# Application of geo-statistics and pairwise established CPT-based correlations for line infrastructure

H.J. Lengkeek
*Delft University of Technology, Delft, The Netherlands, h.j.lengkeek@tudelft.nl*
*Witteveen+Bos, Deventer, The Netherlands, arny.lengkeek@witteveenbos.com*
S.N. Jonkman[1], W. Kanning[2]
*Delft University of Technology, Delft, The Netherlands,*
*s.n.jonkman@tudelft.nl[1], w.kanning@tudelft.nl[2]*

**ABSTRACT:** Parameter determination is the first step in geotechnical engineering. Engineers are often confronted with limited data, large variations and different types of tests, both in-situ and laboratory tests. Within this complex setting, Codes require cautious estimates, so called characteristic values, preferable substantiated with observations and statistical methods. In this study, these statistical methods for populations and trend-functions are elaborated. Most codes and standards only refer to population statistics, whereas the reality is that,with the use of CPTs, trend functions such as correlations or transformation functions are more relevant. The aim of this paper is to provide a method how to use CPT and laboratory tests in practice in order to calculate characteristic values, on the basis of pairwise established CPT-based correlations, typically applicable for line infrastructure projects such as levees.

**Keywords:** Correlation; Variability; Geotechnical parameters; Characteristic value; Cone Penetration Test

## 1. Introduction

The first step in geotechnical engineering is parameter determination. It is also one of the most complex steps due to the geotechnical variability encountered in the soil, variability in tests and models, accuracy of the measurements and most often the limited number of tests.

This paper discusses the application of pairwise established CPT-based correlations in geo-engineering. Selecting and pairing data requires careful examination of test data using geotechnical experience. The method requires both derived parameters from laboratory test and in-situ measurements from CPT tests. The method requires that the laboratory test samples taken from boreholes adjacent to the CPT are paired with the CPT measurements over the same height interval and level. Both statistical analyses and regression analyses are used to transform the measured CPT results to geotechnical design parameters. These correlations are often called transformation models, as in [1, 2].

Most single variant correlations are described by basic mathematical relations. This paper focusses on three basic relations. The trendline and confidence limits, equivalent to a characteristic value with 5% confidence level, are addressed. For more complex correlations an alternative method is presented. The methods are applied on a database with paired laboratory test parameters and CPT measurements, taken at levees in The Netherlands.

This paper will finally present guidance how to determine the characteristic value for various situations along line infrastructure projects such as levees, where CPTs and laboratory test are combined.

## 2. Uncertainty in parameter determination

In geotechnical design, the predominant sources of uncertainties are the soil properties and the calculation model uncertainty. The overall uncertainty underlying a geotechnical parameter results from different sources of uncertainties. There are four primary sources of geotechnical uncertainties, as reported by [3]: (a) inherent variability, (b) measurement error, (c) transformation uncertainty and (d) statistical uncertainty. Inherent variability results primarily from the natural geologic processes, type of soil and state. Inherent variability is categorized as aleatoric in nature, because it cannot be reduced by the quantity of tests. In fact, it can worsen if the quality of selecting and pairing is insufficient causing additional variation. Measurement error is caused by equipment, procedural/operator, and random testing effects. Transformation uncertainty is introduced when measurements are transformed into geotechnical design parameters using empirical or other correlation models. Statistical uncertainty involves the assessment of the probability distribution, sample size and regression methods. These are categorized as epistemic in nature, and can be reduced by quantity of pairs and the quality of models (regression, transformation) and instrumentation.

Section 3 summarizes Code provisions in an attempt to define the characteristic value, taking into account local versus global failure, and regional variation. Section 4 deals with population statistics and probability distribution functions used in geotechnical engineering. These concepts are presented and applied on direct measurements of for example the unit weight of soil. Section 5 addresses regression methods for transformation models, and how to define the characteristic confidence limits. Section 6 presents a

method how to apply this into practice followed by the conclusions in section 7.

## 3. Codes and standards on geotechnical design parameters

Basic principles and rules concerning the structural resistance are given in the Eurocode [4, 5] and ISO 2394 [6]. Additional information can be found in [7]. In the paragraphs below a selection of relevant clauses of these standards will be cited and discussed.

### 3.1. ISO 2394

The following clauses from [6] are relevant to cite. The clauses below apply to materials and soils:

*Clause 2.4.3 characteristic value of a material property: "priori specified fractile of the statistical distribution of the material property in the relevant supply"*

*Clause 9.3.2: " For soils and existing structures, the values should be estimated according to the same principle and so that they are representative of the actual volume of soil or the actual part of the existing structure to be considered in the design."*

Annex D of [6] on "Reliability Based Design", describes the state of the art on uncertainties in parameter determination, statistical characterization and models. Furthermore, it is not prescriptive. Three uncertainties are mentioned (inherent soil variability, measurements errors, transformation uncertainty).

### 3.2. Eurocode 1997-1:

The following clauses from [5] are relevant to cite. The five clauses all apply to geotechnical parameters and cover the various aspects that will be addressed further on:

*Clause 2.4.5.2.(2)P: " The characteristic value of a geotechnical parameter shall be selected as a cautious estimate of the value affecting the occurrence of the limit state."*

*Clause 2.4.5.2.(10): "If statistical methods are employed in the selection of characteristic values for ground properties, such methods should differentiate between local and regional sampling and should allow the use of a priori knowledge of comparable ground properties."*

*Clause 2.4.5.2(11):" If statistical methods are used, the characteristic value should be derived such that the calculated probability of a worse value governing the occurrence of the limit state under consideration is not greater than 5%."*

*Clause 2.4.5.2(11) Note:" In this respect, a cautious estimate of the mean value is a selection of the mean value of the limited set of geotechnical parameter values, with a confidence level of 95%; where local failure is concerned, a cautious estimate of the low value is a 5% fractile."*

Note that these clauses without (P) are Application Rules, examples of generally recognized rules, which follow the Principles clauses (P) and satisfy their requirements (should). It is permissible to use alternatives to the Application Rules, provided that the alternative rules accord with the relevant Principles.

The ISO and Eurocode are conceptually similar, but the Eurocode is more prescriptive. The synthesis of both Codes is given below.

### 3.3. Synthesis

The approach of taking the characteristic value of the mean for global failure and the characteristic value for local failure are in fact the limiting cases, the two extremes. The actual characteristic mean value for global failure depends on more factors, such as scale of fluctuation, extent of failure surface relative to the limit state, and is likely to be in between these values. Various methods are reported in literature [8-14], essentially referring back to the research of Vanmarcke [15, 16] based on the concept of variance reduction due to spatial averaging.

Based on the review of codes, standards and literature the following definitions will be used:

- The characteristic value is statistically defined by the 90% confidence interval. The characteristic value of the population ($X_{kar}$) determines the value with 5% confidence level or 95% probability of exceedance (PoE). The characteristic value of the mean ($X_{m;kar}$) of the population determines the mean value with 5% confidence level or 95% probability of exceedance.
- The representative value takes into account the extent of the ground volume involved in the limit state, the effects of stress, state, time, structure, anisotropy.
- The representative value is generally a value between the characteristic value of the population ($X_{kar}$) and of the mean ($X_{m;kar}$).

