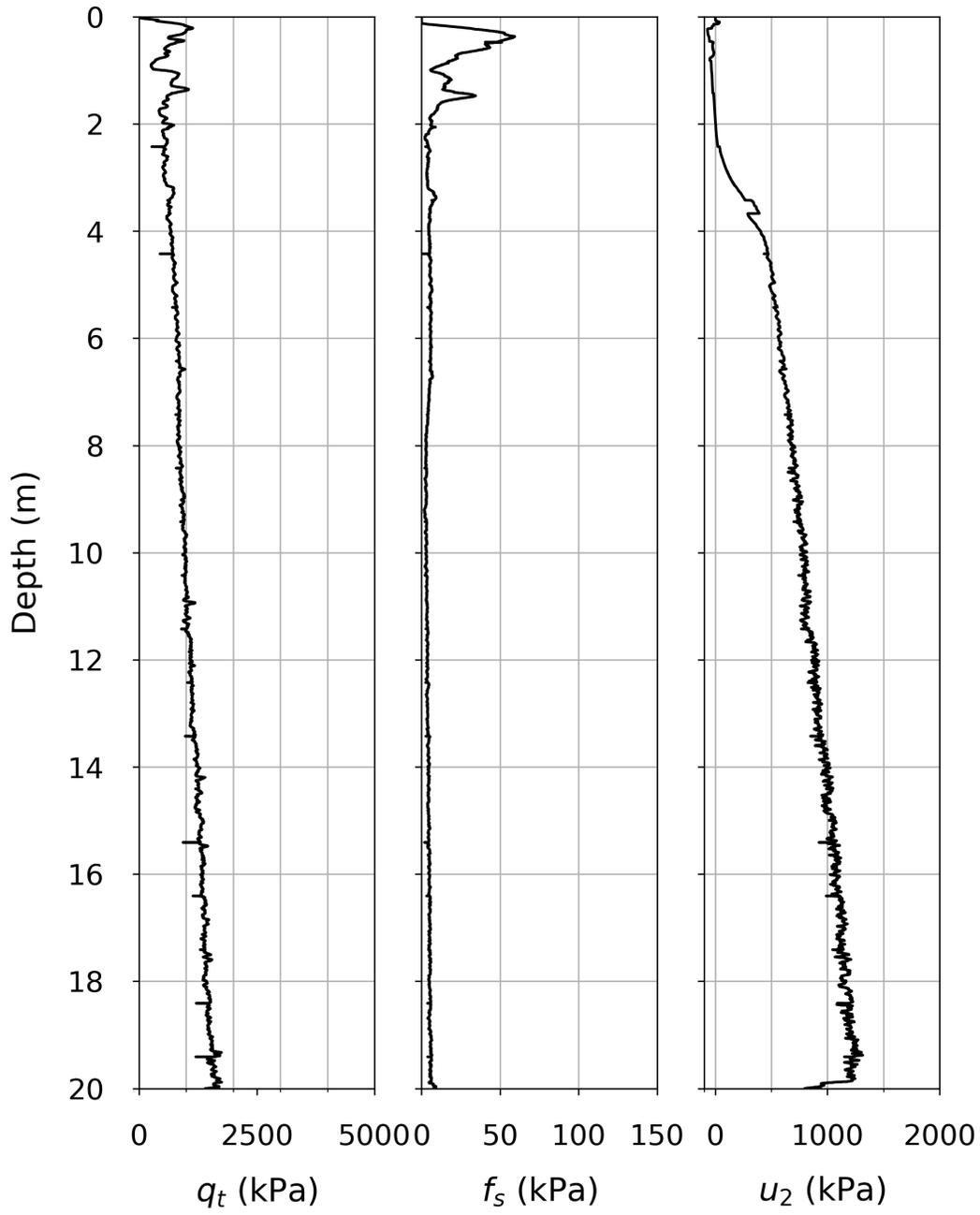
NGTS CPTu C07
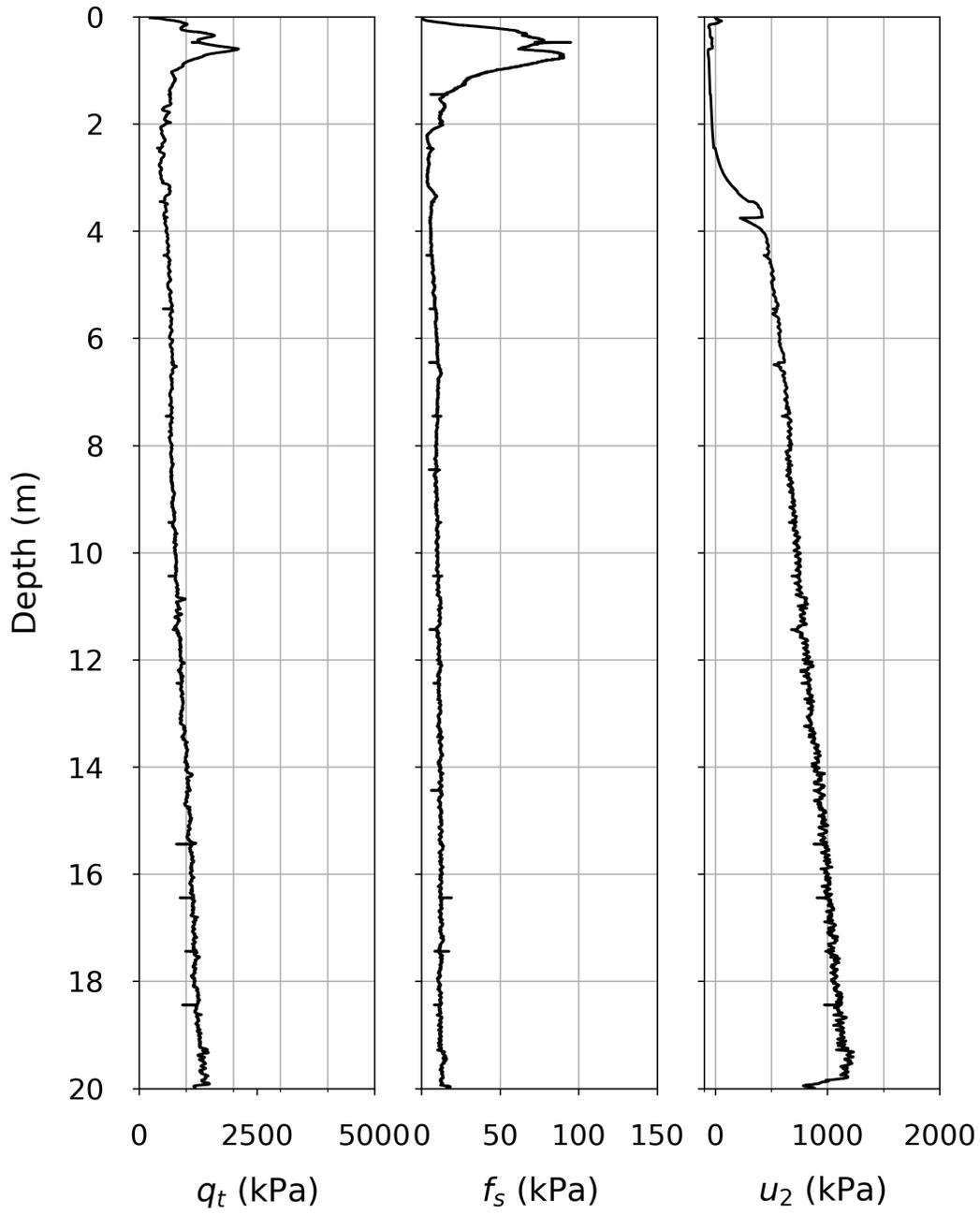
NGTS CPTu C08
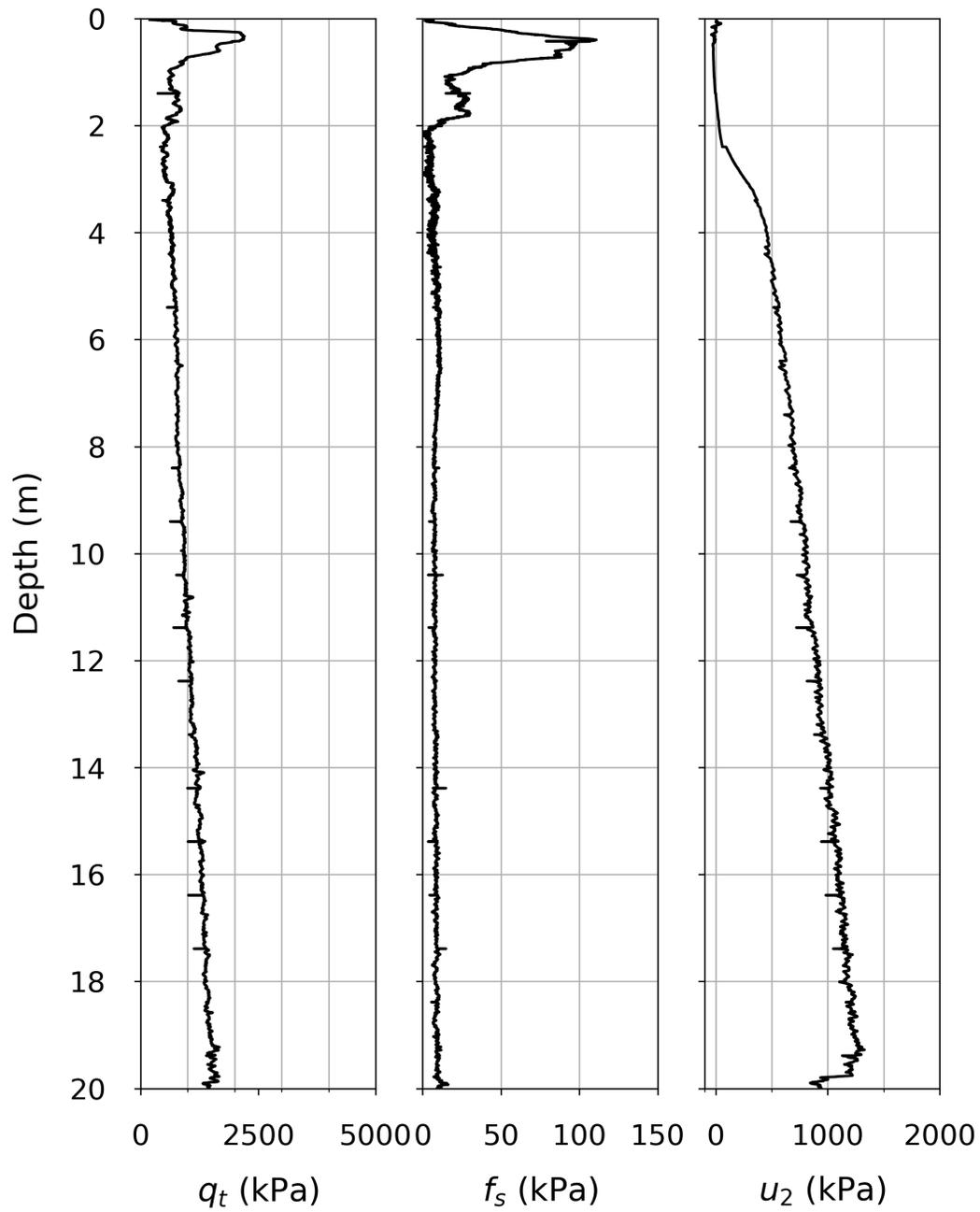
NGTS CPTu C09
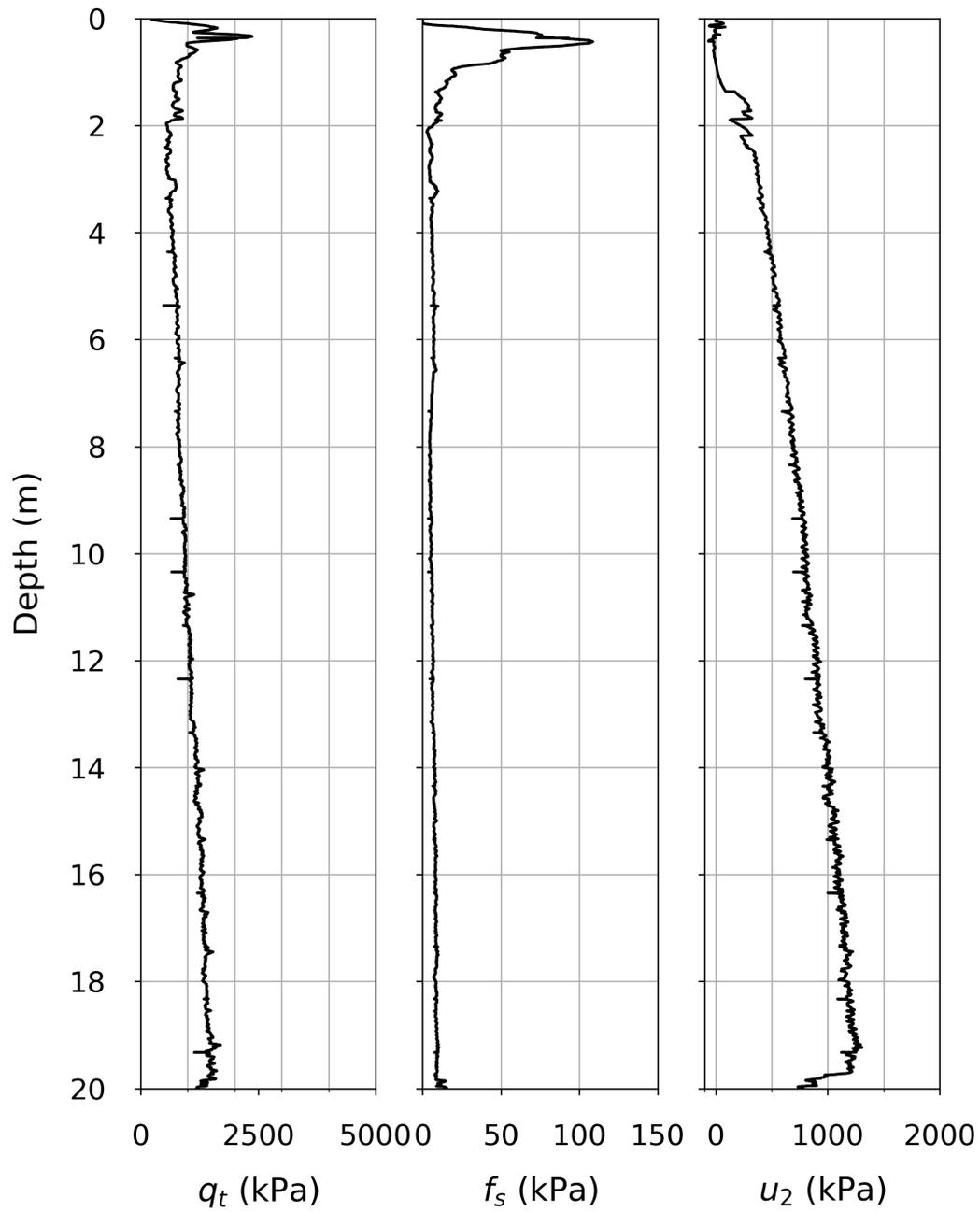
NGTS CPTu C10
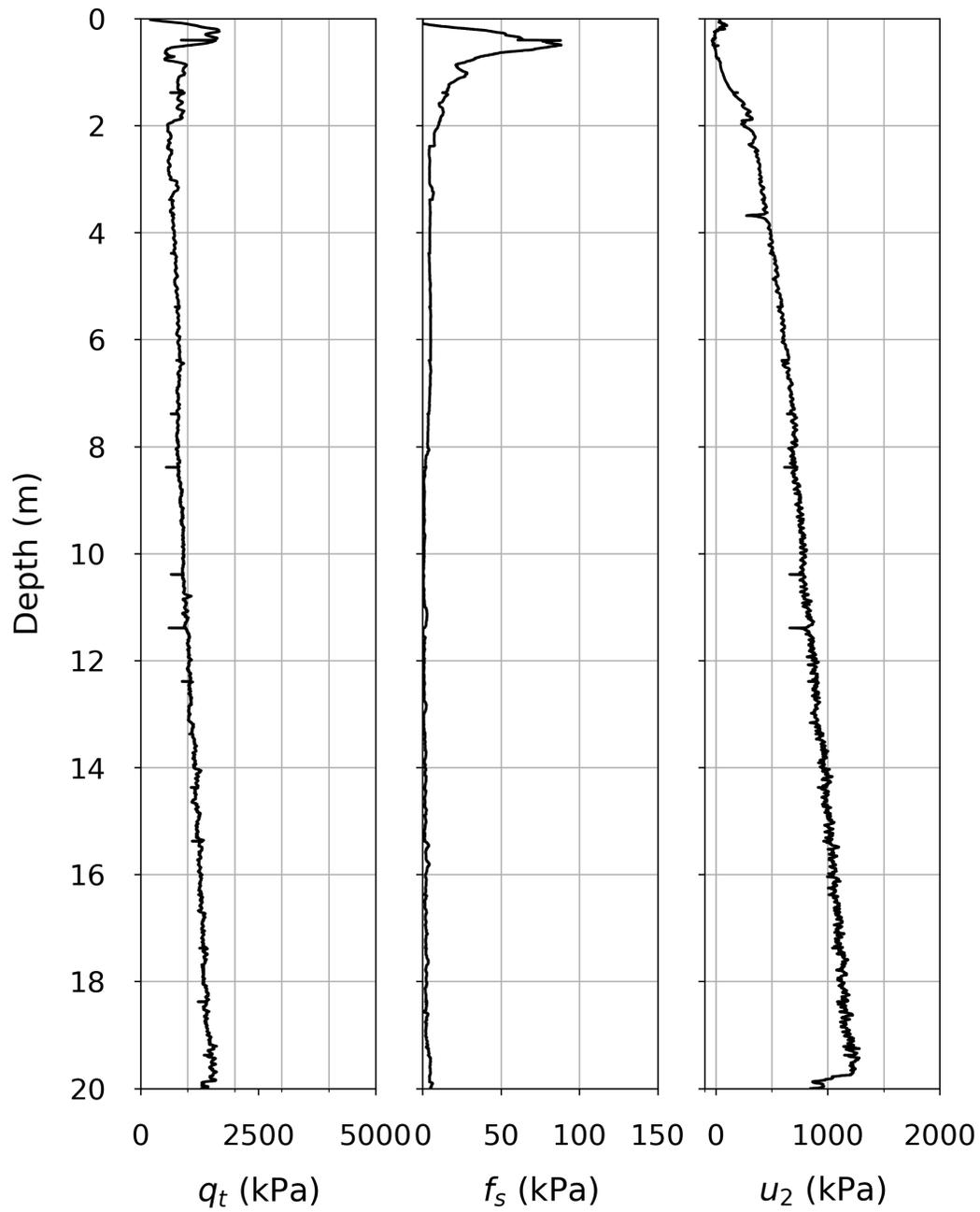
NGTS CPTu C11
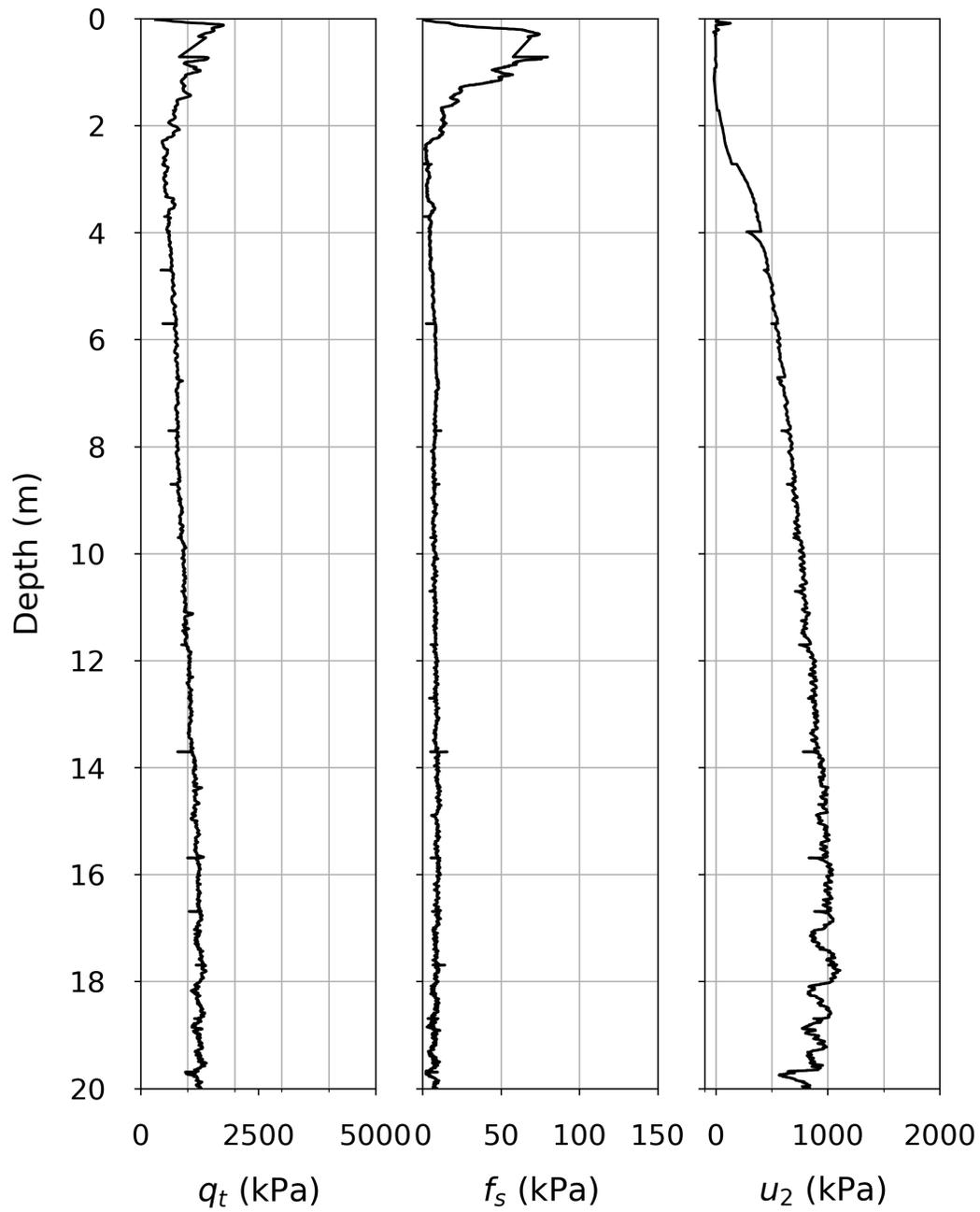
NGTS CPTu C12
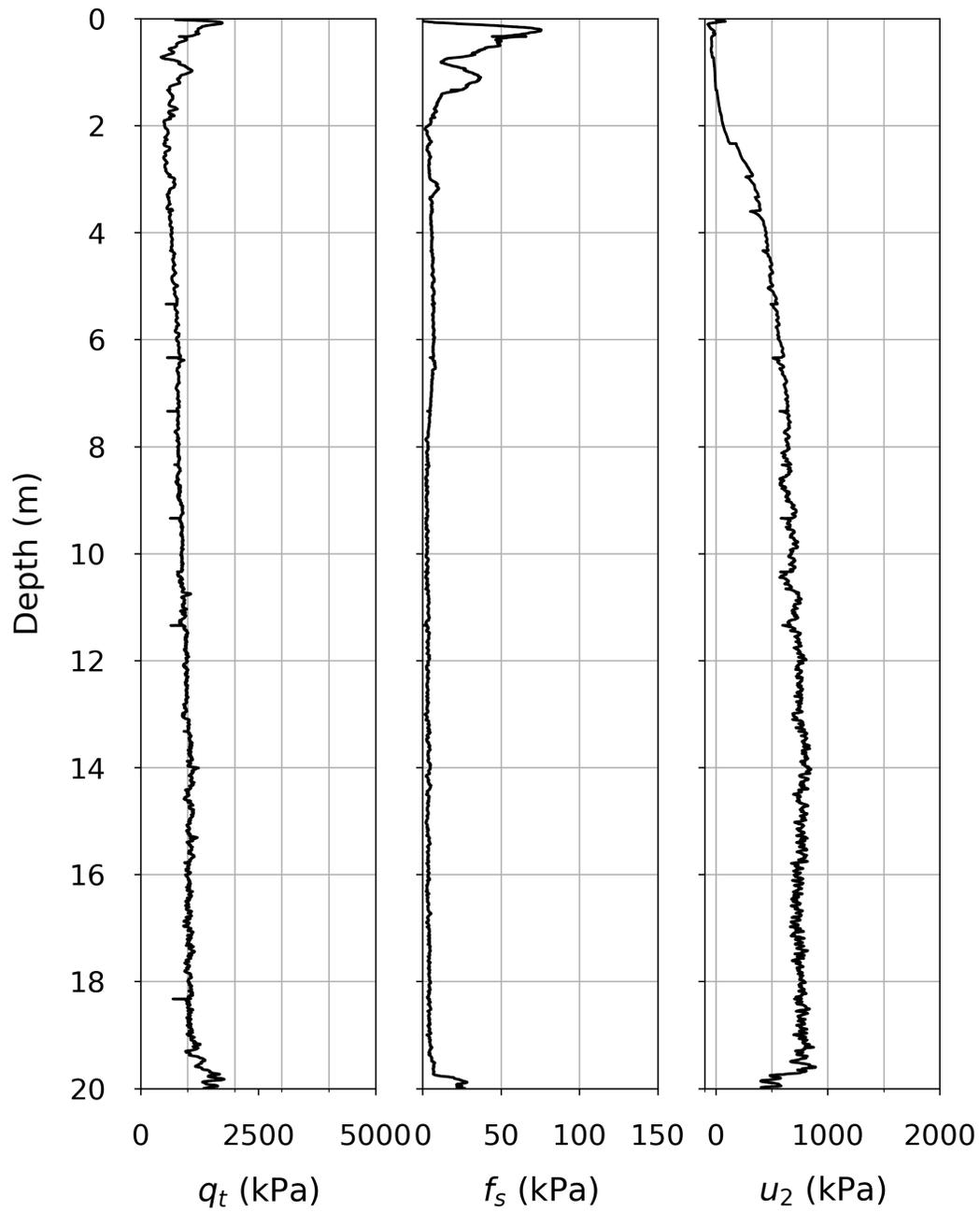
NGTS CPTu C13
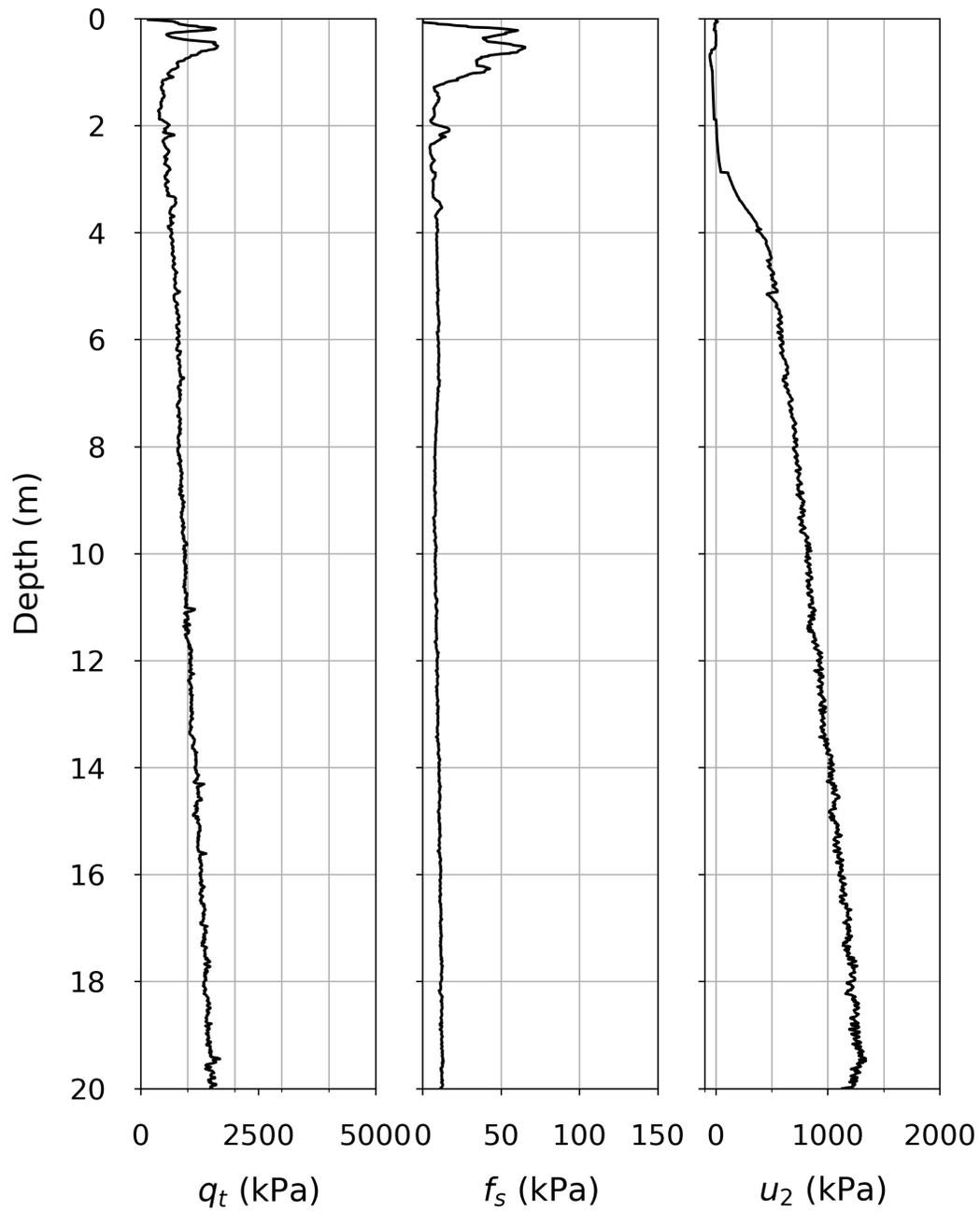
NGTS CPTu C14
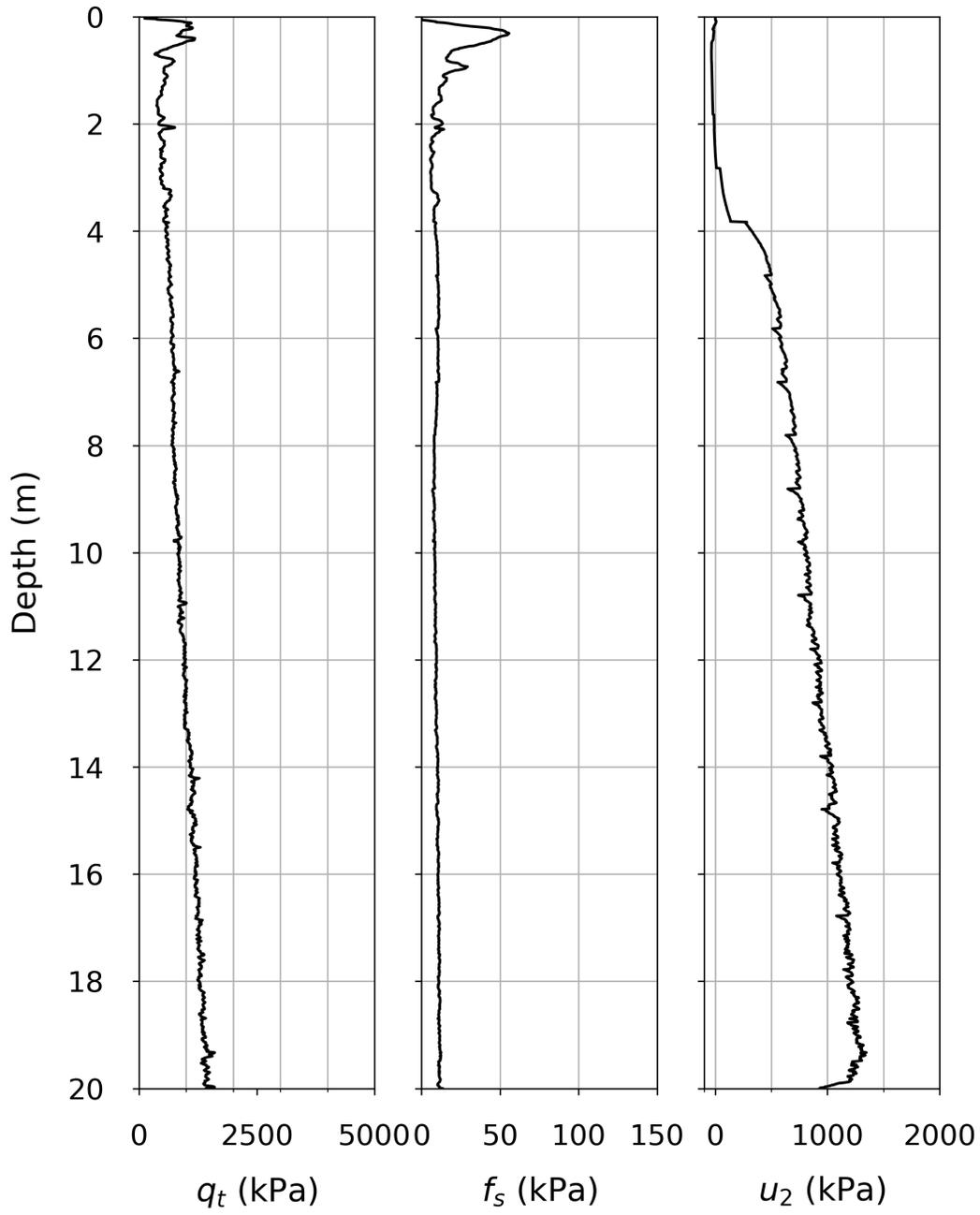
NGTS CPTu C15
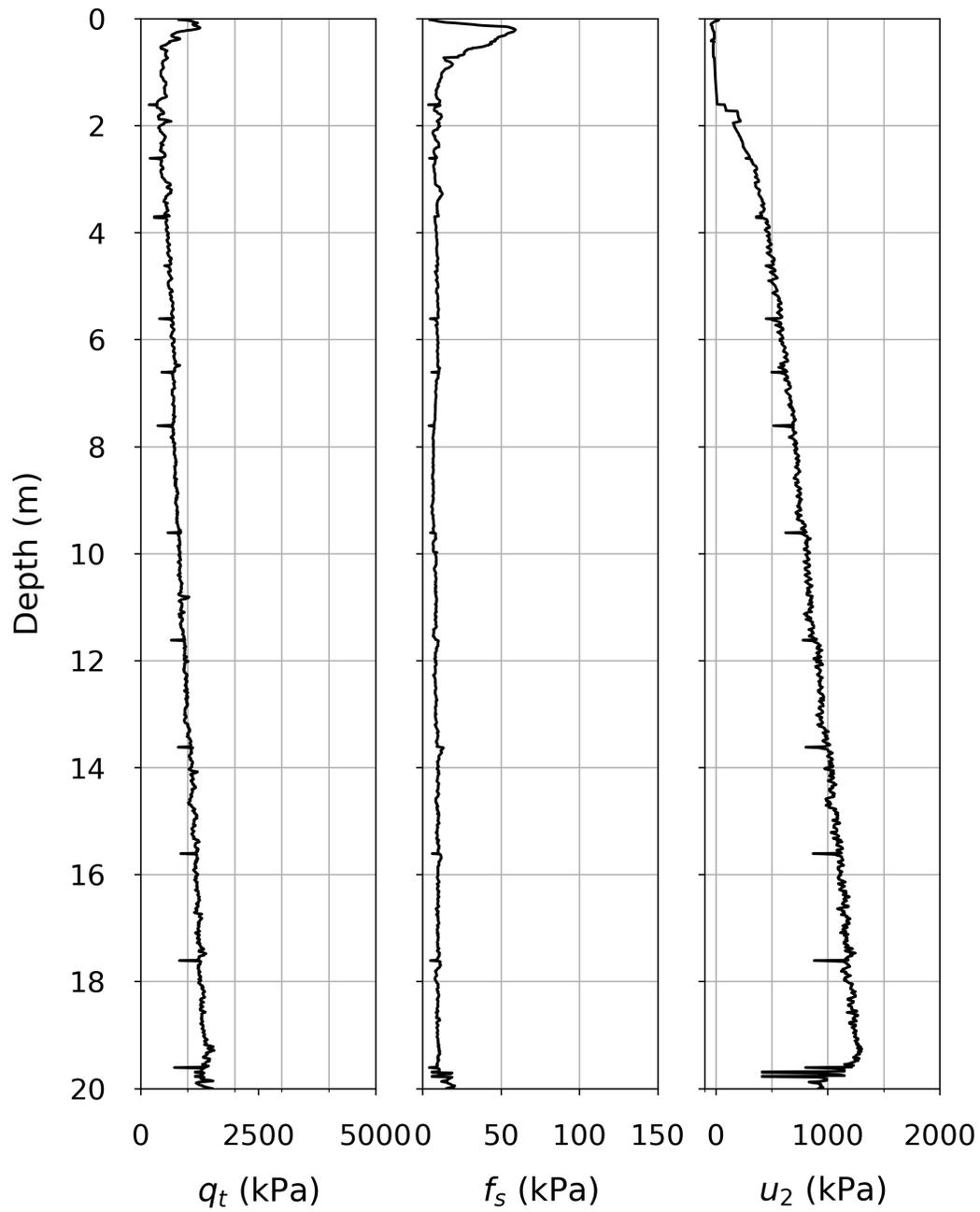
NGTS CPTu C16
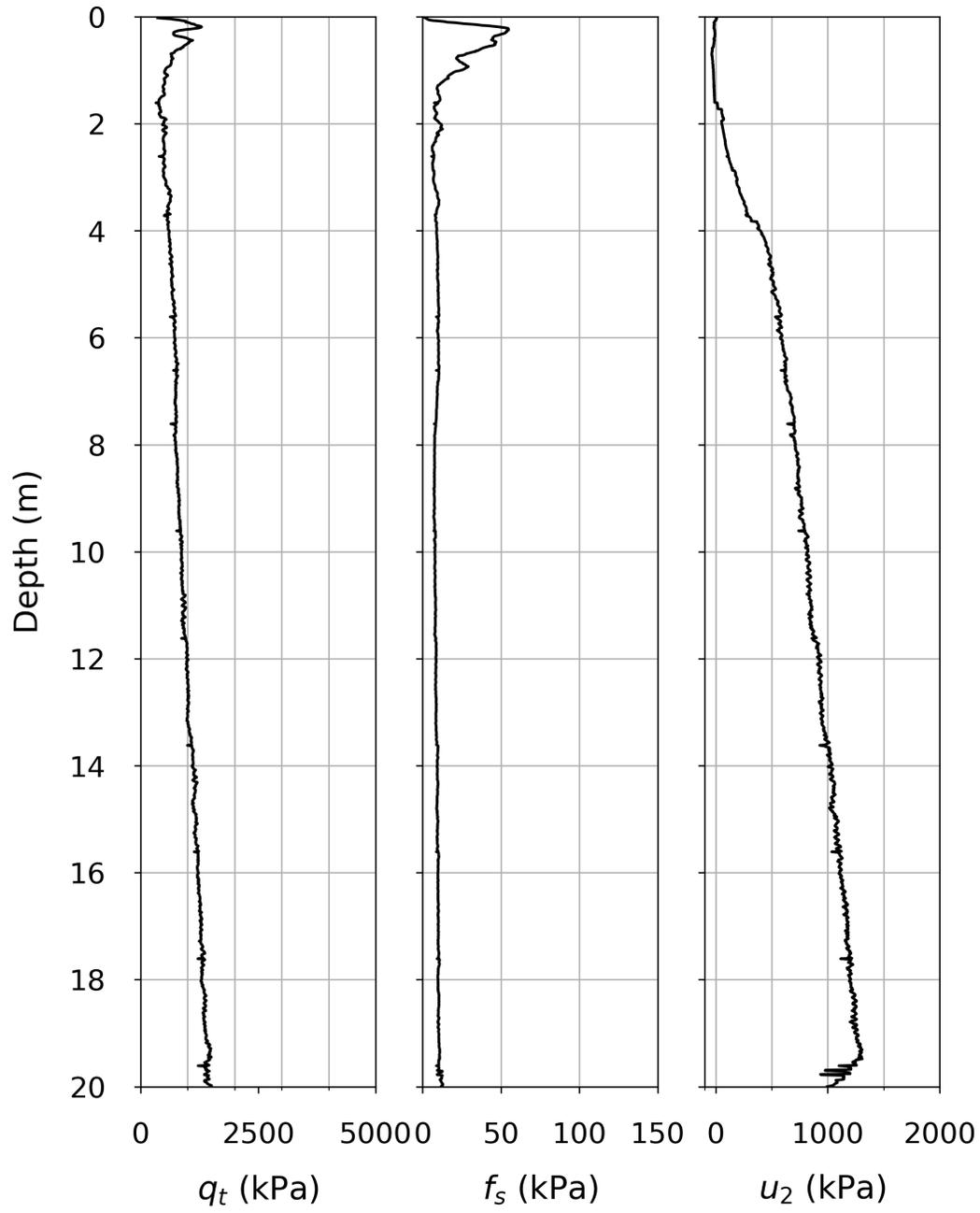
NGTS CPTu C17

NGTS CPTu C19

