Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards

General and local estimation of local average and their application in geotechnical parameter estimations

Yusuke Honjo a, Budhi Setiawan a
a Department of Civil Engineering, Gifu University, Japan

Online Publication Date: 01 September 2007

To cite this Article: Honjo, Yusuke and Setiawan, Budhi (2007) 'General and local estimation of local average and their application in geotechnical parameter estimations', Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 1:3, 167 - 176

To link to this article: DOI: 10.1080/17499610701745960
URL: http://dx.doi.org/10.1080/17499610701745960

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
General and local estimation of local average and their application in geotechnical parameter estimations

YUSUKE HONJO* and BUDHI SETIAWAN

Department of Civil Engineering, Gifu University, Japan

Two kinds of estimation variance functions for estimating a local average (LA) of a stationary (homogeneous) random field (RF) are derived. One is local estimation (LE) and the other is general estimation (GE) of LA. The former is for estimating LA at the observation location, and the latter is for obtaining LA at any arbitrary location within the RF. The geotechnical implications of these two estimations are that LE is for estimating LA of geotechnical parameters at the spot where the investigations are made, whereas GE is for estimating LA at any arbitrary location within the same layer. The behavior of the two estimation variance functions differs greatly, controlled by the number of observations (i.e. sample size) and the normalized layer thickness (layer thickness divided by autocorrelation distance of RF). Based on the derived estimation variance functions, methods for determining reliable characteristic values of geotechnical parameters and necessary sample size are proposed. The methods are based on the same framework as that of the traditional statistical theory, i.e. confidence interval of estimated parameters. However, the assumption of independently and identically distributed (i.i.d.) samples in the traditional statistical theory is replaced by the assumption of correlated samples from a stationary RF. The results obtained from the proposed methods for LE and GE differ from each other as well as from the traditional results, which has significant implications for geotechnical parameter estimation in geotechnical engineering practice.

Keywords: Geotechnical parameter estimation; Characteristic values; Sample size; Random field; Correlation structure; Variance functions

1. Introduction

1.1 Background and purpose of the study

In traditional geotechnical design, there is no specifically defined method for determining geotechnical parameter. In modern code drafting, however, this poses a problem for code writers. A representative value for a geotechnical parameter in a geologically identical soil layer is often called a characteristic value, and ways to determine this characteristic value have attracted the interest of code writers, especially those involved in the development of Eurocode 7 (Ovesen 1989; Simpson and Driscoll 1998; Orr and Farrell 1999; Bauduin 2002) and others (Honjo and Kusakabe 2002; Becker 2006; Watabe et al. 2006, etc.).

In connection with this problem, a sufficient number of samples (i.e. sample size) to obtain a reliable characteristic value are an issue (Schneider 1997; Orr and Farrell 1999). A classic example in geotechnical engineering can be seen in Lumb (1974), who employed the idea of confidence interval for determining sample size, which is based on Student's t-value (standard normal distribution can be used if the variance is known, i.e. z-value). The sample size, n, depends on the t-value (or z-value), which is determined based on the chosen significance level, α.

* Corresponding author. E-mail: honjo@gifu-u.ac.jp
It should be noted that, in the determination of both the soil parameter characteristic values and the sample size, the methods based on the traditional statistical theory, such as employed by Lumb (1974), depend on the assumption of independently and identically distributed (i.i.d.) samples. On the other hand, as originally proposed by Lumb (1974), Vanmarcke (1977) and others, the random field (RF) models spatial variations of soil properties. Many researchers, such as Honjo and Kuroda (1991), Phoon and Kulhawy (1999), Fenton and Griffiths (2002), Griffiths and Fenton (2004) and others employ this approach. In RF, the correlation structure of spatially varying geotechnical parameters plays a major role. The i.i.d. assumption introduced above differs completely from the samples obtained from RF. It is, therefore, very important to establish a theory to determine characteristic values and sample size when samples are obtained from RF.