Local failure refers to the case when a local weak spot can result in failure of the structure. No spatial averaging of variation takes places. An example is the failure of a pile tip with typical failure dimensions of 1 m at the pile tip, see Figure 1. In case the limit state involves local failure, the representative value can be based on the predicted value with 5% confidence level (90% confidence interval of the predicted value):

$$Y_{kar} = Y_m + t_{n-1}^{\alpha} \cdot s_y \sqrt{1 + \frac{1}{n}} \quad (1)$$

In equation (1) is ($s_y$) the standard deviation (σ) and $Y_m$ the average (μ) given a limited sample size.

Global failure refers to the case when the failure surface is relatively large compared to fluctuations and averaging of uncertainty occurs. An example is slope failure with typical dimensions of 5 m deep and 50 m wide sliding plane, see Figure 1.

In case the limit state involves global failure and it can be assumed that all variations are levelled out, the representative value (characteristic estimate of the mean) can be based on the average value with 5% confidence level (90% -confidence interval of the predicted mean):

$$Y_{m;kar} = Y_m + t_{n-1}^{\alpha} \cdot s_y \sqrt{\frac{1}{n}} \quad (2)$$

The degrees of freedom are equal to n-1. Eurocode 0 [4] uses slightly different terminology, the statistical uncertainty is described by the "confidence limit multiplier k":

$$Y_{rep} = Y_m + k \cdot s_y \quad (3)$$

$$Y_{rep} = Y_m\left(1 + k \cdot CV_y\right) \quad (4)$$

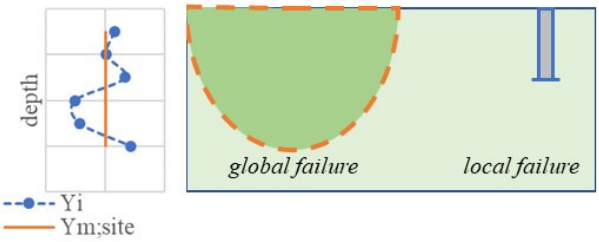


**Figure 1.** Illustration of global and local failure.

## 4. Population geo-statistics

### 4.1. General

This section presents the population statistics on geotechnical data. These are direct measurements of the required design parameter. Section 5 focusses on the relation between indirect measurements such as CPT parameters and direct measurements such as laboratory tests parameters. Most codes and textbooks consider point values. It is useful to introduce the basic principles that also apply to correlations in the next section.

When there is enough data available, statistical analyses can be used to estimate the mean (μ), the sum squared error (SSE), the standard deviation (σ) and the coefficient of variation (CV) of geotechnical design parameters, see equation (5) to (8).

$$\mu_y = \frac{\sum_{i=1}^{n}(Y_i)}{n} \quad (5)$$

$$SSE_y = \sum_{i=1}^{n}\left(Y_i - \mu_y\right)^2 \quad (6)$$

$$\sigma_y = \sqrt{\frac{SSE_y}{n}} \quad (7)$$

$$CV_y = \frac{\sigma_y}{\mu_y} \quad (8)$$

The characteristic value of a material property is generally defined as the value with 95% confidence level ($\alpha = 5\%$). The lower bound or inferior value is the value with 5% probability of adverse value or the value with 95% probability of exceedance (PoE). For a Normal distribution the following equation applies:

$$Y_{kar} = \mu_y + u_\alpha \cdot \sigma_y = \mu_y\left(1 + u_\alpha \cdot CoV_y\right) \quad (9)$$

With factor $u_{\alpha=5\%} = -1.645$.

When the coefficient of variation is large, the characteristic value can become negative. Most physical and empirical models do not allow negative soil parameters, hence the Lognormal distribution is recommended. This semi-infinite distribution allows for mitigation of negative values and is frequently used in these situations. For a Lognormal distribution with mean ($\lambda$) and standard deviation ($\zeta$) and no shift the following equation applies:

$$\zeta_y = \sqrt{ln\left(1 + \left(\frac{\sigma_y}{\mu_y}\right)^2\right)} \quad (10)$$

$$\lambda_y = ln\left(\mu_y\right) - \frac{1}{2}\zeta_y^{\,2} \quad (11)$$

$$CV_y \cong \sqrt{\left(e^{(\zeta_y{}^2)} - 1\right)} \quad (12)$$

$$\ln(Y)_{kar} = \lambda_y + u_\alpha \cdot \zeta_y \quad (13)$$

$$y_{kar} = e^{\ln(Y)kar} \quad (14)$$

Statistical analysis on soil investigation is always based on a sample and not on the population of a geotechnical property. To account for this statistical uncertainty the Student-t distribution is often used. In the following equations $u_\alpha$ is replaced by $t_{n-1}^{\alpha}$, depending on the confidence level ($\alpha = 5\%$) and the degrees of freedom, generally the sample size minus one.

For statistical analysis on a sample the standard error of the mean (SEM: $s_{y,m}$) is defined based on the following equations:

$$Y_m = \frac{\sum_{i=1}^{n} Y_i}{n} \quad (15)$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_m)^2}{n-1}} \quad (16)$$

$$CV_y = \frac{s_y}{Y_m} \quad (17)$$

$$s_{y,m} = s_y\sqrt{\frac{1}{n}} \quad (18)$$

The standard error of the mean is a measure of the dispersion of sample means around the population mean. Note that it is inversely proportional to the square root of the sample size, so it tends to decreases as the sample size increases. The standard error of mean is not often referred to in geotechnical engineering, but as shown later it is in fact the parameter that directly relates to the characteristic mean value used for global failure limit states.

The characteristic values according to Eurocode to be used in a semi-probabilistic approach are defined in equation (1) and (2). The stochastic parameters to be used in a full-probabilistic approach, are the same mean and a corrected standard deviation as in equation (19). The term between [] is the correction for the statistical uncertainty on the standard deviation used in probabilistic analyses, as a consequence of the limited number of tests.

$$s_{y;prob} = s_y \cdot \left[\frac{k}{u_\alpha}\right] \quad (19)$$

With factor $u_{\alpha=5\%} = -1.645$ for value with 5% confidence level or 95% probability of exceedance (PoE). The confidence limit factors (k) for local and global failure are defined as:

$$k_{kar} = t_{n-1}^{0.05}\sqrt{1 + \frac{1}{n}} \quad (20)$$

$$k_{m;kar} = t_{n-1}^{0.05}\sqrt{\frac{1}{n}} \quad (21)$$

### 4.2. Regional versus local dataset

A difference can be made between local and regional datasets. Local datasets refer to dataset that are all taken within the influence zone of the failure mechanism. Hence, partial averaging of local data can occur. Regional datasets are datasets that cover much larger distances, also outside the failure mechanism's influence, but still refer to the same layer and property. Hence, no or limited averaging can occur as only the part of the variance that reflects local uncertainty is subject to averaging.

