Computers and Geotechnics 70 (2015) 159–168

Contents lists available at ScienceDirect

Computers and Geotechnics

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

Research Paper

Site-specific probability distribution of geotechnical properties

Yu Wang [a,*], Tengyuan Zhao [a,1], Zijun Cao [b,2]

[a] Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[b] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, No. 8 Donghu South Road, Wuhan, China

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form 8 July 2015
Accepted 20 August 2015
Available online 8 September 2015

Keywords:
Variability of geotechnical properties
Probability and statistics
Mixture model
Bayesian method
Markov Chain Monte Carlo simulation

ABSTRACT

Although the site-specific nature of soil variability has been well-recognized, it is difficult to obtain the site-specific probability distribution of geotechnical properties. Previous studies on soil variability were usually based on a large number of data that have been collected from many different sites in a large region, or even from different parts of the world. For geotechnical engineering practices in a specific project, it is the variability of geotechnical properties within this specific site, not the variability from many different sites, that geotechnical engineers are interested in and require. This leads to the questions of how to model the site-specific variability of geotechnical properties and how to estimate the site-specific probability distribution of geotechnical properties. This paper aims to address these questions using a statistic concept called mixture model and the Bayes' theorem. It is shown that the site-specific probability distribution of geotechnical properties can be considered as a weighted summation of a number of normal or lognormal distributions with different distribution parameters. Then, estimating site-specific probability distribution of geotechnical properties is equivalent to finding a suitable group of normal or lognormal distributions and their respective weights, based on the data available. The formulation and implementation procedure are illustrated and validated through an effective cohesion example and an example of estimating site-specific probability distribution of effective friction angle, based on a limited number of standard penetration test data (i.e., SPT N values) obtained for the project. The proposed methods perform satisfactorily in the examples.

1. Introduction

Uncertainties or variability are unavoidable in geotechnical engineering, and they arise in loads, geological site interpretation, geotechnical properties, calculation models, etc. (e.g., [1–4]). Among these geotechnically-related uncertainties, variability of geotechnical properties is one of the most important uncertainties. Since these uncertainties significantly affect geotechnical designs and analyses, how to properly model and deal rationally with such uncertainties is an important issue in geotechnical engineering (e.g., [5–7]).

Probability and statistics have been used in previous studies to model the variability of geotechnical properties (e.g., [8–13]). As summarized in Table 1, various soil properties have been shown to follow either normal or lognormal distributions (e.g., [8,14–16]). For example, Lumb [8] studied various properties of four types of soils in Hong Kong, such as a soft marine clay deposited in shallow coastal water, an alluvial sandy clay, a residual silty sand, and a residual clayey silt. The soil properties included the Atterberg limits, compression index, void ratio, drained angle of shearing resistance, and so on. He concluded that variability of these soil properties can be modeled by a normal or lognormal distribution. Hoeksema and Kitanidis [14] collected and studied the data from aquifers in 31 different regions, such as Chicago, East Central Illinois, and Iowa, and showed that hydraulic conductivity and storage coefficient of the aquifers can be considered to be lognormally distributed. Lacasse and Nadim [15] reviewed some soil property test results in Norwegian Geotechnical Institute's files and extensive literature, and provided the probability distributions for different soil properties, such as cone resistance from cone penetration tests in sand and clay, undrained shear strength, plastic limit. They also concluded that these soil properties follow either a normal or lognormal distribution.

It should be clarified, however, that these previous studies were based on a large number of data that have been collected from many different sites in a large region, or even from different parts of the world. For geotechnical engineering practices in a specific

* Corresponding author. Tel.: +852 3442 7605; fax: +852 3442 0427.
E-mail addresses: yuwang@cityu.edu.hk (Y. Wang), tengyzhao2-c@my.cityu.edu.hk (T. Zhao), zijuncao@whu.edu.cn (Z. Cao).
[1] Tel.: +852 3442 6492; fax: +852 3442 0427.
[2] Tel.: +86 027 68774036.

http://dx.doi.org/10.1016/j.compgeo.2015.08.002
0266-352X/

Table 1
Probability distribution of soil properties (after [8,15]).

Soil property	Soil type	PDF
Undrained shear strength	Clay (triaxial tests)	LN
	Clay (index S_u)	LN
	Clayey silt	N
Cone resistance (from cone penetration test)	Sand	LN
	Clay	N/LN
Normalized undrained shear strength	Clay	N/LN
Plastic limit	Clay	N
Liquid limit	Clay	N
Plastic index	Clay	N
Submerged unit weight	All soils	N
Friction angle	Sand	N
Void ratio, porosity, initial void ratio	All soils	N
Overconsolidation ratio	Clay	N/LN
Compression index	Sandy clay	N

Note: "N" and "LN" stand for normal and lognormal distributions, respectively.

project (e.g., reliability-based geotechnical design or risk assessment at a given site), it is the variability of geotechnical properties within this specific site, not the variability from many different sites, that geotechnical engineers are most interested in and require. Please note that, although the soil properties collected from many different sites follow normal or lognormal distributions, the soil properties collected within a specific site might not follow a normal or lognormal distribution. The normal or lognormal distribution shown in the previous studies might be just a result of the central limit theorem, which indicates that lumping data from many different sources (i.e., different sites in this case) tends to result in a normal or lognormal distribution (e.g., [17]). This leads to the questions of what is the site-specific probability distribution of geotechnical properties and how to model the site-specific variability of geotechnical properties. Although the site-specific nature of soil variability has been well-recognized (e.g., [18]), it is a challenging task to obtain the site-specific probability distribution of geotechnical properties. This has been recognized as one of the major criticisms to probability-based methods in geotechnical engineering, as quoted "another criticism is that soil statistics are not readily available because of the site-specific nature of soil variability [18]."