As suggested by Vanmarcke (1977), it is the local averages (LA) of soil properties that are important in controlling the behavior of geotechnical structures, such as piles, shallow foundations and slopes. This point is also taken into account in this study. Characteristic values should be determined from this viewpoint, and it is also stated, for example, in Eurocode 7 (CEN, 2004) that:

The zone of ground governing the behavior of a geotechnical structure at a limit state is usually much larger than a test sample or the zone of ground affected in an in situ test. Consequently, the value of the governing parameter is often the mean of a range of values covering a large surface or volume of the ground. The characteristic value should be a cautious estimate of this mean value. (CEN EN (997-1, 2.4.5.2 (7))

It is important to point out the necessity to distinguish between two situations in the problem of LA estimation: the need to determine LA at a certain location in the ground, for example right under the spot where a structure is to be built, or at any arbitrary location in a certain area, for example a reclaimed area where a container yard is to be located. We define the former case as local estimation (LE) and the latter as general estimation (GE).

The two situations differ greatly in geotechnical engineering practice as well as in statistical estimation. In the LE, we can reduce estimation error considerably by obtaining samples at the site, for example carrying out sounding at the spot. In the GE, on the contrary, a large amount of sampling is required to accurately estimate the LA at any arbitrary points.

1.2 Statement of the problem

As the problem identified above is quite diverse and complex, it is simplified here, in order to highlight the basic idea for estimating LA in both GE and LEs. The main focus in this study is the accuracy of estimating LA; therefore, characteristics of the estimation variances are the main issue to be considered.

The simplifying assumptions made in this study are listed as follows (see also figure 1):

1. A geologically identical soil layer of thickness L is assumed. One-dimensional LA over the whole thickness of the layer of a geotechnical parameter (e.g. undrained shear strength) is estimated. The LA of a geotechnical parameter for vertical direction over a soil layer of length L is defined:

\[
X_L = \frac{1}{L} \int_0^L X(z)dz
\]

(1)

2. The value of the geotechnical parameter at a point is described by the superposition of a constant mean value and a random component. The mean value is unknown, but is a constant, whereas the random component has zero mean, and known standard deviation (or coefficient of variation; Cov) and correlation structure (i.e. type of autocorrelation function and autocorrelation distance [related to the scale of fluctuation (Vanmarcke, 1977)]) are known.

3. LA is estimated by the average of equally spaced samples of size n over length L. Thus, sample interval can be simply calculated as \( \frac{L}{n} \).

The task tackled in this study is the estimation of the unknown local mean value of RF from n samples taken from the whole thickness of the layer at equal intervals of \( \frac{L}{n} \).

It is possible to relax some of the assumptions made. However, as the purpose here is to emphasize the essence of the idea, generalization and extension of the idea will be discussed elsewhere.

Finally in this study, it is important to emphasize that the standard deviation and autocorrelation structure (i.e. figure 1. Modeling soil profile by random field.
autocorrelation function form with its autocorrelation distance) are assumed already known, and thus only the mean value is estimated based on the samples. This assumption is made because the number of samples in geological engineering is rarely sufficient to estimate all the second-order statistics (i.e., variance and autocorrelation function form with the autocorrelation distance). Conservative values of the second-order statistics can be assumed from the previous studies and experience. For the sake of completeness of theoretical development, it is desirable to extend this theory to a case where second-order statistics are also estimated from the observation data, which is similar to t-tests in the traditional statistical theory. This can be a future extension of the present study.

2. Estimation of LA

2.1 Local average and variance function

A geotechnical parameter within a layer is described as \( X(z) \), where \( z \) is the co-ordinate in the vertical direction. As stated in the introduction, only a one-dimensional RF is considered in this study. Since second-order stationarity (i.e., stationarity in a weak sense) is assumed, the following equations hold:

\[
E[X(z)] = \mu \tag{2}
\]

\[
\text{Var}[X(z)] = E[(X(z) - \mu)^2] = \sigma^2 \tag{3}
\]

\[
\text{Cov}[X(z), X(z + \Delta z)] = E[(X(z) - \mu)(X(z + \Delta z) - \mu)] = \sigma^2 \rho(\Delta z) \tag{4}
\]

where \( \mu \) is the mean, \( \sigma^2 \) is the variance, and \( \rho \) is the autocorrelation function.

