Cone penetration data classification by Bayesian inversion with a Hidden Markov model

Cone penetration data classification by Bayesian inversion with a Hidden Markov model

A Krogstad[1], I Depina[2,3], H Omre[1]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[2]Department of Rock and Geotechnical Engineering, SINTEF Building and Infrastructure, Trondheim, Norway
[3]Faculty of Civil Engineering, Architecture and Geodesy, University of Split, Split, Croatia

Abstract. This study examines the application of the Hidden Markov model (HMM) to the soil classification based on Cone Penetration Test (CPT) measurements. The HMM is formulated in the Bayesian framework and composed of a Markov chain prior and a Gaussian likelihood model. The application of the Bayesian framework is considered as suitable because it allows for the integration of different sources of information commonly available in a CPT-based soil classification. The occurrence of different soil classes along a CPT profile is modeled with the Markov chain, while the Gaussian likelihood model establishes a relation between the different soil classes and CPT measurements. Preliminary performance of the HMM is examined on the classification of CPT measurements from the Sheringham Shoal Offshore Wind Farm.

1. Introduction

The Cone Penetration Test (CPT) is an in-situ test that is frequently applied to estimate subsurface stratigraphy, soil parameters, and parameters for a direct geotechnical design [1]. The CPT is highly applicable for the estimation of subsurface stratigraphy or soil classification because it provides continuous and reliable penetration data. The penetration data are commonly recorded as a cone is being pushed into a soil profile in terms of the cone resistance, q_c, sleeve friction, f_s, and pore pressure behind the cone, u_2. Soil classification based on CPT data is commonly conducted by comparing CPT measurements with a set of predefined soil classes, defined in so-called classification charts [2] and [3]. The classification charts assign a soil class to a given CPT measurement based on the values of the normalized cone resistance, Q_t, and the normalized friction ratio, F_r:

$$Q_t = \frac{q_t - \sigma_v}{\sigma'_v}; \quad F_r = \frac{f_s}{q_t - \sigma_v} \tag{1}$$

where q_t is the corrected cone tip resistance, σ_v is the vertical stress, $\sigma'_v = \sigma_v - u_0$ is the effective vertical stress, and u_0 is the in-situ pore pressure. The corrected cone tip resistance, $q_t = q_c + u_2(1 - a)$, accounts for the unequal projected area and area of the shaft [1], with $a = 0.75$ in this study.

Due to the influence of various processes in a soil profile (e.g., stress changes, mineralogy), the classification charts [2] are often considered as indicative and a guide for classification. In order to investigate the uncertainties associated with the CPT-based classification, several

studies implemented more advanced statistical models. For example, fuzzy subset methods were employed in [4] to analyze uncertainties in soil composition. A regression neural network was used in [5] to determine soil stratigraphy. The advantages of the Bayesian framework in integrating different sources of information in the classification process were examined in [6]. The Bayesian framework was also applied in [7] to model the effects of stress normalization on the CPT classification. The application of Bayesian Mixture Analysis to the CPT classification problem was examined in [8]. Additionally, an approach was formulated in the Bayesian setting that in addition to the soil classification identifies the most probable number of soil layers and their thicknesses [9].

This study follows the development path of Bayesian approaches and investigates the application of the Hidden Markov Model (HMM) to the CPT classification problem. The HMM is a model where there are unobserved categorical states that follow a Markov process [10]. Such an approach has previously been used to classify data into geological facies for mapping petroleum reservoirs [11, 16, 17, 18]. From a statistical perspective, the CPT classification is a problem where a set of CPT measurements is used to identify "hidden or unobserved" soil classes [8]. The attributes hidden or unobserved are assigned to soil classes as they are not directly observed and have to be interpreted from CPT measurements. The HMM is considered to be applicable to the CPT classification problem because it provides a joint model for the spatial distribution of soil classes and the relation between CPT measurements and different soil classes. The spatial distribution of soil classes along a CPT profile is modeled with a Markov chain, while the relation between CPT measurements and different soil classes is modeled with a Gaussian bivariate distribution. The solution to the CPT classification problem with the HMM is found in the Bayesian framework, which allows the incorporation of additional information on soil classes. This is considered as favorable for the CPT classification problem due to the commonly available calibration information (e.g., soil classes interpreted from nearby boreholes). The performance of the HMM is evaluated on a set of CPT measurements from the Sheringham Shoal Offshore Wind Farm.

2. HMM for CPT classification

2.1. Discretization and notation

Consider a CPT profile with measurements at depths $\mathcal{L}_{\mathcal{Z}} = \{1, ..., Z\}$, expressed in terms of Q_t and F_r values. For an easier statistical treatment, the analysis in this study is performed in terms of the the logarithm of the normalized measurements with no loss of generality:

$$Q_n = \frac{\log_{10} Q_t}{3}; \quad F_n = \frac{\log_{10} F_r + 1}{2} \tag{2}$$

The $2 \times Z$-vector of CPT measurements is denoted $\mathbf{d} : \{F_{n1}, ..., F_{nZ}, Q_{n1}, ..., Q_{nZ}\}$, while a measurement at depth z is denoted $\mathbf{d}_z = \{F_{nz}, Q_{nz}\}$. The actual soil stratigraphy or the soil class profile at the location is denoted $\boldsymbol{\kappa} : \{\kappa_z; z = 1, ..., Z\}$, where soil class at depth z, κ_z, belongs to a set of different soil classes, $\kappa_z \in \Omega_\kappa : \{1, ..., K\}$. It is important to note that soil classes do not have to follow the predefined soil classes in the classification charts and can be arbitrarily defined to describe different geological features (e.g., combinations of several soil classes in the classification charts). In addition to the CPT measurements it is assumed that some additional knowledge on the soil stratigraphy is available at the considered site. The additional knowledge can be often obtained from various sources, including soil samples from nearby boreholes, geophysical investigations, engineer's prior knowledge and experience.