This paper aims to address this challenge using a statistic concept called mixture models (e.g., [19]) and the Bayes' theorem. It starts with the basic concept of mixture model, followed by implementation of the mixture model concept under two different scenarios: (1) when extensive measurement data on geotechnical properties are available at a specific site; and (2) when the measurement data from a specific site are limited. For the first case, the implementation of the mixture model method is formulated and solved as a nonlinear programming problem. For the second case, the Bayes' theorem, total probability theorem, and Markov chain Monte Carlo (MCMC) simulation are used to obtain the site-specific probability density function (PDF) for geotechnical properties.

2. Concept of mixture model

Mixture model in statistics is a convenient semiparametric framework for modeling unknown distributional shapes (e.g., [19]). It states that an arbitrary PDF, $f(x)$, of a random variable X can be approximated well by a weighted summation of a number n of component density functions $f_i(x)$, $i = 1, 2, \ldots, n$. The component density function is frequently taken as some commonly-used density functions, such as normal distribution. The concept of mixture model can be expressed as:

$$f(x) = \sum_{i=1}^{n} \omega_i f_i(x) \tag{1}$$

where $f_i(x)$ represents a component of the mixture; ω_i is the weight or mixing proportion of the mixture. ω_i is non-negative, and all ω_i sum up to one; that is:

$$\begin{aligned} &0 \leqslant \omega_i \leqslant 1 \ for\ i = 1, 2, \ldots, n \\ &\sum_{i=1}^{n} \omega_i = 1 \end{aligned} \tag{2}$$

If the mixture contains n components (see Eq. (1)), the PDF $f(x)$ is referred to as a n-component mixture density function, and the corresponding cumulative distribution function (CDF), $F(x)$, is referred to as a n-component mixture distribution function.

Fig. 1 illustrates the concept of mixture model. An arbitrary PDF is plotted by a bold solid line in Fig. 1. This PDF does not follow any commonly used PDF (e.g., normal, lognormal, or exponential distributions). However, it can be properly portrayed by a weighted summation of several normal distributions with different weights (e.g., ω_i, $i = 1, 2, 3$). Fig. 1 plots these three normally-distributed component PDFs with different parameters (i.e., mean and standard deviation μ_i, σ_i, $i = 1, 2, 3$) by lines with triangles, crosses, and squares, respectively. According to the concept of mixture model, the value of $f(x^*)$ at a given x^* value for the arbitrary PDF in Fig. 1 is obtained by a weighted summation of their corresponding values at $f_1(x^*|\mu_1, \sigma_1)$, $f_2(x^*|\mu_2, \sigma_2)$, and $f_3(x^*|\mu_3, \sigma_3)$, i.e., $f(x^*) = \sum_{i=1}^{3} \omega_i f_i(x^*|\mu_i, \sigma_i)$. Note that $\sum_{i=1}^{3} \omega_i = 1$, and this weighted summation is applicable to all x values in $f(x)$.

Using the mixture model concept above, a PDF, $f(x)$, of a geotechnical property, X, at a location within a specific site can be considered as a weighted summation of a number of normal or lognormal distribution functions with different distribution parameters (i.e., mean μ and standard deviation σ). The normal and lognormal distribution functions are used as component density functions in this study for constructing the site-specific probability distribution of geotechnical properties, since they are commonly used in mixture model and previous studies (see Table 1) showed that the soil property data collected from different sites generally follow either normal or lognormal distribution. In addition, lognormal distribution has an advantage in modeling non-negative random variable. Because many geotechnical properties are non-negative due to their physical meaning, lognormal distribution is also adopted as one of the possible component density functions. Eq. (1) can be rewritten as:

$$f(x) = \sum_{i=1}^{n} \omega_i f_i(x|\mu_i, \sigma_i) \tag{3}$$

where $f(x)$ now represents the site-specific PDF of a geotechnical property X; ω_i represents the weight for the i-th normally or lognormally distributed component PDF; $f_i(x|\mu_i, \sigma_i)$ is the i-th normally or lognormally distributed component PDF with parameters μ_i and σ_i. The component PDFs, $f_i(x|\mu_i, \sigma_i), for\ i = 1, 2, \ldots n$, can be sorted in a descending order according to their respective weights, ω_i. Therefore, the component $f_i(x|\mu_i, \sigma_i)$ with a small i value has a large value of ω_i and significant effects on the site-specific PDF of the geotechnical property.

Based on the mixture model concept, estimating the site-specific $f(x)$ for a geotechnical property X becomes a problem of estimating the component densities $f_i(x|\mu_i, \sigma_i)$ and their corresponding weights ω_i. A mixture model is able to model a complex and site-specific $f(x)$ for the geotechnical property through a proper identification of its components to represent accurately the local variations of $f(x)$. It not only provides a smooth fit to the overall distribution, but also allows clear demonstration of the multi-modal nature of the distribution, when such multi-modal nature exists, as shown in the illustrative example in the following section.