The advantage of a model based on regional database is that it covers a whole range of measurements and the statistical error is reduced. The disadvantage is that it can be biased from location to location and regional variation is included in the variance. In case of a specific local structure, site specific soil investigations are available and hence there is no uncertainty about the local conditions. In case of a line infrastructure project, many sites or sections need to be checked which do not all have local soil investigations. An example of such a regional database over multiple sites along a line infrastructure project is illustrated in Figure 2.

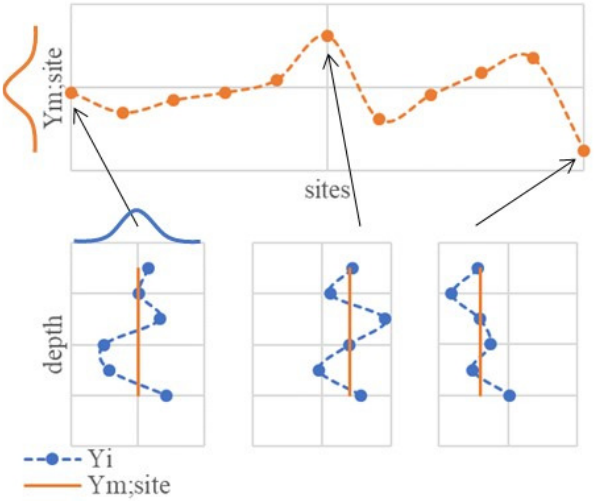


**Figure 2.** Illustration of regional dataset.

For the sites with soil investigation the local average and local standard deviation can be used to determine the characteristic value. For sites without local soil investigation one needs to make a cautious estimate. In practice the scale of fluctuation along the line infrastructure is equal or larger than the failure length in the same direction. Hence, failure is more likely to occur at a location where the local average is lowest. For this situation one needs to define the characteristic value, for both local and global failure. An example of such a model is presented in [17]. In this paper a pragmatic approach is presented how a regional database can be used on the basis of a regional model.

The total variance in a regional database can be related to the local variance at a site, and regional variation of the mean per site. The characteristic value for a line infrastructure with regional database can be determined on the basis of the following model assumptions:

- The total variance is equal to the sum of the regional variance of site means and the local variance at the sites.
- The variances are independent.
- The local variances are constant at each site of the region.
- The regional variance term $\alpha$ is the ratio of local variance and the total variance.

$$\sigma_{tot}^2 = \sigma_{reg}^2 + \sigma_{loc}^2 \quad (22)$$

$$\alpha = \frac{\sigma_{loc}^2}{\sigma_{tot}^2} = 1 - \frac{\sigma_{reg}^2}{\sigma_{tot}^2} \quad (23)$$

The total mean, total standard deviation and regional standard deviation can be derived directly from a regional dataset. The local standard deviation follows from the regional variance term $\alpha$, see equation (22) and (23).

For a line infrastructure with a regional database, basically four situations exist. Either the location is known and local investigations exist, or this is not the case. Furthermore the failure can be local or global. The following combinations in Table 1 should be considered for the determination of the characteristic value:

**Table 1.** Practical application for characteristic values of measured geotechnical parameters applied to line infrastructure, four cases

| | Known location | Unknown location |
|---|---|---|
| | Local dataset | Regional dataset |
| Global failure | case I (eq. 2) | case III (eq. 24) |
| Local failure | case II (eq. 1) | case IV (eq. 25) |

For case I and case III a global failure is expected, so in general equation (2) applies. For case II and case IV a local failure is expected so in general equation (1) applies. The differences between the four cases is elaborated below.

For case I it is recommended to use equation (2), using the local data, local average, local standard deviation, and Student-t and number of samples is based on the local database. For case II it is recommended to do the same based on equation (1).

For case III and case IV the local average is unknown. In that case a more cautious estimate is appropriate. For case IV the approach is to apply equation (25), which is in fact the same as equation (2). The difference is that the regional average, regional standard deviation, and Student-t and number of samples is based on the regional database.

For case III (unknown locations with no local site investigation) the main uncertainty is in the local average. A cautious estimate can be derived from characteristic value of the site means. The characteristic value for case III can be set equal to this value. As the mean follows from the number of sites, the statistical uncertainty in equation (24) is based on the number of sites too. The equations for case III and IV are presented below:

$$Y_{m;kar;reg} = Y_{m;tot} + t_{n_{sites}-1}^{0.05} \cdot s_{reg} \sqrt{1 + \frac{1}{n_{sites}}} \quad (24)$$

$$Y_{kar;reg} = Y_{m;tot} + t_{n_{tot}-1}^{0.05} \cdot s_{tot} \sqrt{1 + \frac{1}{n_{tot}}} \quad (25)$$

In the example as shown in Figure 3 to Figure 5 simulations with a regional mean is 20 kN/m$^3$ with a total standard deviation of 1 kN/m$^3$. For the first site (1) also the characteristic value of the mean according to equation (2) and of the population according to equation (1) is shown.

The local and regional standard deviation vary with the regional variance term ($\alpha$ *is* 0.2 *to* 0.8). Figure 3 to 5 present results from random generations. In Figure 3 the within variation of each site is lager than the variation between the sites. This is also reflected in the regional characteristic values. The characteristic value of case III and IV differ significantly. In Figure 5 the within variation of each site is smaller than the variation between the sites and Figure 4 provides intermediate results. Hence it is concluded that for regions with high regional variation the characteristic value for global

failure and local failure almost coincide in case on local data is available.

