Structural Safety 70 (2018) 128–138

Contents lists available at ScienceDirect

Structural Safety

journal homepage: www.elsevier.com/locate/strusafe

# A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data

Xiangrong Wang [a], Hui Wang [b,*], Robert Y. Liang [c], Hehua Zhu [d], Honggui Di [a]

[a] Department of Civil Engineering, The University of Akron, Akron, OH 44325-3905, USA
[b] Graduate School AICES, RWTH Aachen University, Schinkelstraße 2, Aachen 52062, Germany
[c] Department of Civil and Environmental Engineering and Engineering Mechanics, The University of Dayton, Dayton, OH 45469-0243, USA
[d] Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, Shanghai 200092, China

ARTICLE INFO

*Article history:*
Received 25 April 2017
Received in revised form 22 August 2017
Accepted 28 October 2017
Available online 4 November 2017

*Keywords:*
Cone penetration test
Site investigation
Geostatistics
Subsurface stratification
Hidden Markov random field

ABSTRACT

This paper presents a new probabilistic site characterization approach for both soil classification and property estimation using sounding data from multiple cone penetration tests (CPTs) at a project site. A hidden Markov random field (HMRF) model based Bayesian clustering approach is developed, which can describe not only the heterogeneity of properties in statistically homogeneous soil layers, but also the correlation between spatial distributions of different soil layers. The latter has not been well considered in the existing CPT interpretation methods. A Monte Carlo Markov chain based expectation maximization (MCMC-EM) algorithm is adopted to calibrate the established HMRF model, so that both the subsurface soil/rock stratification and the pertinent soil properties can be estimated in a probabilistic manner. The proposed CPT interpretation approach is validated and demonstrated using a series of numerical examples, including using real CPT data. It is shown that the proposed method is able to accurately identify soil layers, pinpoint their boundaries, and provide reasonable estimates of the associated soil properties. In addition, comparative studies show that combining analysis of CPT data from multiple soundings, rather than interpreting them separately, can significantly enhance the accuracy of interpretation and simplify the subsequent task of interpreting stratigraphic profiles.



## 1. Introduction

Cone penetration test (CPT) is one of the most commonly used geotechnical site investigation methods for profiling subsurface stratigraphy (i.e., identify numbers of underground soil layers and their boundaries) and inferring pertinent soil properties [1]. However, retrieving soil sample for visual inspection is not available in CPTs and the sounding records (i.e., cone resistance and sleeve friction) are typically highly variable due to the subsurface soil heterogeneity; therefore, constructing subsurface stratification and the interpreting relevant soil properties of CPT data based on soil behaviors type charts [2,3] may still be challenging [4]. To address this difficulty, considerable efforts have been conducted in the past for developing more accurate and effective CPT interpretation methods [4–12].

Generally speaking, each CPT sounding sample is a direct reflection/representation of the local physical and mechanical behaviors of the detected subsurface soil. The cause of the spatial differences of these local soil properties (manifested as fluctuation in CPT sounding record) is twofold: 1) different soil formations or lithological units would possess distinctive soil properties along the penetration path of a CPT; 2) in each presumably statistically homogeneous layer (i.e., a soil layer consisting of same soil formation) detected by a CPT, the soil properties may still be varying due to the inherent spatial variability. Hence, interpretation of CPT sounding data could be considered as a two-layer, spatial-constrained segmentation (i.e., delineate the boundaries of each soil layer) problem, in which the statistics of the soil properties, referred as the statistical characteristics in the feature space, are spatially dependent on the stratigraphic distribution of the soil formation, which are governed by certain spatial correlations and referred as the spatial characteristics in the physical space. In a sense, the latter aspect seems more critical for the accuracy of interpretation, since misclassification of the soil formations will consequently lead to misestimating of the associated soil properties. Although each of the existing methods has its own theoretical

* Corresponding author.
*E-mail addresses:* xwang2@uakron.edu (X. Wang), wang@aices.rwth-aachen.de (H. Wang), rliang1@udayton.edu (R.Y. Liang), zhuhehua@tongji.edu.cn (H. Zhu), hdi@uakron.edu (H. Di).

https://doi.org/10.1016/j.strusafe.2017.10.011
0167-4730/

background for representing the statistical characteristics of the material properties, few of them can be viewed as a complete model, since the spatial information of the CPT sounding containing spatial characteristics in the physical space is not fully taken into account. In addition, most of the existing research has focused on subsurface characterization using soundings from a single CPT. Nevertheless, interpreting multiple CPT sounding data at a project site by one sounding at a time may not be the best approach, especially for those subsurface regions containing thin soil layers, since CPT data of these thin layers from a single sounding record may not be sufficient to support accurate soil classification and properties estimation. Such misclassification and biased estimates may lead to additional difficulties in distinguishing or determining soil types, and interfering accurate stratigraphic profiling of the entire subsurface region of a project site. Therefore, there is a need for a methodology that allows for using multiple CPT sounding records at a project site to interpret subsurface soil stratification and relevant soil properties, with the accompanying ability to quantify uncertainties associated with such interpretation.