In this study, the term $p(\cdot)$ is a generic way to define probability. Such that, $p(\kappa_z)$ defines the probability of any soil class, $\kappa_z \in \Omega_\kappa$, at the depth z, while $p(\boldsymbol{\kappa})$ defines the probability of the entire stratigraphy at the location of a CPT profile. $p(\cdot)$ defines the probability mass function (pmf) for discrete variables and the probability density function (pdf) for continuous variables.

Estimated parameters are denoted such that $\hat{x}$ is to be understood as the estimate of the true parameter value x.

2.2. Model definition

The goal of the CPT classification with the HMM is to calculate the probability of any profile of soil classes given the CPT measurements, $p(\boldsymbol{\kappa}|\boldsymbol{d})$. In the Bayesian setting, this probability is denoted as posterior because it incorporates the measurements with the additional or prior knowledge. The posterior probability is defined according to the Bayes law as follows:

$$p(\boldsymbol{\kappa}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})}{p(\boldsymbol{d})}, \tag{3}$$

where $p(\boldsymbol{\kappa})$ is the prior model, $p(\boldsymbol{d}|\boldsymbol{\kappa})$, is known as the likelihood model, and $p(\boldsymbol{d})$ is the probability of CPT measurements, which also serves the role of a normalizing constant. The prior model is constructed by assessing a distribution for the soil class profile $\boldsymbol{\kappa}$ from the additional knowledge (e.g., borehole samples). The likelihood model depends on the data acquisition procedure, with the parameters of the likelihood model assessed by using the actual soil class vector of the training well, $\boldsymbol{\kappa}$, along with the CPT data. With these two distributions in place the full posterior is defined. The evaluation of the normalizing constant is usually unfeasible and avoided in the majority of the implementations.

2.3. Likelihood model

In the context of the CPT-based classification, the likelihood model, $p(\boldsymbol{d}|\boldsymbol{\kappa})$, provides a statistical model that relates CPT measurements with different soil classes. The development of the likelihood model, in this study, is based on the following two assumptions. The first assumption is conditional independence between the CPT data vector at each step, $\boldsymbol{d}_z$, given the soil class profile vector $\boldsymbol{\kappa}$. The second assumption is single site dependence between the CPT data vector at each step, $\boldsymbol{d}_z$ and the soil classes at each step κ_z. These two assumptions lead to the following relation:

$$p(\boldsymbol{d}|\boldsymbol{\kappa}) = \prod_{z=1}^{Z} p(\boldsymbol{d}_z|\boldsymbol{\kappa}) = \prod_{z=1}^{Z} p(\boldsymbol{d}_z|\kappa_z). \tag{4}$$

The second assumption ignores any possible convolution (i.e., averaging) effects that might occur during the data acquisition. A Gaussian bivariate likelihood model is selected to model the aforementioned relations in this study. It is important to note that alternative and more complex relations can be also considered. The Gaussian bivariate model requires the assessment of mean parameters, $\mu|\kappa \quad \forall\kappa \in \Omega_\kappa$, and covariance matrices, $\Sigma|\kappa \quad \forall\kappa \in \Omega_\kappa$. These parameters can be estimated using the CPT data, $\mathbf{d}$ and the actual soil class profile $\boldsymbol{\kappa}$ vector commonly available from calibration boreholes to find the mean and covariance matrix for the data point subset corresponding to each soil class. Alternatively, the parameters can be estimated alternative information sources or engineer's prior knowledge or experience.

2.4. Prior model

It is assumed that the true soil class profile, $\boldsymbol{\kappa}$, follow a first order Markov chain, [10], and by that it fulfills the Markov property, which is defined as follows

$$p(\kappa_z|\kappa_{z-1}, ..., \kappa_1) = p(\kappa_z|\kappa_{z-1}) \qquad \forall z \in \{1, ..., Z\}, \tag{5}$$

where $p(\kappa_1|\kappa_0)$ is defined as $p(\kappa_1)$ for notational ease. The Markov property states that the probability of transitioning from any current soil class to any other soil class, including the

probability of transitioning to the same soil class, only depends on the current soil class. Here $p(\kappa_z|\kappa_{z-1})$ is the probability of transitioning from any soil class κ_{z-1} to any soil class κ_z. These probabilities for all pairs of κ_{z-1} and κ_z define the $(K \times K)$ matrix $\boldsymbol{P}$, with K being the number of separate soil classes in the model. The matrix element $[P]_{i,j}$ corresponds to $p(\kappa_z = j|\kappa_{z-1} = i)$. The initial marginal prior $p(\kappa_1)$ is initially defined as the stationary pdf of $\boldsymbol{P}$. Our Markov chain prior is assumed to be homogeneous, meaning $\boldsymbol{P}$ is the same matrix for all values of z or that $p(\kappa_z|\kappa_{z-1})$ is invariant of z. The prior probability of any soil class vector, $\boldsymbol{\kappa}$, is given by the following expression

$$p(\boldsymbol{\kappa}) = p(\kappa_1) \prod_{z=2}^{Z} p(\kappa_z|\kappa_{z-1}), \tag{6}$$

In practice, an estimator $\hat{P}$ of the transition matrix P is calculated by counting the transitions between different soil classes of the $\boldsymbol{\kappa}$ vector in the available data. The counts are then placed in a $K \times K$ matrix and normalized by rows. This estimator is denoted as the strict transition matrix estimator and forms the basis for other alternative transition matrix estimators. An alternative is to add a small constant to certain matrix elements where a transition is thought possible although not observed in the training data. The starting probability vector $p(\hat{\kappa_1})$ is estimated by the proportions of each state in $\boldsymbol{\kappa}$. Note that the assumptions made for the prior model, assuming conditional dependence only on the previous point is a simplification of reality. However, this simple assumption allows for a model to take into account important prior information regarding soil stratigraphy, ordering, and common or non-common transitions between layers.