Both exponential (EAF) and Gaussian type autocorrelation functions (GAF) are introduced in this study as they are often used in geotechnical applications (e.g., Vanmarcke 1977, Phoon and Kulhawy 1999, Fenton and Griffiths 2001, Griffiths and Fenton 2004, etc.).

Exponential type autocorrelation function (EAF):

\[
\rho(\Delta z) = \exp[-(\Delta z/\theta)] \tag{5}
\]

Gaussian type autocorrelation function (GAF):

\[
\rho(\Delta z) = \exp[-(\Delta z/\theta^2)] \tag{6}
\]

where \( \theta \) is the autocorrelation distance. (Note that the scale of fluctuation defined by Vanmarcke (1977) is \( 2\theta \) in the former, and \( \sqrt{\pi}\theta \) in the latter.)

It is apparent that the mean of LA coincides with the original mean of RF, \( \mu \). Furthermore, Vanmarcke (1977, 1982) extensively studied the variance reduction of the LA from the original variance of RF, and derived the so-called variance function, \( \Gamma^2(L) \). The variance of LA is given as:

\[
s_L^2 = E\left[ \left( \frac{1}{L} \int_0^L X(z)dz - \mu \right)^2 \right] = \sigma^2 \Gamma^2(L) \tag{7}
\]

The variance function, \( \Gamma^2(L) \), for EAFs and GAFs are given, respectively, as:

**Variance function for EAF:**

\[
\Gamma^2(L) = \left( \frac{\theta}{L} \right)^2 \left[ 2 \left( \frac{L}{\theta} - 1 + \exp\left(-\frac{L}{\theta}\right) \right) \right] \tag{8}
\]

**Variance function for GAF:**

\[
\Gamma^2(L) = \left( \frac{\theta}{L} \right)^2 \frac{\sqrt{\pi}}{\theta} \text{erf}\left( \frac{L}{\theta} \right) + \exp\left(-\frac{L}{\theta}^2 \right) - 1 \tag{9}
\]

Note that

\[
\text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2} dt
\]

LA can differ significantly from the mean value of RF, because \( L \) may not be long enough to accommodate all possible fluctuations of RF. Since the property of LA is greatly affected by the relative magnitude of layer thickness, \( L \), and the autocorrelation distance of RF, \( \theta \), we introduce normalized layer thickness, \( L_n \), which is given as:

\[
L_n = \frac{L}{\theta} \tag{10}
\]

Since the layer thickness and the sampling length is assumed to be the same in this paper, \( L_n \) can also be termed as normalized sample line length.

2.2 General and local estimation of local average

LA is estimated by sample average, which is an averaging value of equally spaced samples of size \( n \) over length \( L \):

\[
\hat{X}(n) = \frac{1}{n} \sum_{i=1}^{n} X(z_i) \tag{11}
\]

It is assumed for simplicity that samples are taken at equal intervals of \( \Delta z = L/n \), where \( z_1, \ldots, z_n \) indicates the location of these sample points and the co-ordinates are conveniently defined as:
\[ z_i = (i - 0.5) \frac{L}{n} = (i - 0.5) \Delta z \quad (i = 1, \ldots, n) \]  

(12)

The estimation variances for the GE\s and LE\s are evaluated, respectively. The difference in evaluating the estimation variances for these two cases lies in the true mean value for which the estimators are evaluated; that is to say, the true mean is set to the original mean value of RF for the GE, whereas it is set to LA along the sampling line for the LE.