To overcome the aforementioned limitations of the existing methods, we have developed a Bayesian clustering approach based on hidden Markov random field (HMRF) models, where multiple CPT sounding data at a project site are taken as input for simulation and modeling purpose. Compared with the existing hard clustering approaches (i.e., the clustering results of which are deterministic) using distance-based similarity measures [11,12], the proposed approach has several advantages. To be more specific, first, the proposed approach is essentially a Bayesian method and hence it is likelihood based and soft clustering results (i.e., which are probabilistic) can be achieved. An appealing advantage derived from this probabilistic clustering approach is the ability of uncertainty quantification as the interpretation of the clustering result is also uncertain in nature. Second, in the proposed approach, the correlations in the feature space and the physical space are linked together using HMRF model, which is a two-layer model ideally suited for the CPT sounding data interpretation. The spatial characteristics of the stratigraphic sequence (i.e., stratigraphic profile detected along the penetration path of a CPT) in the physic space are described in the hidden layer of soil states (i.e., soil types) based on Markov models, which has been proven very successful for modeling subsurface stratigraphic profile [13–16]. Taking such spatial constraints in physic space into consideration can greatly improve the accuracy and robustness of the proposed approach, especially for CPT data of highly heterogeneous subsurface regions (i.e., clusters of certain soil type are uniformly spread or highly overlapped in the feature space). The observation (i.e., sampled soil properties) layer, with the statistical characteristics of material properties in the feature space, is handled by a commonly used Gaussian mixture model (GMM) [17]. As will be shown later, such two-layer model framework enables the proposed method to identify boundaries between different lithological units effectively and accurately, especially for thin layers and adjacent soil layers possessing very close material properties. Moreover, the proposed method allows one to interpret multiple CPT sounding records at a project site simultaneously. It will be shown in this paper that combining all the available sounding records from multiple CPTs can provide a more comprehensive and reliable description of the statistical characteristics of soil formations, thereby greatly enhancing the accuracy of soil classification and soil properties estimation in a project site with complex geological settings.

This paper is organized in the following manner. Section 2 provides a mathematical description of HMRF and the related sampling algorithm. In Section 3, a series of numerical examples of an artificial case are used to illustrate the application of the proposed method, as well as to examine its performance. In Section 4, the proposed methodology is applied to real CPT sounding data to demonstrate its applicability in handling actual project site conditions. Conclusions are then presented in Section 5.

## 2. HMRF model

### 2.1. General concepts and definitions

Within a framework of a random field, subsurface domain along the vertical penetration path of a CPT can be discretized into a set of depths $\boldsymbol{S} = \{1, 2, 3, \ldots, s\}$, according to the resolution of the CPT sounding samples. For each depth $i \in \boldsymbol{S}$, its soil label (representing a soil type) $X_i$ is considered as a random variable, which takes a value $x_i$ in the set of soil labels $\boldsymbol{L} = \{1, 2, 3, \ldots, k\}$, denoted as $X_i = \{x_i | x_i \in \boldsymbol{L}\}$. Then a configuration of the soil labels of all the depths can be denoted as $\boldsymbol{x} = (x_1, x_2, \ldots, x_s)$, and that all the possible configurations can be represented as $\boldsymbol{X} = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_s) | x_i \in \boldsymbol{L}, i \in \boldsymbol{S}\}$. Similarly, the CPT sounding sample of each depth is denoted as $Y_i = \{y_i | y_i \in \boldsymbol{R}^d\}$, where $\boldsymbol{R}^d$ is a $d$ dimensional feature space corresponding to the number of soil properties sampled, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_s)$ is a realization of a sequence of sounding samples. The CPT sounding samples are fundamentally dependent on the material properties of the soils being penetrated by the cone. Therefore, given a certain soil label $X_i = l$, the corresponding sounding samples $Y_i$ will follow a conditional probability distribution:

$$p(y_i | l) = f(y_i; \theta_l), \quad l \in \boldsymbol{L}, \quad i \in \boldsymbol{S} \tag{1}$$

in which $\theta_l$ is a set of parameters dependent on soil label $l$, and that we denote the set of all the parameters as $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_l)$, $f(\cdot; \theta_l)$ is a function family with same analytic form.

The object of site characterization using CPT data actually is to estimate the unknown true stratigraphic profile $\boldsymbol{x}^*$ and the corresponding soil properties $\boldsymbol{\Theta}^*$ based on given sounding samples $\boldsymbol{y}$. Herein we follow the *Maximum a posteriori* (MAP) criterion, which implies that the best estimation of $\boldsymbol{x}^*$ and $\boldsymbol{\Theta}^*$, denoted as $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{\Theta}}$, lead to the maximum posterior probability:

$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\Theta}}) = \underset{\boldsymbol{x} \in \boldsymbol{X}}{\operatorname{argmax}} P(\boldsymbol{x}, \boldsymbol{\Theta} | \boldsymbol{y}) \tag{2}$$

According to Bayes' theory, the posterior probability in Eq. (2) can be further expressed as:

$$P(\boldsymbol{x}, \boldsymbol{\Theta} | \boldsymbol{y}) \propto P(\boldsymbol{x}) P(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\Theta}) \tag{3}$$

where $P(\boldsymbol{x})$ is referred as the prior probability, which reflects the spatial characteristics (i.e., correlation) of the distribution of soil labels (i.e., subsurface stratigraphic profile); $P(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\Theta})$ denotes the likelihood probability, which provides a statistical description of the sounding samples in the feature space. The mathematic description of prior probability and likelihood probability is introduced in the following sections.

### 2.2. Markov random field (MRF) and prior probability

MRF is a powerful and widely used model for simulating the distribution pattern of the subsurface lithological units [13–15,18]. In an MRF, the random variable at depth $i \in \boldsymbol{S}$ is correlated with other corresponding variables of certain spatially adjacent depths $\partial i$, which are referred as neighbors of depth $i$. Particularly, for the depths along the one-dimensional CPT penetration path, the neighborhood system can be defined as $\boldsymbol{N} = \{\partial i = (i - \gamma, \ldots, i - 1, i + 1, \ldots i + \gamma)\}$, where $\gamma$ is an integer. Under this condition, the soil stratification $\boldsymbol{x}$ is said to be an MRF with respect to neighborhood system $\boldsymbol{N}$, if and only if the joint probability of each possible configuration $\boldsymbol{x} \in \boldsymbol{X}$ is strictly positive, i.e.,