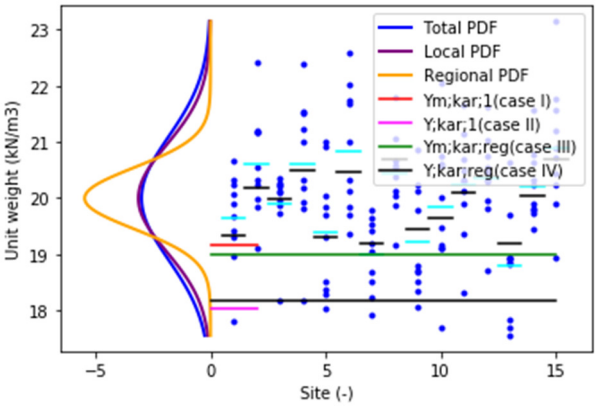


**Figure 3.** Simulation of 15 sites with low regional variance ($\alpha = 0.8$), 10 samples per site.

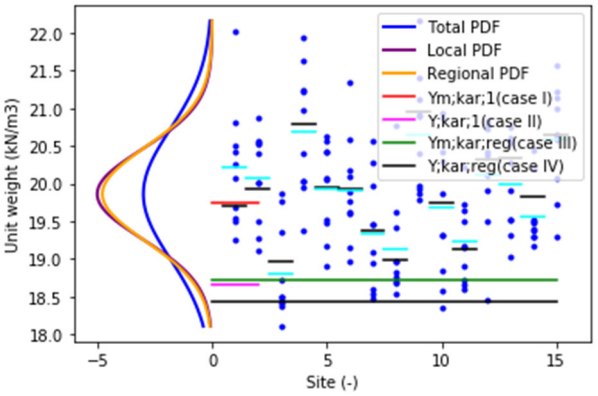


**Figure 4.** Simulation of 15 sites with average regional variance ($\alpha = 0.5$), 10 samples per site.

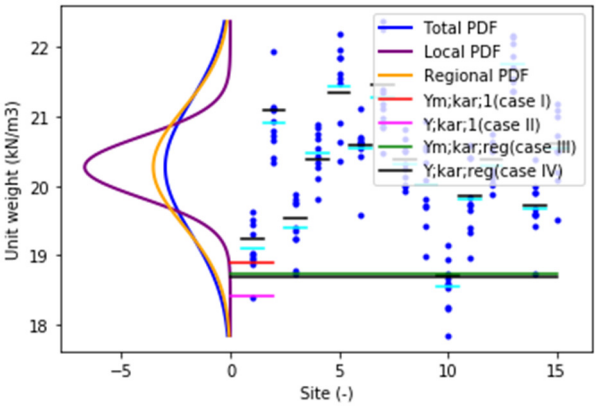


**Figure 5.** Simulation of 15 sites with high regional variance ($\alpha = 0.2$), 10 samples per site.

In the Dutch National Annex of Eurocode 7 [5] regional geotechnical parameters are presented including the coefficient of variation. This coefficient of variation is both applicable for local variations and for regional variations. This implies that $\alpha = 0.5$. This can be a useful a priori assumption in case no information is available about the regional variation.

## 5. Trend function geo-statistics

Transformation models relate test measurements to appropriate design properties. Uncertainty is introduced by fitting procedures and inaccurate paired data. Most transformation models were developed for a specific geomaterial type and/or a specific location. Site specific models are generally more precise than "global" models calibrated from data covering many sites. However, site-specific models can be significantly biased when applied to another site [2].

Regression methods are statistical tools to determine the best fit, and uncertainties associated with this best fit. Different regression methods and regression lines are available. In engineering practice mainly Ordinary Least Squares (OLS) regression is used.

The variation of the data about the trendline is modelled as a zero-mean random variable ($\varepsilon$). The standard deviation of $\varepsilon$ is an indicator of the magnitude of transformation uncertainty, but it also includes measurements errors and statistical uncertainty. Combining the basic statistical methods with the regression analyses provide the confidence limits for a target confidence level, such as $\alpha = 5\%$ equivalent to the characteristic value. Various regression analyses exist, generally based on the least square method. For correlations with CPT measurements three basic types of trendlines are selected, based on one independent variable, see Table 2.

The three trendlines are selected to cover most of the trends found in CPT measurements. The stochastic parameters are based on a Normal or Lognormal distribution. For the selection of the best trendline other statistical tools are available. Most important is to align the selected trendline with the empirical or physical relations or existing correlation model used in geotechnical engineering. Reference is made to [5] Clause 2.4.5.2.(1)P: " The selection of characteristic values for geotechnical parameters shall be based on results and derived values from laboratory and field tests, complemented by well-established experience.". In addition, the standard deviation and coefficient of determination ($R^2$) can be used to select the best trendline.

The first method is based on a linear trendline with free intercept. The regression method is Ordinary Least Squares (OLS) assuming homoscedasticity. The stochastic parameters are the intercept and standard deviation on regression of independent variable $(a_m, s_y)$. The accompanying regression parameter is the slope $(b_m)$.

The second method is based on a linear trendline with "no intercept". The preposed regression method used is OLS through the origin. The method requires homoscedasticity, whereas in practice often the variance of the error increases with the dependent variable (Y). As shown later, this is no limitation for practical application. The stochastic parameters are the slope and standard deviation of the slope parameter $(b_m, s_b)$. There are two alternatives based on statistics of the slope (ratio) of each data pair, either assuming a Normal or Lognormal distribution.

The third method is based on a power function. The trendline can be derived directly or after a Lognormal transformation is applied to both X and Y. The trendline after Lognormal transformation provides a similar linear trendline as in the first method. The transformation also implies a Lognormal distributed error term. The

equivalent stochastic parameters assuming for normality are the multiplier and standard deviation of the multiplier ($a_m$, $s_a$). The accompanying regression parameter is the exponent ($b_m$).

**Table 2.** Regression methods and trendlines

| Regression method | Trendline | Equation | Stochastic parameters |
|---|---|---|---|
| 1 OLS, free intercept | Linear function | $y = a + b \cdot x$ | $a_m, b_m, s_y$ |
| 2 OLS, through origin | Linear function | $y = b \cdot x$ | $b_m, s_b$ |
| 3 LN transformation, OLS | Power function | $y = a \cdot x^b$ | $a_m, b_m, s_a$ |

For all methods the standard deviation is used to determine the confidence limits and the prediction interval, corresponding to a characteristic value with 5% confidence level. The equations for the three regression methods are presented below.

## 5.1. CPT-based correlation undrained shear strength peat

The regression methods will be applied to a paired database, consisting of Direct Simple Shear (DSS) tests on peat and CPTU tests. The pairs are taken from CPTs and adjacent boreholes at 1 m distance, averaging the CPT over 20 cm at the corresponding level as the sample. The classification of peat samples is based on both the borehole logs and the CPT measurements. The tests have been taken from different soil investigation projects for levee strengthening projects in the Netherlands. All peat layers are Holocene layers, without further differentiation to type, structure or origin as classification is not always available. The in-situ stress level varies from 10 kPa in the green field hinterland (polders) to 100 kPa below the levee embankments. The samples are generally slightly overconsolidated due to aging, groundwater variations and man-made activities. The overconsolidation corresponds typically to a pre overburden stress of about 20 kPa.