It should be noted that, in this study, LA in the LE is averaged along the line where sampling is made. A more general formulation of this problem is possible; however, in order to keep the problem simple, this setting is preserved.

The estimation variance in the GE case can be specified as follows:

\[ s_{2G}(n, L_m) = E \left[ \left( \sum_{i=1}^{n} \frac{1}{n} X(z_i) - \mu \right)^2 \right] \]  

(13)

where the notation \( s_{2G} \) implies the estimation variance for the GE case. The variance is taken for the mean value of RF.

On the other hand, the estimation variance for the LE case is specified as:

\[ s_{2L}(n, L_m) = E \left[ \left( \sum_{i=1}^{n} \frac{1}{n} X(z_i) - \frac{1}{L} \int_0^L X(z)dz \right)^2 \right] \]  

(14)

where \( s_{2L} \) denotes the estimation variance for the LE case. In this case, the variance is taken for the local mean at the sampling location.

It may seem strange that both \( s_{2G} \) and \( s_{2L} \) are not only functions of \( n \), but also of \( L_m \), the normalized layer thickness (or the normalized sampling length). This will probably be understood when detailed estimation variances are derived in the next section.

### 2.2.1. Estimator variance in GE.

The estimator variance in the GE is obtained as

\[ s_{2G}(n, L_m) = E \left[ \left( \sum_{i=1}^{n} \frac{1}{n} X(z_i) - \mu \right)^2 \right] \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{ X(z_i) - \mu \} \{ X(z_j) - \mu \} \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} E\{ X(z_i) - \mu \} \{ X(z_j) - \mu \} \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[X(z_i), X(z_j)] \]  

(15)

As samples are taken at equal intervals, \( \Delta z = L/n \), equation (15) can be rewritten as follows by knowing that \( \rho(k \Delta z) = \rho(\Delta z)^k \):

\[ s_{2G}(n, L_m) = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho(\Delta z)^{k_{ij}} \]

\[ = \frac{\sigma^2}{n^2} \left\{ n + \sum_{i=1}^{n-1} 2(n-i) \rho(\Delta z)^i \right\} \]  

(16)

where

In the case of EAF:

\[ \rho(\Delta z) = \exp \left[ -\frac{\Delta z}{\theta} \right] = \exp \left[ -\frac{L}{n \theta} \right] = \exp \left[ -\frac{L_m}{n} \right] \]  

(17)

In the case of GAF:

\[ \rho(\Delta z) = \exp \left[ -\left( \frac{L_m}{n} \right)^2 \right] = \exp \left[ -\frac{(L_m)^2}{n^2} \right] \]

(18)

We propose writing the GE variance by using the GE variance function, \( \Lambda^2(n, L_m) \), as follows:

\[ s_{2G}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[X(z_i), X(z_j)] = \sigma^2 \Lambda^2(n, L_m) \]  

(19)

Therefore, \( \Lambda(n, L_m) \) is defined as

\[ \Lambda^2(n, L_m) = \frac{s_{2G}^2}{\sigma^2} = \frac{1}{n^2 \sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[X(z_i), X(z_j)] \]

\[ = \frac{1}{n^2} \left\{ n + \sum_{i=1}^{n-1} 2(n-i) \rho(\Delta z)^{i} \right\} \]  

(20)

It is interesting to note the following relationships:

\[ \lim_{L_m \to 0} \Lambda^2(n, L_m) = 1.0 \]

(21)

\[ \lim_{L_m \to \infty} \Lambda^2(n, L_m) = \frac{1}{n} \]

(22)

### 2.2.2. Estimator variance in LE.

The estimator variance in the LE is obtained as:

\[ s_{2L}(n, L_m) = E \left[ \left( \sum_{i=1}^{n} \frac{1}{L} \int_0^L X(z)dz \right)^2 \right] \]

\[ = E \left[ \left( \sum_{i=1}^{n} \frac{1}{n} X(z_i) - \frac{1}{L} \int_0^L X(z)dz - \mu \right)^2 \right] \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} E([X(z_i) - \mu] [X(z_j) - \mu]) \]