2.5. Posterior model

A first order Markov chain is selected as the prior model and a likelihood model on a first order factorial form, which results in a posterior model that is a Hidden Markov Model, or a HMM, [10]. In an HMM, the states or the soil classes of the Markov chain are hidden, but at each step (i.e., depth) the hidden soil class has a corresponding observation (i.e., CPT measurement). This fits very well to the CPT classification problem, as the observations $\boldsymbol{d}_z$ are available only at certain intervals, and the hidden states would be the true unknown soil classes. The structure of the dependencies in the HMM is displayed in Figure 1.

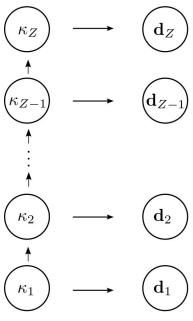


Figure 1: The graph indicates dependencies in the posterior model, with dependencies illustrated by arrows.

The expression for the posterior model is found by using Equation 3, and inserting the expression for the likelihood model in Equation 4 and the expression for the prior model in

Equation 6. The resulting posterior model is given as

$$p(\boldsymbol{\kappa}|\boldsymbol{d}) = \frac{1}{p(\boldsymbol{d})} p(\kappa_1) \prod_{z=2}^{Z} p(\kappa_z|\kappa_{z-1}) \prod_{z=1}^{T} p(\boldsymbol{d}_z|\kappa_z)$$
$$= \frac{1}{p(\boldsymbol{d})} p(\kappa_1) p(\boldsymbol{d}_1|\kappa_1) \prod_{z=2}^{Z} p(\kappa_z|\kappa_{z-1}) p(\boldsymbol{d}_z|\kappa_z), \tag{7}$$

where $p(\boldsymbol{d})$ is a normalizing constant. From Equation 7 one can derive the following expression for the posterior model on a first order Markov chain form.

$$p(\boldsymbol{\kappa}|\boldsymbol{d}) = p(\kappa_1|\boldsymbol{d}) \prod_{z=2}^{Z} p(\kappa_z|\kappa_{z-1}, \boldsymbol{d}). \tag{8}$$

Unlike the Markov chain chosen for the prior model, this posterior Markov chain does not have a stationary transition matrix, meaning that in the prior model $p(\kappa_z|\kappa_{z-1})$ is the same for all $z \in \{1, ..., Z\}$, while this is not case in the posterior model for $p(\kappa_z|\kappa_{z-1}, \boldsymbol{d})$. It is important to note that the Gaussian bivariate distributions, defining the likelihood model, are not updated.

2.6. Posterior model inference

To asses the model by brute force would require computation of all possible vectors of soil classes, which would be practically unfeasible and instead the recursive Forward-Backward algorithm is used (e.g. [12, 13]). This algorithm calculates the posterior distribution $p(\boldsymbol{\kappa}|\boldsymbol{d})$ without explicitly calculating the constant $p(\mathbf{d})$. The Forward-Backward algorithm calculates $p(\kappa_z|\kappa_{z-1}, \boldsymbol{d})$ for all combinations of κ_z and κ_{z-1}, and for all values of z thereby fully defining the posterior model $p(\boldsymbol{\kappa}|\boldsymbol{d})$. This provides a basis to simulate soil class profiles from the posterior model and to predict the true soil class profile $\boldsymbol{\kappa}$. Depending on the interpretation of the algorithm results, there can be several predictions of the true stratigraphy, $\boldsymbol{\kappa}$. In this study, the maximum a posteriori prediction, (MAP), and the marginal maximum a posteriori prediction (MMAP) are the predictors examined. We will also look at the marginal probabilities for the classes as well as simulated soil class profiles.

The MAP predictor refers to the most likely soil class profile given the observations and it is defined as follows:

$$\hat{\boldsymbol{\kappa}}_{\text{MAP}} = \arg\max_{\boldsymbol{\kappa}}(p(\boldsymbol{\kappa}|\boldsymbol{d})). \tag{9}$$

To compute the MAP predictor, the implementation of the Viterbi algorithm, (e.g. [12, 14]) is needed. This recursive algorithm exploits the Markov property of the posterior model to find the most probable vector of soil classes sequentially which is very computationally effective.
The marginal probabilities, $p(\kappa_z|\boldsymbol{d})$, for each state $\kappa_z \in \Omega_\kappa$ at step z is found from the previously calculated $p(\kappa_z|\kappa_{z-1}, \boldsymbol{d})$ by summing $p(\kappa_z|\kappa_{z-1}, \boldsymbol{d})p(\kappa_{z-1}, \boldsymbol{d}) \quad \forall \kappa_{z-1} \in \Omega_\kappa$. Thus the marginal probability for all states at each step is calculated by recursion. For each state we can then plot a vector of that states marginal probability along the well. These can be an indication of the flexibility of the model. Marginal probability vectors that only takes values very close to zero or one, and that shifts between these two possibilities in one or very few steps may indicate a rigid model.

The MMAP predictor is defined by finding at each step z, the soil class with the highest marginal probability, $p(\kappa_z|\boldsymbol{d})$, and is defined by:

$$\hat{\boldsymbol{\kappa}}_{\text{MMAP}} = \{\hat{\kappa_z} = \arg\max_{\kappa_z}(p(\kappa_z|\boldsymbol{d})); z \in \mathcal{L}_\mathcal{Z}\}. \tag{10}$$

The MMAP predictor can be easily found from the marginal probabilities.

The predictors are compared to the actual soil class profile where that is available or some other sensible independent prediction if not. For comparison reasons we also compute a Naive Bayesian predictor. This naive Bayesian MMAP predictor is defined as:

$$\hat{\boldsymbol{\kappa}}_{NB} = \{\arg\max_{\kappa_z} p(\boldsymbol{d}_z|\kappa_z); z \in \mathcal{L}_{\mathcal{Z}}\}. \tag{11}$$

It is a special case of the usual MMAP predictor where we assume a non coupled prior model, meaning we assume no spatial dependence in the model. In practice, it is found at each step as the soil class with the highest likelihood. As this predictor utilizes the likelihood model but not the prior model it gives an indication of how much the prior model influences the results. As the prior model incorporates spatial correlation between the data points into the model this comparison also reflects on the importance of that.