q_t (kPa) f_s (kPa) u_2 (kPa)
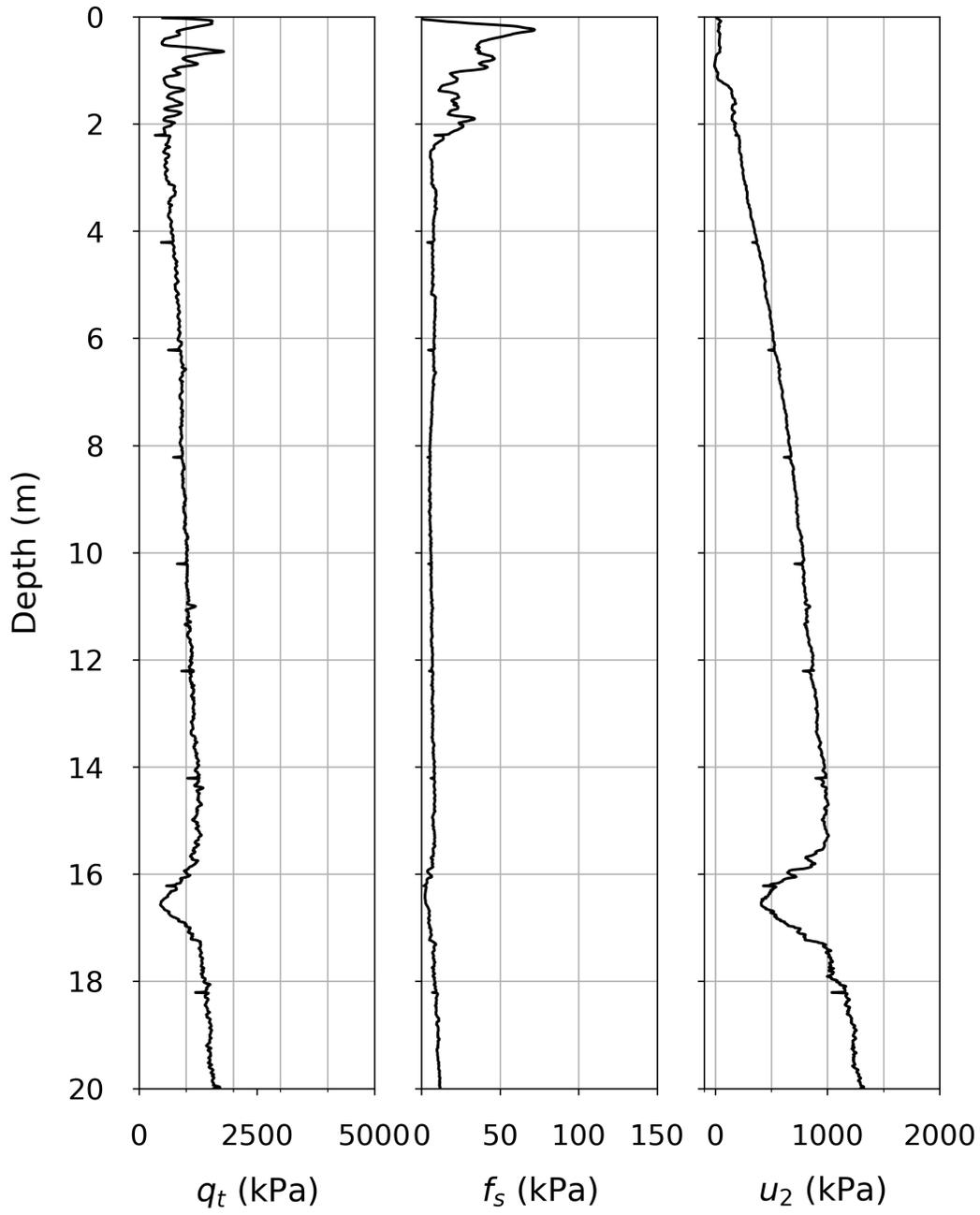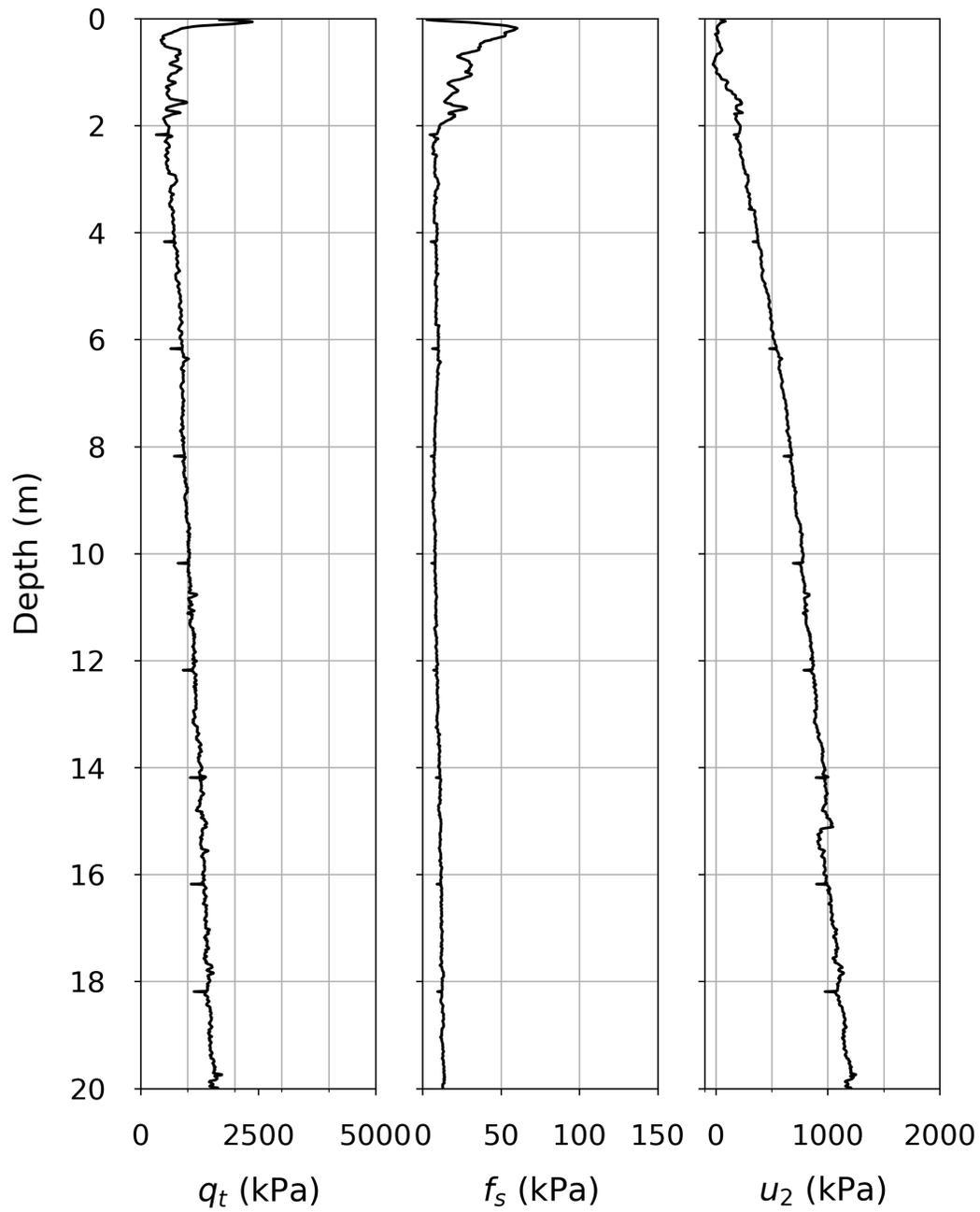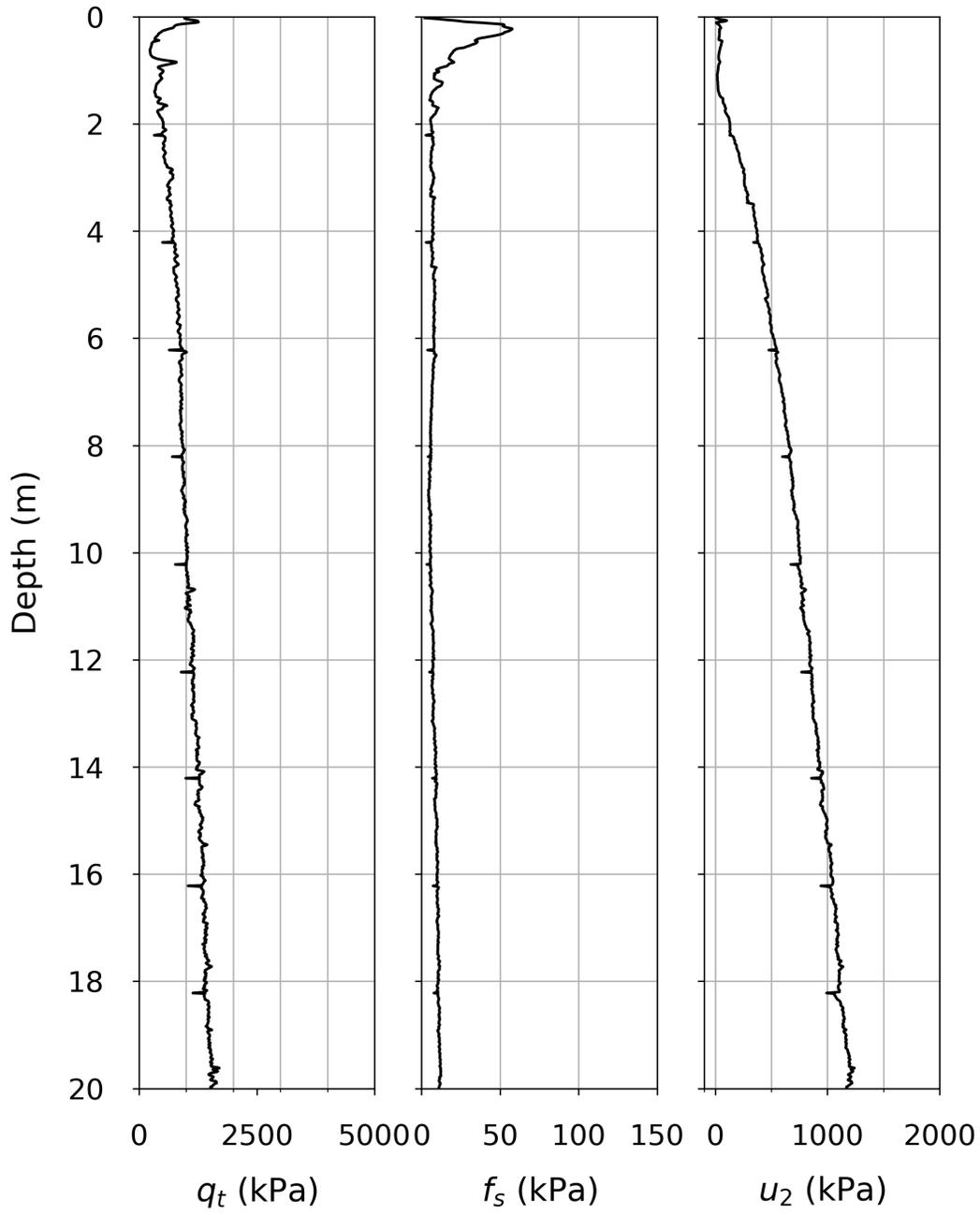
NGTS CPTu C20

NGTS CPTu C22
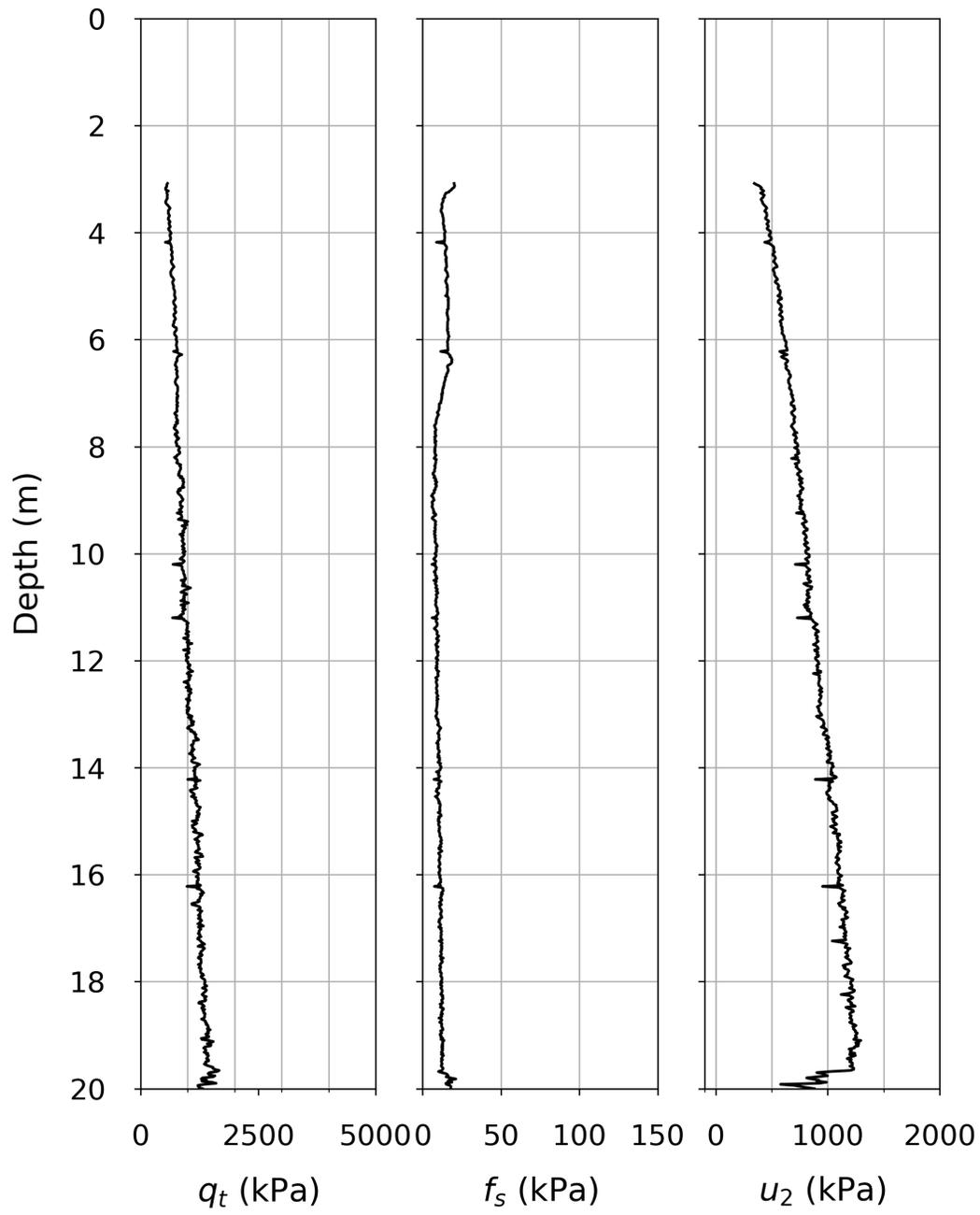
NGTS CPTu C23
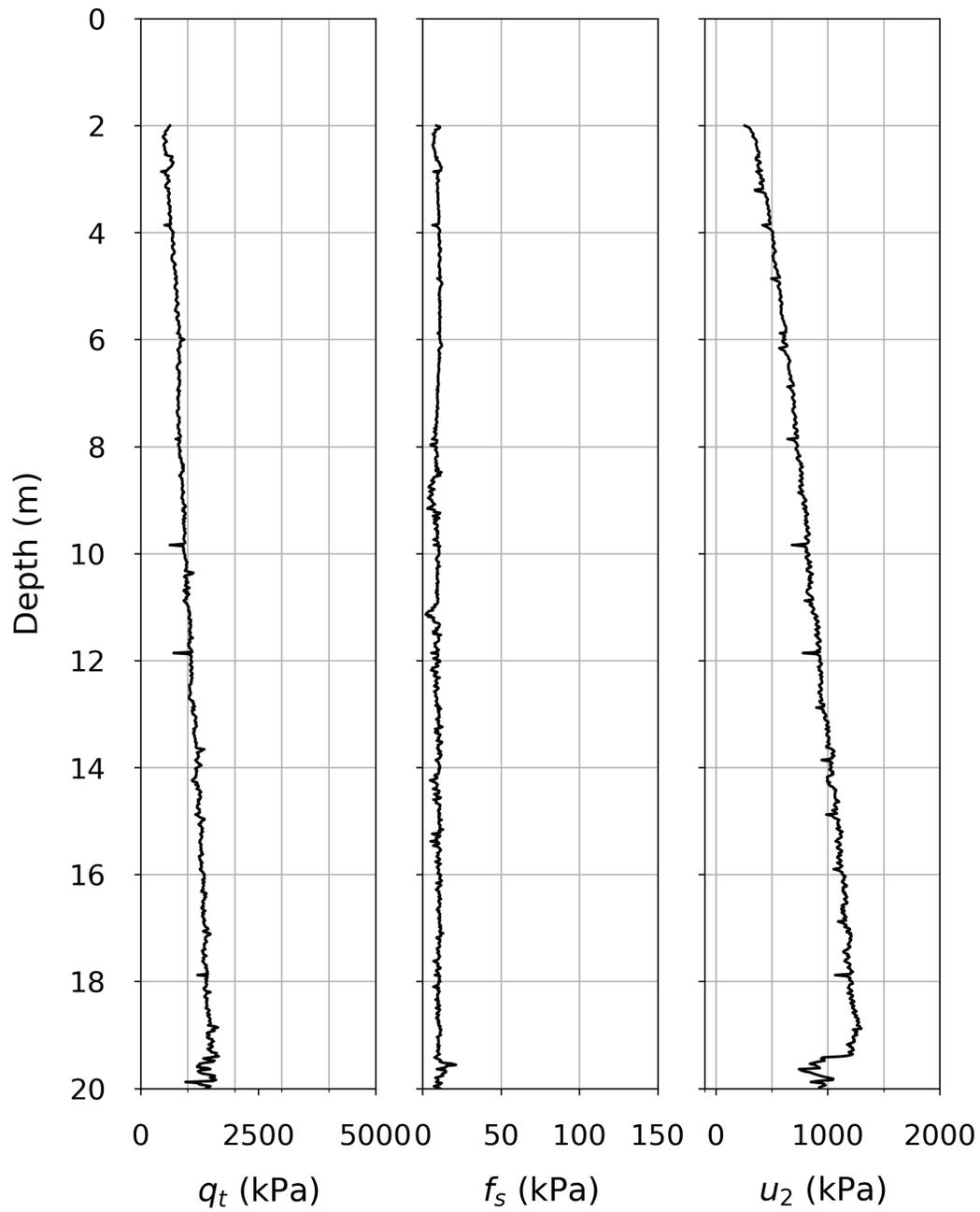
NGTS CPTu C24
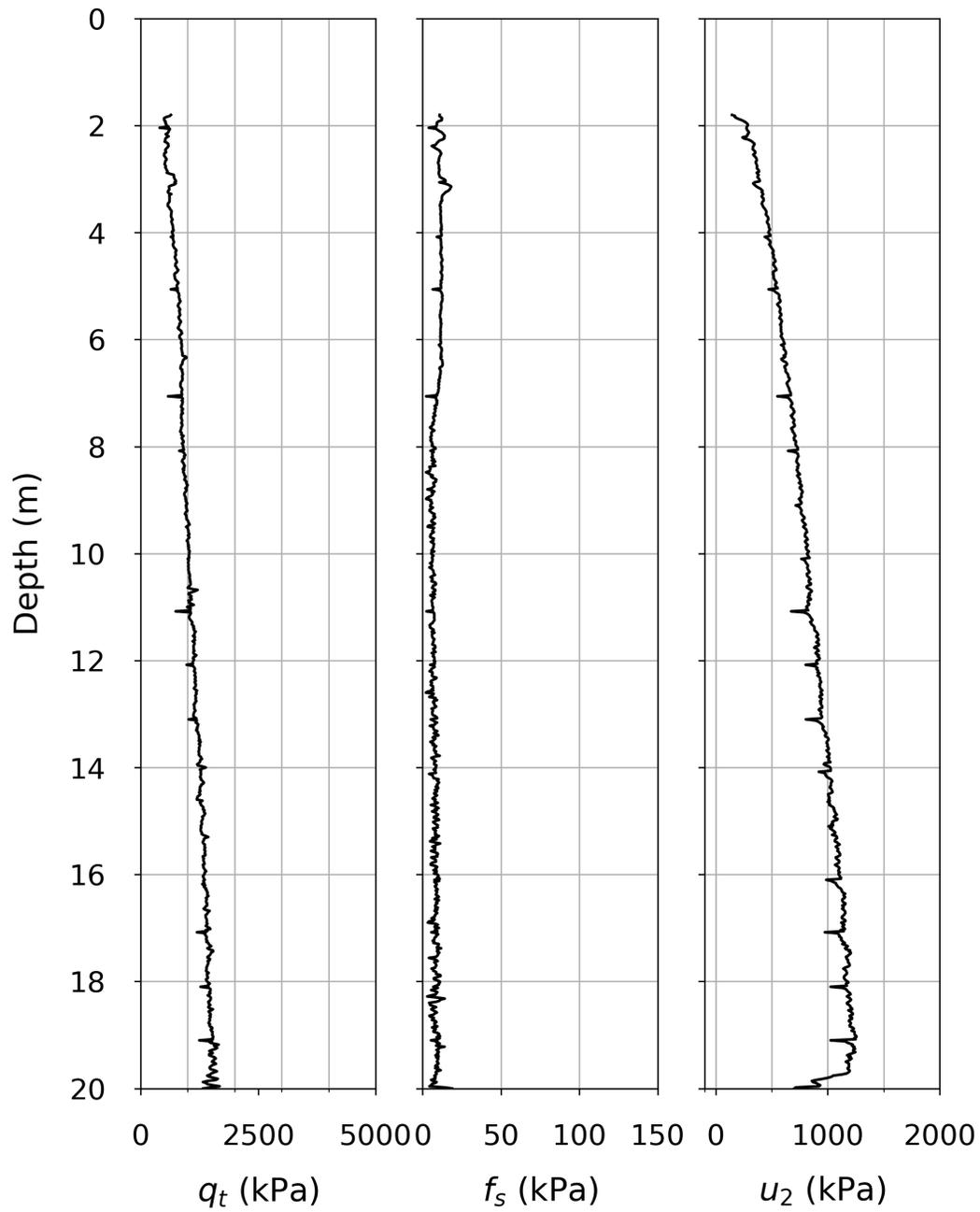
NGTS CPTu C25
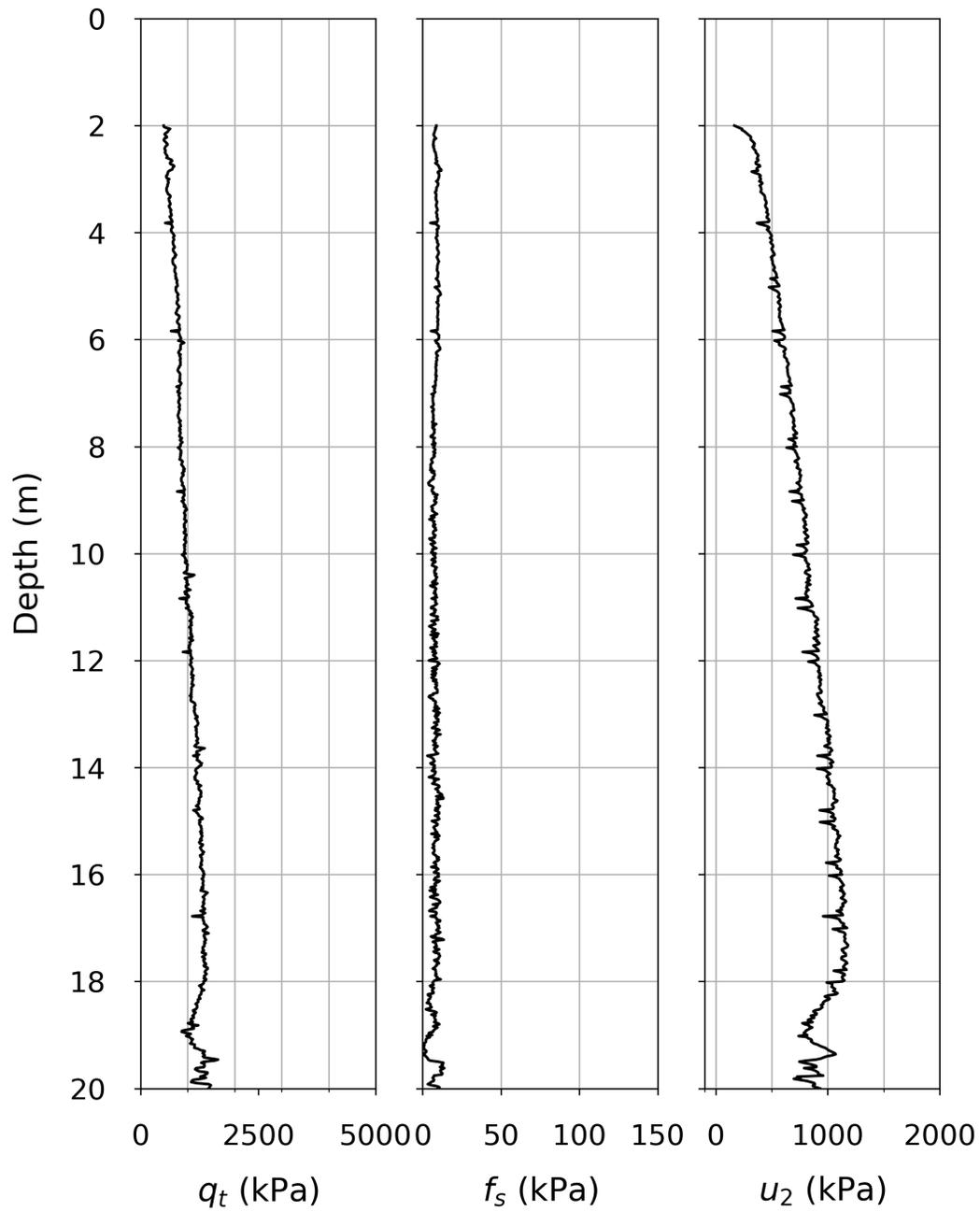
NGTS CPTu C26
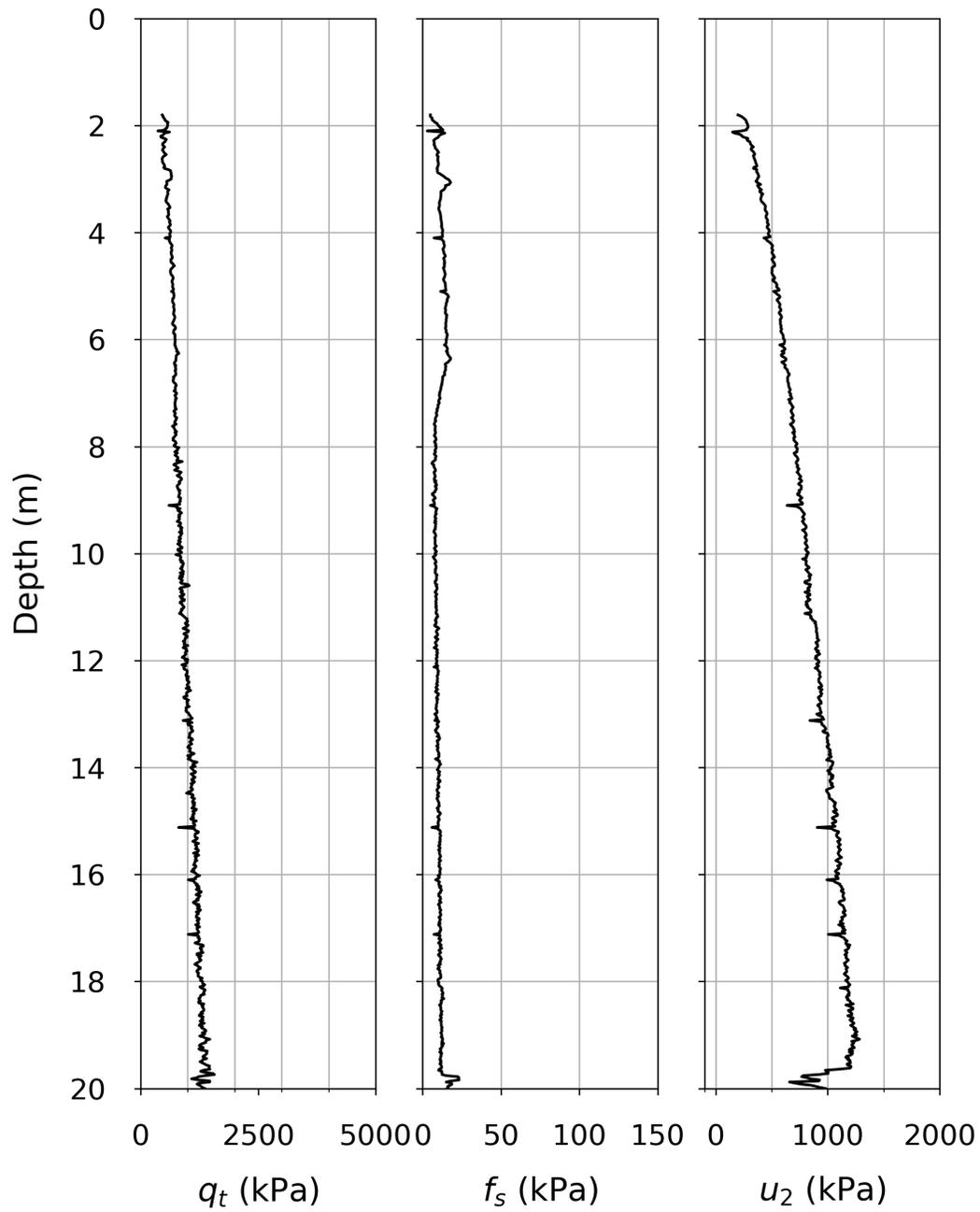
NGTS CPTu C27
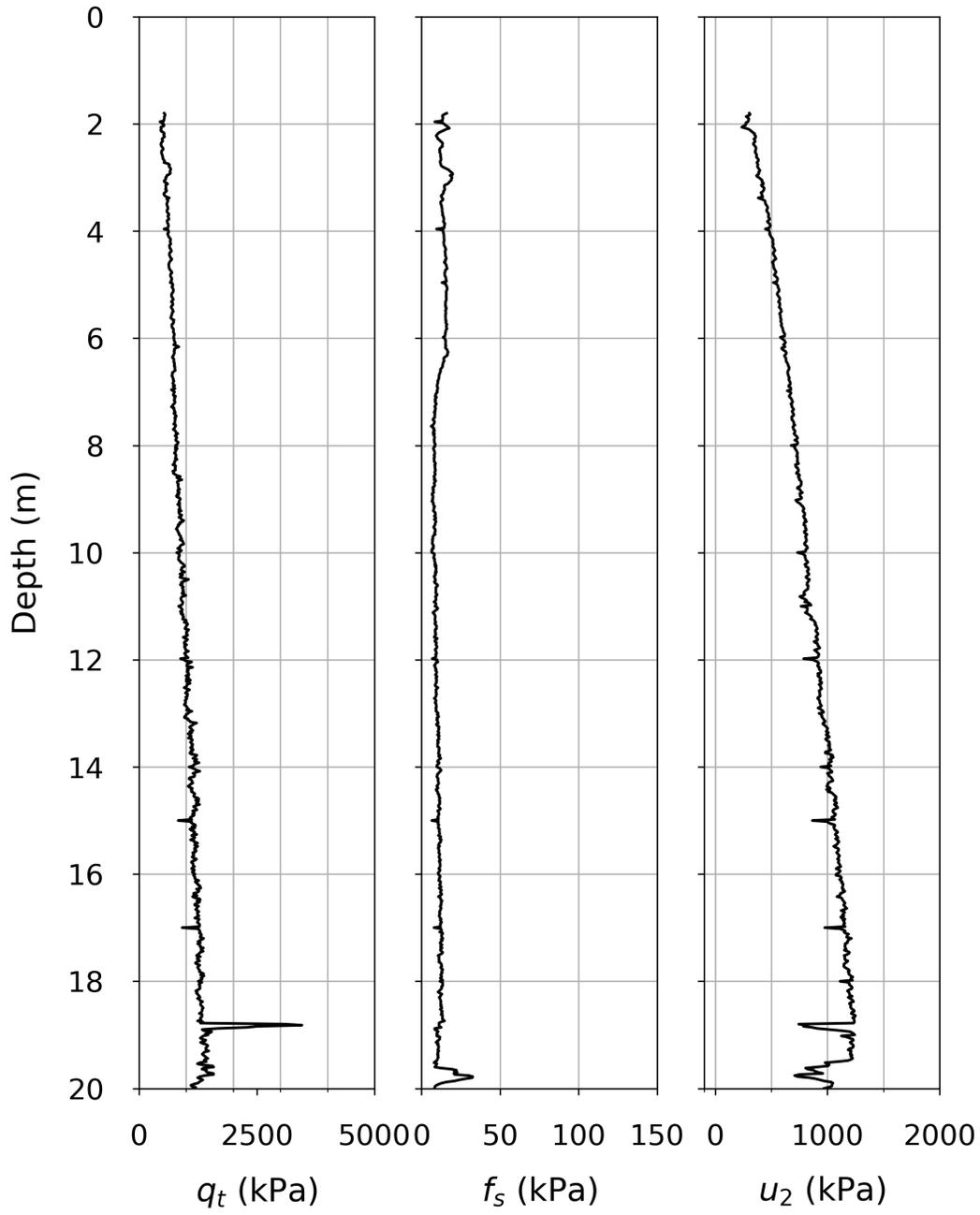
NGTS CPTu C28
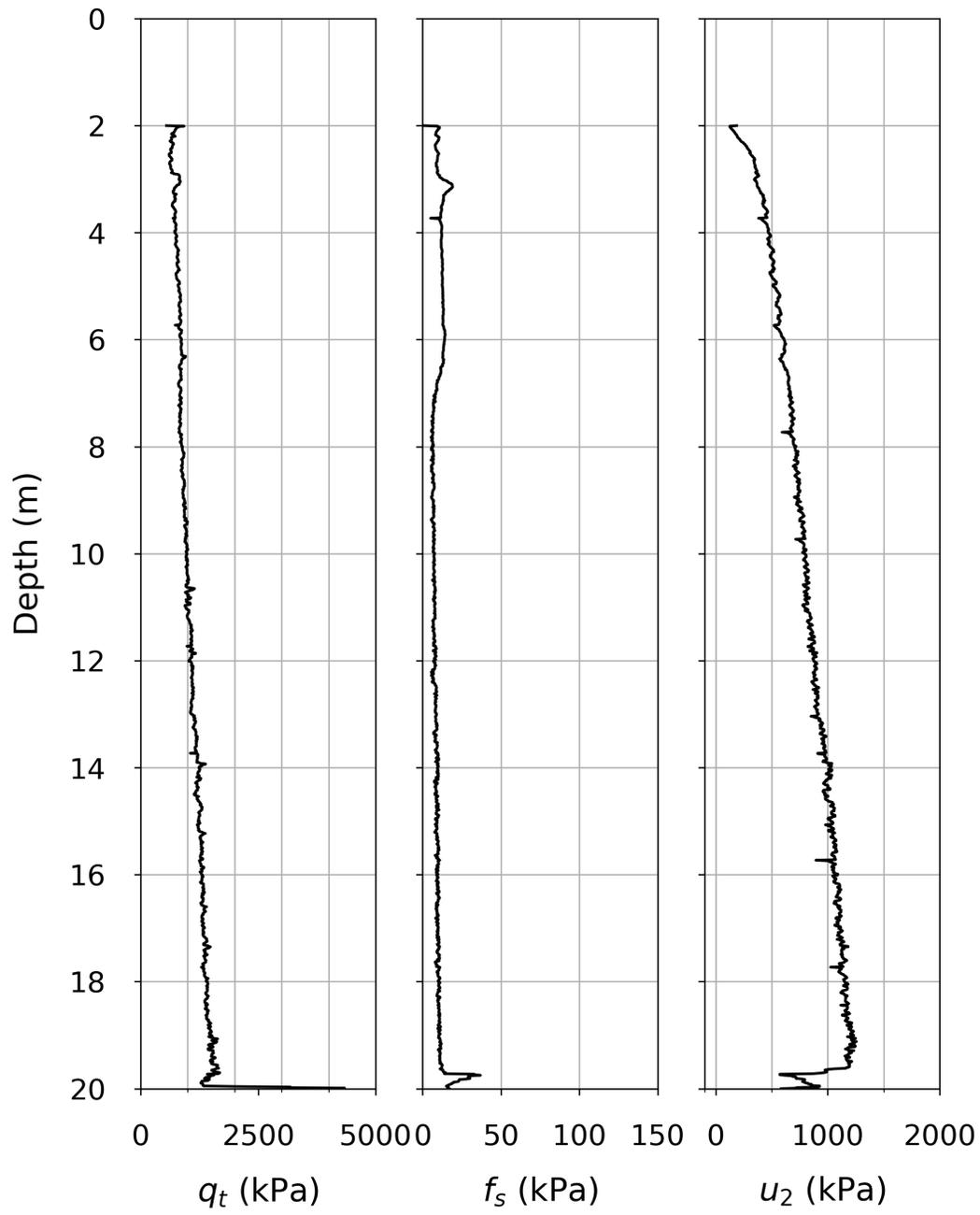
NGTS CPTu C29
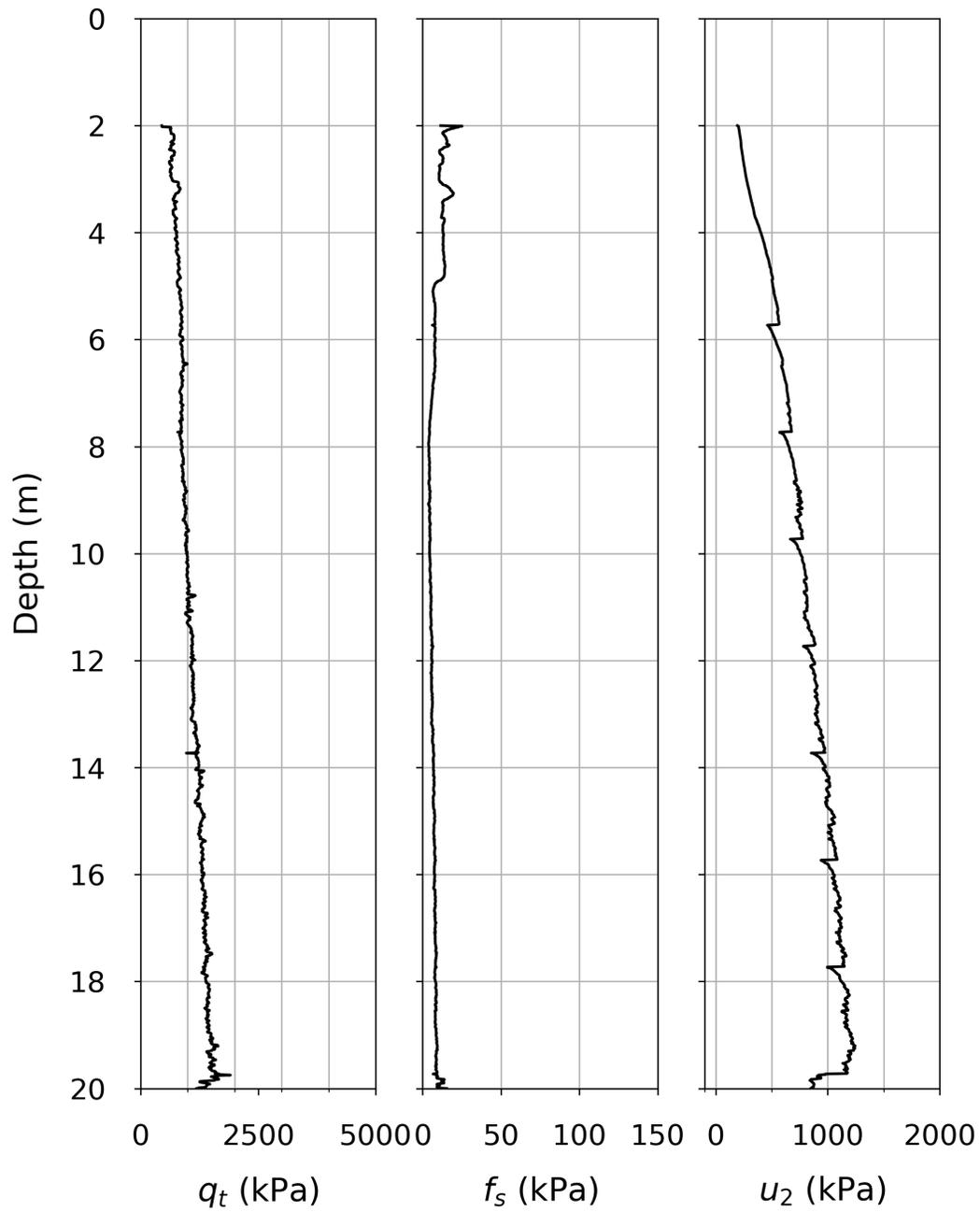
NGTS CPTu C30