\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{nL} \int_0^L E[(X(z_i) - \mu)(X(z_j) - \mu)]dz \]

\[ -2 \sum_{i=1}^{n} \frac{1}{nL} \int_0^L E[(X(z_i) - \mu)(X(z) - \mu)]dz \]
3. Application examples

3.1 Determination of a characteristic value

3.1.1. Method for determining a characteristic value. A typical method for determining a characteristic value based on the traditional statistical theory applies the confidence interval (Orr and Farrell 1999). The same method is employed in this study.

However, it is assumed here that the standard deviation and autocorrelation distance are already known, and that mean value is estimated based on the samples. This assumption is made because the number of samples in geotechnical engineering is rarely sufficient to estimate all the second-order statistics. Conservative values of second-order statistics (i.e. COV and autocorrelation distance) can be assumed from previous studies and experience (e.g. Lumb, 1974; Phoon and Kulhawy 1999).

Suppose $X_i (i = 1, \ldots, n)$ are i.i.d. samples from $N(\mu, \sigma^2)$. It is assumed that the variance is known and the mean value must be estimated from the sample mean.

The one-sided confidence interval can be written as follows:

$$ \Pr \left( \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha} \right) = 1 - \alpha$$

where $z_{\alpha}$ is the $\alpha$ percentile point of the standard normal random variable, i.e. $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z_{\alpha}} e^{-x^2/2} dx = \alpha$.

Thus, the lower boundary of the mean value is obtained as:

$$ \mu \geq \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha} \tag{31}$$

The characteristic value is obtained as this lower boundary:

$$ x_k = \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha} = \bar{X} \left( 1 - z_{\alpha} \cdot \text{Cov} \right) \tag{32}$$

If we apply the same logic to determining the characteristic value of LE and GE of $L_A$, $1/\sqrt{n}$ in equation (31) should be replaced by $\Lambda_{LE}(n, L_A)$ and $\Lambda_{GE}(n, L_A)$, respectively:

$$ x_k = \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot \text{Cov} \Rightarrow x_k = \bar{X} \left( 1 - \Lambda_{LE}(n, L_A) \cdot \text{Cov} \right) \tag{33}$$

$$ x_k = \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot \text{Cov} \Rightarrow x_k = \bar{X} \left( 1 - \Lambda_{GE}(n, L_A) \cdot \text{Cov} \right) \tag{34}$$

3.1.2. An example. The example presented here is based on Orr and Farrell (1999) in explanation of determining a characteristic value in connection with Eurocode 7. The characteristic value of the angle of shearing resistance, $\phi'$, is determined for a 10-m depth of ground consisting of sand

\[ \text{Figure 3. Distribution of } \phi' \text{ value for a homogeneous layer (modified from Orr and Farrell 1999).} \]

where the following ten $\phi'$ values at the equal intervals were assumed to be obtained from triaxial tests: $33.0^\circ$, $35.0^\circ$, $35.5^\circ$, $32.5^\circ$, $37.5^\circ$, $34.5^\circ$, $36.0^\circ$, $31.5^\circ$, $37.0^\circ$ and $33.5^\circ$. The LA of $\phi'$ for the whole layer thickness (i.e. 10 m) is estimated. The data is presented in figure 3.

Arithmetic mean of the sample is $34.4^\circ$ and the standard deviation is $1.97^\circ$ (or COV = 0.057). In this example, we assume COV of $\phi'$ obtained from triaxial tests to be 0.08 (i.e. slightly conservative value from sample COV). Phoon and Kulhawy (1999) collected numerous test results on various soil parameters. According to their recommendation, COV of $\phi'$ of sand ranges between 5 and 11% with an average of 9%. The result is consistent with the data given in this example, and it is not unreasonable to set COV of 8% as a conservative assumption.