3. Case study

3.1. Geological information

The implemented model is applied to the classification of CPT profiles at the Sheringham Shoal Offshore Wind Farm (SSOWF). The SSOWF is located around 20 km North of the coast of Norfolk in the UK. The geological history of the area is generally known and gives a good indicator of the present geological formations (e.g., [15]). The geological formations consist mainly of the Quaternary sediments with Cretaceous chalk underneath. On some locations Holocene sand layers can be found atop of the Quaternary formation. The Quaternary sediments can be divided into the four formations, the Swarte Bank, the Egmond Ground, the Bolders Bank, and the Botney Cut formation (e.g., [15]). These four formations along with the layer of Cretaceous chalk make up the main five geological formations that are encountered from the seabed to roughly 50 meters below the seabed. They are denoted as follows, BCT for Botney cut, BDK for Bolders Bank, EG for Egmond Ground, SBK for Swarte Bank and CK for cretaceous chalk. In addition, the layer of Holocene sand will be denoted HS and treated as a geological formation. Descriptions of the typical soil properties for these six geological formations are presented in Table 1. In addition to the general description of the geological formations in Table 1, sand layers can be found within the BDK and SBK layers, while the BCT unit is an infill into the BDK unit and not present at some parts of the SSOWF area. All the main geological formations have a relatively consistent thickness across the SSOFW area except the SBK, which varies in thickness from several to hundreds of meters thick. Based on the flexibility of the implemented HMM model in the definition of soil classes, the soil classes in this case study are selected to correspond to the geological formations in Table 1. An extensive soil investigation procedure was conducted at the SSOWF site to determine the soil stratigraphy and derive soil parameters for the design of foundations for offshore wind

Table 1: Descriptions of soil classes.

Soil class	Typical description
HS	Loose fine to medium sand with shell fragments, or slightly clayey, gravelly, shelly sand
BCT	Very soft to firm sandy clay with sand seams
BDK	Stiff to very stiff gravelly clay. The gravel fraction being fine to medium size subrounded chalk.
EG	Very dense fine sand, locally with seams and layers of silt and clay.
SBK	Hard to very hard gravelly clay. The gravel fraction comprising fine to medium size rounded to subrounded particles of chalk
CK	Very weak to weak low to medium density off-white to white chalk.

turbines. The soil investigation consists of a series of CPT soundings, conducted at the location of each of the offshore wind turbines, and boreholes in the proximity of several CPTs across the SSOWF site. Soil samples were taken from the boreholes to calibrate the interpretation of soil stratigraphy and geotechnical design parameters for the design of offshore wind turbine foundations. Figure 2 shows an example of a soil class profile interpreted from a borehole and a nearby CPT profile. Given that the borehole is very close to the CPT profile, it assumed that the borehole soil stratigraphy can be used as the actual soil class profile. In the context of the statistical model implemented in this study, the information obtained from boreholes can be used for model training to determine the prior model parameters. Once the model parameters have been determined on a training CPT profile, the model can be applied to other locations at the SSOWF site to interpret soil stratigraphy from the CPT measurements.

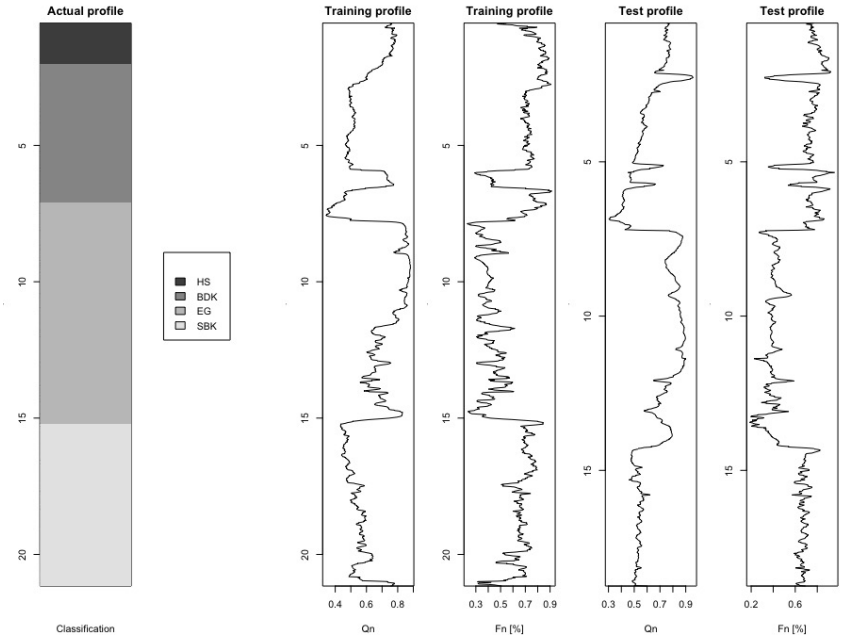


Figure 2: All data used in the case study: left to right: the actual soil class profile, the log-transformed normalized cone resistance and the log-transformed normalized friction ratio for the training profile and the log-transformed normalized cone resistance and log-transformed normalized friction ratio for the testing profile.

In this case study we will train the model on data from a borehole and a CPT profile. These are taken from the same location such that the borehole profile is used as the actual soil class vector, $\boldsymbol{\kappa}$. When all model parameters are estimated the model is tested on a second CPT profile. This test profile does not have a corresponding borehole so there is no true soil class profile available, instead the results from the test profile are compared to an interpreted soil class vector. This interpreted soil class vector is an engineering interpretation of the same testing CPT profile and gives a realistic result to compare our results to, though not as precise as the borehole profile. In Figure 2 the data used in the case study is presented. The training profile has a depth of 21.15 m and the testing profile has a depth of 18.75 m.