## 5.2. Method 1: Linear function and OLS regression

Method 1 is used in case both the measurements and empirical or correlation models are best described by a linear function with free intercept $Y = b \cdot X + a$. Excel Linest function can be used to derive slope $b_m$, the intercept $a_m$ and the standard deviation $s_y$. The following equations follow from statistical textbooks.

$$Y_i = b_m \cdot X_i + a_m + \varepsilon \quad (26)$$

$$b_m = \frac{\sum_{i=1}^{n}(X_i - X_m)(Y_i - Y_m)}{\sum_{i=1}^{n}(X_i - X_m)^2} \quad (27)$$

$$a_m = Y_m - b_m \cdot X_m \quad (28)$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (b_m \cdot X_i + a_m))^2}{n-2}} \quad (29)$$

$$CV_y = \frac{s_y}{a_m} \quad (30)$$

The coefficient of variation ($CV_y$) is not constant with X, the presented definition above is normalized at the intercept (X=0). When comparing values from literature this should be carefully examined. The $s_y$ is the standard deviation of the estimate along the trendline. The characteristic values of Y can be determined by the following equation. The degrees of freedom are equal to n-2 for method 1.

$$Y_{kar;i} = b_m \cdot X_i + a_m + k_{kar} \cdot s_y \quad (31)$$

There is relatively greater uncertainty for values of X that are farther from its mean value. This is taken into account by the following leverage term, as in [11]:

$$\frac{(X_i - X_m)^2}{SSE_x} = \frac{(X_i - X_m)^2}{\sum_{i=1}^{n}(X_i - X_m)^2} \quad (32)$$

The leverage term is relevant for the characteristic mean value, and in particular for low or high values of X. The effect on the prediction interval is negligible and therefore often ignored in practice and not even considerd in a probabilistic analyses. The equation of the confidence limit multiplier for the prediction interval and confidence interval of the mean are:

$$k_{kar} = t_{n-2}^{0.05}\sqrt{1 + \frac{1}{n} + \frac{(X_i - X_m)^2}{\sum_{i=1}^{n}(X_i - X_m)^2}} \quad (33)$$

$$k_{m;kar} = t_{n-2}^{0.05}\sqrt{\frac{1}{n} + \frac{(X_i - X_m)^2}{\sum_{i=1}^{n}(X_i - X_m)^2}} \quad (34)$$

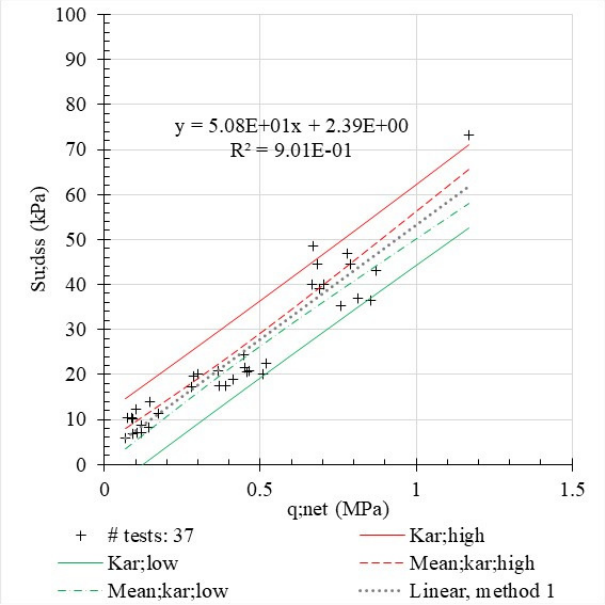


**Figure 6.** Method 1, OLS regression with free intercept.

Figure 6 shows paired data of net cone resistance versus undrained shear strength from Direct Simple Shear tests on peat samples. The intercept is just above the origin. Consequently, the lower characteristic value for low cone resistances is negative. This is physically not possible so in those cases another method would be preferable.

## 5.3. Method 2: Linear function and OLS regression through origin

Method 2 is used in case both the measurements and empirical or correlation models are best described by a linear function through the origin $Y = b \cdot X$. Different methods are available in practice and three of them will be described and compared:

- OLS through the origin
- Classical statistics on ratio Y/X
- Lognormal transformation

## Method 2a:

Method 2a is based on OLS regression through the origin. Excel Linest function can be used to derive slope $b_m$, the standard deviation $s_y$ and the standard error of mean (SEM) $s_{b,m}$. It should be noted that in this approach standard deviation of the slope parameter (b) is used as stochastic parameter. This standard deviation can be determined from the standard error of mean. The equations follow from statistical textbooks.

$$Y_i = (b_m + \varepsilon)X_i \quad (35)$$

$$b_m = \frac{\sum_{i=1}^{n}(X_i \cdot Y_i)}{\sum_{i=1}^{n}(X_i{}^2)} \quad (36)$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - b_m \cdot X_i)^2}{n-1}} \quad (37)$$

$$s_{b,m} = \frac{s_y}{\sqrt{\sum_{i=1}^{n}(X_i{}^2)}} \quad (38)$$

$$s_b = s_{b,m}\sqrt{n} = \frac{s_y\sqrt{n}}{\sqrt{\sum_{i=1}^{n}(X_i{}^2)}} \quad (39)$$

$$CV_b = \frac{s_b}{b_m} \quad (40)$$

The characteristic values of Y can be determined by the following equation. The degrees of freedom are equal to n-1 for method 2 and no leverage term is included.

$$Y_{kar;i} = (b_m + k_{kar} \cdot s_b)X_i \quad (41)$$

$$k_{kar} = t_{n-1}^{0.05}\sqrt{1+\frac{1}{n}};\ k_{m;kar} = t_{n-1}^{0.05}\sqrt{\frac{1}{n}} \quad (42)$$

## Method 2b:

Method 2b is based classical statistics of the ratio $\frac{Y_i}{X_i}$, which is in fact not a regression method. Consequently, the average slope $b_m$ can deviate from the one determined in method 2a by regression. The equations follow from statistical textbooks.

$$Y_i = (b_m + \varepsilon)X_i \quad (43)$$

$$b_m = \frac{\sum_{i=1}^{n}\left(\frac{Y_i}{X_i}\right)}{n} \quad (44)$$

$$s_b = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{Y_i}{X_i} - b_m\right)^2}{n-1}} \quad (45)$$

$$Y_{kar;i} = (b_m + k_{kar} \cdot s_b)X_i \quad (46)$$

## Method 2c:

Method 2c is based on Lognormal transformation of both X and Y. The method is comparable with method 2b, because $\ln Y - \ln X = \ln\frac{Y}{X}$. The average slope $b_m$ deviates from the one determined in method 2a and 2b. The confidence limits are not symmetrical due to the Lognormal transformation. This method is also described in [7].