The autocorrelation distance (or scale of fluctuation) must be assumed, where the EAF is assumed in this example. Unfortunately, there is no data available for the autocorrelation distance in Phoon and Kulhawy (1999) for the friction angle of sand. However, the scale of fluctuation for cone penetration resistance, $q_c$, in sand is available for several cases with the recommended range between 0.1 and 2.1 m. It is also possible to guess the scale of fluctuation

\[ ^1 \text{There are opinions that tan } \phi' \text{ should be used in obtaining the characteristic value of the internal friction angle instead of } \phi'. \text{ The authors just borrowed the example from Orr and Farrell (1999), and are not making any judgments on what quantities should be used.} \]
from the data provided in figure 3 based on a method proposed by Vannarcke (1977), who recommends evaluation using the average distance between crossing intersections (about 2 m in this case), which results in $\theta = 0.8$ m.

Based on this information, the autocorrelation distance, $\theta$, for this case was set in a parametrically in order to observe the difference in results. In this example, $\theta$ is assumed to be either 0.4, 0.8 or 1.67 m. These give the normalized sample length, $L_m$ of 25, 12.5 and 6, respectively. Whether this is a conservative assumption or not depends on the purpose of the estimation. It will be seen later that a longer autocorrelation distance is a conservative assumption for estimating GE (i.e. the true mean value of RF), but is not conservative for LE.

The characteristic value is determined by either of the equations below:

$$\phi_c = \phi(1 - z^2 \cdot \Lambda_c(n, L_n) \cdot \text{Cov})$$

$$\phi_c = \phi(1 - z^2 \cdot \Lambda_c(n, L_n) \cdot \text{Cov})$$

For GE, the first equation should be used, whereas for LE, the second one should be employed. By taking a significance level of 5%, the following values are substituted to the equations:

$$\phi = 34.4^\circ, \quad \text{Cov} = 0.08, \quad z_{0.05} = 1.645$$

$$\Lambda_c(10, 6.0) = 0.538, \quad \Lambda_c(10, 6.0) = 0.100$$

$$\Lambda_c(10, 12.5) = 0.411, \quad \Lambda_c(10, 12.5) = 0.142$$

$$\Lambda_c(10, 25.0) = 0.341, \quad \Lambda_c(10, 25.0) = 0.194$$

$$\Lambda_c(10, \infty) = \sqrt{10} = 0.316, \quad \Lambda_c(10, \infty) = 0.316$$

The calculated results are presented in table 1. LE gives the larger value especially for longer autocorrelation distance. On the other hand, GE results in a smaller value, especially for longer autocorrelation distance. Further geotechnical interpretation of the results is that, assuming $\theta = 0.8$ m as most likely, if designing a structure to be built on the spot where the investigation was made, it is recommended that the LA of $\phi_c = 33.8^\circ \approx 34^\circ$ be taken. On the other hand, if the structures are to be built in a relatively wide area, and the investigation is carried out to obtain a representative characteristic value for the whole area, then $\phi_c = 32.5^\circ$ may be employed. The difference of 1.5° in friction angle cannot be considered a small difference in geotechnical design.

### 3.2 Sample size determination

The required sample size for estimating the LA for the GE and LE is considered.

Here, it is assumed again that the standard deviation and autocorrelation distance are already known, and that the mean value of LA is estimated based on the sample mean. Therefore, conservative values of Cov and autocorrelation distance must be assumed in applying this method in practice.

As in the case of the traditional sample size determination, sample size can be determined based on the confidence interval as applied to soil sampling by Lumb (1974) and others. This method is explained first, and then it is extended to determine the sample size for estimating LA in a RF.