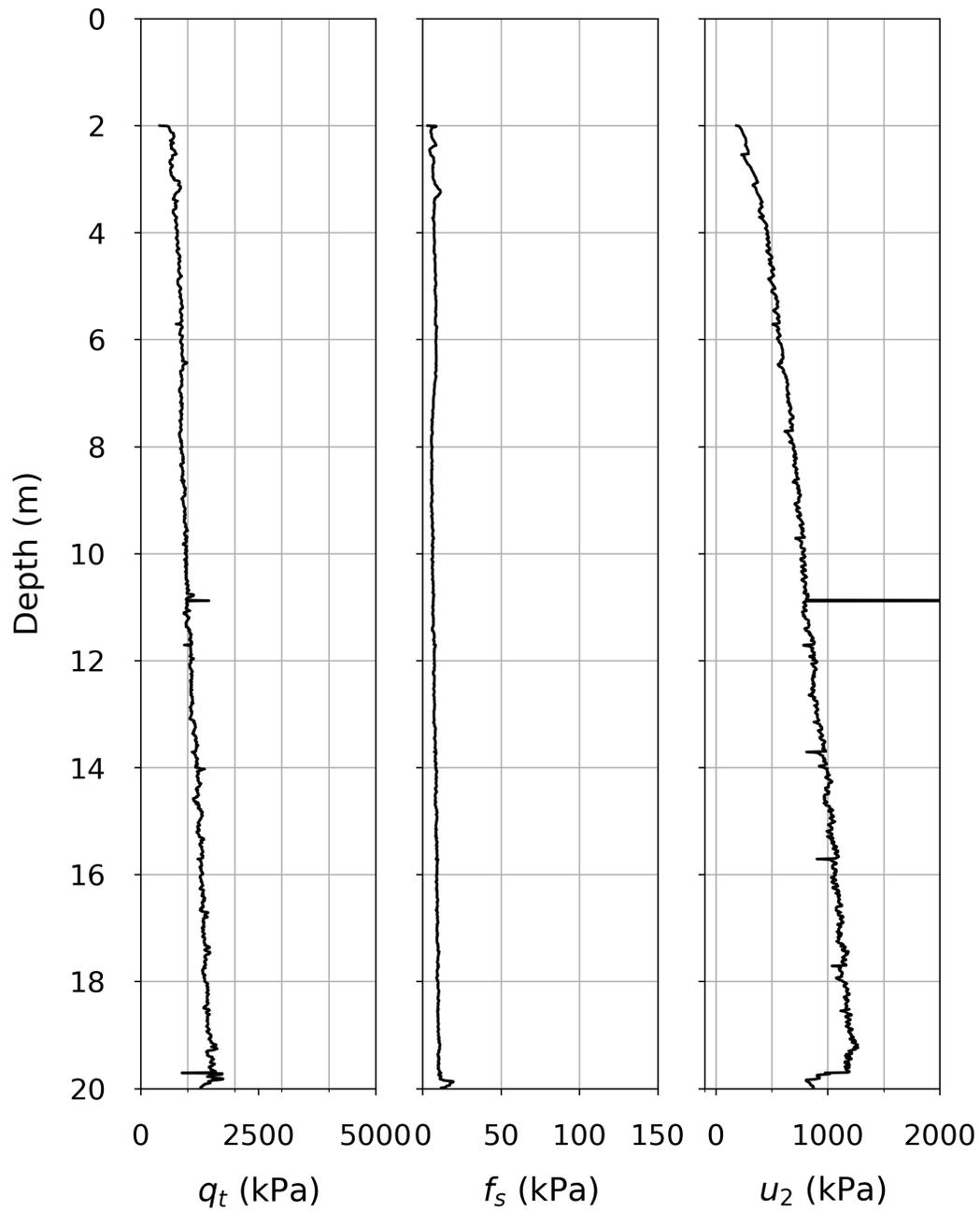
NGTS CPTu C31
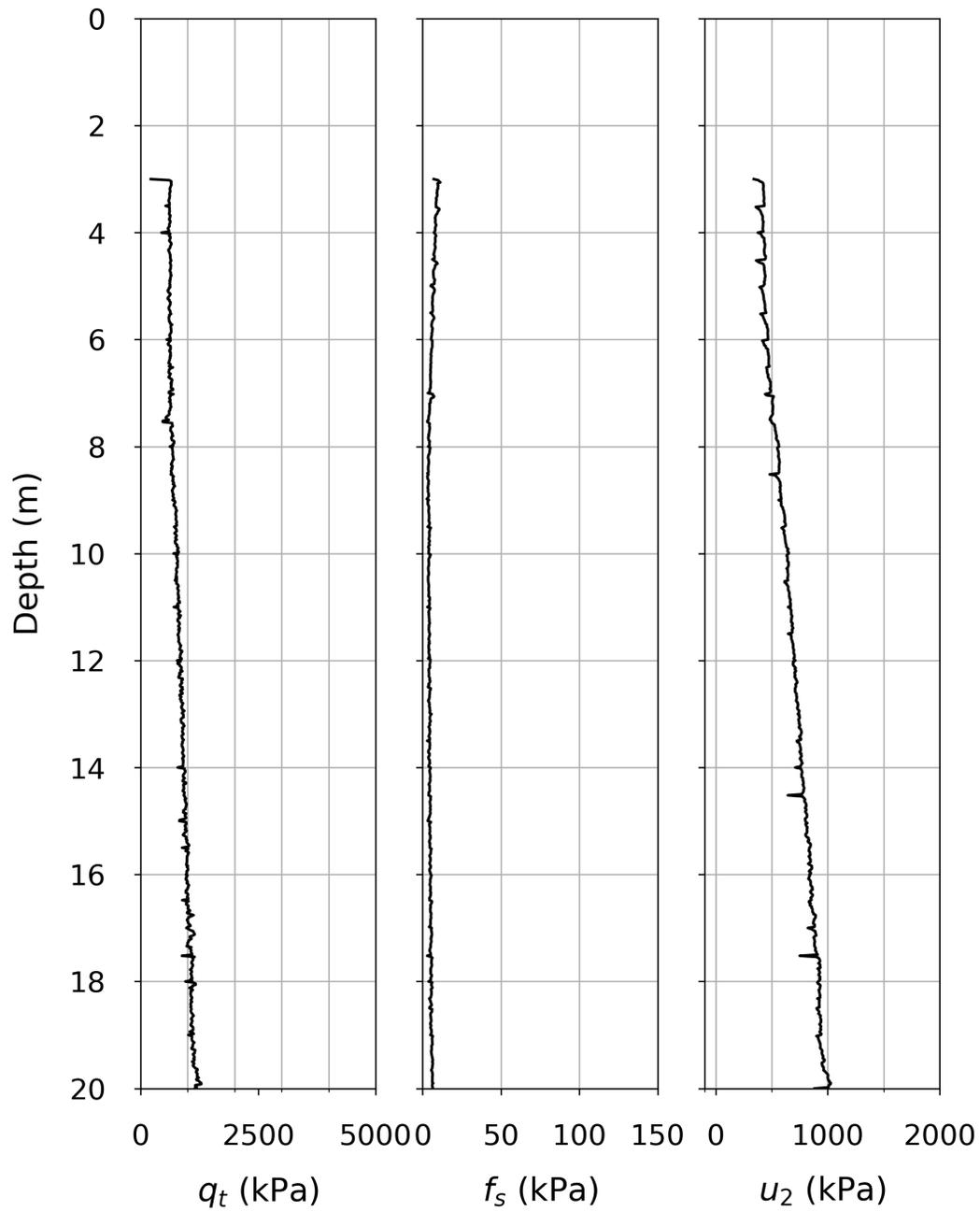
NGTS CPTu C32
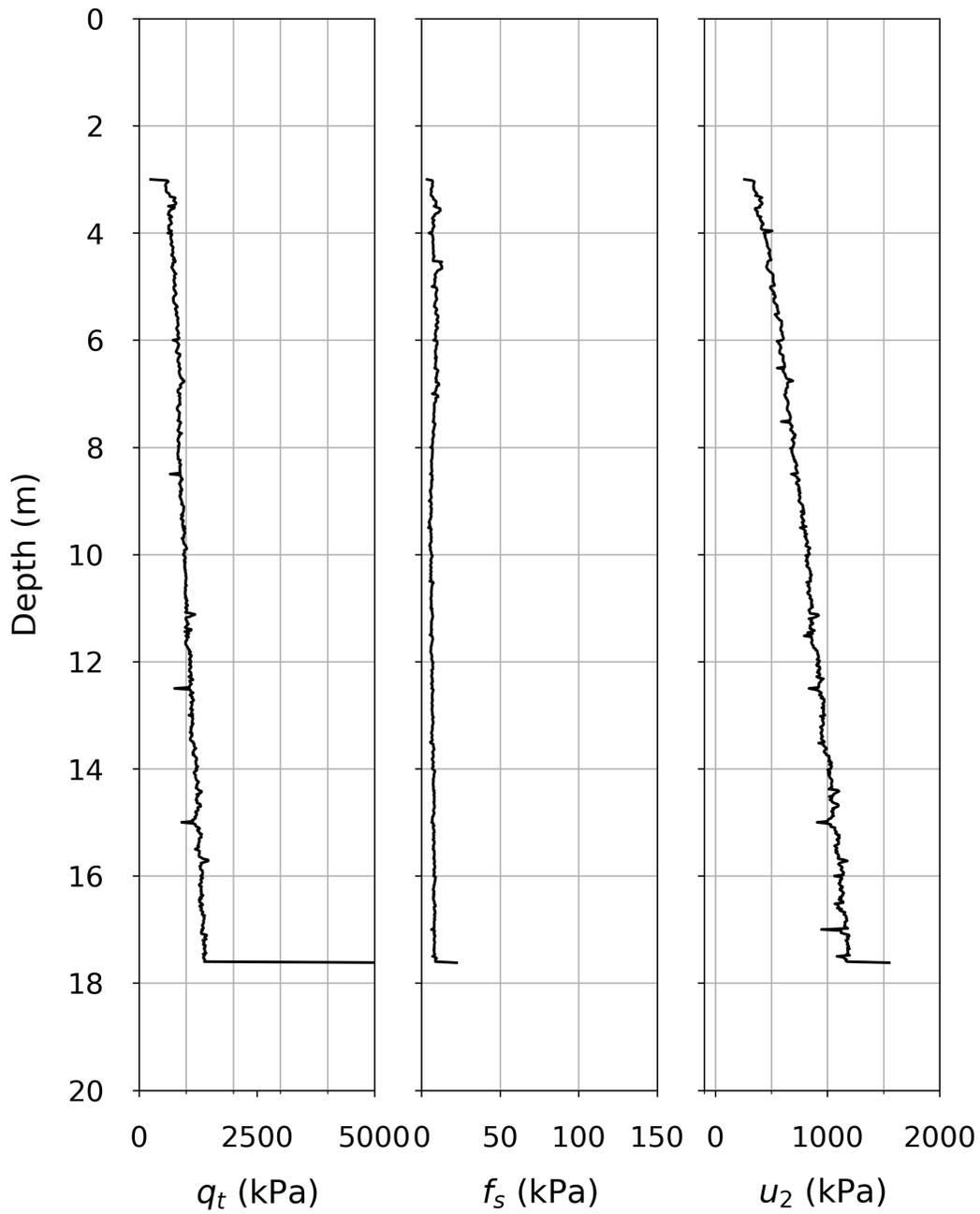
NGTS CPTu C33
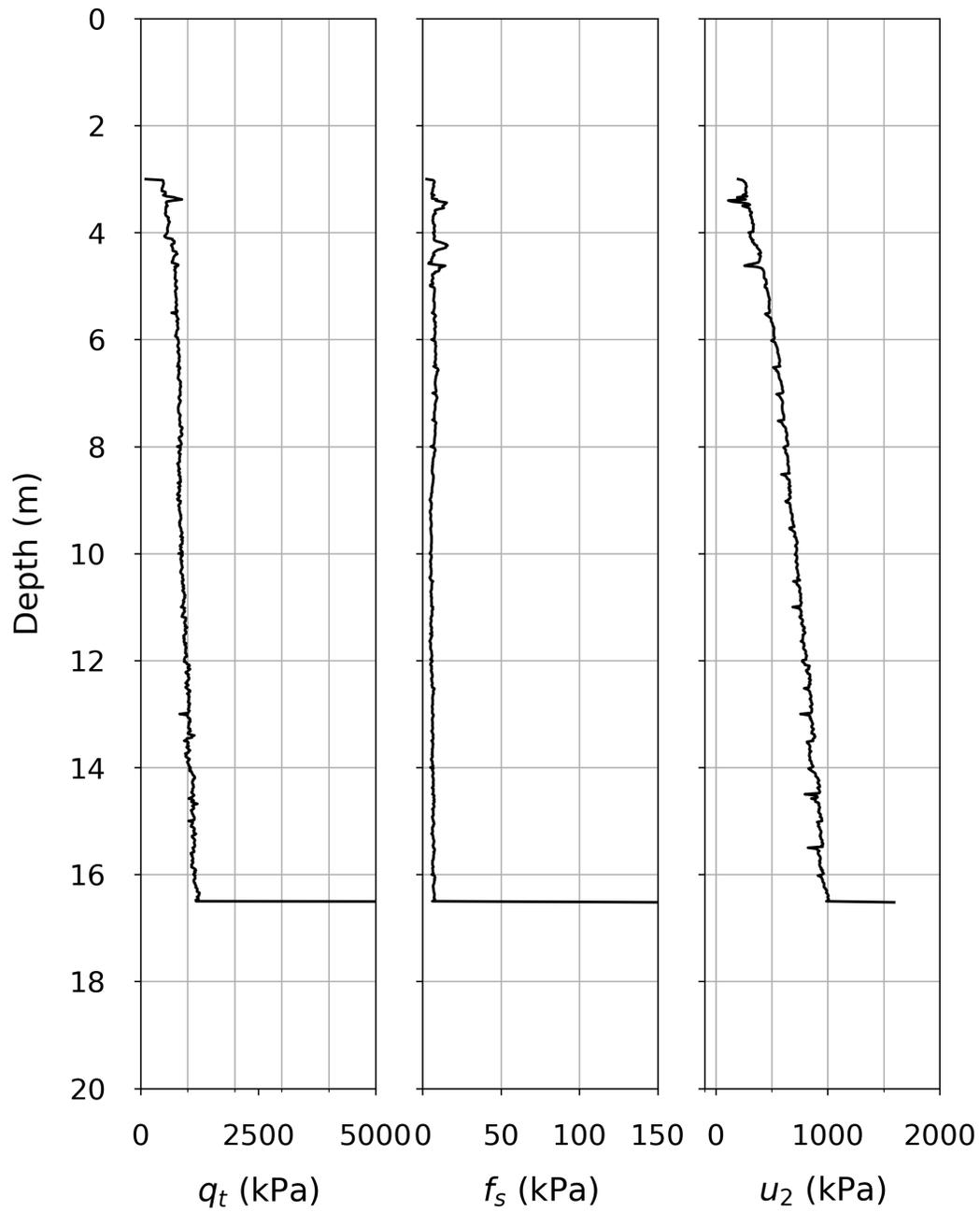
Vegvesen CPTu 100
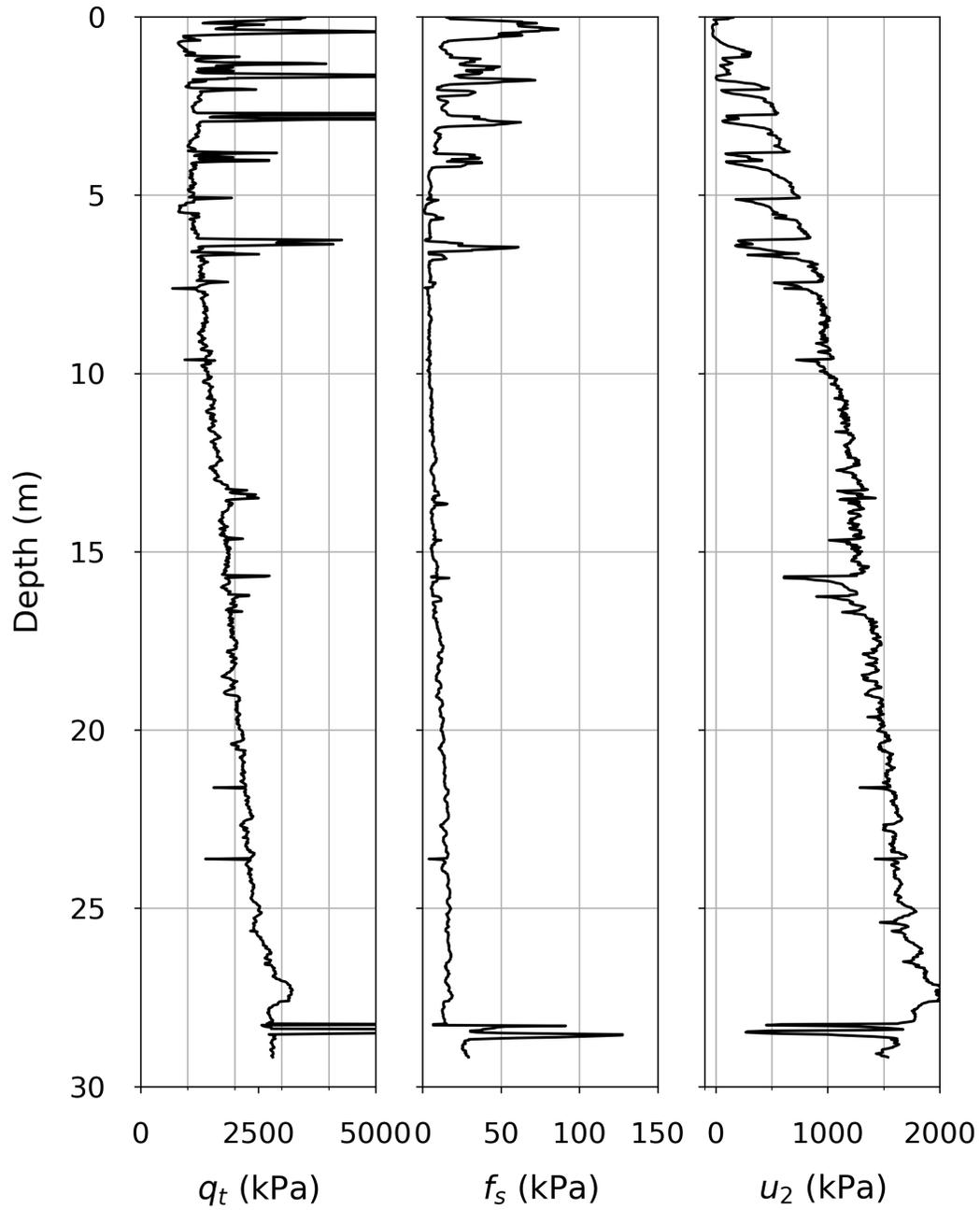
Vegvesen CPTu 102
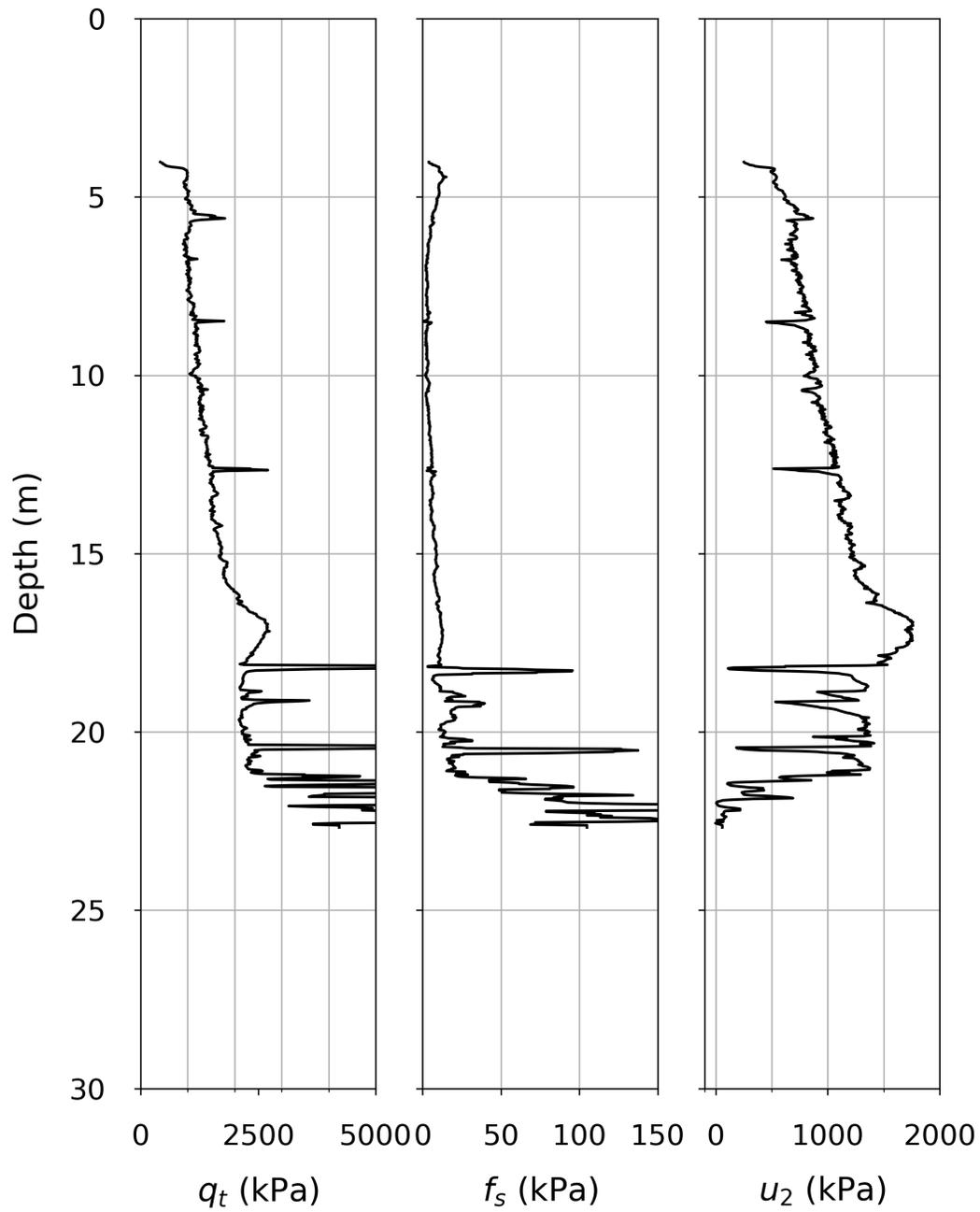
Vegvesen CPTu 106
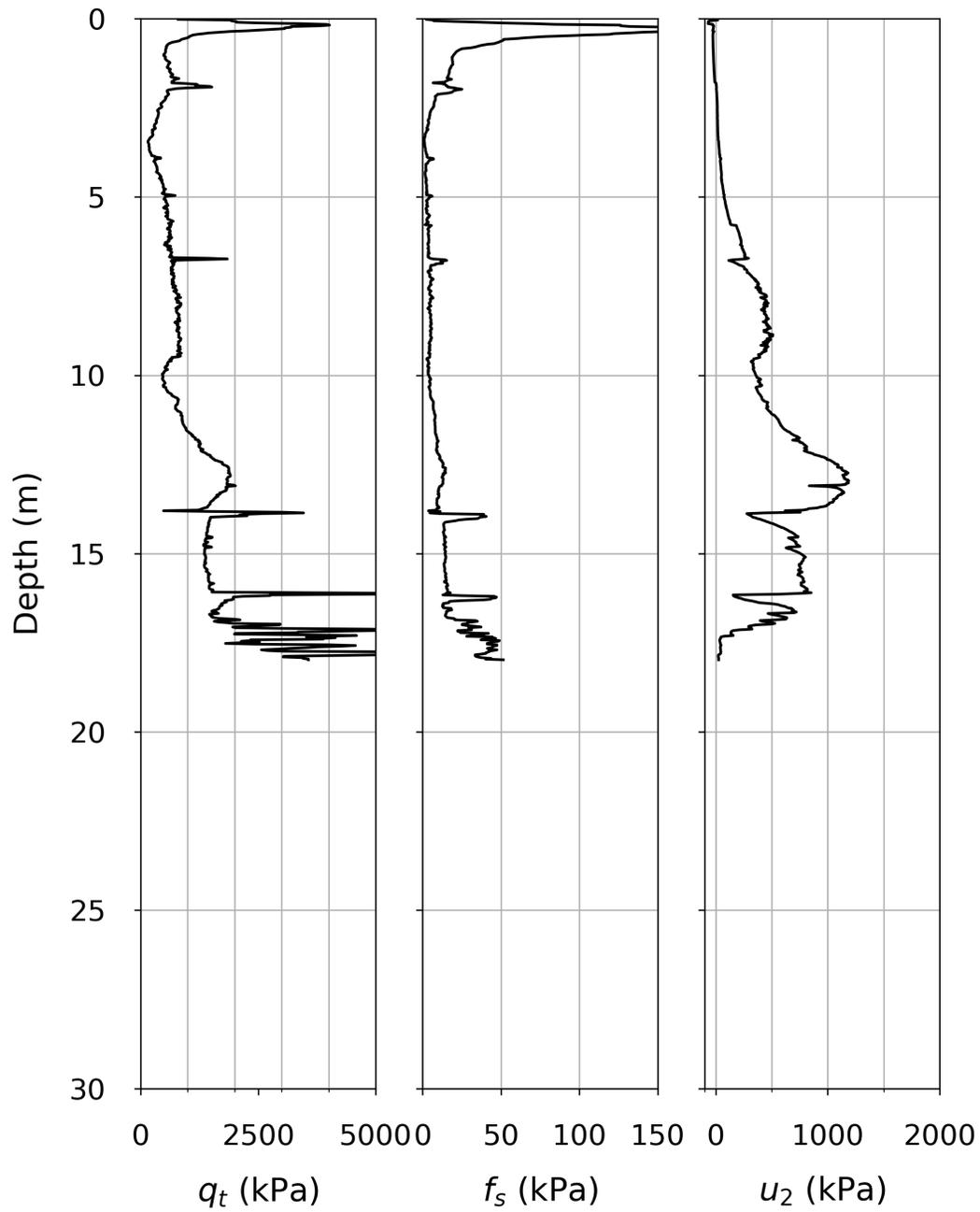
Vegvesen CPTu 107
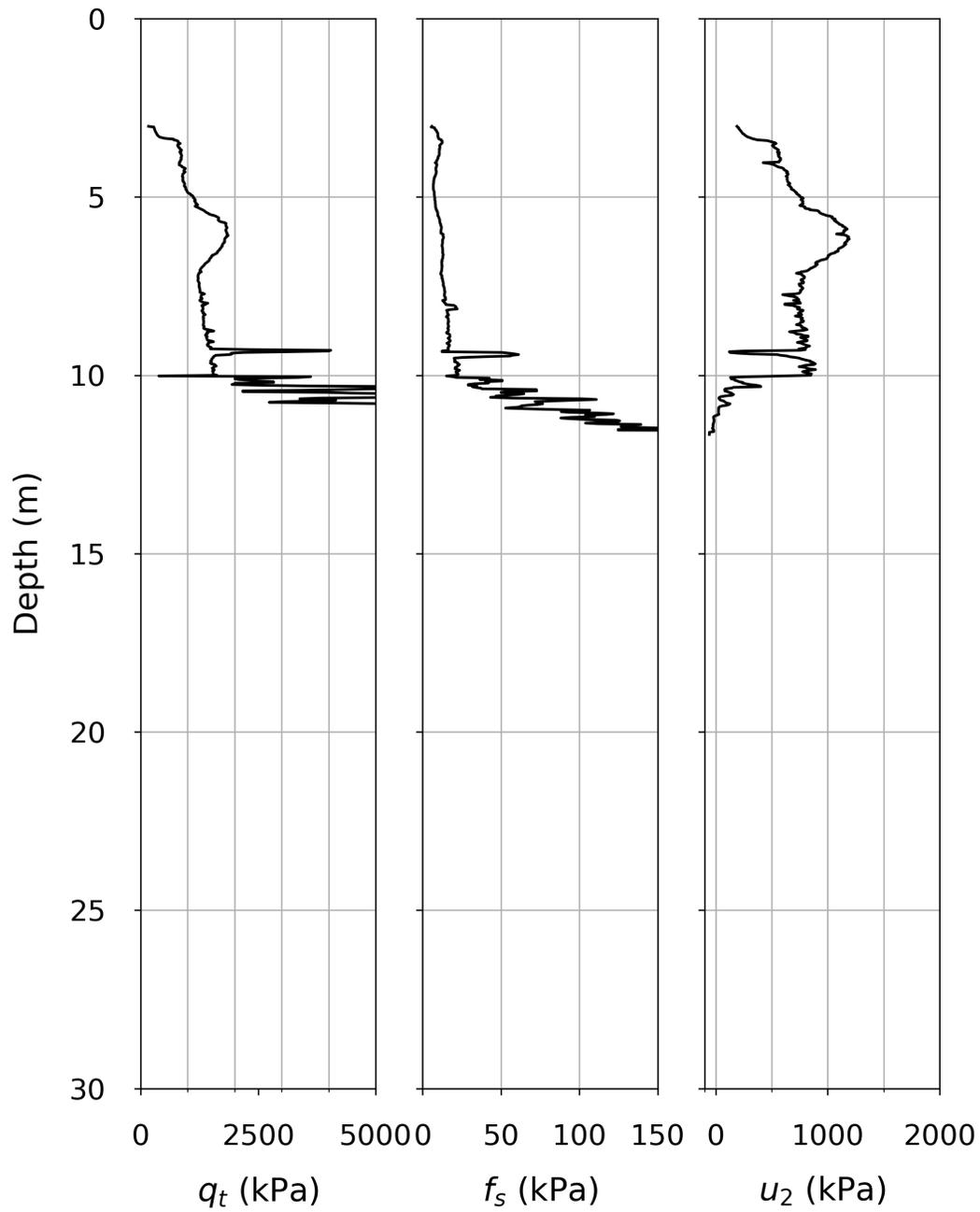
Vegvesen CPTu 149
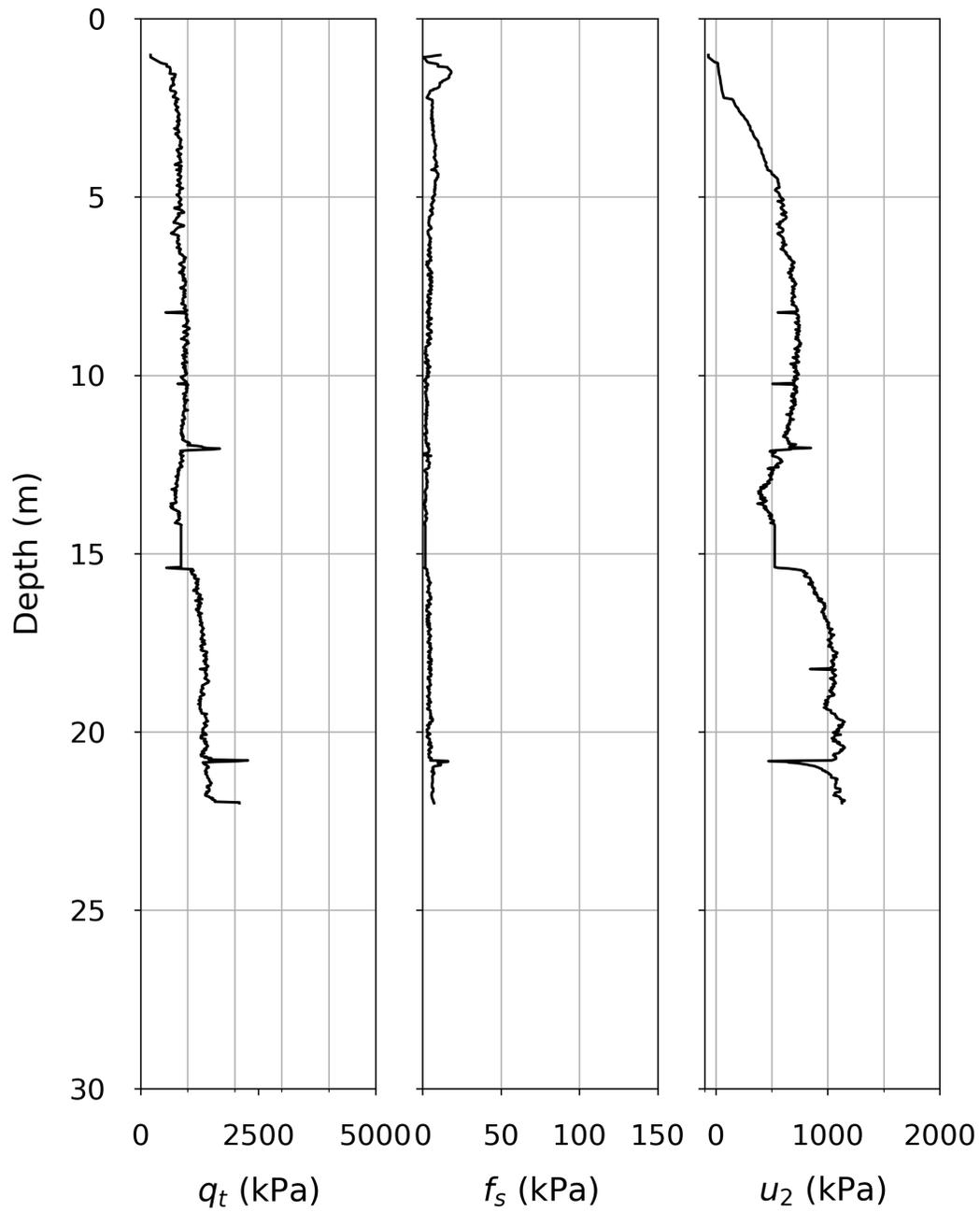
Vegvesen CPTu 154
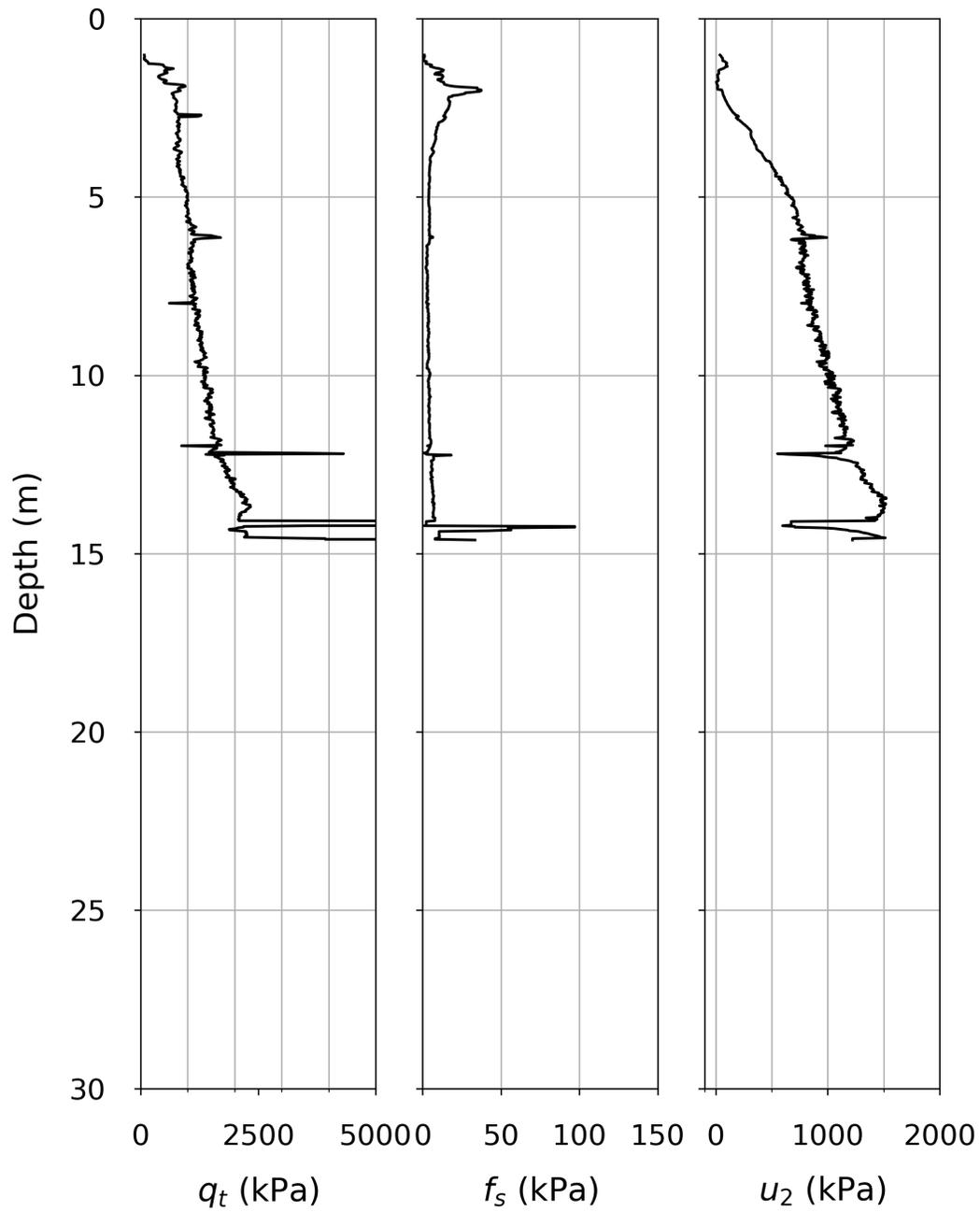
Vegvesen CPTu 155
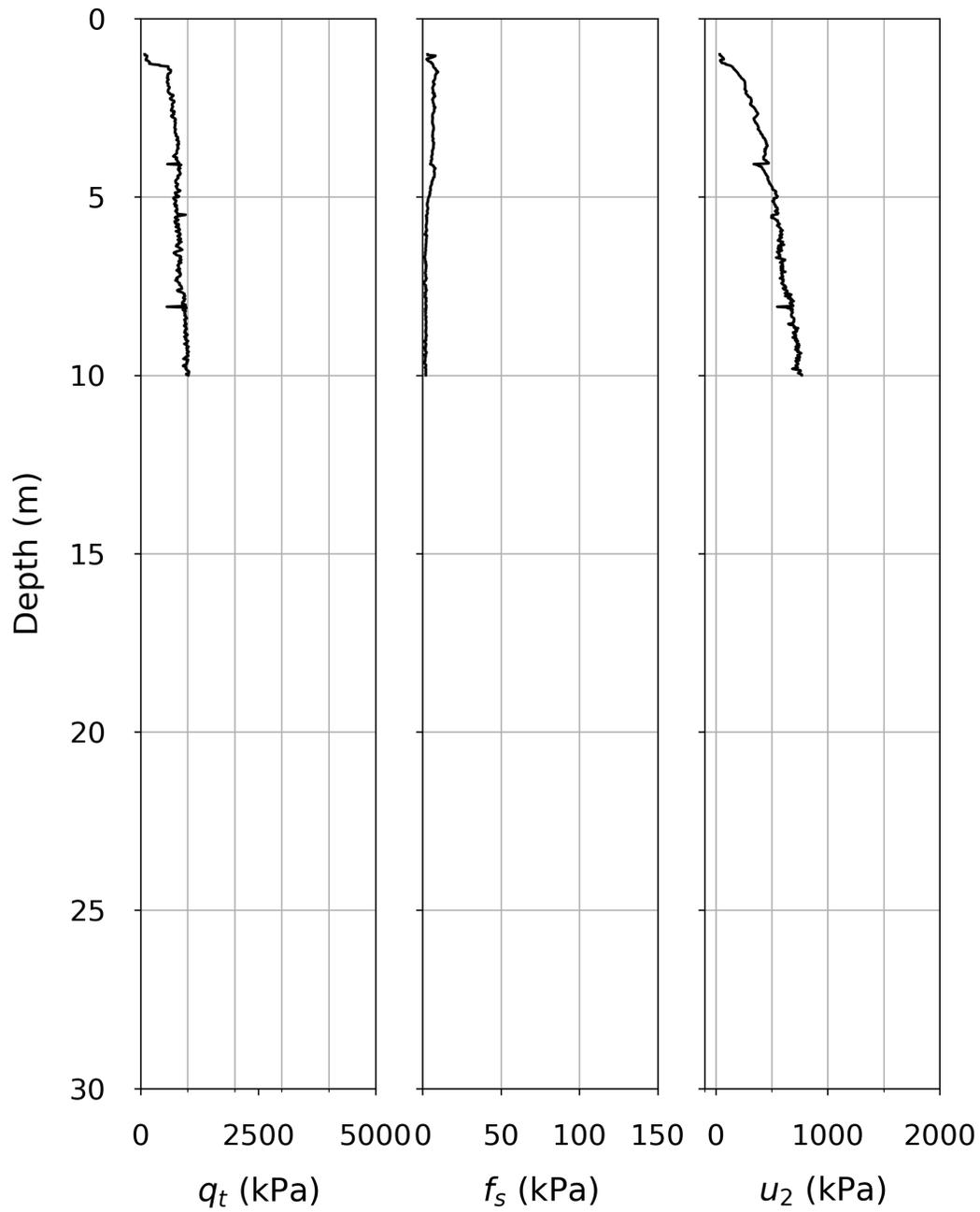