3.2. Model parameters

The model parameters are estimated based on the training CPT profile in Figure 2 and presented in Figure 3. The model parameters include the transition matrix, the starting probability, and the likelihood parameters. The estimated transition matrix $\hat{P}_1$ denotes the matrix that was estimated by strictly counting the transitions between the different soil classes and then

normalizing by rows. The estimated transition matrix $\hat{P}_2$ is found by adding a small constant to all elements in the matrix. A small constant is added to provide additional flexibility to the model, such that the model takes into consideration transition from any soil class to any other soil class even if such a transition is not yet observed. Note that the estimate of the strict transition matrix has many elements equal to zero, which means that no transition between these soil classes were observed. This estimate limits the prediction profiles of the model to only already observed transitions. However, the non-strict transition matrix allows for transitions from any soil class to any other soil class with very low probabilities for unobserved transitions. Figure 3b shows the CPT data from the training well classified based on the information provided from the borehole samples. Note that the points corresponding to the same soil class are to a large degree clustered. The occurrence of smaller sub-clusters can be also observed. For example, the EG soil class has a sub-cluster located to the bottom right far away from the main cluster at the upper left. The likelihood model parameters are presented in Figure 3c with the location of the mean for the four geological units in the model and the 90% probability ellipse contour of the Gaussian bivariate distribution. It is also important to observe that the SBK and the BDK soil classes represent similar soil properties with their means close to each other.

The compatibility between the information provided by the training CPT profile and the measurements in the test CPT profile is examined in Figure 4. Figure 4a shows a comparison between CPT measurements at the training and the test CPT profiles. The main clusters appear to be relatively close with some differences in the placement of sub-clusters and the spread of the points further away from the main clusters. This indicates a relatively good compatibility of the data at the two CPT profiles. Figure 4b shows the data points from the test CPT profile plotted with the covariance matrix ellipses estimated from the training CPT profiles. This plot indicates the applicability of the relation between the CPT measurements and the soil classes established by the likelihood model determined from the training CPT profile. The comparison of the likelihood model with the measurements from the test CPT profile reveals a relatively good agreement.

3.3. Results and discussion

The model predictions on the training and the test CPT profiles are presented in Figure 5. The performance of the model on the training CPT profile in Figures 5a to 5d is presented to demonstrate the effects of parameter selection on the estimated soil class profiles. The evaluation of the model characteristics on the test CPT profile only would be more challenging as the actual soil class profile is unknown and only an engineering prediction exists.

The left side sub figures contain: the soil class profile observed from the borehole samples in the case of the training CPT profile, an engineering interpretation of the soil class profile using classification charts [2] in the case of the test CPT profile, and the predictors of the actual soil class profile with the marginal probabilities along the CPT profile. The predictors are presented from left to right as MAP, MMAP and the Naive Bayesian. The right side figures contain simulated profiles from the posterior model that will help to illustrate the effects of the selection of different model parameters.

The predictors and the marginal probabilities for the training CPT profile with the non-strict and the strict transformation matrices are presented in Figures 5a and 5c, respectively. Additionally, Figures 5b and 5d show several simulated profiles from the training CPT profile with the strict and the non-strict prior transition matrices, respectively. From the comparison of Figures for the training CPT profile with the strict and the non-strict transition matrices it can be observed that the strict prior transition matrix enforces the same ordering of transitions as found in the training CPT profile. The predictions for the training CPT profile with the strict transition matrix fit most closely with the actual soil class profile. The Naive Bayesian predictor is the same for the both cases as this predictor does not account for the spatial ordering between different

soil classes. The selection of a non-strict transition matrix leads to a less rigid model. This can be observed from the simulations in Figure 5b, which show more variation than those in Figure 5d. Similar can be observed in the marginal probability plots, where the marginal probability of the strict model show almost only clear jumps between the soil classes. The simulations differ little, with the transition from the EG to the SBK layer being approximately at the same point in all of the realizations. This is due to abrupt transitions in the CPT measurements in that range, see Figure 2, which severely influences the likelihood model that again makes the posterior somewhat rigid in that region. These observations indicate both that the model is somewhat rigid in itself as well as that five simulated profiles might be too few to detect the underlying variation. When simulating more profiles, more deviations appear but mostly in the form of the original layers containing internal bands of a layer not found there in the training data. The actual transitions found in the training data always appear in the simulations and usually at the same depth.

In addition to the visualizations, the results are presented in terms of a table with accuracy scores in percentages in Table 2 and as confusion matrices in Tables, 3, 4, 5 and 6. In the confusion matrices, each column represents a true class and each row represents a predicted class. So element i, j represents how many points of true class j were classified as class j by the predictor. Each matrix element contains three values representing the three predictors MAP, MMAP and Naive Bayesian MMAP in that order.

The non-strict model displays both jumps and transition that were not observed and zones of uncertainty where more than one soil class has a considerable marginal probability. It is worth noting that these differences occur also when changing the small constant added to each element during the estimation of the prior transition matrix. The size of the constant does not seem to matter as long as it is of a lower order of magnitude than the lowest probability of any observed transition.

Figures 5e and 5g present: an engineering interpretation of the soil stratigraphy based on the classification charts [2], the predictors and the marginal probabilities for the test CPT profile. It is important to note that the engineering interpretation of the soil class profile is presented only to provide a reference and it should not be considered as the actual soil class profile to directly compare the results to. Several simulated profiles of the test CPT profile are presented in Figures 5f and 5h for the strict and the non-strict transition matrices. Most of the observations made for the training CPT profile can be translated to the test CPT profile. The non-strict prior transition matrix gives a less rigid model with a few transitions in the upper part of the CPT profile that are not present in the predicted profiles using the strict prior transition matrix.