$$Y_i = (b_m \cdot \varepsilon)X_i \quad (47)$$

$$\lambda_b = \frac{\sum_{i=1}^{n}\ln\left(\frac{Y_i}{X_i}\right)}{n} \quad (48)$$

$$b_m = e^{\lambda_b} \quad (49)$$

$$\zeta_b = \sqrt{\frac{\sum_{i=1}^{n}\left(\ln\left(\frac{Y_i}{X_i}\right) - \lambda_b\right)^2}{n-1}} \quad (50)$$

$$CV_b = \sqrt{\left(e^{(\zeta_b{}^2)} - 1\right)} \quad (51)$$

$$Y_{kar;i} = e^{\lambda_b + k_{kar}\cdot\zeta_b} \cdot X_i \quad (52)$$

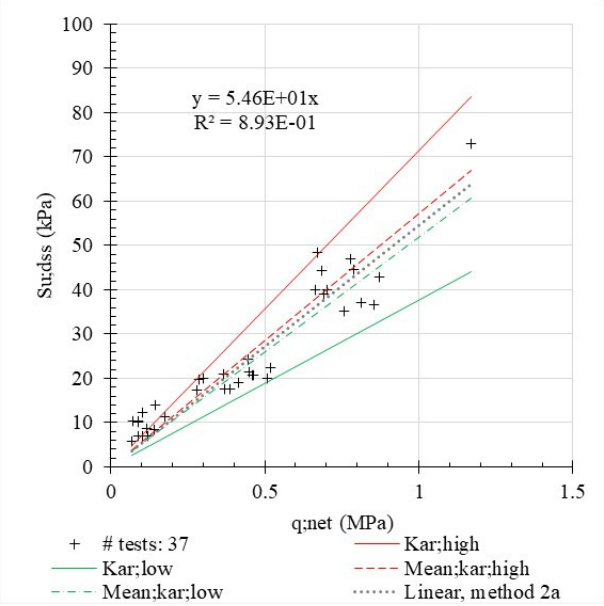


**Figure 7.** Method 2b, OLS through origin.

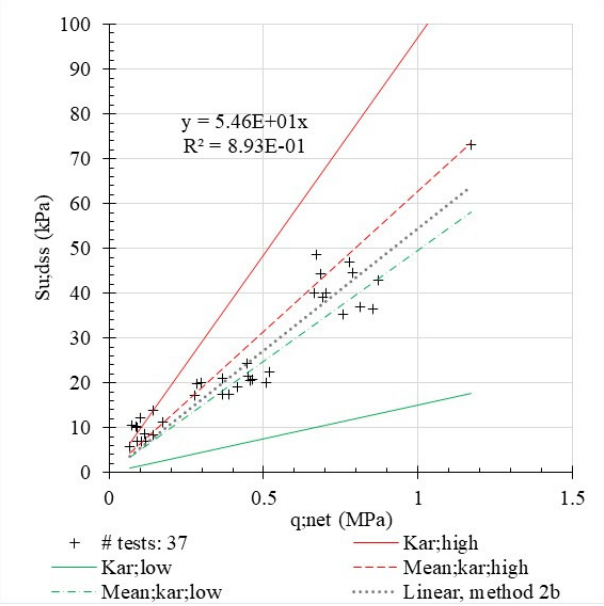


**Figure 8.** Method 2b, classical ratio estimation.

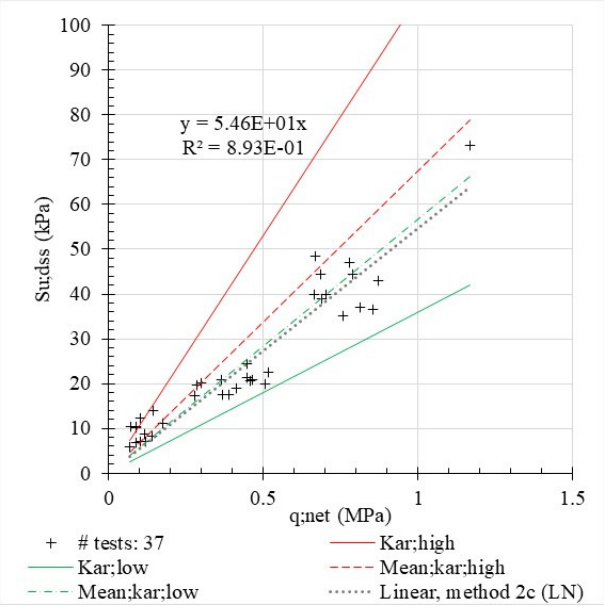

**Figure 9.** Method 2c, LN transformation and ratio estimation.

Three methods are presented for a linear function through the origin. Method 2a is based on OLS regression. Method 2b is a ratio estimator method and not a regression method. Method 2b is based on Lognormal transformation of both X and Y. This method is also a ratio estimator. Method 2a is less sensitive to outliers than method 2b and 2c. This is illustrated in Figure 7 to Figure 9. The relative higher undrained shear strength measured at low net cone resistance has a significant effect on the standard deviation and the average slope. The Lognormal transformation slightly reduces this effect as can be expected. Method 2b is advised in the particular case of heteroscedasticity, when the error increases proportionally with the independent variable X. Method 2c has been applied in [1, 2] to compare the result of a correlation with measured values. This should theoretically be a 1:1 plot with slope $b_m$ as bias factor. Method 2c can thus also be used to plot determined parameters from complex or multivariate correlations versus measured laboratory test results. It should be noted that this is not the same as pairwise correlating.

The advantage of method 2a is that is works for all cases. Another advantage is that the basic statistical parameters, mean ($b_m$) and standard error of the mean (SEM) ($s_{b,m}$), are automatically determined in Excel by the Linest function (with intercept through origin).

## 5.4. Method 3: Power function, Lognormal transformation and OLS regression

Method 3 is used in case correlation models are best described by a power function though the origin, $Y = a \cdot X^b$. The first step is a Lognormal transformation of both X and Y. After Lognormal transformation, the applied method is similar to method 1, based on a linear function with free intercept. Excel Linest function can be used to derive slope $b_m$, intercept $\ln(a_m)$ and standard deviation $s_{y'}$ (based on Lognormal values). The confidence limits are asymmetric on normal scale because of the Lognormal transformation. This is worked out in method 3a. The equations follow from statistical textbooks.

### Method 3a:

$$Y'_i = \ln(Y_i) = b_m \cdot \ln(X_i) + \ln(a_m) + \varepsilon \quad (53)$$

$$s_{y'} = \sqrt{\frac{\sum_{i=1}^{n}\left(\ln(Y_i) - (b_m \cdot \ln(X_i) + \ln(a_m))\right)^2}{n-2}} \quad (54)$$

$$Y_i = a_m \cdot e^{\varepsilon} \cdot {X_i}^{b_m} \quad (55)$$

The characteristic values of Y can be determined by the following equation. The degrees of freedom are equal to n-2 for method 3 and no leverage term is included.

$$Y_{kar;i} = a_m \cdot e^{k_{kar} \cdot s_{y'}} \cdot {X_i}^{b_m} \quad (56)$$

$$k_{kar} = t_{n-2}^{0.05}\sqrt{1+\frac{1}{n}};\ k_{m;kar} = t_{n-2}^{0.05}\sqrt{\frac{1}{n}} \quad (57)$$