Appendix B

Machine Learning Profiles

Machine Learning Classification CPTu C01
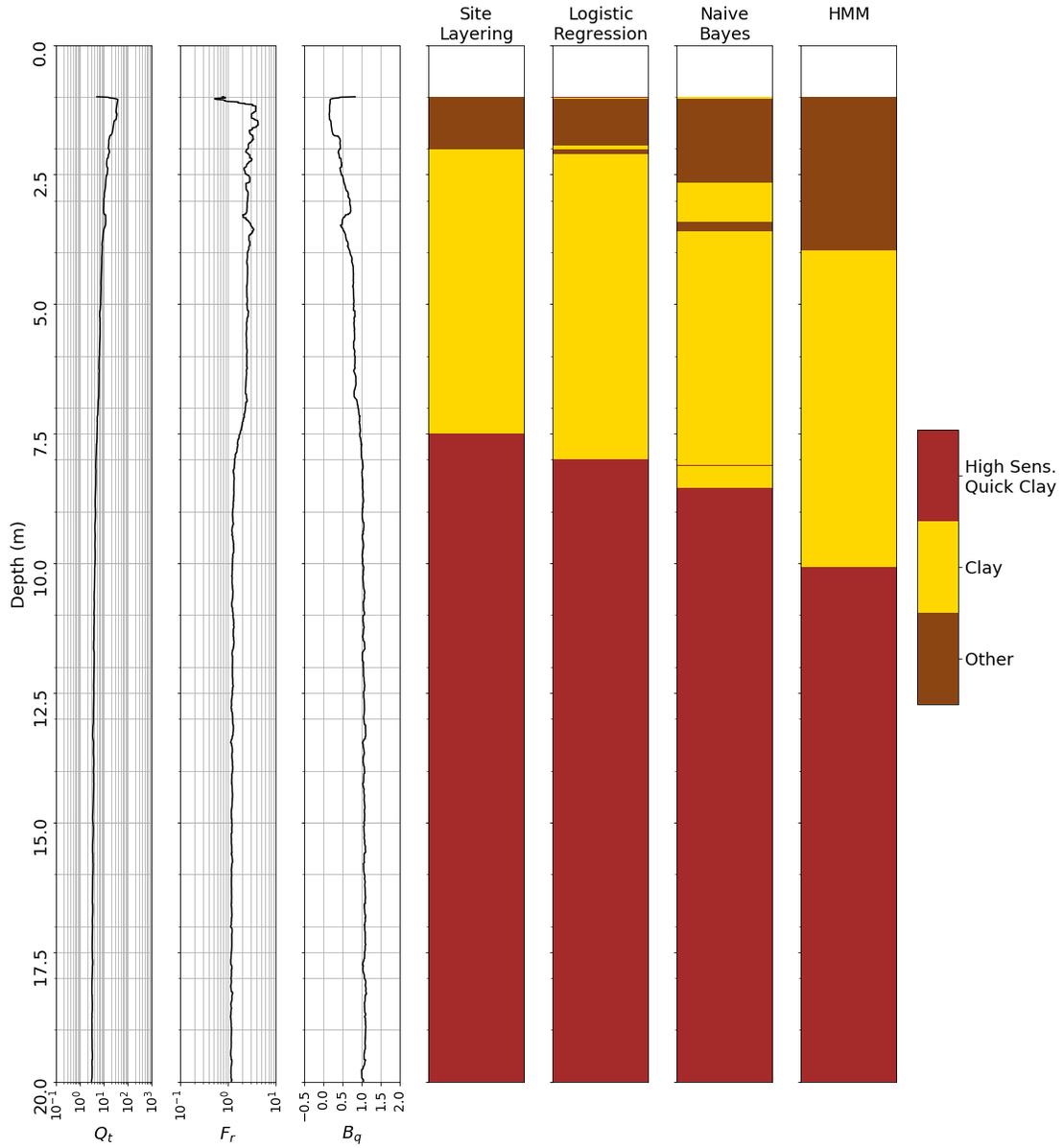
Machine Learning Classification CPTu C02
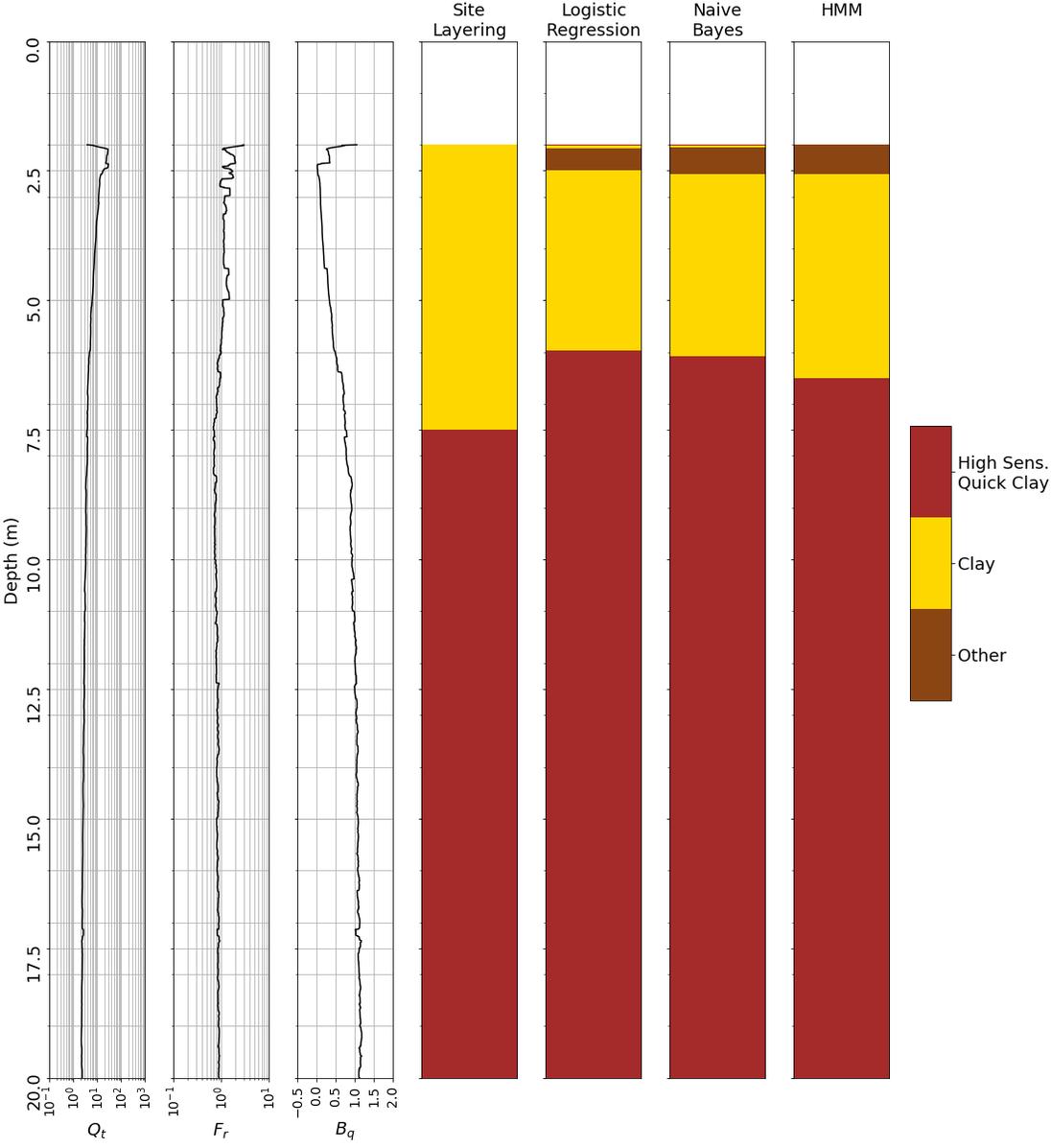
Machine Learning Classification CPTu C03
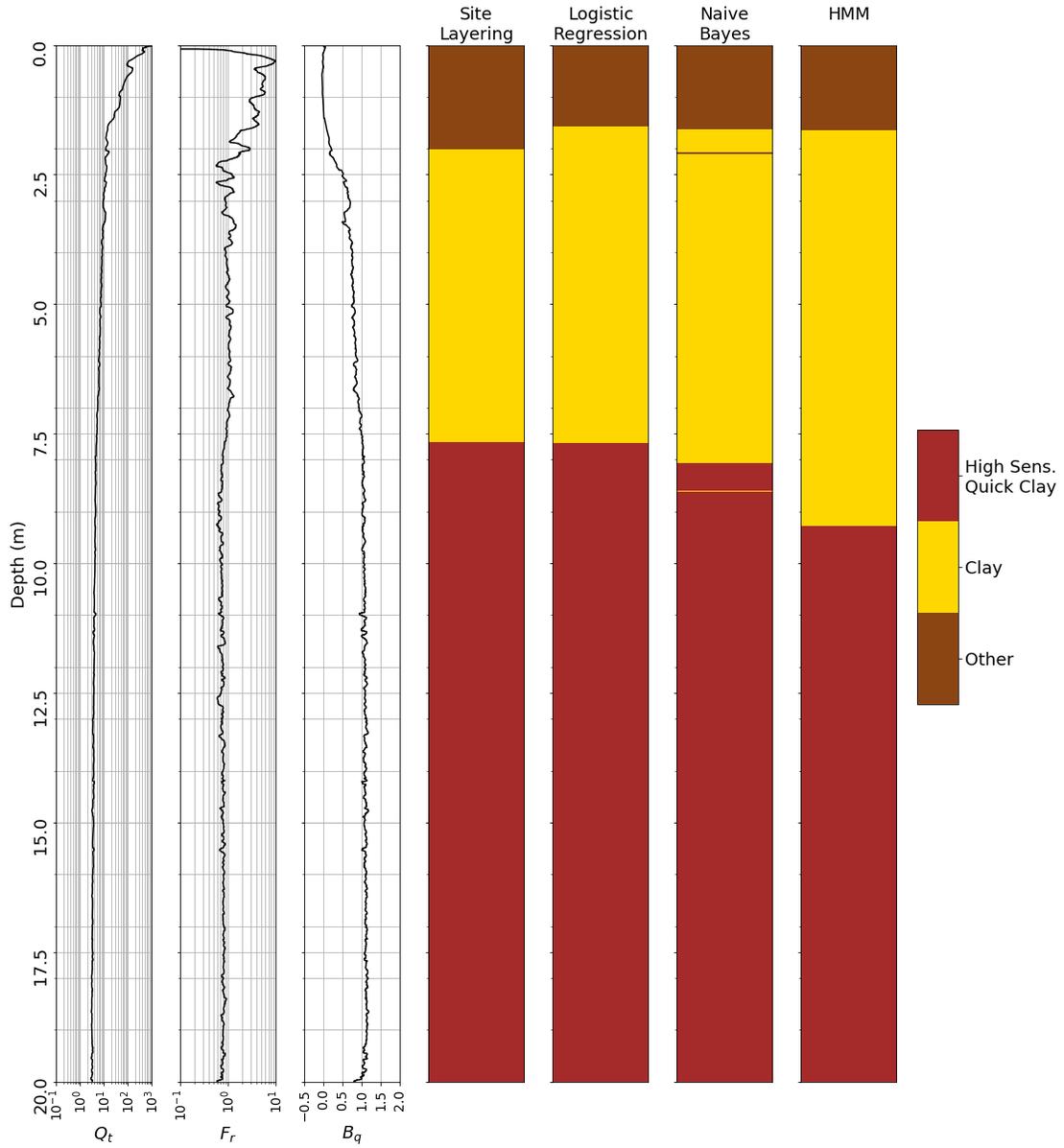
Machine Learning Classification CPTu C04
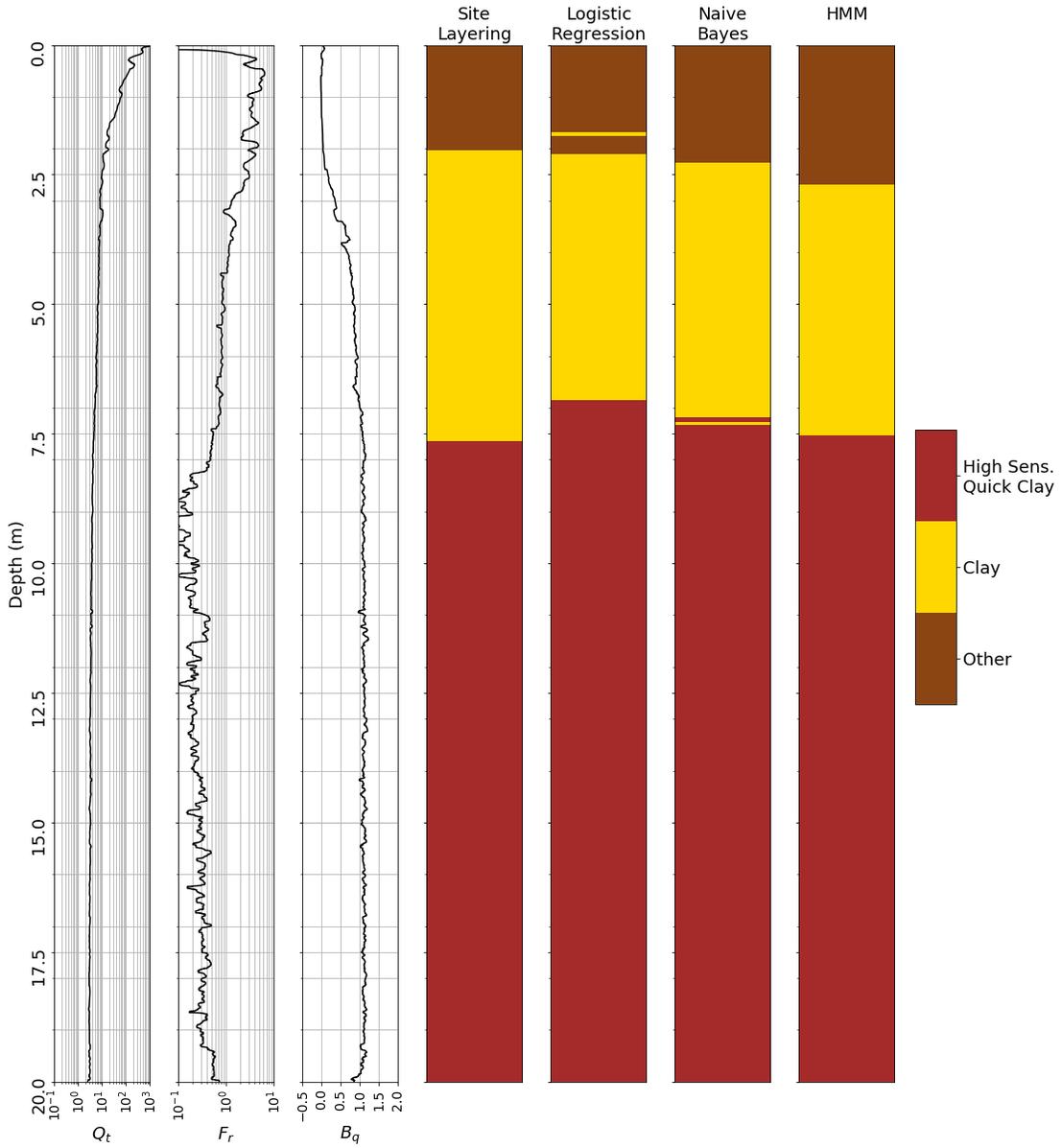
Machine Learning Classification CPTu C05
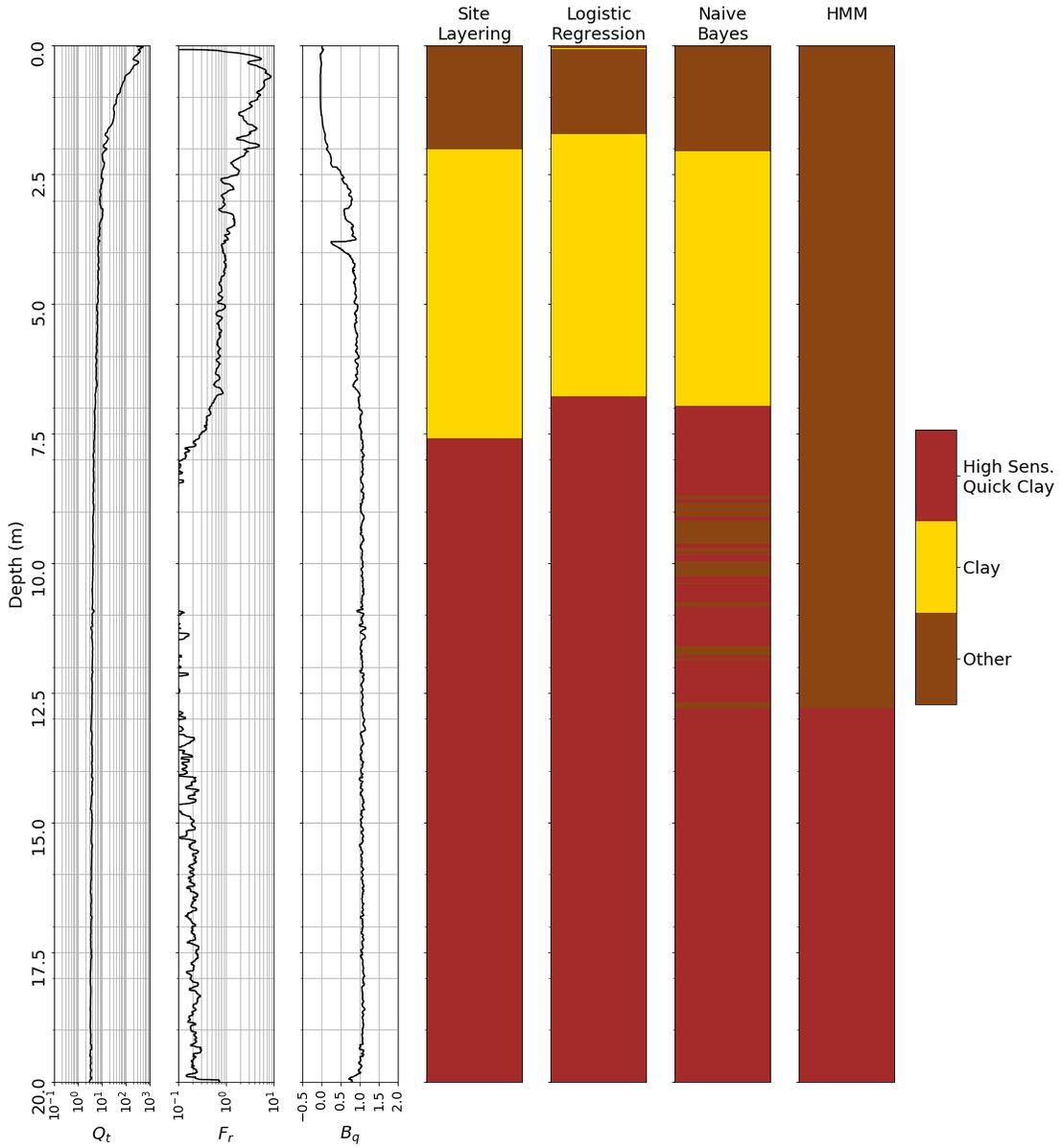
Machine Learning Classification CPTu C06
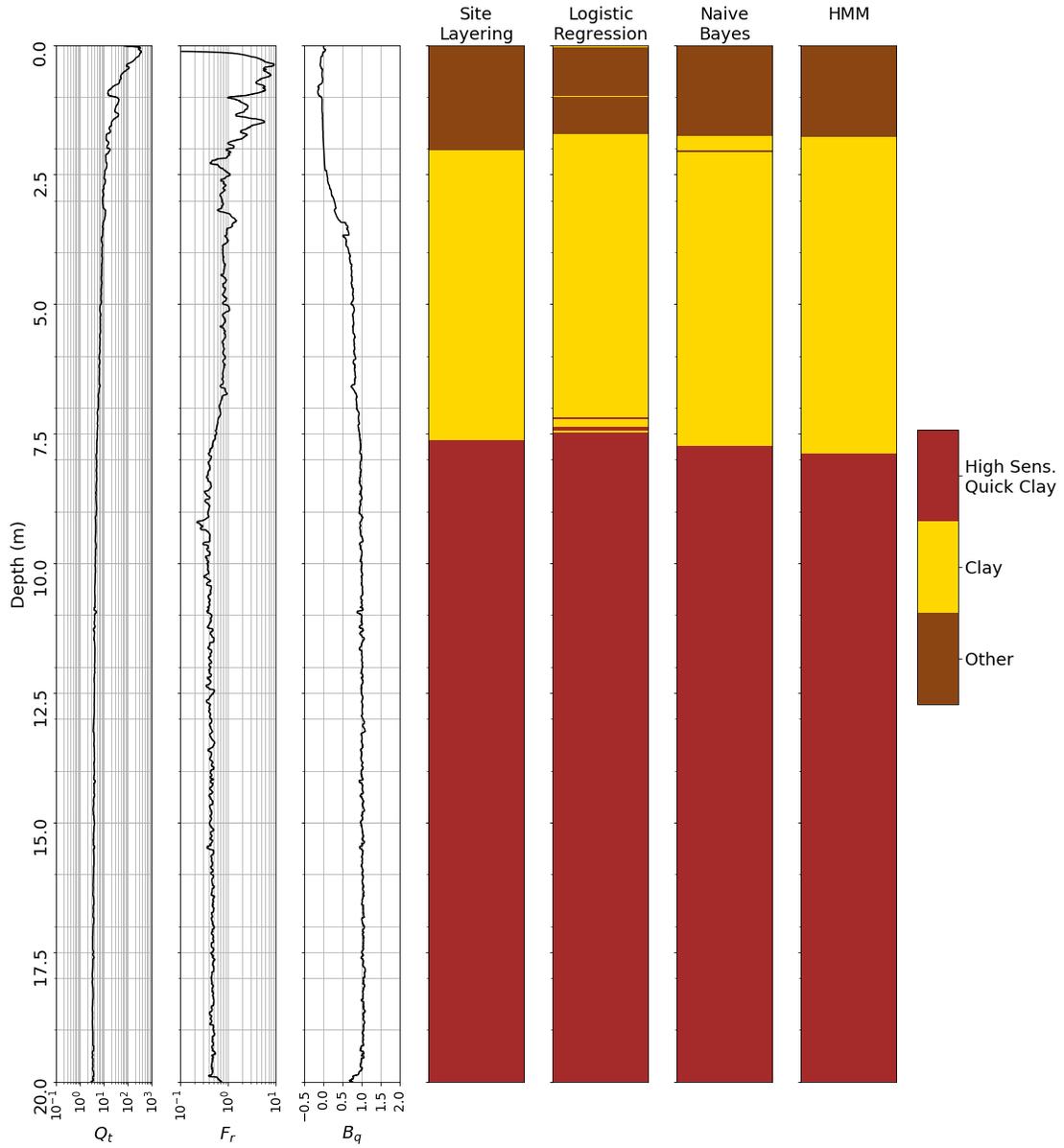
Machine Learning Classification CPTu C07
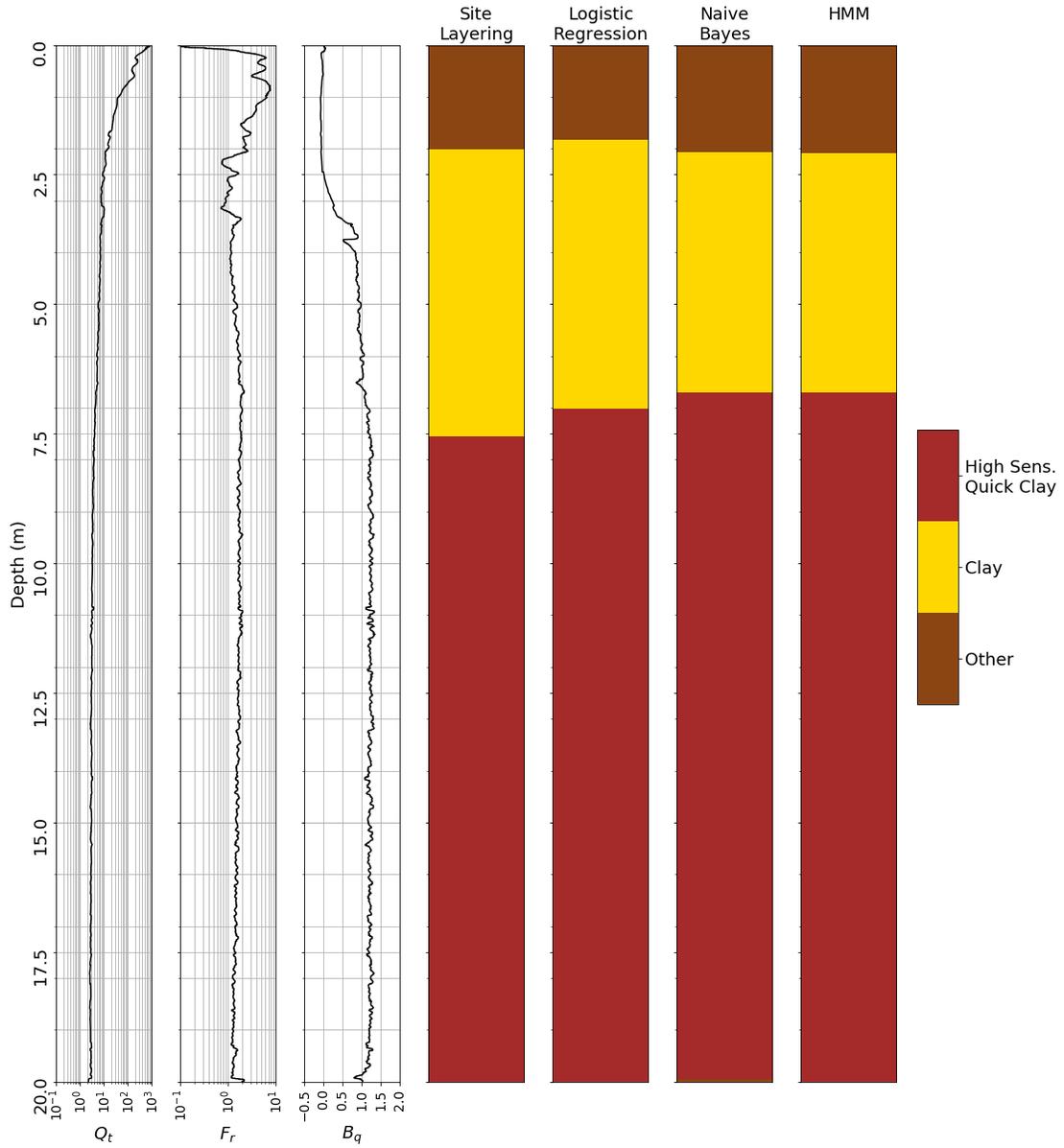