The comparison between the predicted profiles and the soil profiles available from the borehole samples or the engineering interpretation using classification charts reveals that the strict prior transition matrix model results in a better fit. This is also true for the prediction accuracy in both the training and testing well. The MAP and the MMAP predictors both with the strict and the non-strict prior transition matrices give more realistic results than the Naive Bayesian predictor. This demonstrates the advantage of adding spatial ordering to the model. It is worth to notice that the MAP and the MMAP estimator gives almost identical results, at most varying by on a few points. This is an interesting result that is likely connected to the strict ordering of our classes.

It can be observed that the model with the non-strict prior transition matrix appears to misclassify between the BDK and the SBK layers. This is due to the similarity between the properties of the respective soil classes, which can be observed from Figure 3c and Table 1. This can be also detected in the confusion matrices.

In this paper the soil classes are based on geological formations. The reason for this choice was that the background geological report used in the construction of $\boldsymbol{\kappa}$ was most easily interpreted this way. Another and more intuitive option would be to define the soil classes after the clusters

of CPT measurements. This interpretation would make the soil classes correspond more closely to their soil properties. An effect of changing the basis of soil classes would be a shift in importance from the prior model to the likelihood model. In our paper, the successive nature of the geological formations used as a basis for the soil classes leads to a prior transition matrix that has almost all weight along the diagonal. If not altered up by setting all elements to a non zero constant, this matrix does put strict constraints on what transitions are possible and by that also which soil class profiles are possible. When setting the classes corresponding to geological layers, which often follow a strict ordering, this will always be a problem. If the classes are set based on soil properties this is less likely to be a problem and in this case study it would not be an issue. A classification based on soil properties would also not have the same degree of overlap between classes in the likelihood model. This might improve the performance of the naive Bayesian approach and overall put more emphasis on the likelihood model in the results. In this case study, only two alternatives were evaluated for the prior model, the strict counting estimation and the flexible estimation. Several other could have been attempted, for instance in the case of geological formations as soil classes, an upper triangular flexible transition matrix could be useful. This is a compromise between the alternatives used here that modifies the strict counting matrix by adding constants only to the upper diagonal, thus making it possible for the model to consider transition from a state to any state corresponding to a geological formation that is beneath the current state. Such a transition matrix would make the model incorporate realistic transitions though they have not been observed. Considering the presented results, the choice of prior model does not seem to be the biggest challenge for model implementations. However, the consequences of the prior models assumptions with more complex formulations should be examined in more detail in future studies due to their effects on the results.
Finally, the results of the implemented study should be considered as preliminary due to only a single training and one test sample used. Although the presented results indicate good performance of the model, more extensive validations are required to draw general conclusions. However, the method of data classification with the HMM has been tested extensively on a similar classification problem in petroleum engineering with successful results [11, 16, 17, 18] that motivated the application to CPT classification.

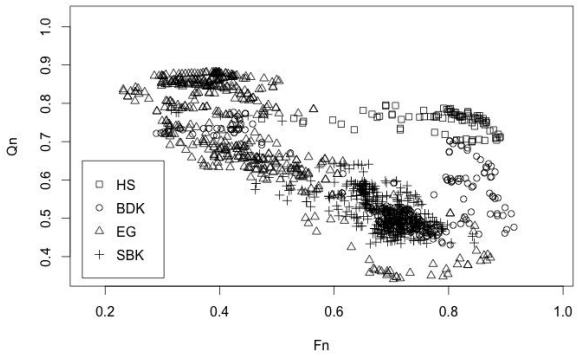
4. Conclusions

This study examined the application of the Hidden Markov Model to the soil classification based on CPT measurements. The model is composed of a Markov chain that models spatial ordering of soil classes along a CPT profile and a Gaussian likelihood model that links CPT measurements with different soil classes. The Bayesian formulation of the model is considered as advantageous for the considered problem as it allows the model to integrate additional sources of information, commonly available in a CPT-based soil classification. Additional advantages, when compared to the CPT classification based on classification charts, include arbitrary definitions of soil classes supported by the Gaussian likelihood model. The probabilistic framework of the model allows it to account from some of the uncertainties in the classification process. The Bayesian setting of the model provides a framework for a more consistent treatment of additional sources of information in the CPT-based soil classification.
The model when applied to the classification of CPT profiles from the Sheringham Shoal Offshore Wind Farm achieved good accuracy scores of over 90% for strict priors and around 80% for non strict. The profiles seem realistic when compared with our prior knowledge of the geological conditions. However, additional and more extensive tests are necessary to further validate the model performance. Further extensions of the model are planned to adapt the soil class definitions to data clusters instead of geological formations and to consider Bayesian updating of the relations between soil classes and CPT measurements.

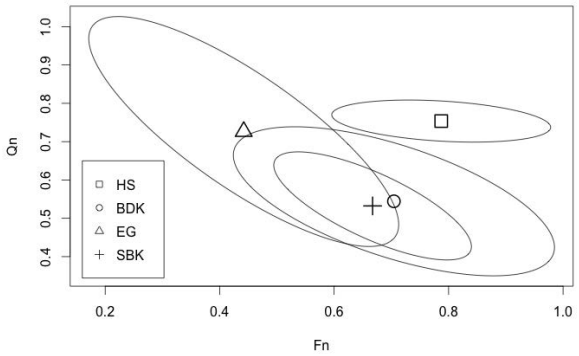
$$\widehat{p(\kappa_1)} = \begin{pmatrix} 0.0734 \\ 0.2462 \\ 0.3870 \\ 0.2934 \end{pmatrix} \quad \widehat{\mathbf{P_1}} = \begin{pmatrix} 0.98851 & 0.01149 & 0 & 0 \\ 0 & 0.99658 & 0.00342 & 0 \\ 0 & 0 & 0.99782 & 0.00218 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\widehat{\mathbf{P_2}} = \begin{pmatrix} 0.98848 & 0.01150 & 0.00001 & 0.00001 \\ 0.00001 & 0.99655 & 0.00343 & 0.00001 \\ 0.00001 & 0.00001 & 0.99779 & 0.00219 \\ 0.00001 & 0.00001 & 0.00001 & 0.99997 \end{pmatrix}$$

(a) Prior model parameters.