### Method 3b:

Method 3b presents an approximation of symmetrical confidence limits. This has been achieved by setting the lower confidence limit equal at one standard deviation offset for both method 3a and 3b. Symmetrical confidence limits and can be approximated by:

$$Y_i = (a_m + \varepsilon){X_i}^{b_m} \quad (58)$$

$$s_a = (1 - e^{-s_{y'}})a_m \quad (59)$$

$$CV_a = \frac{s_a}{a_m} = (1 - e^{-s_{y'}}) \quad (60)$$

$$Y_{kar;i} = (a_m + k_{kar} \cdot s_a){X_i}^{b_m} \quad (61)$$

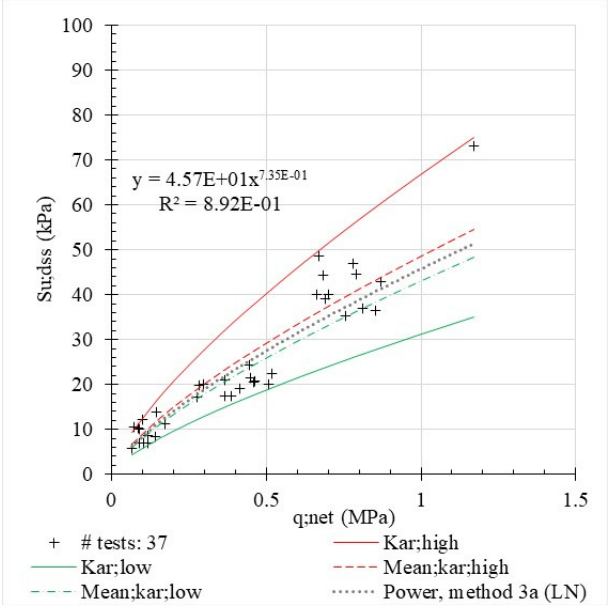


**Figure 10.** Method 3a, power function, OLS regression, LN transformation.

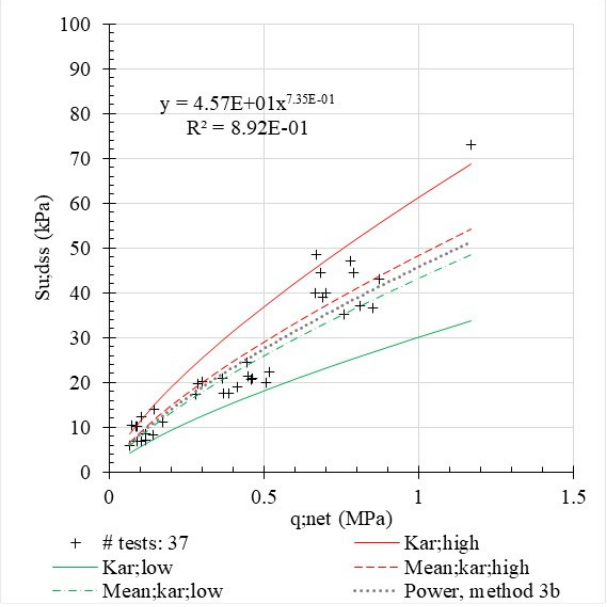


**Figure 11.** Method 3b, power function, OLS regression, LN transformation, symmetric confidence interval.

Method 3a provides symmetrical confidence limits on log-log-scale. On normal scale the confidence limits are asymmetric. In method 3b normality is approximated. A comparison of both methods is shown in Figure 10 and Figure 11. The mean trend is the same in both methods. Furthermore, it can be concluded the methods provides similar confidence limits, but method 3a is asymmetrical.

## 5.5. Synthesis

The approach with the three trendlines is successfully applied to one set of paired data as shown in the

examples. The preferred trendline can be selected based on the evaluation of the coefficient of determination and the coefficient of variation. In most cases there is a general accepted and preferred relation and type of trendline based on physical reasoning which will determine the preferred trendline. The advantage of the approach with the best of three trendlines is that the error is less and consequently the statistics can be kept straight forward using the Normal or Lognormal distribution.

In this example the coefficient of determination is almost identical for all methods. The confidence limits show comparable offset, except for method 2b and 2c. This is mainly related to the fact that statistics on the ratio are used, instead of regression. Applying method 2b and 2c in practice would be more unfavorable. These methods are better applied in another context (see [1, 2]). Method 3b is a good alternative for method 3a in case a standard deviation is requested based on a normal distribution.

For this particular example based on pairwise correlation of undrained shear strength versus net cone resistance method 2a and method 3a and 3b are all good options. Method 1a is not preferred as it yields to negative characteristic values at low net cone resistances. The preferred method would be method 2a, linear regression and OLS through origin as this corresponds also to the empirical correlation used in practice. In the next chapter it is described how these methods can be applied in practice.

## 6. Application in practice

The regression analysis provides the stochastic parameters, confidence limits and characteristic values for correlations. Those parameters are the basis for semi- and full-probabilistic calculations. The final question is how to apply this in design, using local CPTs and regional correlations. The selection of the overall characteristic parameter defined as the parameter with 95% probability of exceedance related to the limit state is not straight forward for point values as shown in section 4. This is even more true for pairwise established CPT-based correlations One should account for regional and local variations, failure extent and availability of local CPTs. The following six combinations as presented in Table 3 and Figure 12 are considered.

**Table 3.** Practical application for characteristic values geotechnical parameters from CPT-based correlations applied to line infra-structure, six cases

| | Local correlation | Regional correlation | Regional correlation |
|---|---|---|---|
| | Local CPT | Local CPT | Regional CPT |
| Global failure | case A (eq. 62) | case C (eq. 64) | case E (eq. 66) |
| Local failure | case B (eq. 63) | case D (eq. 65) | case F (eq. 67) |

A local correlation is locally derived and applied with local CPTs. A regional correlation is regionally derived and applied with local or regional CPTs. For case A and B it is assumed that a pairwise correlation is established based on local CPTs and laboratory tests. This is the most favorable situation. In both cases it is advised to use the characteristic mean value ($T_{m;kar}$) of the correlation. Practically this can be done taking the low characteristic confidence limit of the mean from regression, for example based on equation (34).

For case C, D, E and F a regional correlation is applied which can be biased due to regional variation. In that case a more cautious value is appropriate, and the characteristic value ($T_{kar}$) is advised. Practically this can be done taking the low characteristic prediction limit from regression, for example based on equation (33). The selection of the CPT value is presented per case below.

For case A, local CPT(s) are available and global failure is expected. For this situation the characteristic mean ($X_{m;kar}$) of the measured CPT parameter is advised. Practically this can be done by taking the average from regression analyses, since CPTs have measurements every 2 cm and the vertical scale of fluctuation is typically 0.2 m.

For case B, local CPT(s) are available and local failure is expected. For this situation the characteristic value ($X_{kar}$) of the measured CPT parameter is advised. Practically this can be done taking the lower characteristic prediction limit from regression analyses. Alternatively, the decisive local CPT (with lowest average) can be used.

For case C, the same approach as case A is recommended. The only difference is the transformation part.

For case D, the same approach as case B is recommended. The only difference is the transformation part.

For case E, no local CPT(s) are available and global failure is expected. This situation is comparable to that of case III in clause 4.3. Practically the same approach as in equation (24) is advised, taking into account the regional variation. For this case the regional characteristic value ($X_{m;kar;reg}$) of the measured CPT parameter is advised.