Machine Learning Classification CPTu C08
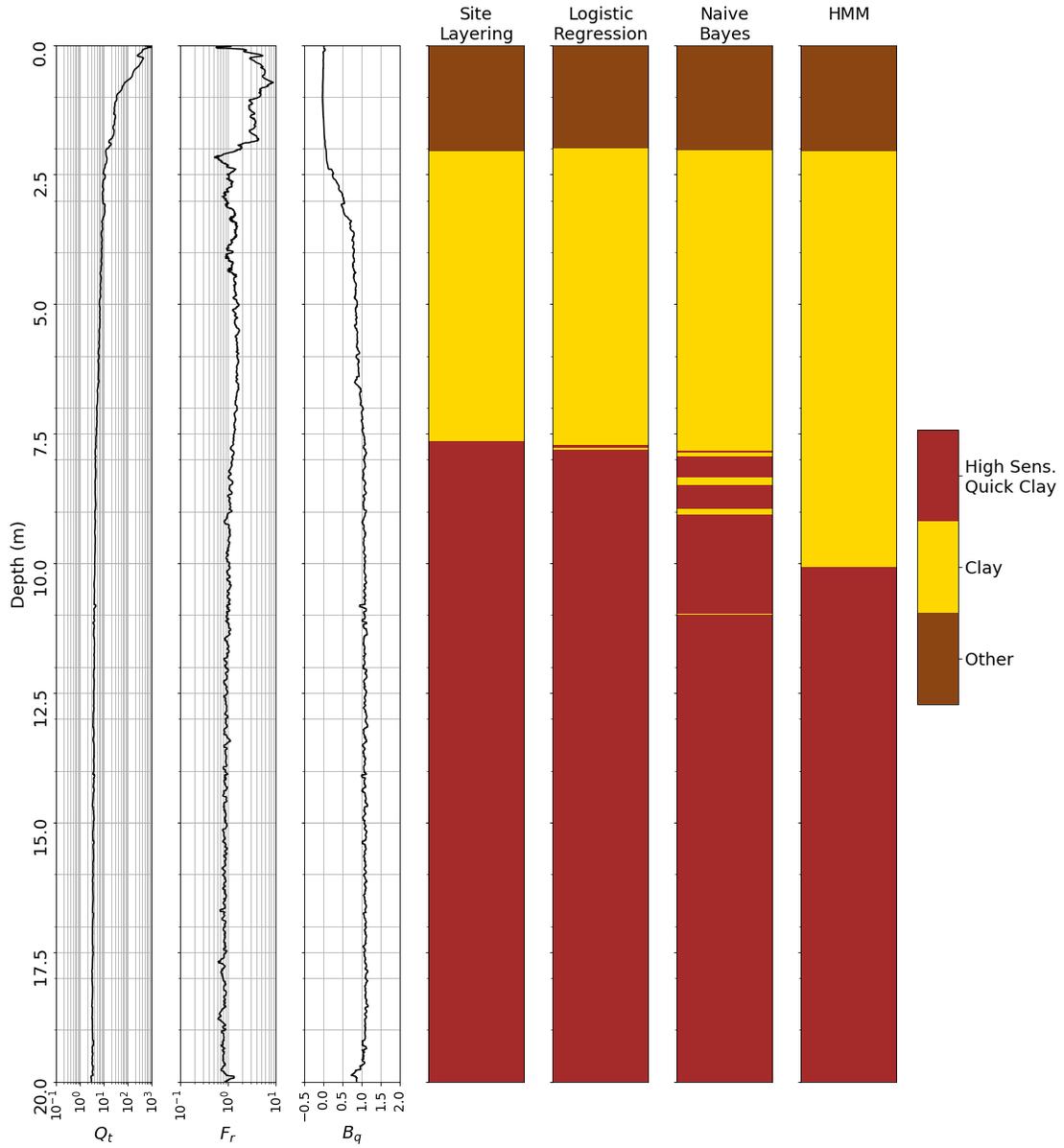
Machine Learning Classification CPTu C09
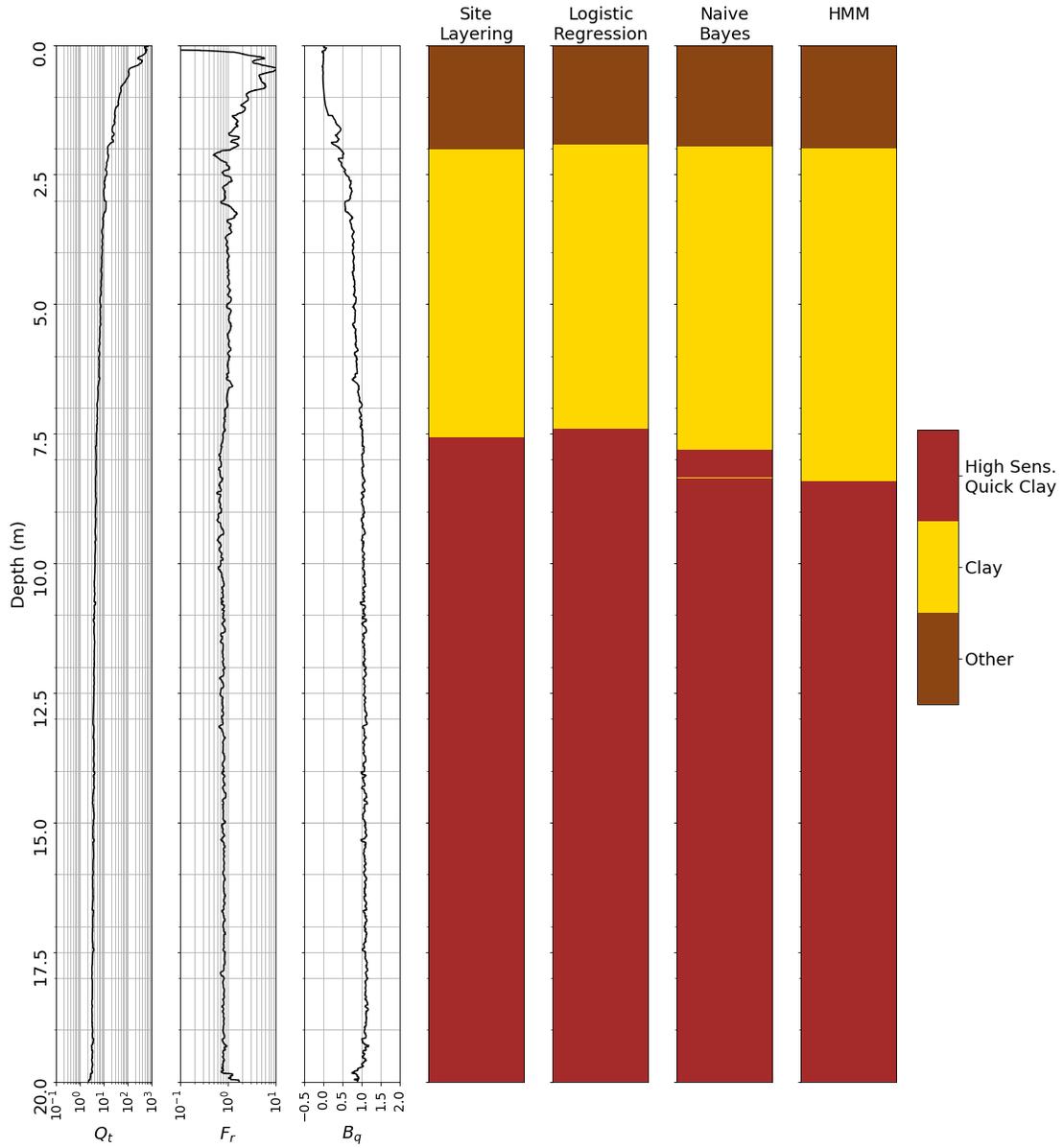
Machine Learning Classification CPTu C10
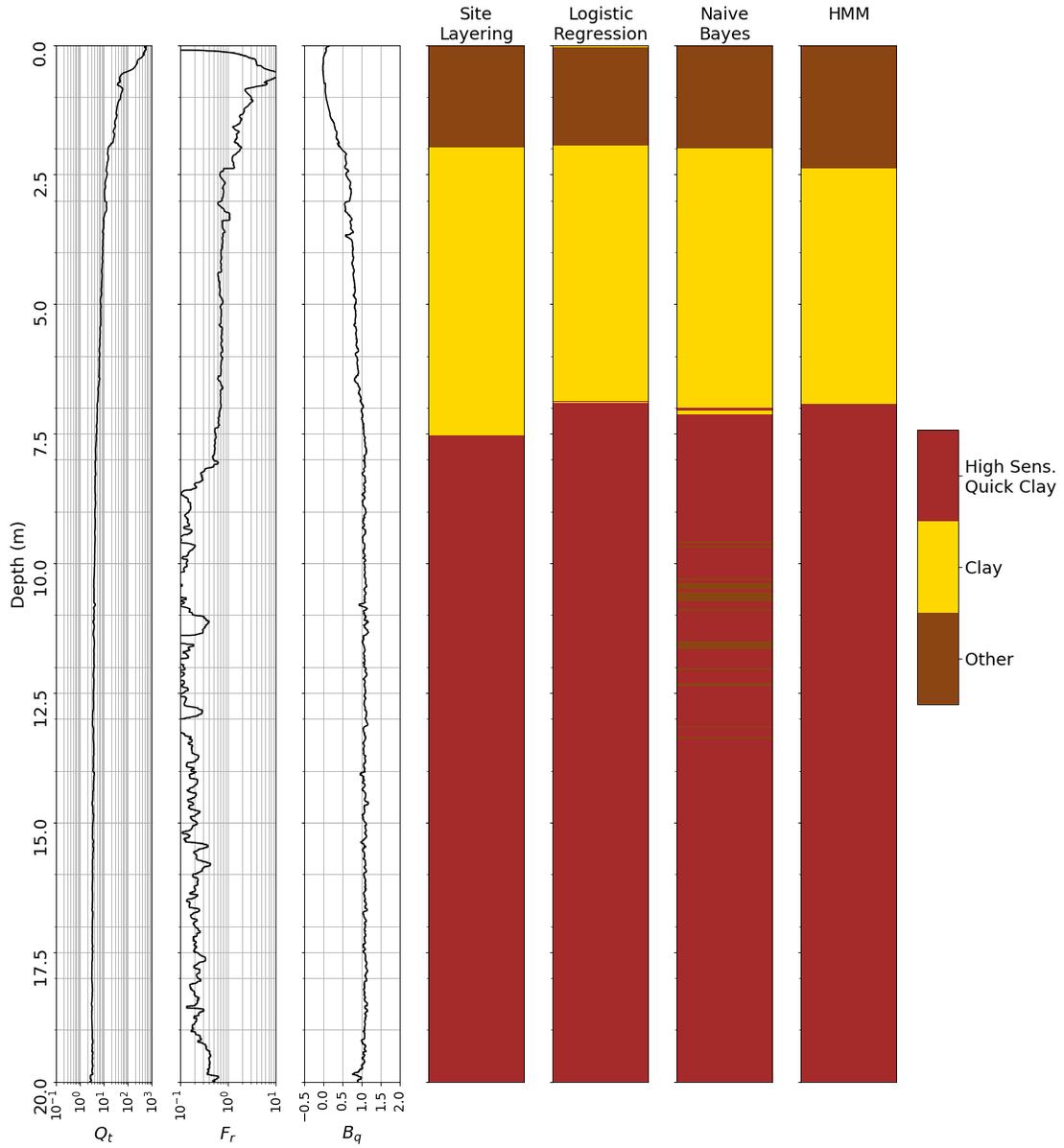
Machine Learning Classification CPTu C11
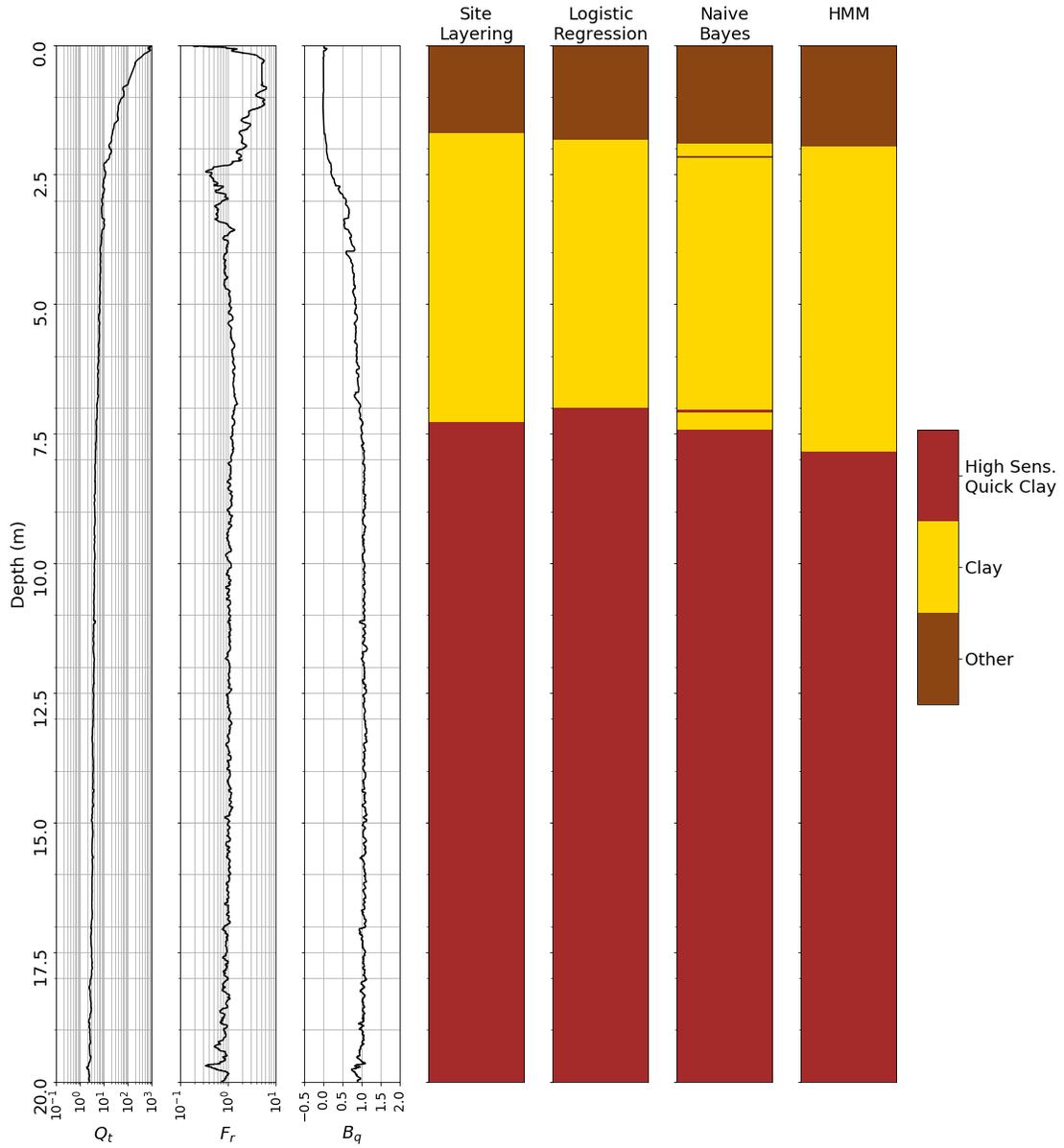
Machine Learning Classification CPTu C12
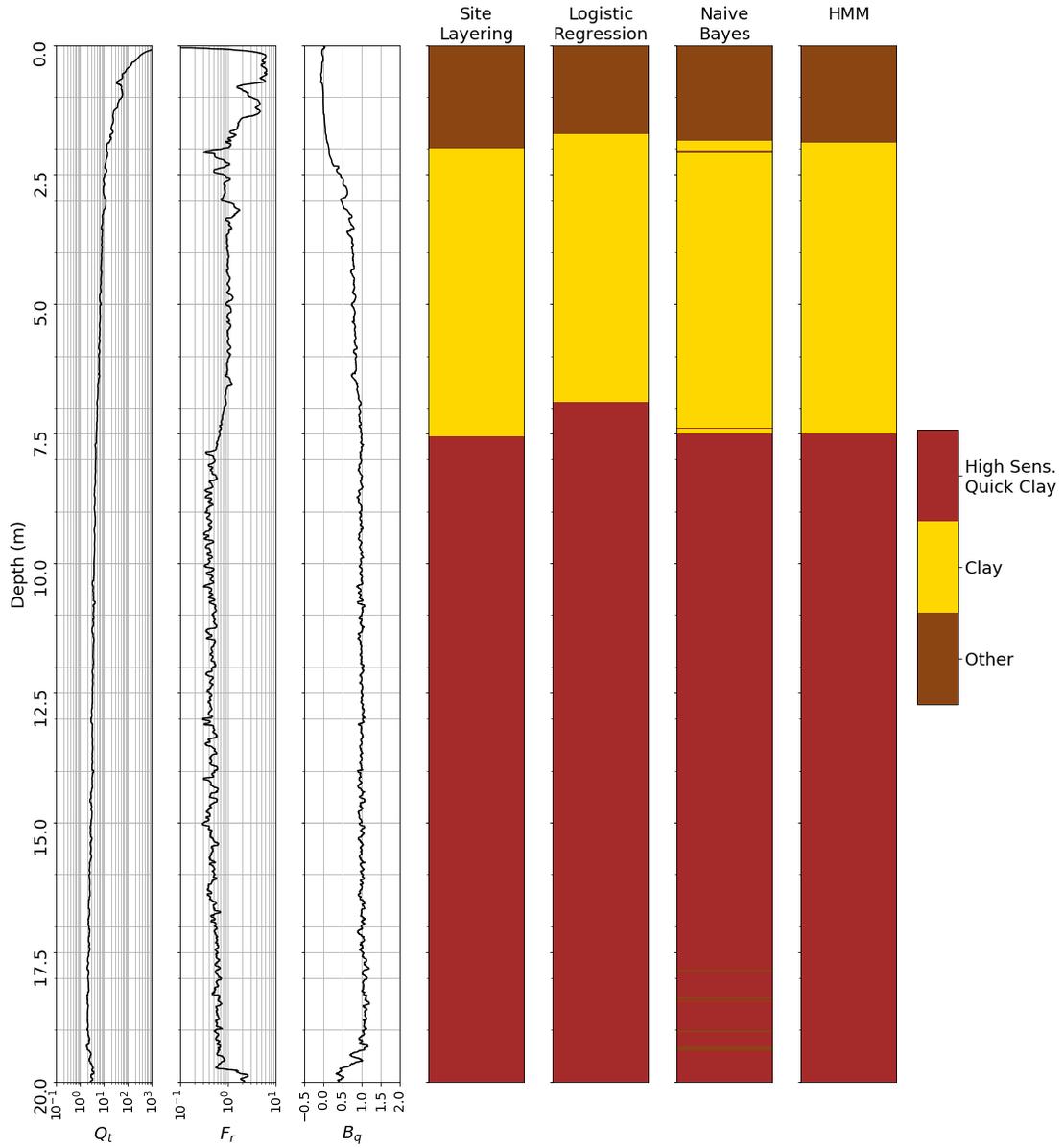
Machine Learning Classification CPTu C13
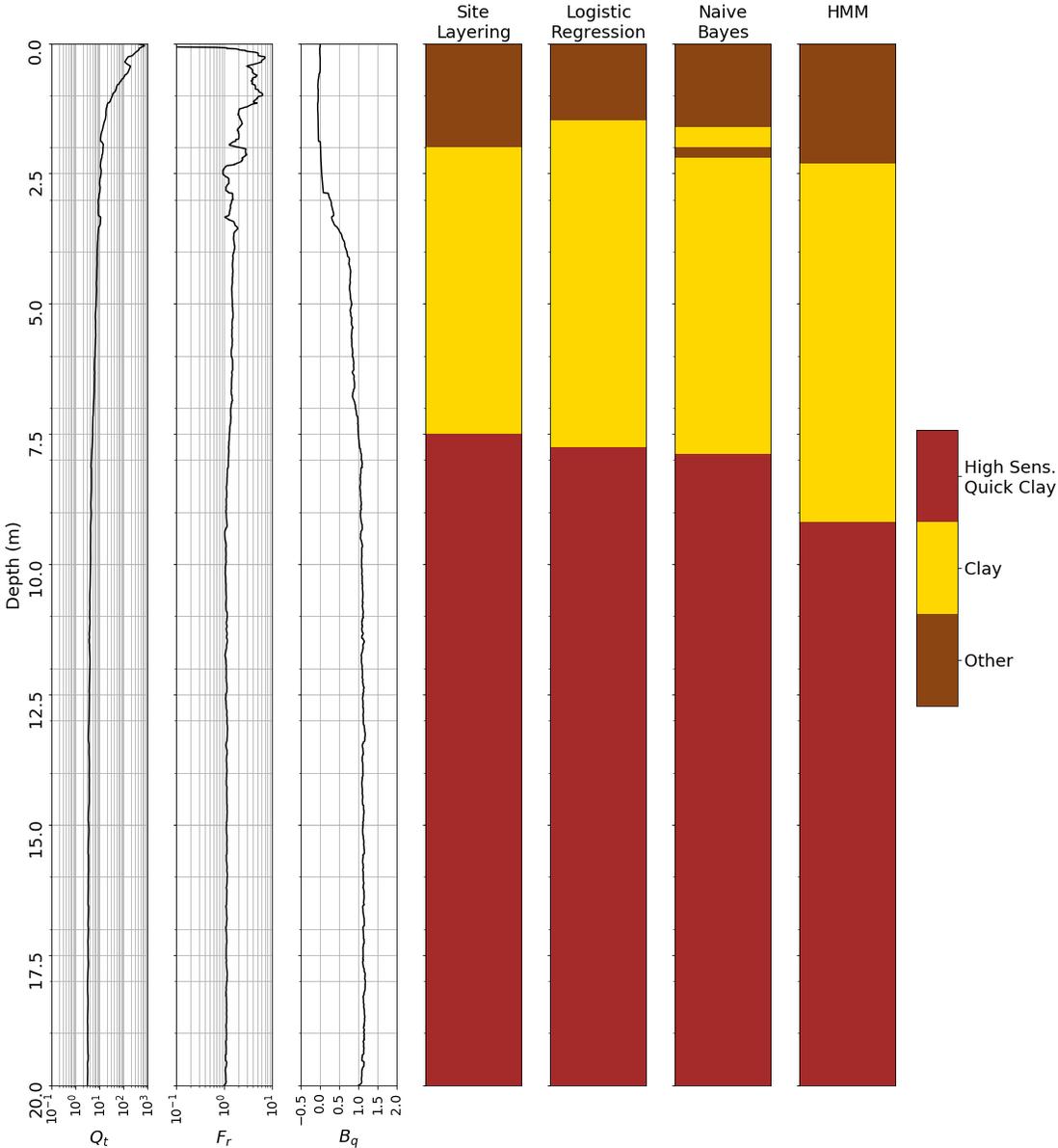
Machine Learning Classification CPTu C14
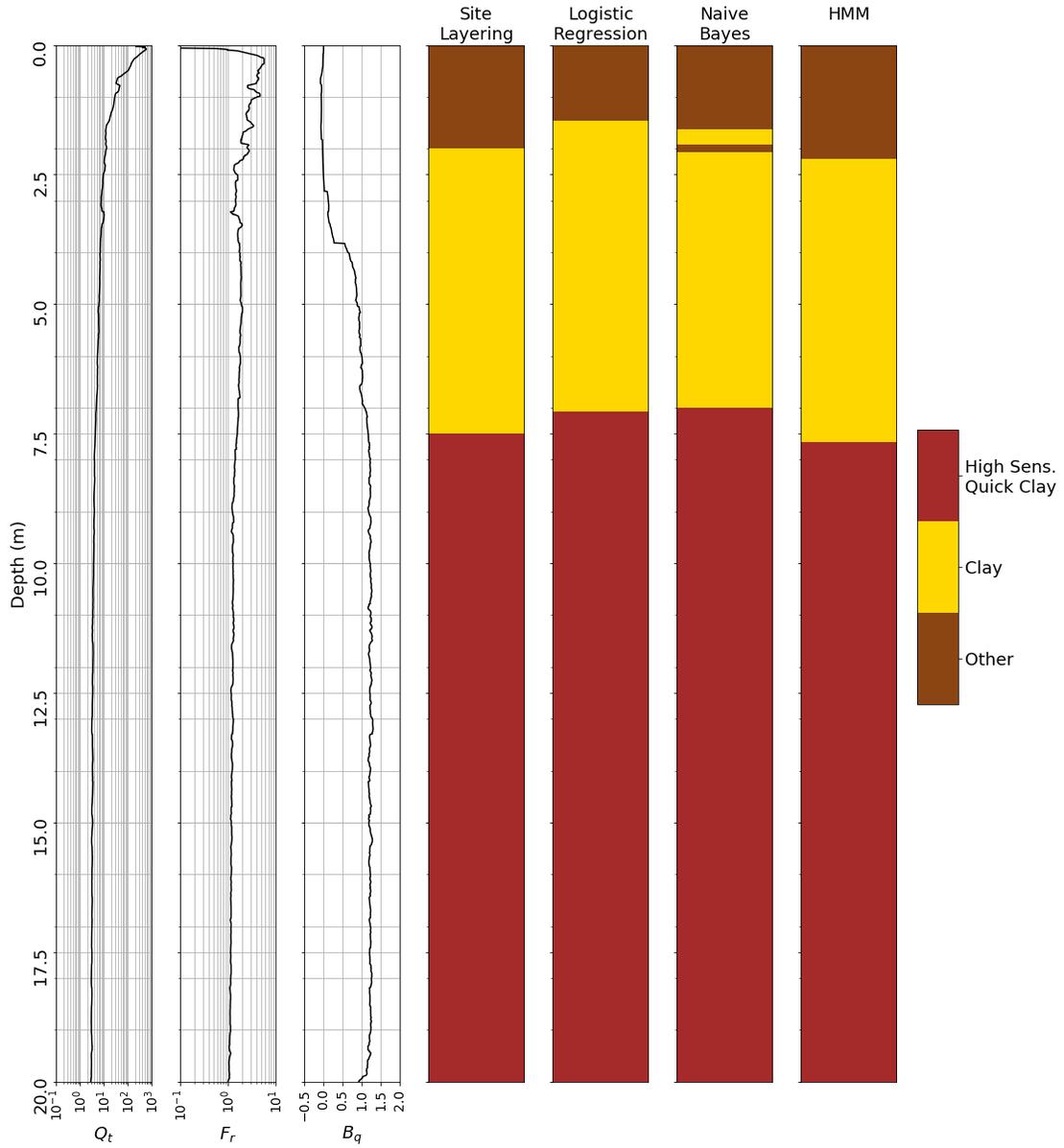
Machine Learning Classification CPTu C15