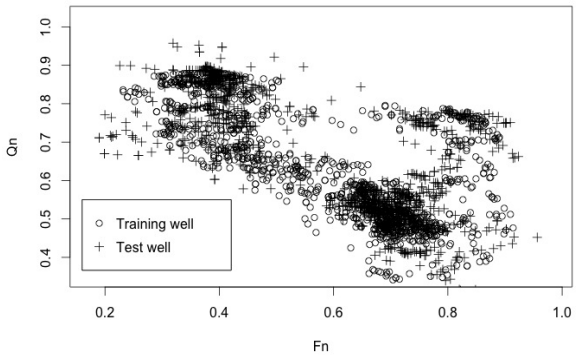


(b) Classified CPT measurements from the training CPT profile.

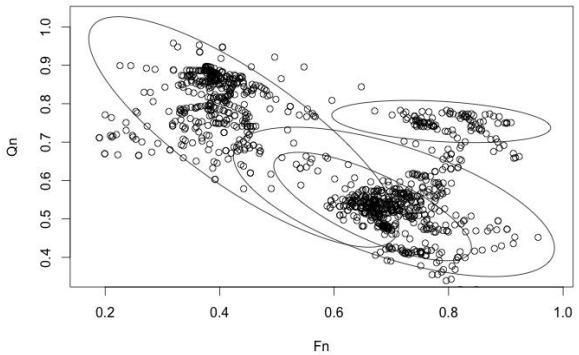


(c) Likelihood parameters visually represented.

Figure 3: Estimated model parameters: (a) Prior transition matrix estimated with and without strict counting, $\widehat{\mathbf{P}}_1$ and $\widehat{\mathbf{P}}_2$ respectively, and stationary probability vector $\widehat{p(\kappa_1)}$; (b) The data points from the training CPT profile plotted with the soil classification; (c) The estimated likelihood functions in the form of the mean and the 90% probability ellipse contour.

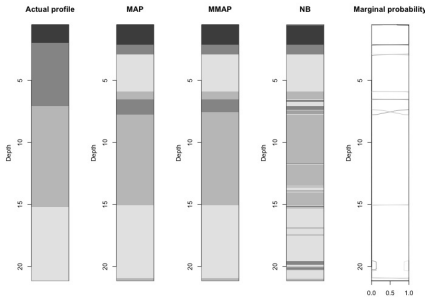


(a) Data points from both CPT profiles.

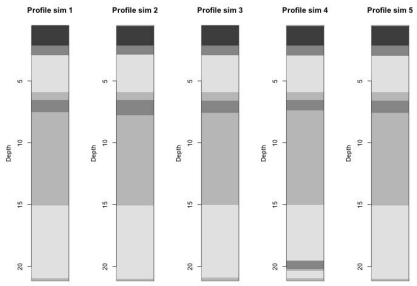


(b) Data points from the test CPT profile with the covariance matrix ellipses estimated from the training CPT profile.

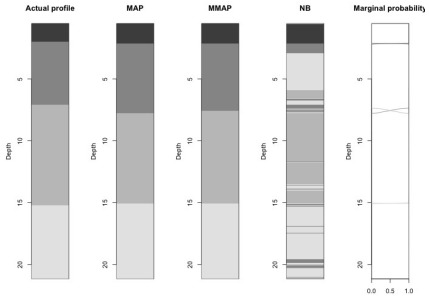
Figure 4: Comparison of CPT measurements from the test CPT profile with: (a) the CPT measurements from the training CPT profile and (b) the estimated likelihood functions from the training CPT profile.



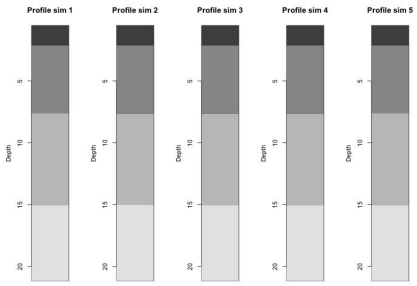
(a) Training CPT profile, non-strict transition matrix: actual soil class profile, model predictions, and marginal probabilities.



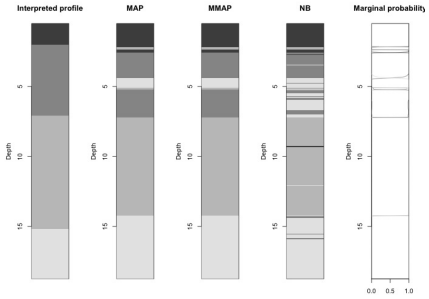
(b) Training CPT profile, non-strict transition matrix: simulated profiles.



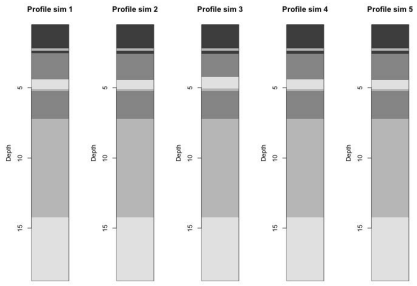
(c) Training CPT profile, strict transition matrix: actual soil class profile, model predictions, and marginal probabilities.



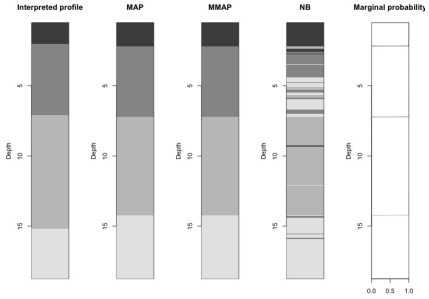
(d) Training CPT profile, strict transition matrix: simulated profiles.