#### 3.2.1 Sample size determination in traditional theory. If a mean value is estimated from $n$ samples by an estimator

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and all samples are i.i.d. samples from a normal distribution with known variance, $\sigma^2$, the sample size can be determined by specifying the confidence interval of $\alpha$ significance level. A one-sided confidence interval is assumed:

$$Pr\left(-z_\alpha \geq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right) = Pr\left(-z_\alpha \geq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right) \leq \alpha \quad (35)$$

where $z_\alpha$ is the $\alpha$ percentile point in the standard normal distribution.

Therefore, to ensure the probability that the estimated mean value is smaller than $\mu - \Delta \sigma$ being less or equal to $\alpha$, where $\Delta = (\bar{X} - \mu) / \sigma$, a sufficient sample size can be obtained. Since $(\bar{X} - \mu) / \sigma < 0$,

$$-z_\alpha \geq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

thus,

$$\Delta \geq \frac{1}{z_\alpha} \frac{\sigma}{\sqrt{n}} \quad (36)$$

$\Delta$ can be called a normalized lower bound of an estimated mean value, the typical value of which could be 0.5, 1.0, etc. (Nagai 2003).

It should be recognized that the formulation taken here differs from that in equation (30), because the tasks to be solved differ between the two cases. In equation (30), it is necessary to obtain a lower boundary of the confidence interval that has a probability of 1 - $\alpha$, with the true mean included in the interval. On the other hand, in equation (35), sufficient sample size is necessary to ensure an
estimated mean value larger than \( \mu - \Delta \sigma \) with probability equal to or more than \( 1 - \alpha \).

3.2.2. Sample size determination in GE and LE. The same theory applies to the problem of sample size determination for GE and LE of LA. The only difference is that \( \sigma/\sqrt{n} \) in equation (35) should be replaced by \( s_{L,G}(n, L_a) \) and \( s_{L,E}(n, L_a) \), respectively.

For the case of GE,

\[
Pr \left( -z_{\alpha} \geq \frac{\bar{X} - \mu}{s_{L,G}(n, L_a)} \right) = Pr \left( -z_{\alpha} \geq \frac{\bar{X} - \mu}{\sigma/s_{L,G}(n, L_a)} \right) = Pr \left( -z_{\alpha} \Delta_{G}(n, L_a) \geq \frac{\bar{X} - \mu}{\sigma} \right) \leq \alpha \quad (37)
\]

Thus, requiring the estimated mean value to be larger than \( \mu - \Delta \sigma \) with the probability of \( 1 - \alpha \), a sufficient sample size can be obtained by finding \( n \) that satisfies the following equation:

\[
-\Delta_{G}(n, L_a) \geq \frac{\Delta}{z_{\alpha}} \geq \Delta_{G}(n, L_a) \quad (38)
\]

In the case of LE, the necessary sample size can be obtained in the same manner, except that \( \Lambda_{L}(n, L_a) \) is to be replaced by \( \Lambda_{L}(n, L_a) \) in equation (38):

\[
-\Delta_{L}(n, L_a) \geq \frac{\Delta}{z_{\alpha}} \geq \Delta_{L}(n, L_a) \quad (39)
\]

3.2.3. Some calculated results. The necessary sample size for various combinations of \( \Delta \) and \( L_a \) are calculated based on equations (38) and (39) for GE and LE in table 2. The following observations are possible:

1. For GE, the necessary number of samples increases as \( L_a \) decreases. (Note that NP in table 2 implies that more than 50 samples are required.) When \( L_a \) is smaller than a certain size (say, <1.0), it is not practically possible to estimate the mean value of a general LA (i.e. mean value of RF) reliably just from observing the given sample line length, \( L_a \).

2. For LE, on the contrary, the necessary sample size increase as \( L_a \) increases. If LA is taken for a relatively small \( L_a \) (for example <1.0), the mean value of the local LA can be estimated reliably with a considerably small sample size.

3. Sample size given in the i.i.d. case gives the same results for GE and LE. However, the given size is conservative for LE, but not conservative for GE.