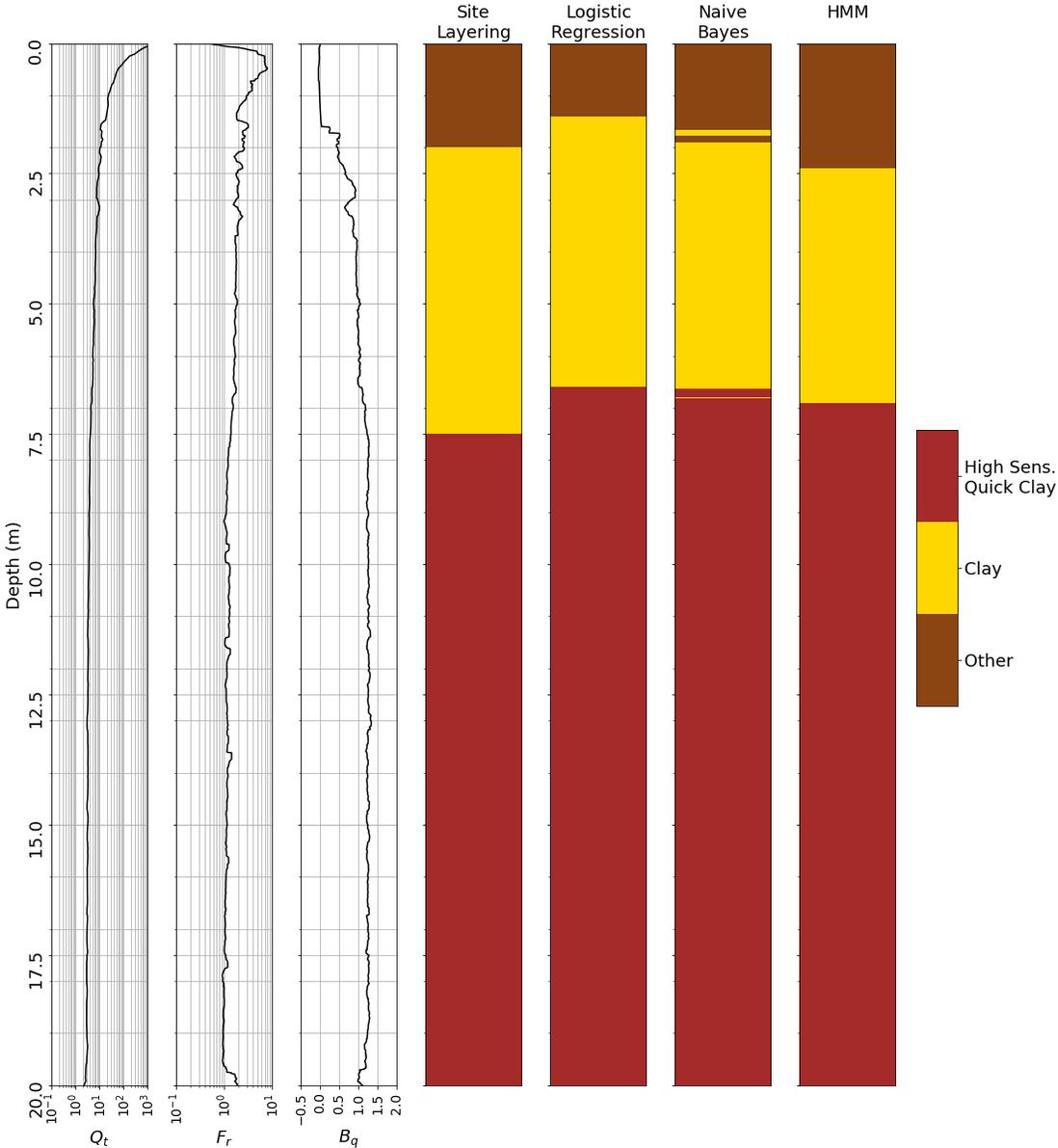
Machine Learning Classification CPTu C16
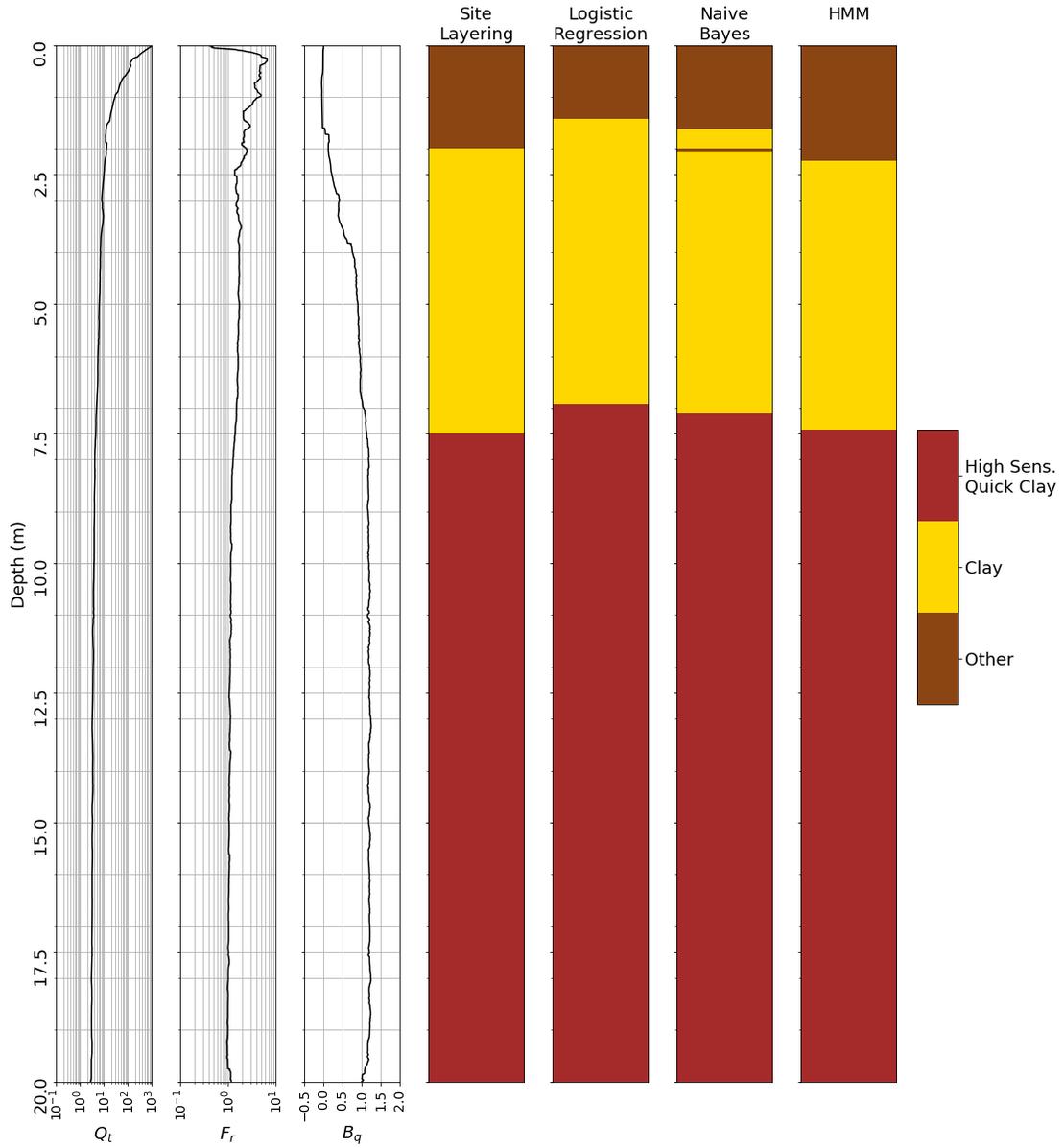
Machine Learning Classification CPTu C17
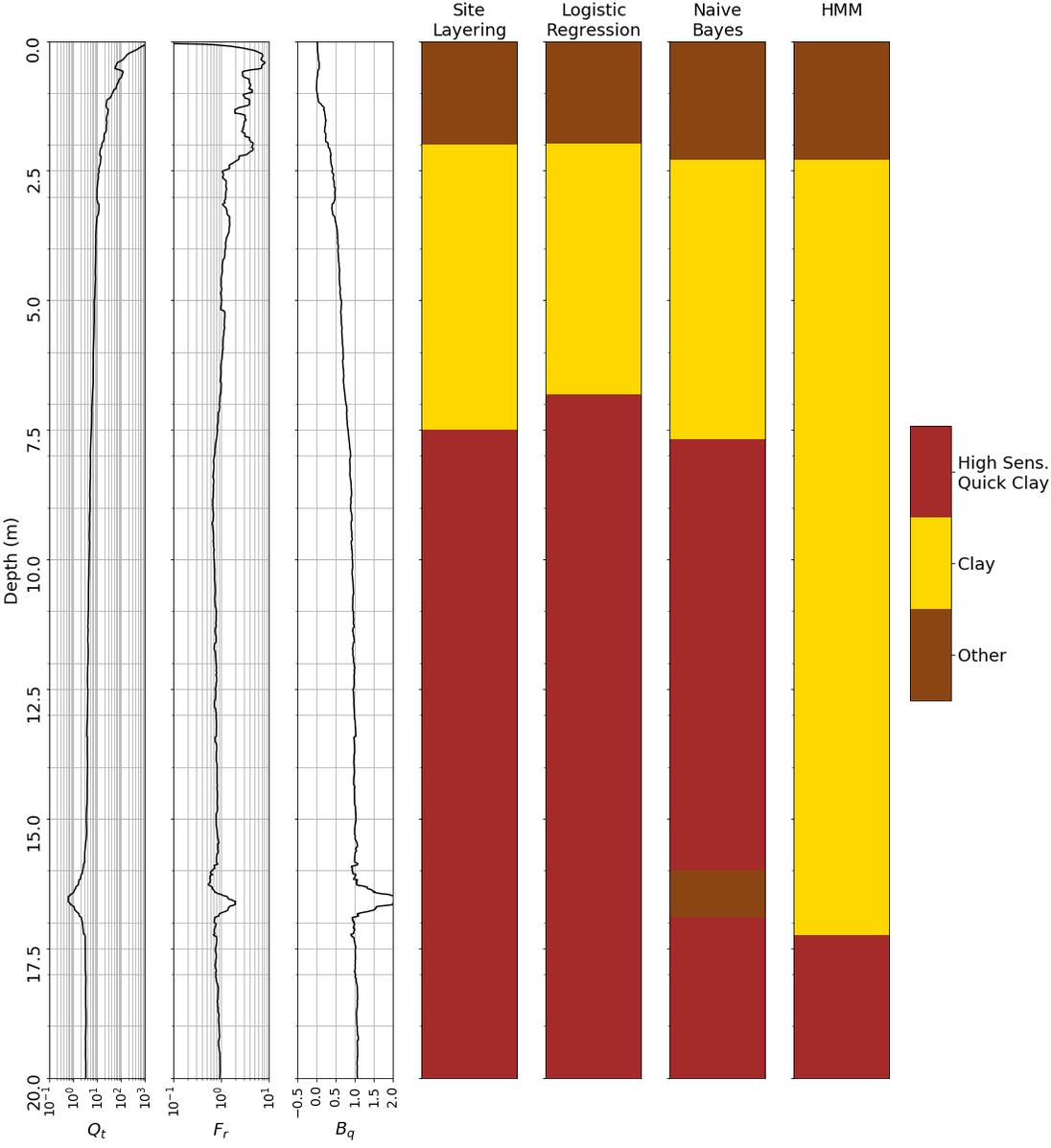
Machine Learning Classification CPTu C19
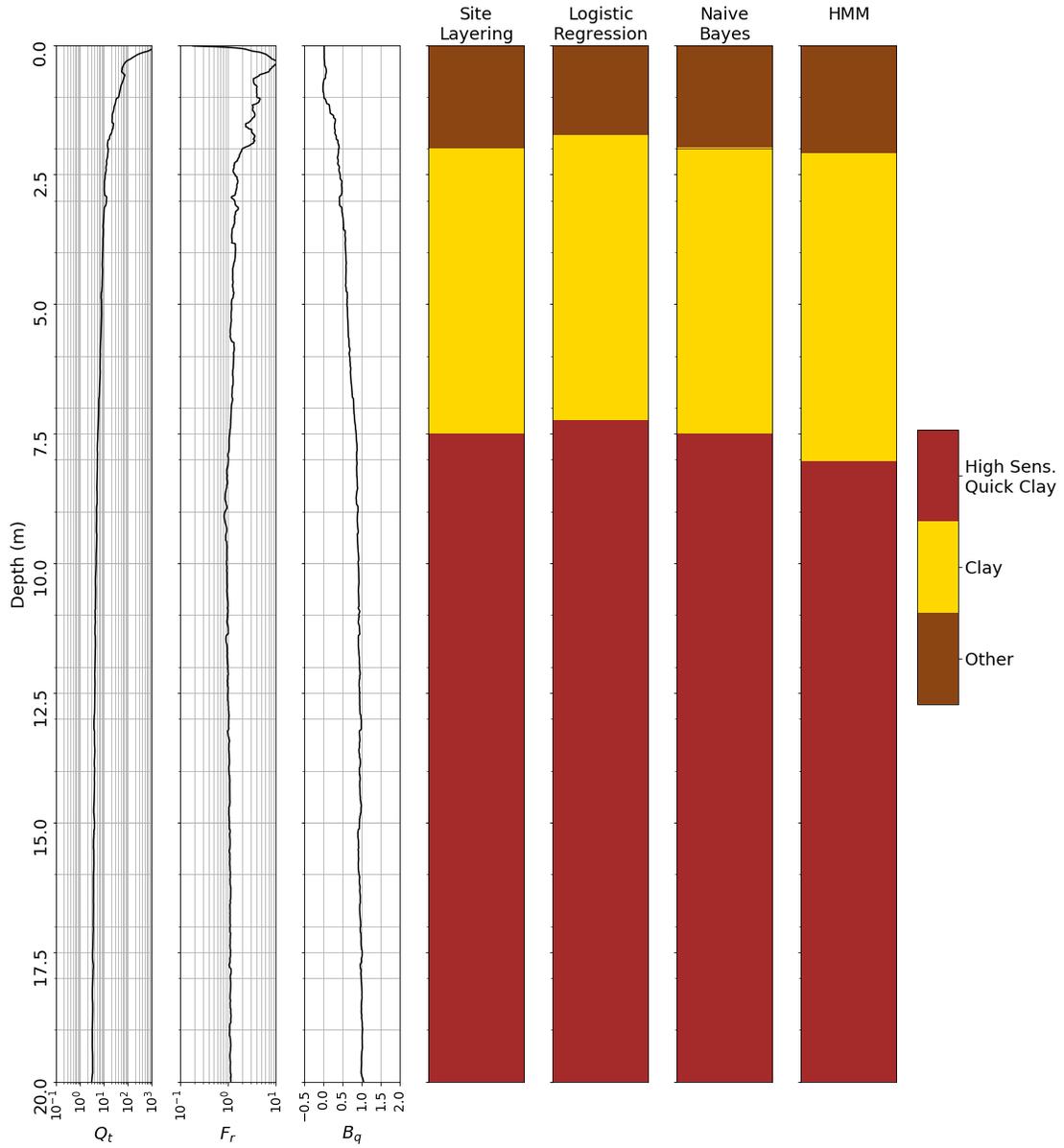
Machine Learning Classification CPTu C20
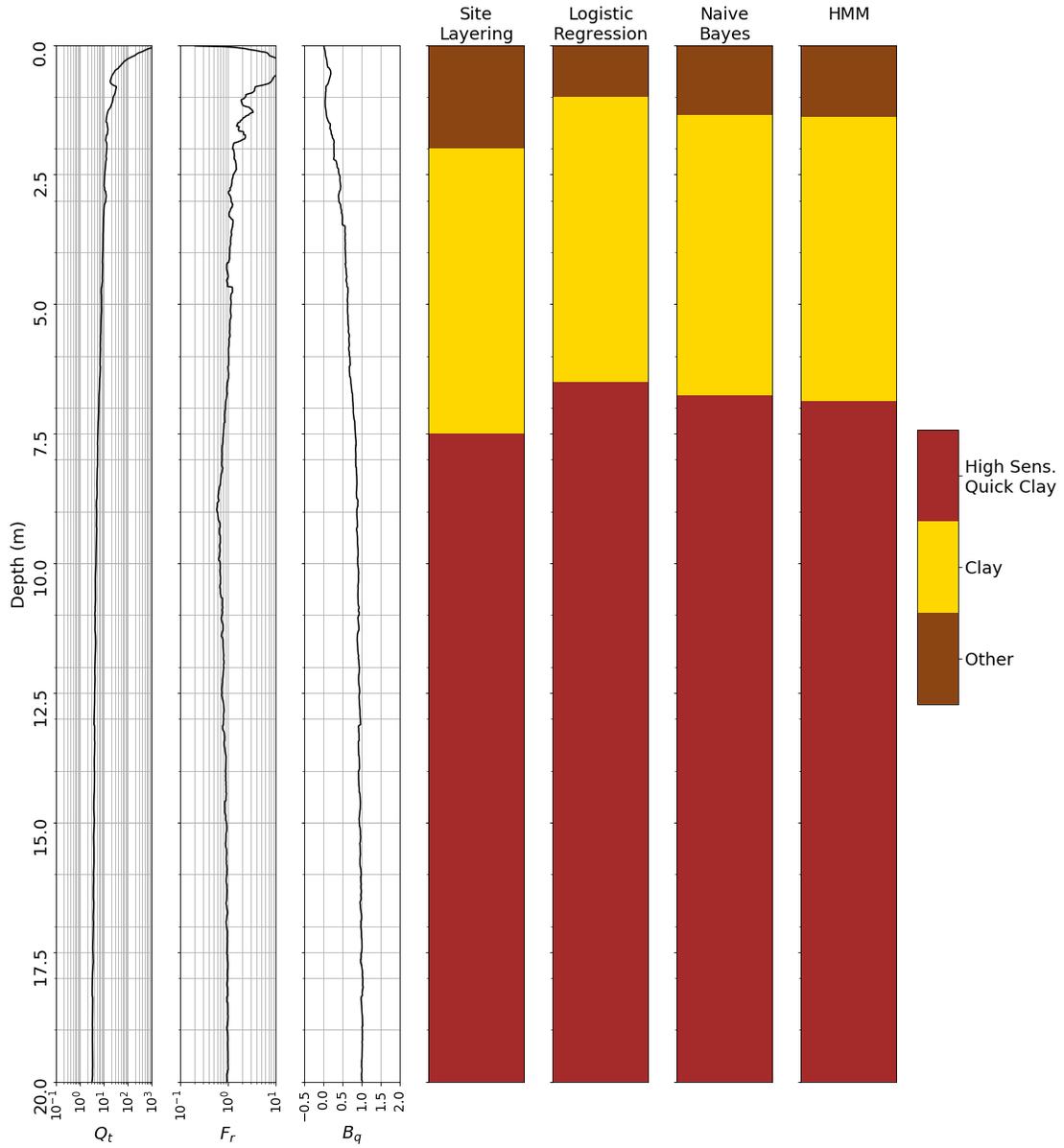
Machine Learning Classification CPTu C22
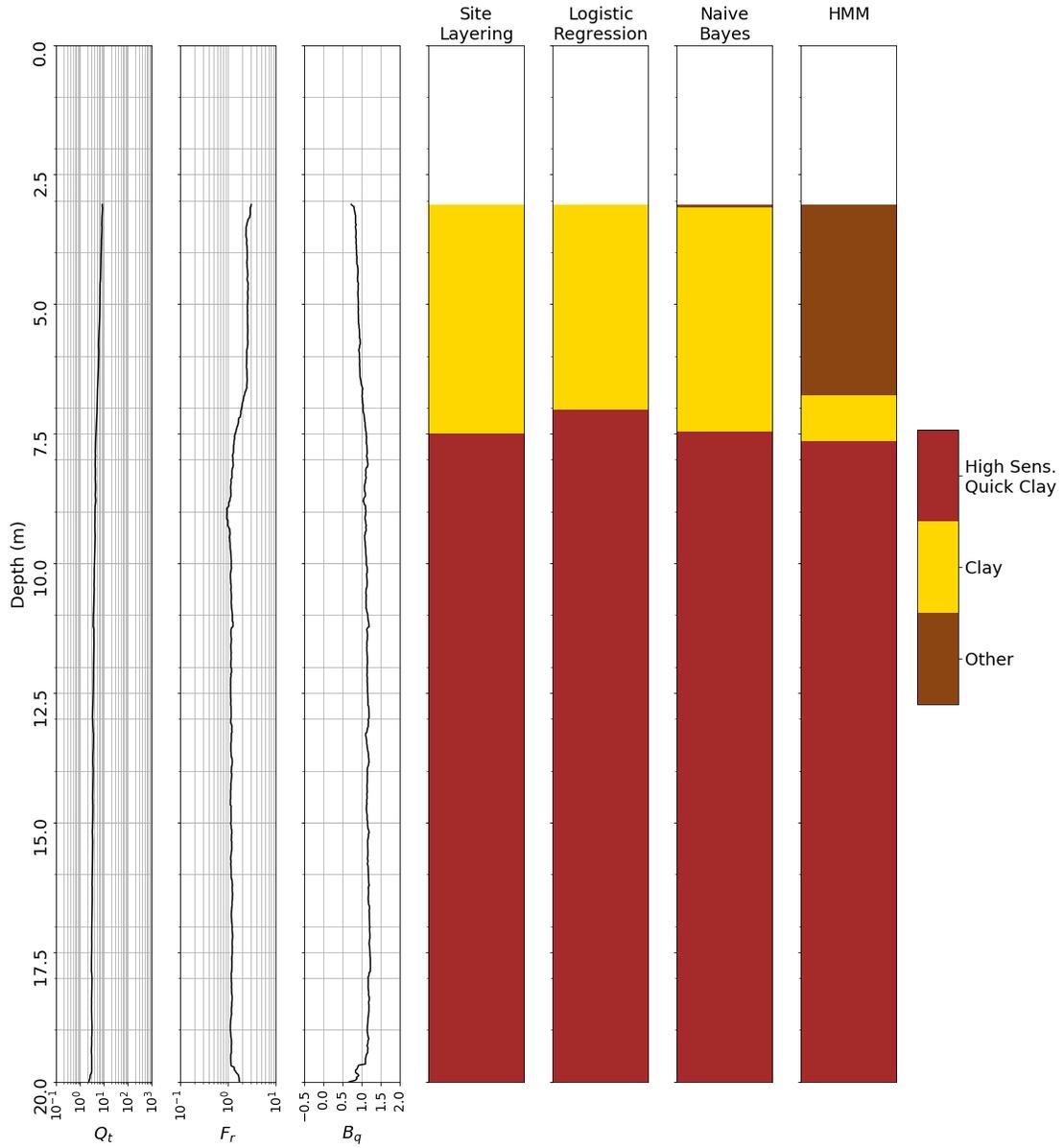
Machine Learning Classification CPTu C23
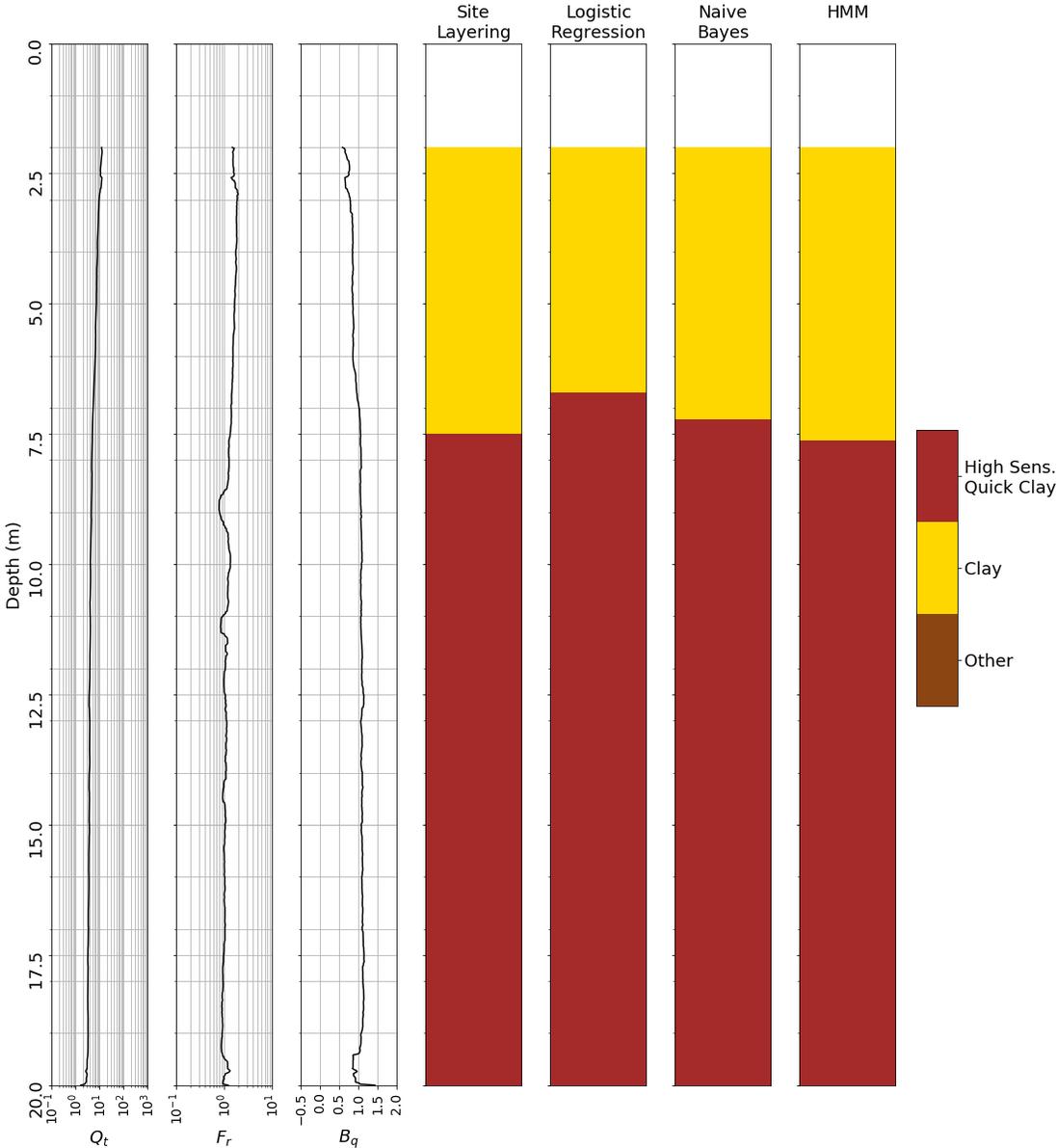
Machine Learning Classification CPTu C24
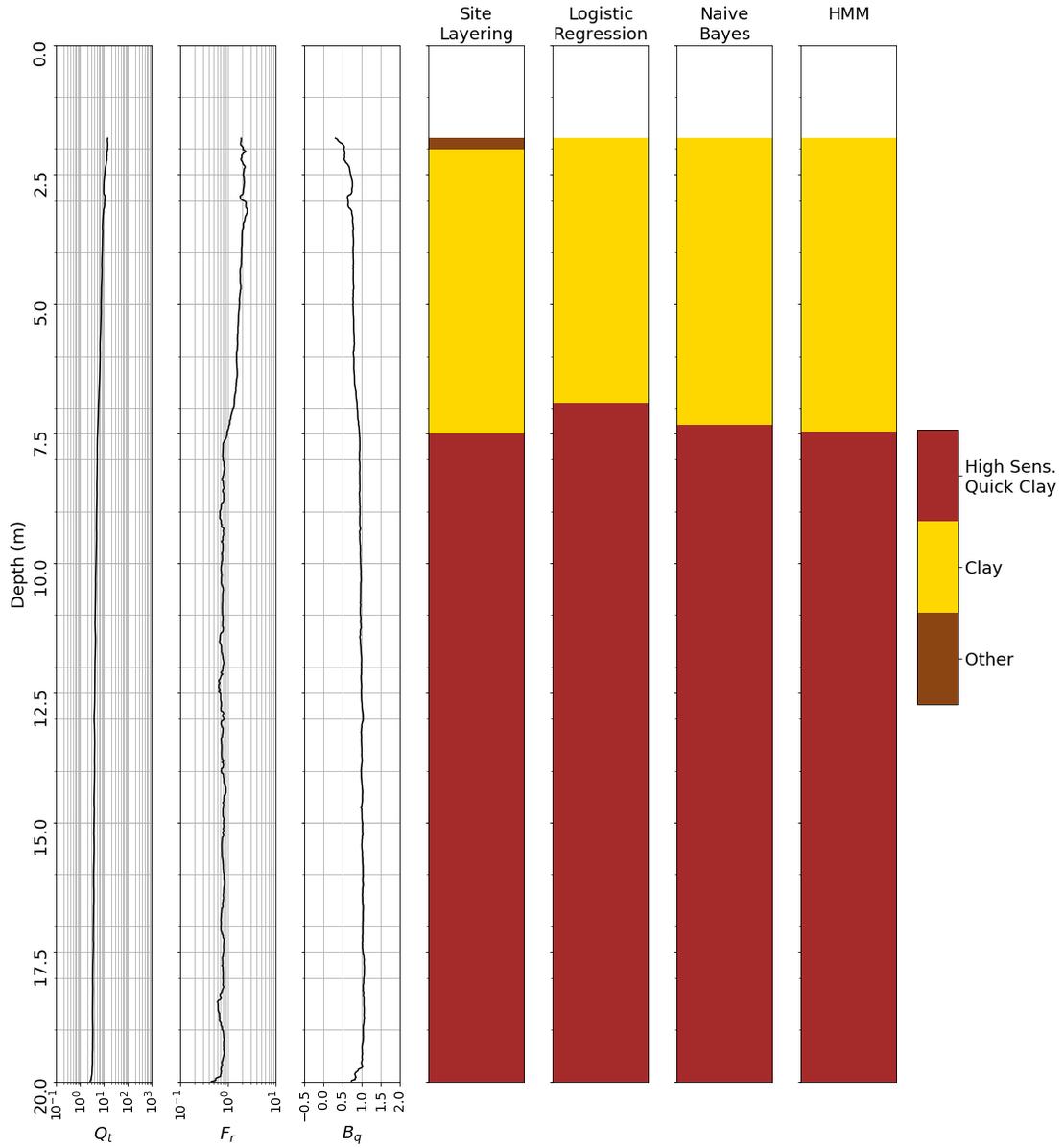