For case F, no local CPT(s) are available and local failure is expected. This situation is comparable to that of case IV in clause 4.3. Practically the same approach as in equation (25) is advised, taking into account the regional variation. For this case the regional characteristic value ($X_{kar;reg}$) of the measured CPT parameter is advised.

The equations for case A to F can be written as a transformation function (T) and independent variable (X). In section 5 it is shown how to derive the characteristic values for a transformation function. Combining this the following generic equations are advised for application in practice:

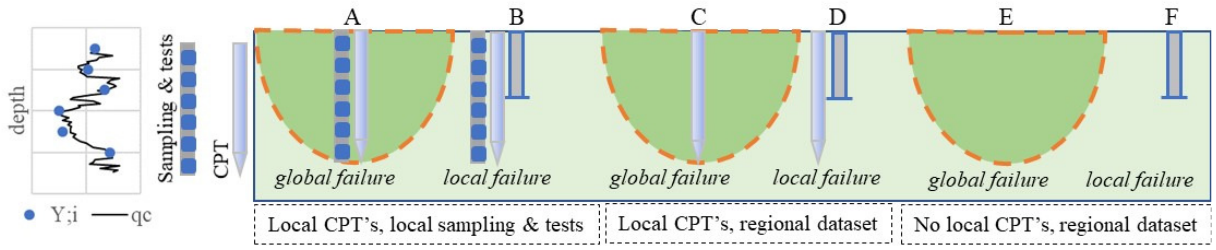


**Figure 12.** Illustration of 6 cases for line infrastructure projects.

$$Y_{T;A;m;kar} = T_{m;kar} \cdot X_{m;kar} \quad (62)$$
$$Y_{T;B;kar} = T_{m;kar} \cdot X_{kar} \quad (63)$$
$$Y_{T;C;m;kar} = T_{kar} \cdot X_{m;kar} \quad (64)$$
$$Y_{T;D;kar} = T_{kar} \cdot X_{kar} \quad (65)$$
$$Y_{T;E;m;kar} = T_{kar} \cdot X_{m;kar;reg} \quad (66)$$
$$Y_{T;F;kar} = T_{kar} \cdot X_{kar;reg} \quad (67)$$

## 7. Conclusions

This paper presents an overview of geo-statistics, both on populations and on trend functions. The aim of this paper is to discuss the application of pairwise established CPT-based correlation. In section 4 the basic concepts are worked out. It is shown that the determination of a characteristic value in case of global failure is subject to different interpretations and more complicated than might be expected from the codes.

For line infrastructure and regional databases, additional uncertainties exist due to regional variation. The model shown at the end of section 4 provides a pragmatic approach how to derive the characteristic value and the equivalent stochastic parameters.

In section 5 the basic regression methods and geo-statistics are presented for trend functions. The approach with the three trendlines is successfully applied in the examples. The preferred trendline can be selected based on the evaluation of the coefficient of determination and the coefficient of variation. The statistical parameters can be determined with Excel Linest function.

In section 6 a pragmatic approach is suggested to apply the geo-statistics of section 4 and 5 into practice. The approach takes into account whether the correlation is local or regional, and whether the failure is local or global. The uncertainties are applied to either the transformation parameter, the CPT measurements, or both.

## Acknowledgement

This work is part of the "Perspectief research programme All-Risk" with project number P15-21, which is (partly) financed by NWO Domain Applied and Engineering Sciences.

## References

[1] Phoon, K.-K., "Role of reliability calculations in geotechnical design," Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, vol. 11, no. 1, pp. 4-21, 2017/01/02 2016, doi: 10.1080/17499518.2016.1265653.
[2] Ching, J. and Phoon, K.-K., "Transformations and correlations among some clay parameters — the global database," Canadian Geotechnical Journal, vol. 51, no. 6, pp. 663-685, 2014/06/01 2014, doi: 10.1139/cgj-2013-0262.
[3] Phoon, K.-K. and Kulhawy, F. H., "Characterization of geotechnical variability," Canadian Geotechnical Journal, vol. 36, no. 4, a, pp. 612-624, 1999/11/22 1999, doi: 10.1139/t99-038.
[4] Eurocode 0: Basis of structural design, EN1990, 2002.
[5] Eurocode 7: Geotechnical design - part 1: General rules, EN1997-1, 2005.
[6] General principles on reliability for structures, ISO2394, 2015.
[7] Gulvanessian, H., Calgaro, J.-A., and Holický, M., Designers' Guide to EN 1990 Eurocode: Basis of Structural Design. Thomas Telford, 2002.
[8] Hicks, M. A., "An explanation of characteristic values of soil properties in Eurocode 7," Modern Geotechnical Design Codes of Practice: Implementation, Application and Development, vol. 1, p. 36, 2012.
[9] Hicks, M. A. and Nuttall, J. D., "Influence of soil heterogeneity on geotechnical performance and uncertainty: a stochastic view on EC7," in Proceedings 10th International Probabilistic Workshop, Universität Stuttgart, Stuttgart, 2012, pp. 215-227, doi: doi.org/10.1016/j.compgeo.2014.05.004.
[10] Orr, T. L. L., "Defining and selecting characteristic values of geotechnical parameters for designs to Eurocode 7," Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, vol. 11, no. 1, pp. 103-115, 2017/01/02 2016, doi: 10.1080/17499518.2016.1235711.
[11] Prästings, A., Spross, J., and Larsson, S., "Characteristic values of geotechnical parameters in Eurocode 7," Proceedings of the Institution of Civil Engineers - Geotechnical Engineering, vol. 172, no. 4, pp. 301-311, 2019, doi: 10.1680/jgeen.18.00057.
[12] Schneider, H. R., "Panel discussion: Definition and determination of characteristic soil properties," in Proceedings, 1997, pp. 2271-2274.
[13] Schneider, H. R. and Schneider, M. A., "Dealing with uncertainties in EC7 with emphasis on determination of characteristic soil properties," Modern Geotechnical Design Codes of Practice (Arnold P, Fenton GA, Hicks MA and Schweckendiek T (eds)). IOS Press, Rotterdam, the Netherlands, pp. 87-101, 2012.
[14] Tietje, O., Fitze, P., and Schneider, H. R., "Slope Stability Analysis Based on Autocorrelated Shear Strength Parameters," Geotechnical and Geological Engineering, vol. 32, no. 6, pp. 1477-1483, 2014/12/01 2013, doi: 10.1007/s10706-013-9693-8.
[15] Vanmarcke, E. H., "Probabilistic modeling of soil profiles," Journal of the geotechnical engineering division, vol. 103, no. 11, a, pp. 1227-1246, 1977.
[16] Vanmarcke, E. H., "Reliability of earth slopes," Journal of the geotechnical engineering division, vol. 103, no. 11, b, pp. 1247-1265, 1977.
[17] Calle, E. O. F., "Statistiek bij regionale proevenverzamelingen (in Dutch), Statistics of regional datasets," Geotechniek, januari, pp. 40-44, 2008.