4. It is understood, even intuitively, that the necessary sample size increases in both GE and LE when \( \Delta \) decreases because by decreasing \( \Delta \), the lower bound of the confidence interval is allowed to come closer to the true mean value.

The most interesting observation is that the estimation error decreases drastically in LE (local LA mean value estimation), especially when \( L_a \) is small. This is because if \( L_a \) is smaller, the values are more correlated within the interval, which makes the observations very effective in estimating LA.

It should be noted that the observations given here on the sample size determination are based solely on statistical and theoretical viewpoints. In practice, there are many other factors to be considered in determining sample size: characteristics and restrictions of soil investigation method, finding soil layer boundaries, seams and weak spots, etc. However, the findings presented here may provide some fundamental insight into this problem, especially the distinction between GE and LE.
4. Conclusion

A methodology is proposed for estimating the variance of LA of geotechnical parameters assuming a soil layer as a homogeneous RF. Although the study is at its primary stage, it presents the basic idea under simple assumptions, i.e., known variance and correlation structure, it gives new insight on determination of characteristic values of soil parameters and necessary sample size in soil investigation.

It is important to distinguish between the GE and LE of LA, because the situation differs significantly between these two cases in both geotechnical and statistical context. It is very important to distinguish between these two conditions, but they have rarely been explicitly recognized in geotechnical engineering practice.

The essence of the findings is contained in the GE and LE estimator variance functions, i.e., \( \lambda_{GE}(n, L_n) \) and \( \lambda_{LE}(n, L_n) \). It is of prime importance to recognize that these are functions of sample size, \( n \), and the normalized layer thickness (or the normalized sample line length), \( L_n \). The following are important features of these functions:

1. When \( L_n \) is small, \( \lambda_{GE}(n, L_n) \) is close to 1.0, whereas \( \lambda_{LE}(n, L_n) \) is close to 0.0, which implies that GE has relatively large statistical uncertainty compared to LE. They both reach 1/n as \( L_n \) becomes a large number, which coincides with the i.i.d. sampling condition.

2. Based on the facts explained above, very small sample size may be sufficient to estimate LE of LA, especially if \( L_n \) is small. On the other hand, GE of LA is impossible when \( L_n \) is too small, as shown in Table 2.

3. Based on the results obtained, by identifying the situation as either GE or LE, the characteristic values of geotechnical parameters can be determined in a more reasonable way, as illustrated in the example of this paper.

The number of samples required for LE and GE differ considerably; therefore, it is very important to distinguish between these two conditions when proposing a site investigation plan for geotechnical projects. This is the main statement of this research. It is also important to realize that the autocorrelation distance (i.e., scale of fluctuation) plays a dominant role in evaluating \( \lambda_{GE}(n, L_n) \) and \( \lambda_{LE}(n, L_n) \). It should be also noticed from the description of this paper that a longer autocorrelation distance is a conservative assumption for estimating GE of LA, but is not conservative for LE of LA.

Numerous research tasks follow along the line proposed in the present paper, some of which are listed as follows:

1. Extending the methodology in a more practical direction, such as to multi-dimensional RFs, multi-layered ground conditions, geotechnical parameters with non-constant trend components, unequal sample line length, and local averaging length, etc.

2. Relating the LE and GE concept to risk assessment and reliability based design of geotechnical structures. One example of such efforts can be seen in Honjo and Kuroda (1991).

3. Storing more practical examples related to LE and GE.

4. The autocorrelation distance of various geotechnical parameters may become increasingly significant. It is advisable to estimate these quantities for cases where abundant data is available.

5. For the sake of completeness of theoretical development, it is desirable to extend this theory to cases where the second-order statistics are also estimated from the observation data, which is similar to t-tests in the traditional statistical theory.

Acknowledgements

The authors are grateful to Dr Makoto Suzuki, Shimizu Corporation, for discussions and variable comments. The authors sincerely appreciate the careful review comments of the managing editor and the reviewers of this paper, which considerably improved the quality of the paper.

References