Machine Learning Classification CPTu C25
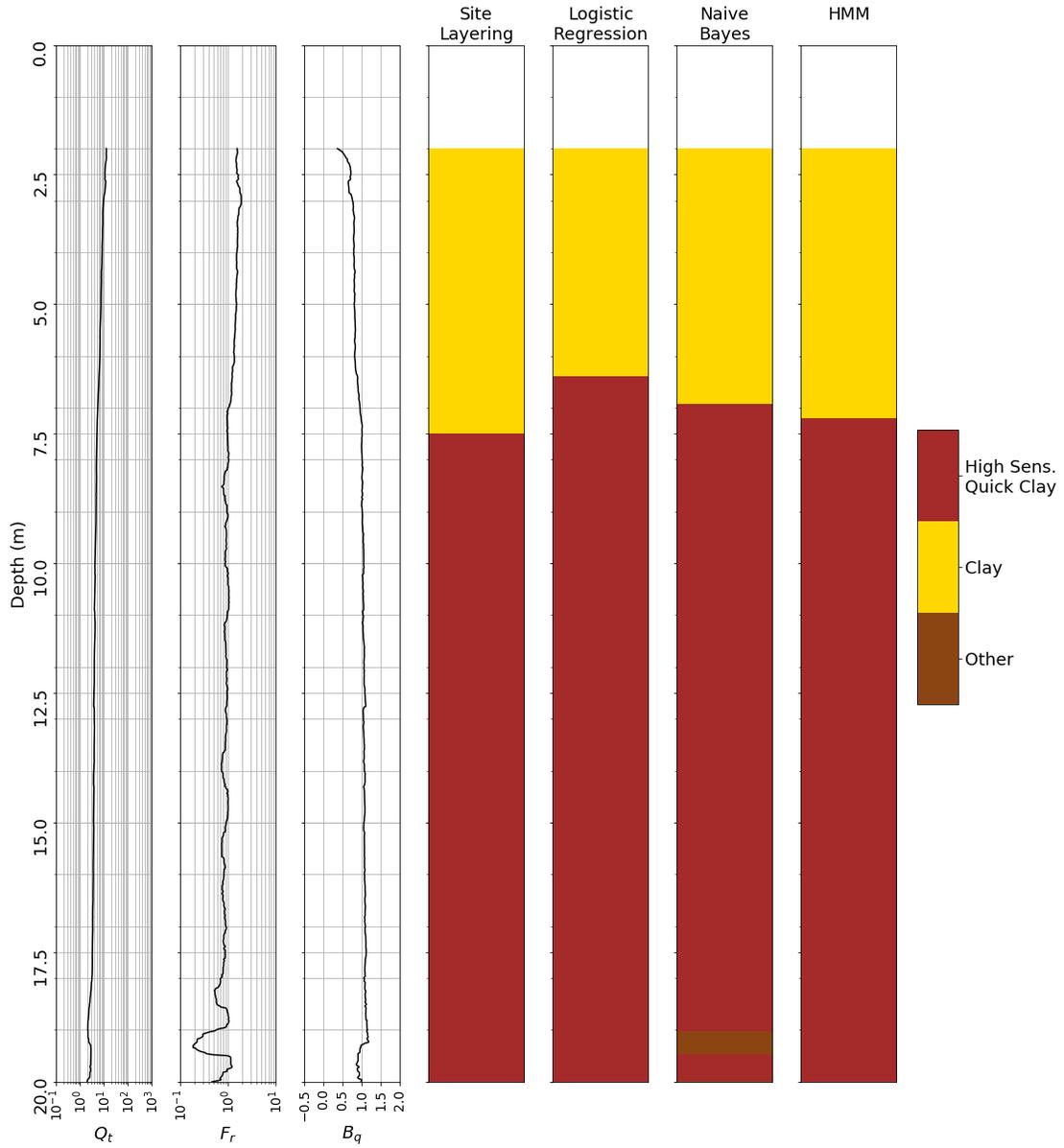
Machine Learning Classification CPTu C26
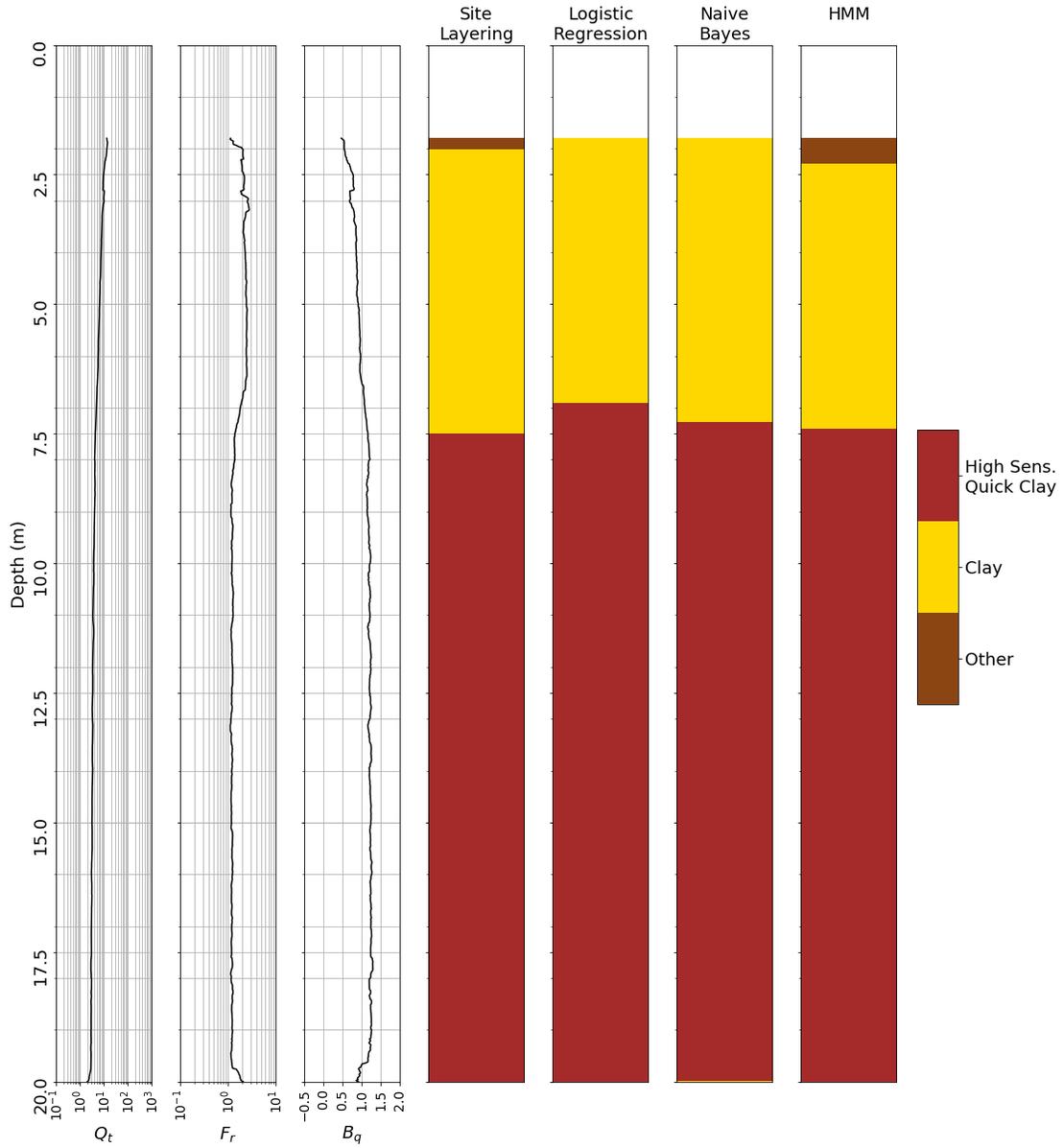
Machine Learning Classification CPTu C27
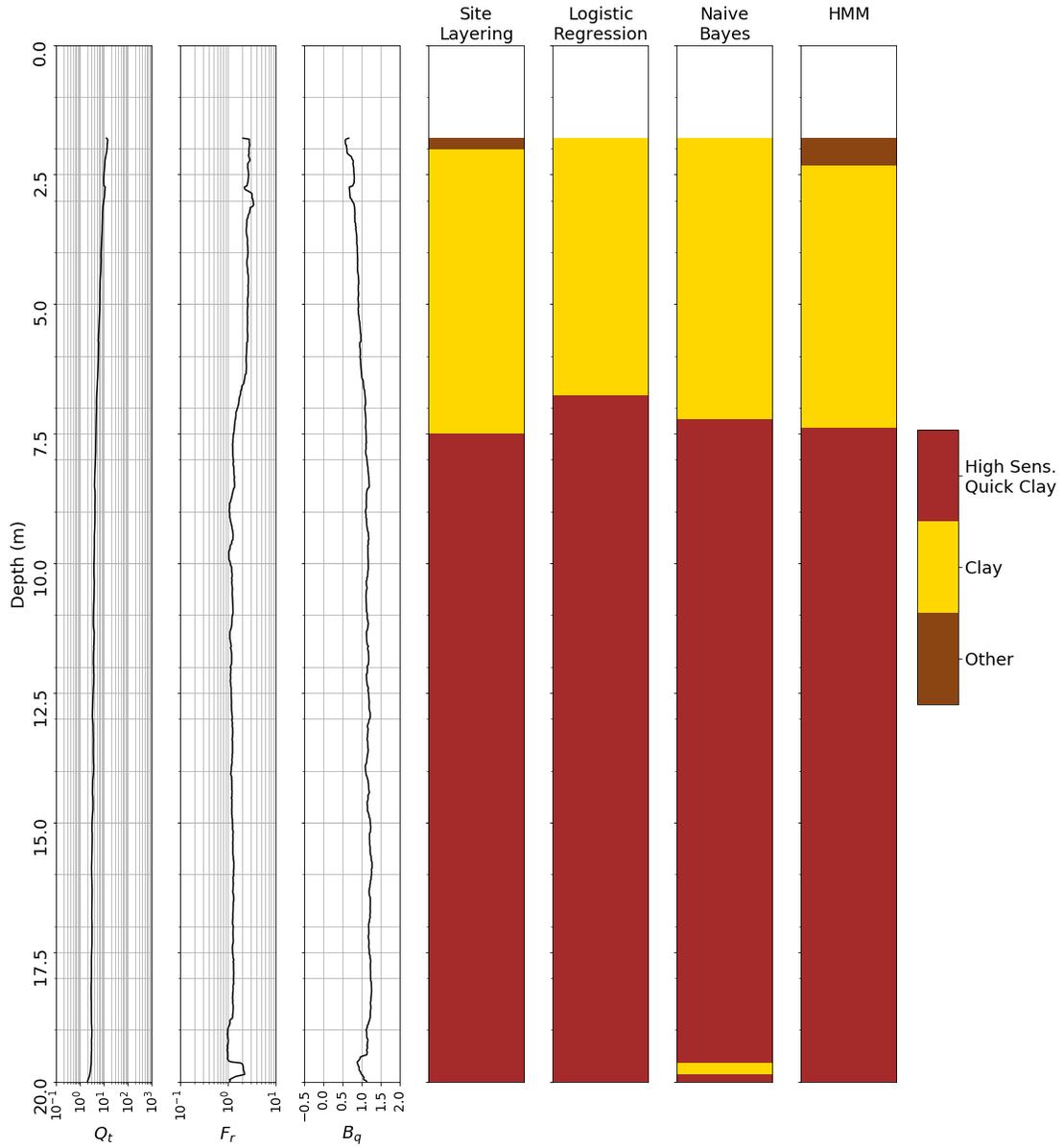
Machine Learning Classification CPTu C28
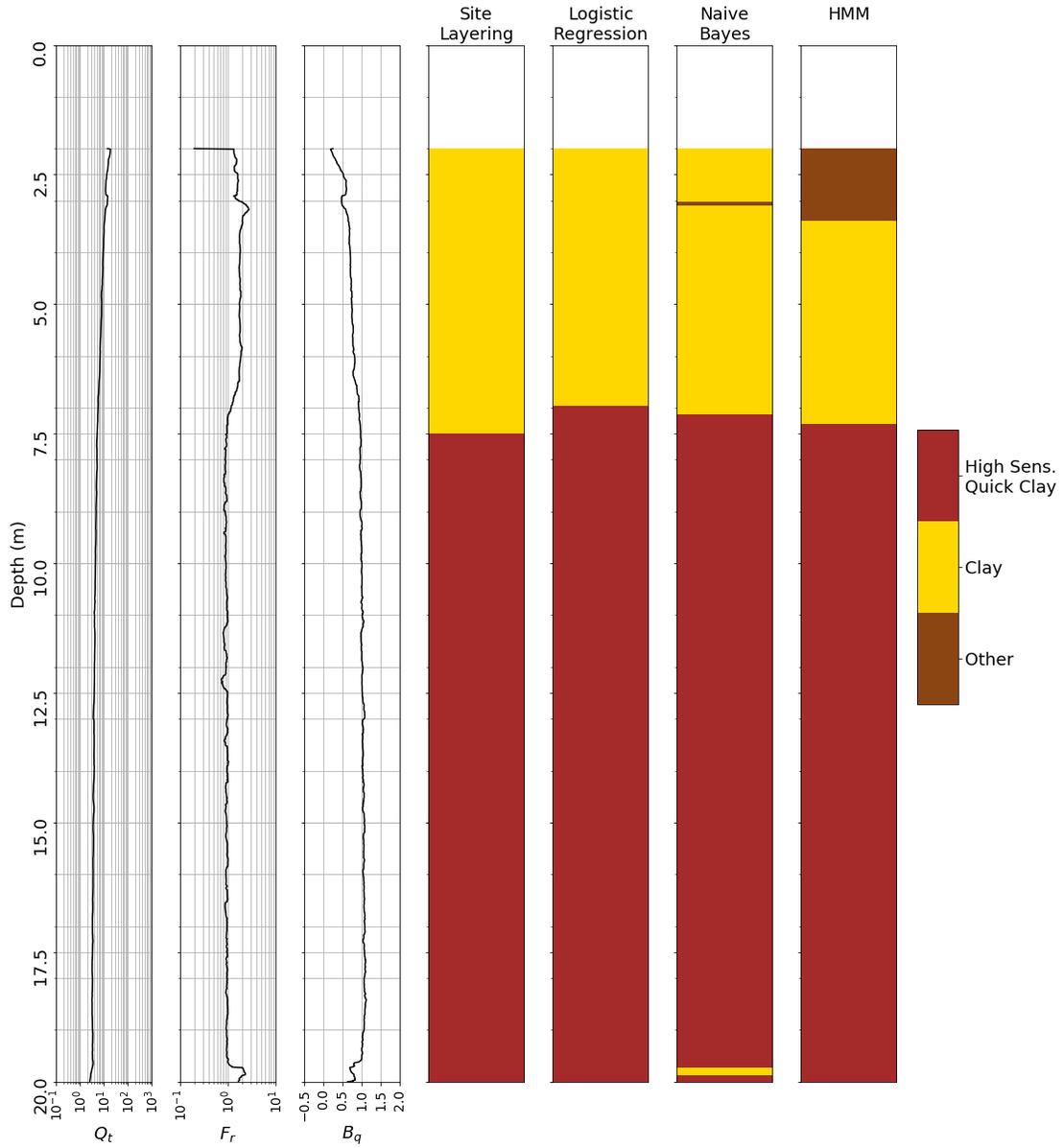
Machine Learning Classification CPTu C29
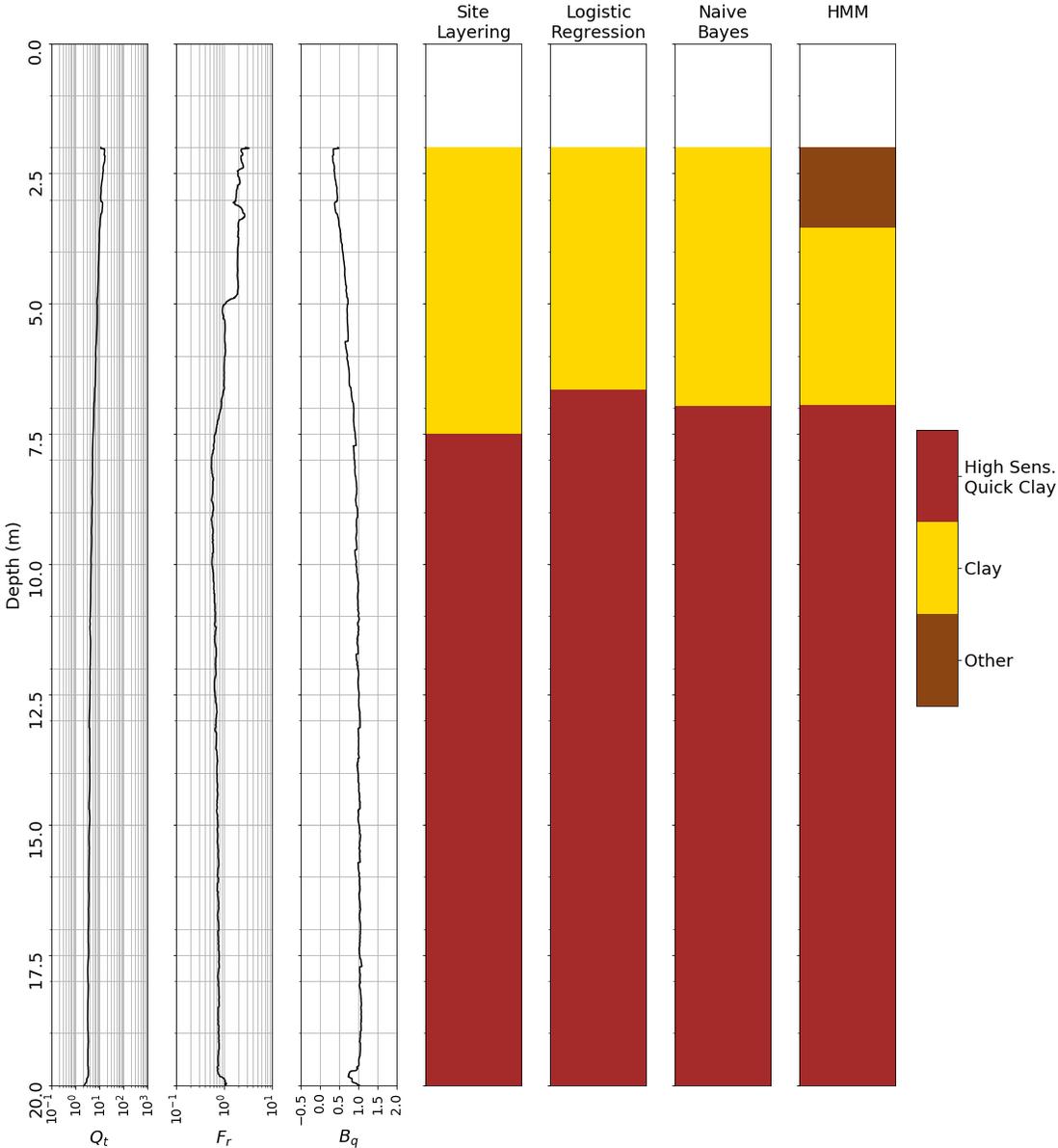
Machine Learning Classification CPTu C30
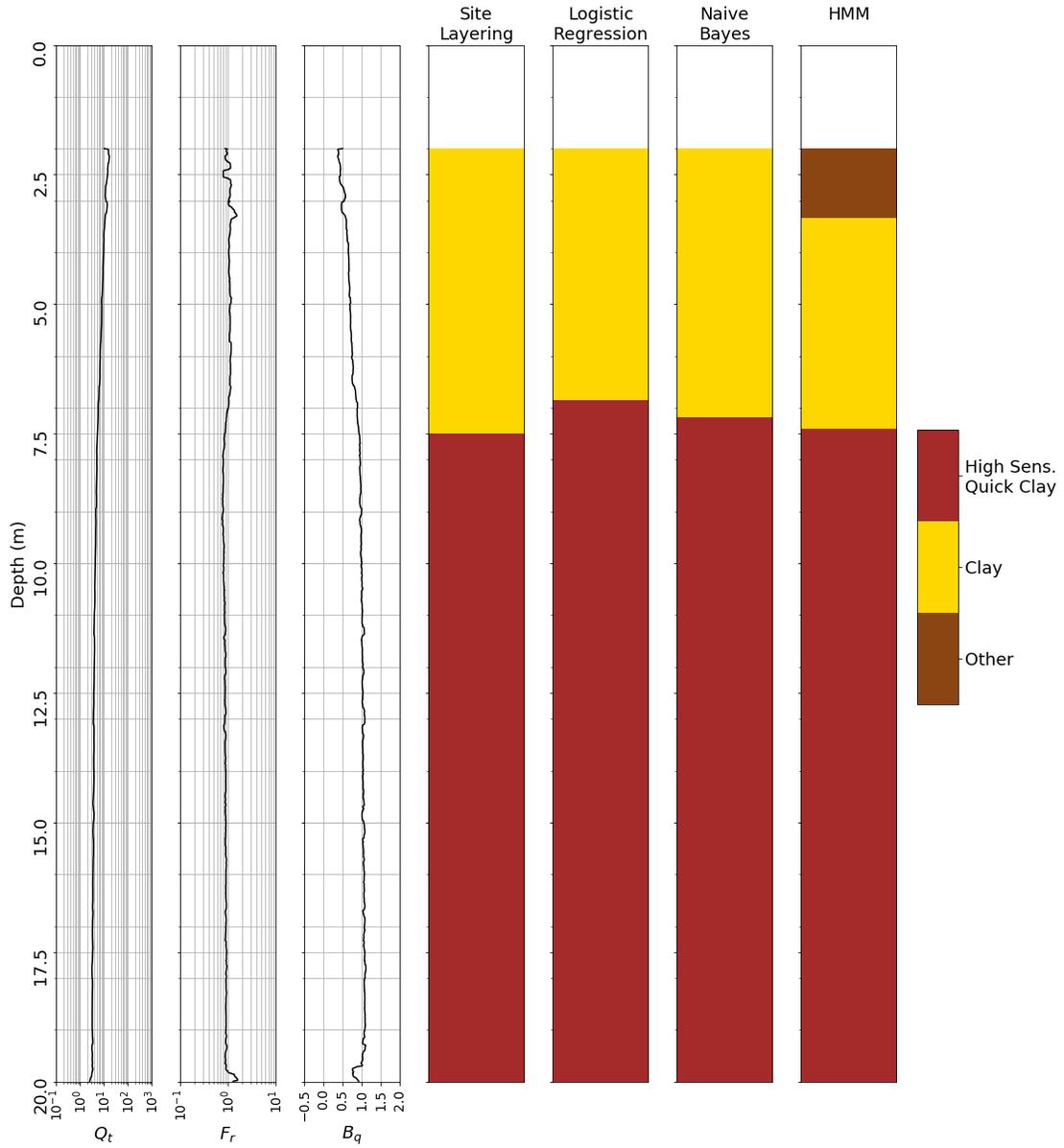
Machine Learning Classification CPTu C31
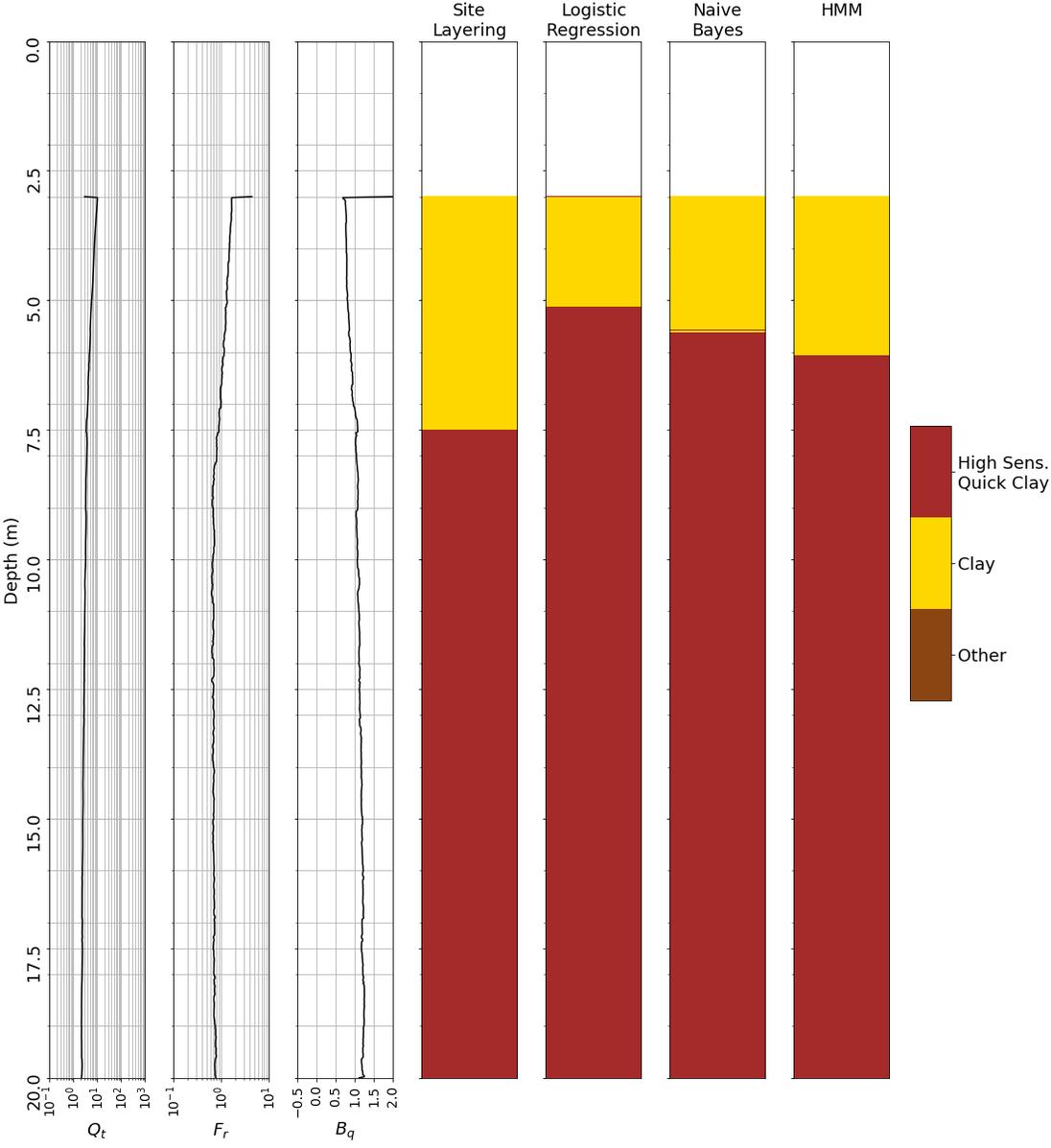
Machine Learning Classification CPTu C32