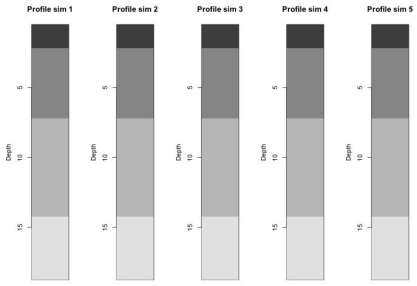
(e) Test CPT profile, non-strict transition matrix: engineering prediction, model predictions, and marginal probabilities.



(f) Test CPT profile, non-strict transition matrix: simulated profiles.



(g) Test CPT profile, strict transition matrix: engineering prediction, model predictions, and marginal probabilities.



(h) Test CPT profile, strict transition matrix: simulated profiles.

Figure 5: Results and simulations from the training and the test CPT profiles: (a), (c), (e) and (g) Actual and engineering prediction of soil class profiles, model predictions, and marginal probabilities; (b), (d), (f) and (h) simulated profiles.

Table 2: Accuracy for all predictors and wells.

Well and set up / Predictor	MAP	MMAP	NB MMAP
Training, non strict prior	77%	78%	69%
Training, strict prior	95%	96%	69%
Testing, non strict prior	86%	86%	78%
Testing, strict prior	93%	93%	78%

Table 3: Confusion matrix, training well, non strict prior. In each matrix element predictors are from left to right MAP, MMAP and NB MMAP.

Predicted class / True class	HS	BDK	EG	SBK	Total
HS	87, 87, 83	8, 8, 8	0, 0, 2	0, 0, 0	95, 95, 93
BDK	0, 0, 0	74, 75, 53	38, 25, 31	0, 0, 34	112, 110, 118
EG	0, 0, 4	38, 38, 38	412, 425, 394	13, 13, 28	463, 476, 464
SBK	0, 0, 0	172, 171, 193	9, 9, 32	335, 335, 286	516, 515, 511
Total	87	292	459	348	1186

Table 4: Confusion matrix, training well, strict prior. In each matrix element predictors are from left to right MAP, MMAP and NB MMAP.

Predicted class / True class	HS	BDK	EG	SBK	Total
HS	87, 87, 83	8, 8, 8	0, 0, 2	0, 0, 0	95, 95, 93
BDK	0, 0, 0	284, 284, 53	38, 25, 31	0, 0, 34	322, 309, 118
EG	0, 0, 4	0, 0, 38	412, 425, 394	0, 0, 28	412, 425, 464
SBK	0, 0, 0	0, 0, 193	9, 9, 32	348, 348, 286	357, 357, 511
Total	87	292	459	348	1186

Table 5: Confusion matrix, testing well, non strict prior. In each matrix element predictors are from left to right MAP, MMAP and NB MMAP.

Predicted class / True class	HS	BDK	EG	SBK	Total
HS	85, 85, 85	20, 20, 21	0, 0, 4	0, 0, 0	105, 105, 110
BDK	0, 0, 0	211, 211, 138	7, 7, 8	0, 0, 6	218, 218, 152
EG	0, 0, 0	19, 19, 22	404, 405, 397	0, 0, 0	423, 424, 419
SBK	0, 0, 0	40, 40, 109	58, 57, 60	206, 206, 200	304, 303, 369
Total	85	290	469	206	1050

Table 6: Confusion matrix, testing well, strict prior. In each matrix element predictors are from left to right MAP, MMAP and NB MMAP.

Predicted class / True class	HS	BDK	EG	SBK	Total
HS	85, 85, 85	10, 10, 21	0, 0, 4	0, 0, 0	95, 95, 110
BDK	0, 0, 0	280, 280, 138	7, 7, 8	0, 0, 6	287, 287, 152
EG	0, 0, 0	0, 0, 22	404, 405, 397	0, 0, 0	404, 405, 419
SBK	0, 0, 0	0, 0, 109	58, 57, 60	206, 206, 200	264, 263, 369
Total	85	290	469	206	1050

References

[1] Lunne T, Robertson P and Powell J 1997 *Geotechnical Practice*
[2] Robertson P 1990 *Canadian Geotechnical Journal* **27** 151–158
[3] Senneset K and Janbu N 1985 *Strength Testing of Marine Sediments: Laboratory and In-situ Measurements* (ASTM International)
[4] Zhang Z and Tumay M T 1999 *Journal of Geotechnical and Geoenvironmental Engineering* **125** 179–186
[5] Kurup P U and Griffin E P 2006 *Journal of Computing in Civil Engineering* **20** 281–289
[6] Jung B C, Gardoni P and Biscontin A 2008 *Geotechnique* **58** 591–603
[7] Cetin K O and Isik N S 2007 *Journal of Geotechnical and Geoenvironmental Engineering* **133** 887–897
[8] Depina I, Le T M H, Eiksund G and Strøm P 2016 *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* **10** 27–41
[9] Wang Y, Huang K and Cao Z 2013 *Canadian Geotechnical Journal* **50** 766–776
[10] MacDonald I and Zucchini W 1997 *Hidden Markov and Other Models for Discrete- valued Time Series* Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis) ISBN 9780412558504
[11] Larsen A, Ulvmoen M, Omre H and Buland A 2006 *Geophysics* **71(5)** 69–78
[12] Lindberg D V 2014 *Inference and categorical Bayesian inversion of convolved hidden Markov models applied to geophysical observations* Ph.D. thesis NTNU
[13] Baum L E, Petrie T, Soules G and Weiss N 1970 *Annals of Mathematical Statistics* **41** 164–171
[14] Forney G D 1973 *Proceedings of IEEE* **61(3)** 268–278
[15] Le T M H, Eiksund G R, Strøm P J and Saue M 2014 *Engineering Geology* **177** 40–53
[16] Rimstad K, Avseth P and Omre H 2012 *Geophysics* **77** B69–B85
[17] Ulvmoen M and Omre H 2010 *Geophysics* **75** R21–R35
[18] Rimstad K and Omre H 2013 *Computational Statistics & Data Analysis* **58** 187–200