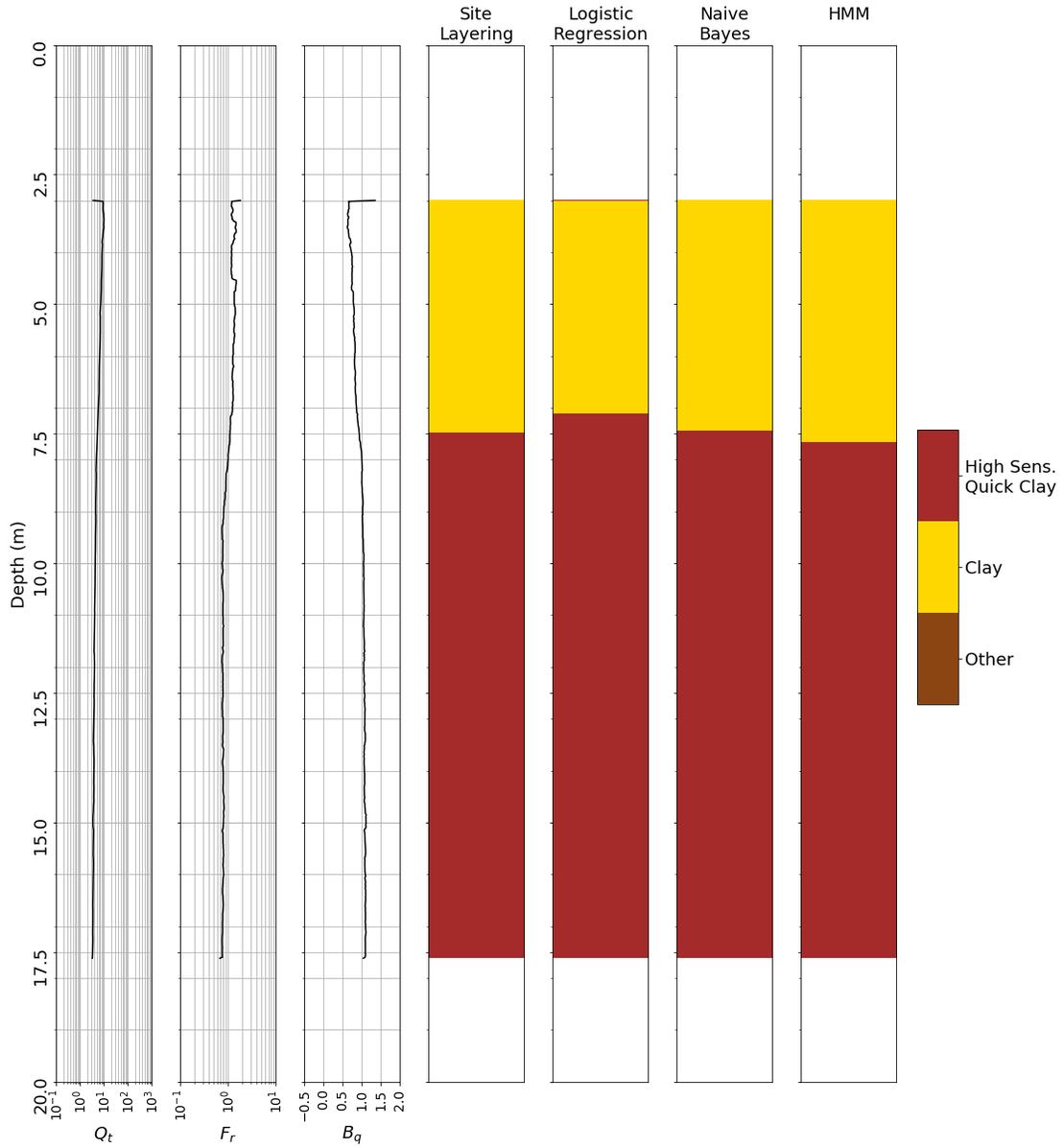
Machine Learning Classification CPTu C33
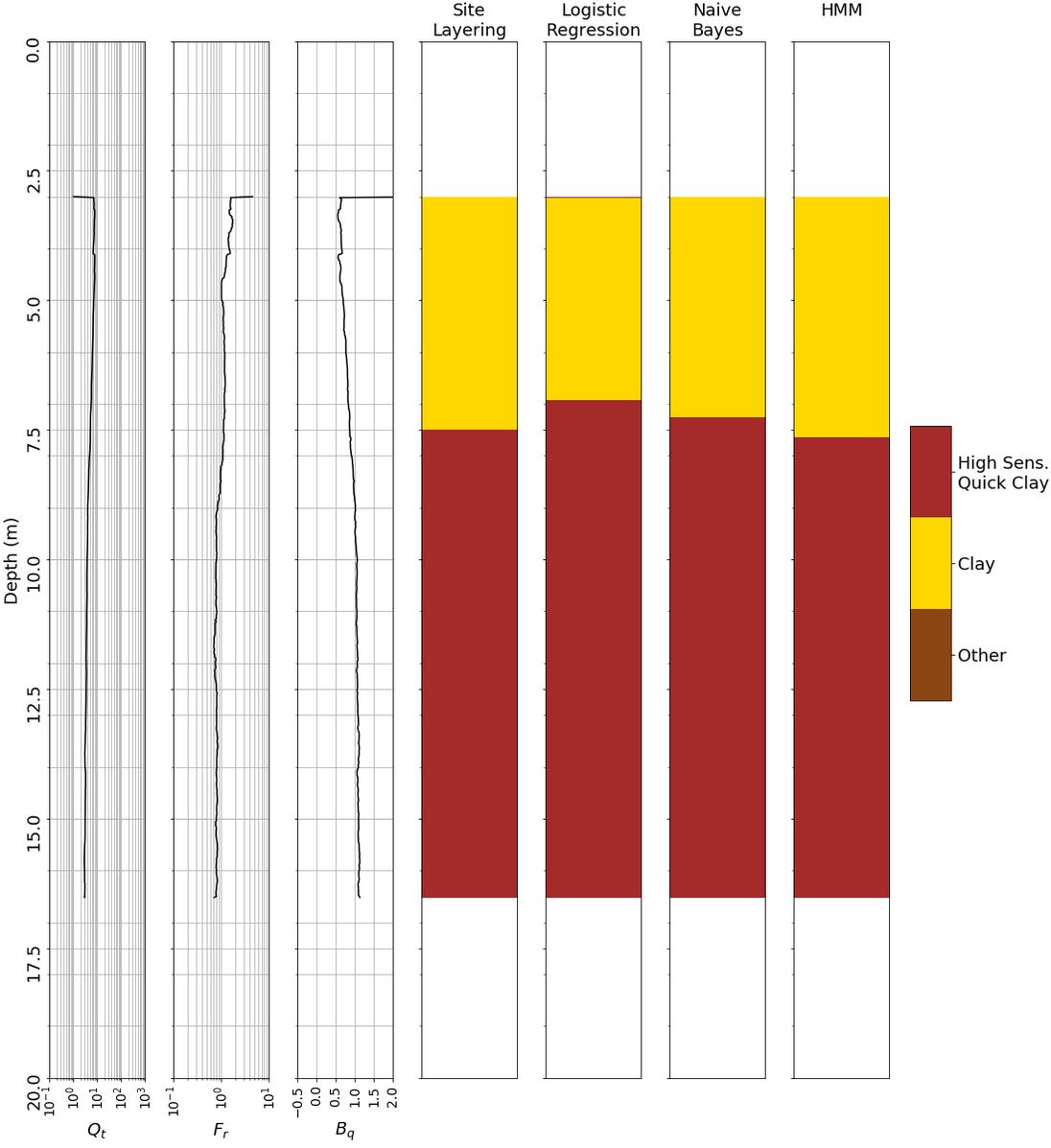
Machine Learning Classification CPTu 100
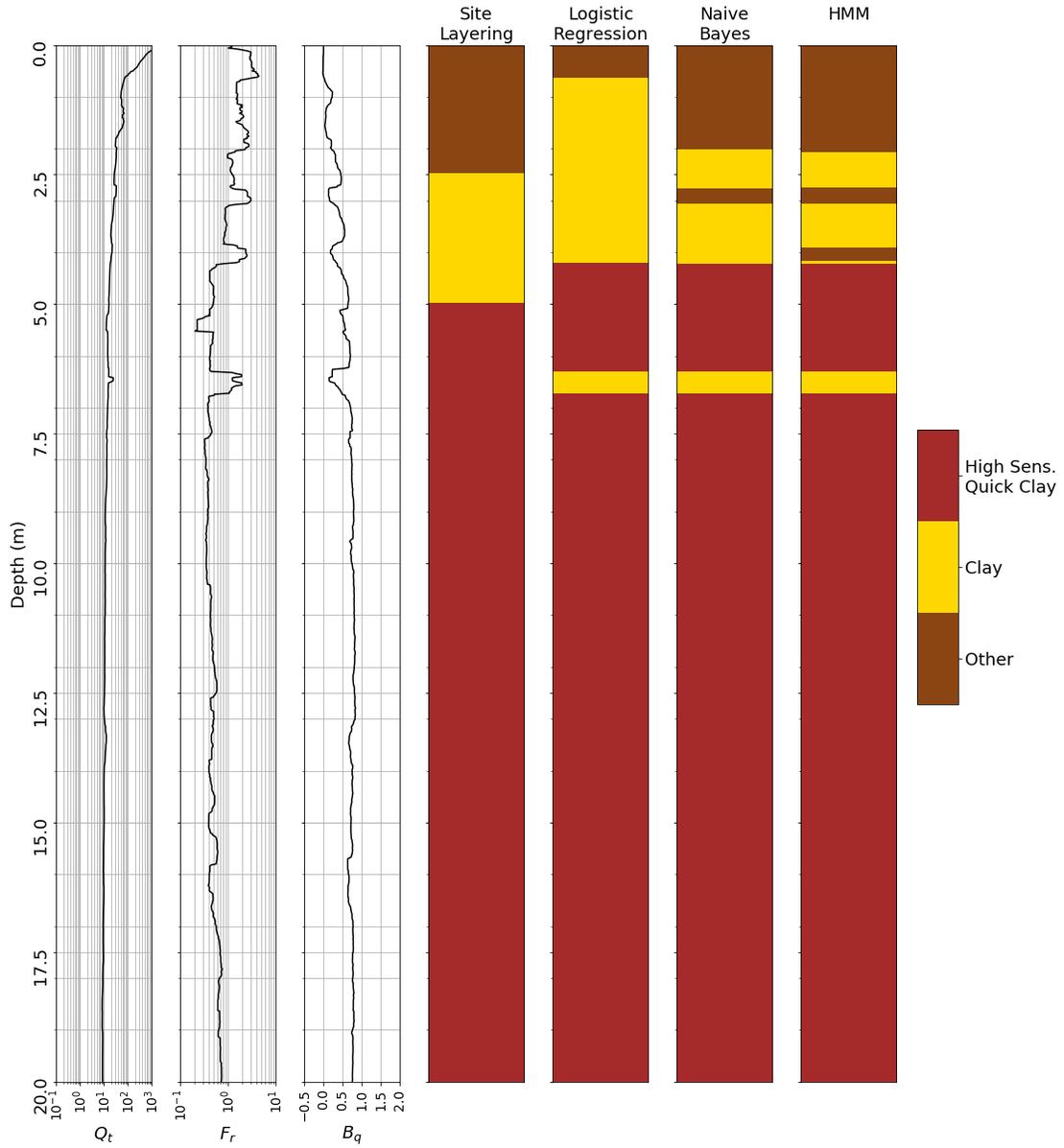
Machine Learning Classification CPTu 102
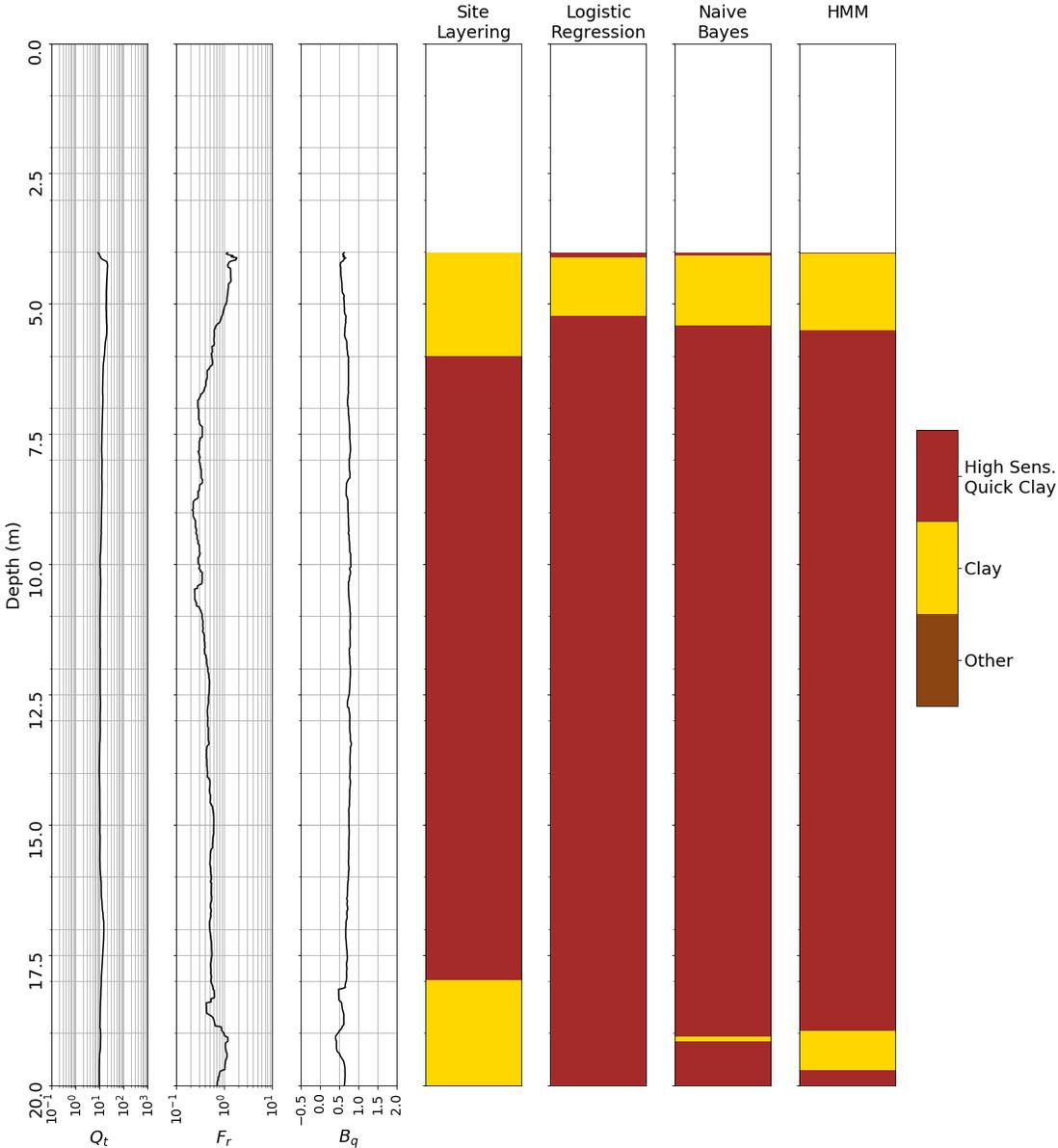
Machine Learning Classification CPTu 106
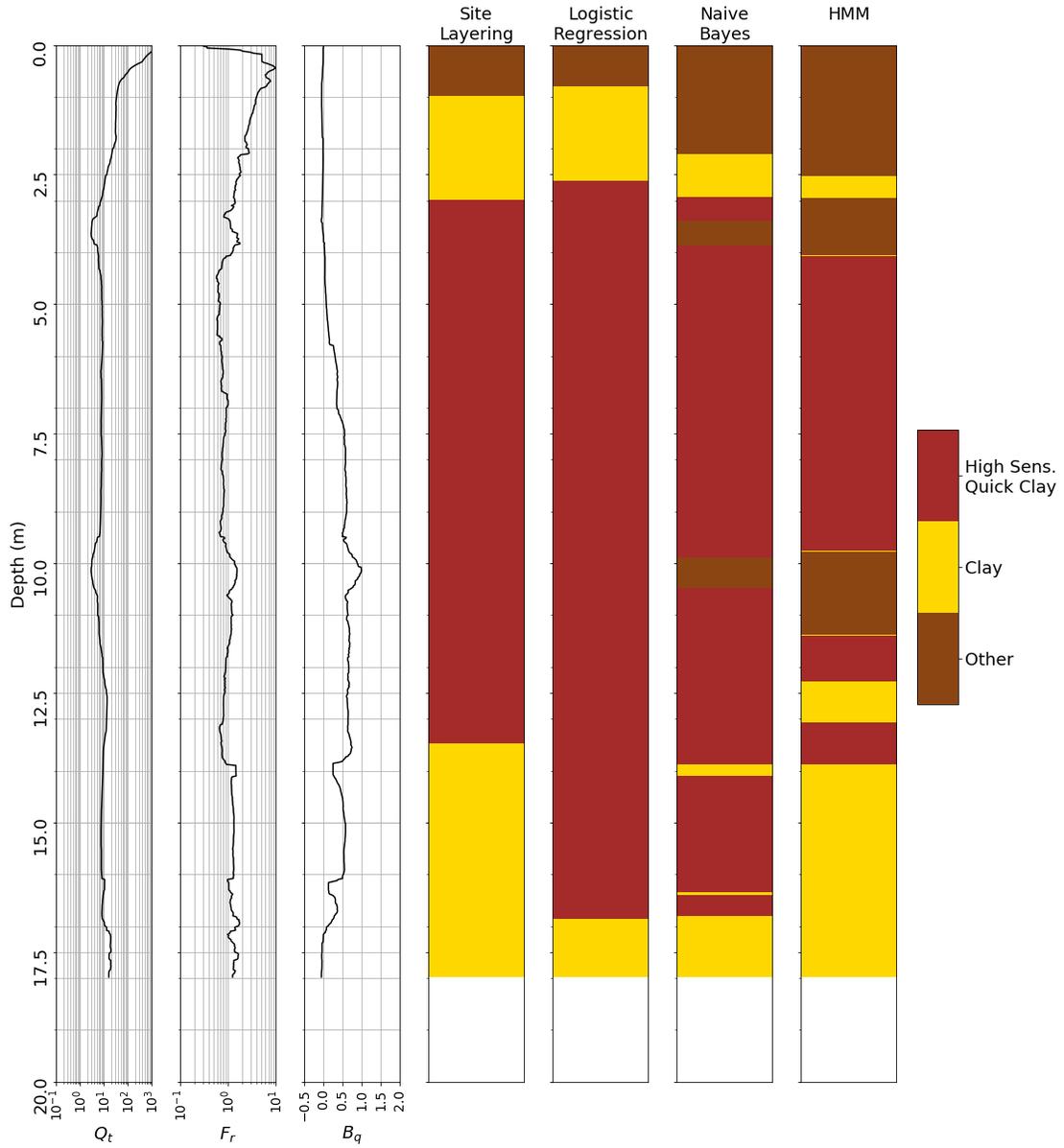
Machine Learning Classification CPTu 107
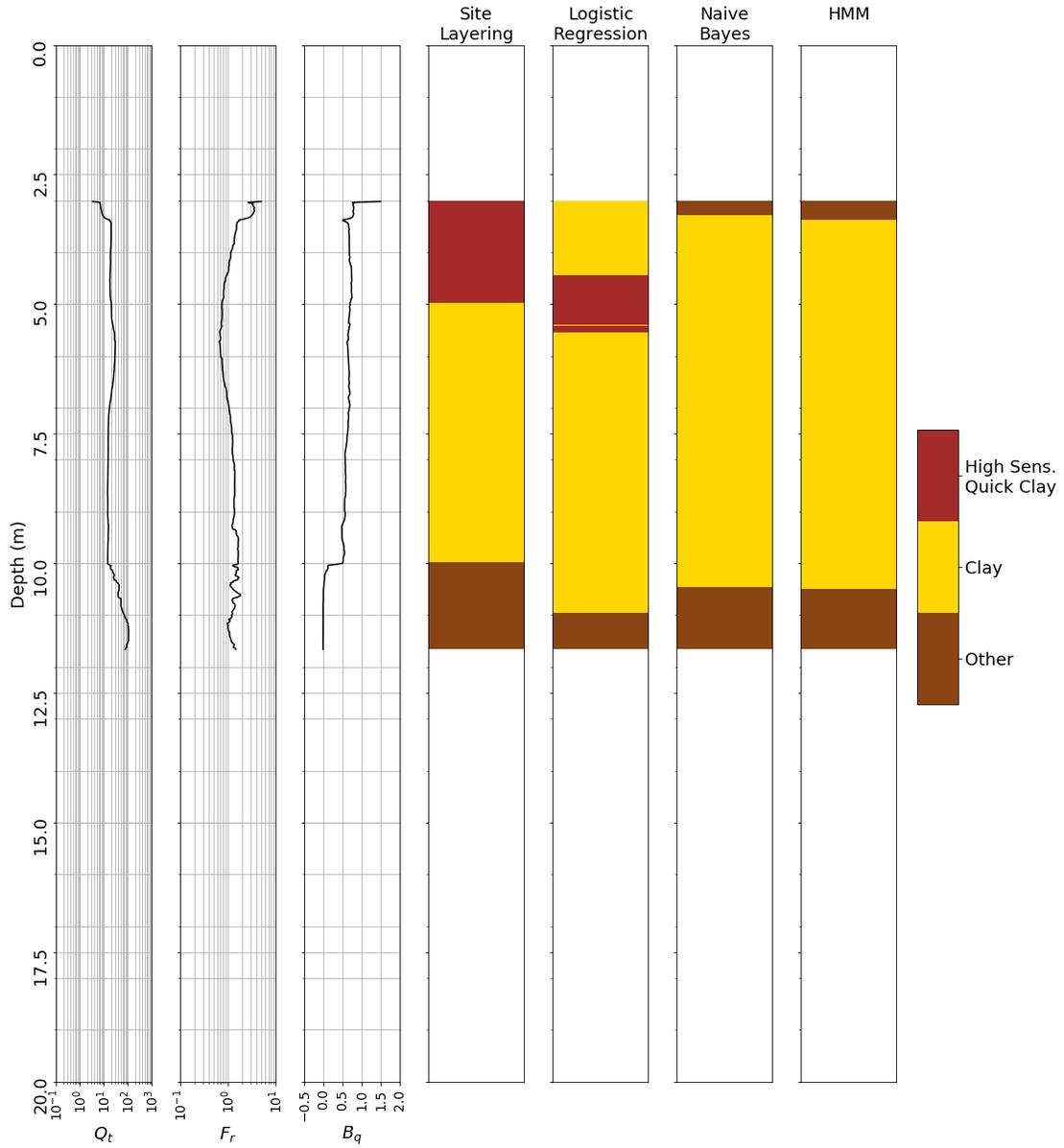
Machine Learning Classification CPTu 149
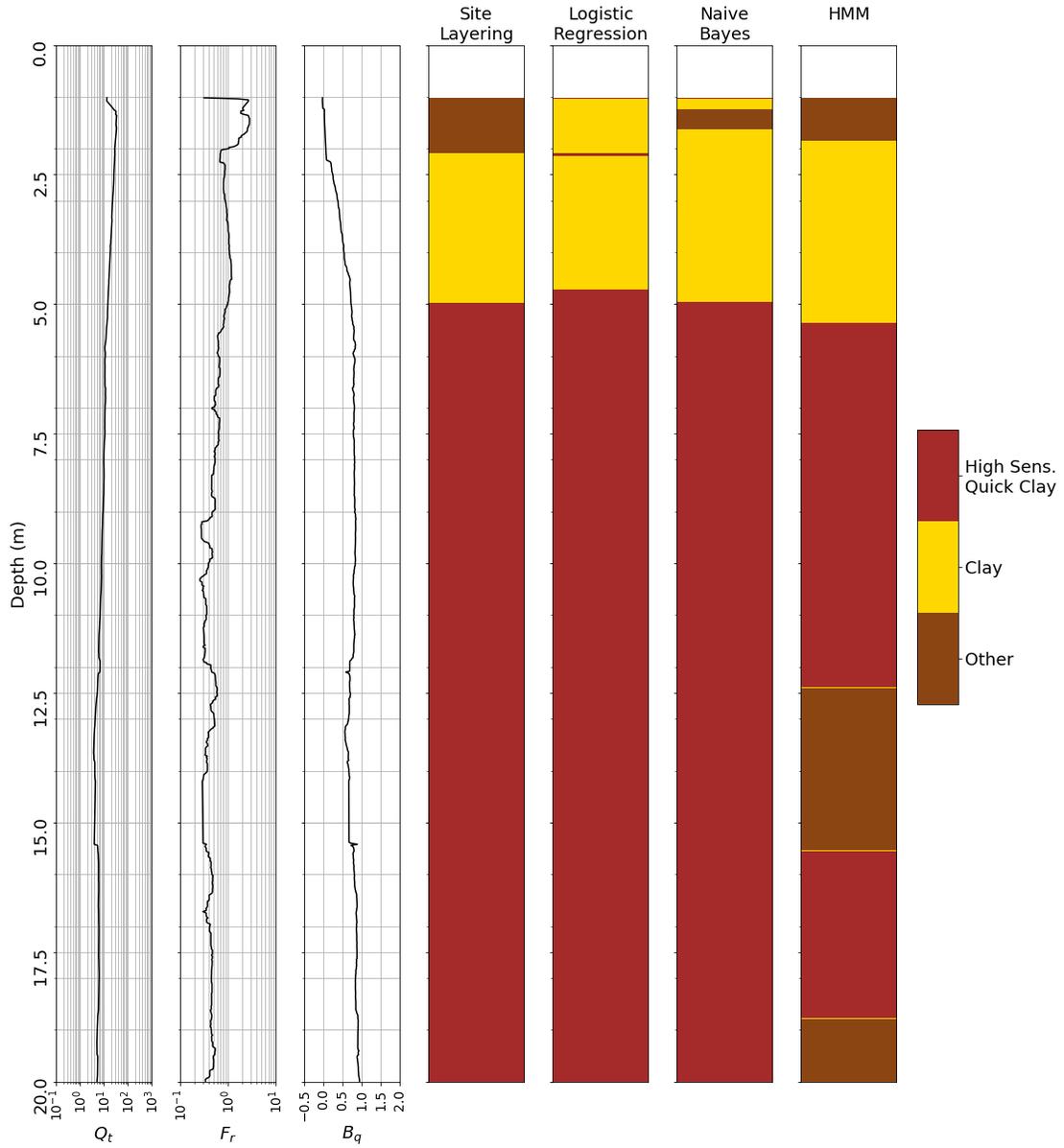
Machine Learning Classification CPTu 154

Machine Learning Classification CPTu 155

Cristian Godoy Leiva