Proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering– Rahman and Jaksa (Eds)

Variability of shear strength and compressibility properties of Glaciolacustrine sediments in Northern Germany

Variabilité des propriétés de résistance au cisaillement et de compressibilité des sédiments glaciolacustres en Allemagne du Nord

Julia Sorgatz, Martin Pohl & Jan Kayser
Geotechnical Engineering, Federal Waterways Engineering and Research Institute, Germany, julia.sorgatz@baw.de

ABSTRACT: A major challenge of most geotechnical engineering projects is soil data scarcity. This paper aims at extending prior knowledge on shear strength and compressibility of Glaciolacustrine sediments of Northern Germany. Based on triaxial, oedometer and complementary laboratory tests on specimens from 13 different locations, the inherent variability of shear strength and compressibility is analysed; typical ranges and coefficients of variation are established. Prior to variability analysis, k-means clustering, a simple machine learning algorithm, is applied to distinguish soil types by their descriptive properties. The procedure presented in this paper accounts for the multivariate character of soil and provides data on variability of strength and compressibility more accurately. It was found that plasticity index and clay content can be considered to distinguish different soil types. Moreover, it can be shown that mean and variability of shear strength and compressibility are clearly affected by the dominant soil type.

RÉSUMÉ: Un défi majeur de la plupart des projets d'ingénierie géotechnique est le manque de données pédologiques. Cet article vise à enrichir les connaissances antérieures de la résistance au cisaillement et de la compressibilité des sédiments glaciolacustres d'Allemagne du Nord. Nous avons analysé la variabilité inhérente de la résistance au cisaillement et de la compressibilité sur la base de tests triaxiaux et oedométriques ainsi que de tests supplémentaires en laboratoire géotechnique d'éprouvettes de sédiments glaciolacustres collectées sur 13 sites différents. Nous avons ainsi établi des marges et coefficients typiques de variation. Avant d'analyser la variabilité, on applique le k-means clustering, un algorithme simple d'apprentissage automatique, pour distinguer différents types de sol par leurs propriétés descriptives. La procédure présentée tient compte du caractère multivarié des données podologiques et fournit des données plus précises de la variabilité des propriétés de résistance au cisaillement et de compressibilité. Les analyses ont permis de constater que, par exemple, l'indice de plasticité et la teneur en argile peuvent être considérés pour distinguer les différents types de sols. En plus, on peut démontrer que la moyenne et la variabilité de la résistance au cisaillement et de la compressibilité sont nettement affectées par le type de sol dominant.

KEYWORDS: soil variability, glaciolacustrine sediments, coefficient of variation, k-means clustering, multivariate data analysis

1 INTRODUCTION

According to the European standard (EC7) characteristic values are fundamental input data for geotechnical designs using a semi-probabilistic design approach. They are "selected as a cautious estimate of the value affecting the occurrence of the limit state" (DIN EN 1997-1:2014-03).

In this context, a major challenge of geotechnical engineering is soil data scarcity or the "the curse of small sample size" (Phoon, 2017) which does not allow to select characteristic values based on sophisticated statistical analyses. Commonly, this issue is tackled by experience, engineering judgement and local data repositories. But even with these resources, as outlined by Bond (2011), engineers may not be well trained at predicting the appropriate degree of caution needed to select the characteristic value of a geotechnical parameter. Thus, although still not integral part of everyday engineering practice, the advantages of more advanced methods such as Bayesian inference in conjunction with prior knowledge are increasingly recognized to account for uncertainty inherent to soil parameters (e.g., Phoon and Kulhawy 1999a, 1999b, Wang et al. 2016, Phoon 2017, Wang 2017).

Despite their local uniqueness (Phoon 2019), point statistics, e. g. mean and standard deviation, of soils have been investigated by various authors, often for particular applications (e. g. Lumb 1966, 1974, Phoon und Kulhawy 1999a, 1999b, Uzielli et al. 2006, Löfman und Korkiala-Tanttu 2019) and summarised in standards (e. g. JCSS, 2006). Engineers may supplement their site-specific data by these values. However, as pointed out by Löfman and Korkiala-Tanttu (2019), typical ranges of soil parameters provided in literature can be improved by accounting for local characteristics such as the materials' genesis. In the case of North German Glaciolacustrine sediments, few information on typical values have been published (Ehlers et al. 2011, Kausch 2020), which, do not cover the materials' inherent variability.

The presented work focuses on the analysis of shear strength and compressibility of Glaciolacustrine sediments of Northern Germany. Based on triaxial, oedometer and complementary laboratory tests from 13 different locations the inherent variability of shear strength and compressibility is analysed; typical ranges and coefficients of variation are established. Prior to variability analysis, a simple machine learning algorithm is applied to distinguish different sediments on the basis of their classification properties.

The paper is organised as follows: Firstly, the employed data and methods are introduced. Secondly, to provide data on the soils' variability more accurately, the machine learning algorithm k-means clustering is applied to distinguish sediments based on their descriptive properties. Finally, after validation of the clusters, typical ranges and the variability of shear strength and compressibility are presented for each cluster. The paper closes with a brief recap of results as well as an outlook.

2 DATA AND METHODOLOGY

2.1 *Characteristics of Glaciolacustrine sediments in Northern Germany*

Besides Marine clay and boulder clay, Glaciolacustrine sediments are typical soils of Northern Germany. They were deposited in reservoirs or lakes which have come from glaciers

Table 1. Classification properties of the studied sites (assessment of inherent variability).

Site	Depth	No of specimens	Clay content	Organic content	Water content w_n	Plasticity index I_P
--	in m	--	in %	in %	in %	--
Brunsbüttel	37 - 40	3	8.0 - 37.0	2.6 -5.6	18.3 - 24.3	0.12 - 0.31
Zerben	6 - 13	2	4.0 - 52.0	5.0	26.2 - 35.9	0.16 - 0.48
Levensau	8 / 37 - 42	4	4.0 - 50.0	1.9 - 4.2	19.2 - 20.3	0.13 - 0.32
Kiel-Holtenau	15 - 19	2	25.0 - 70.0	2.2 - 3.8	18.6 - 21.3	0.31 - 0.32
Kiel - Friedrichsort	17 / 32 - 33	3	14.0 - 26.0	2.4 - 2.5	32.2	0.13
Hunte	5 - 14	2	64.0 - 76.0	7.3 - 7.7	32.2 - 34.3	0.43 - 0.51
Steinhavel	4 - 10 / 21	3	13.0 - 38.0	1.3 - 7.9	24.5 - 29.7	0.16 - 0.53
Ahse	6	2	31.0-36.0	4.0 - 5.1	21.9 - 28.0	0.22 - 0.31
Lauenburg	5 - 45	42	5.0 - 65.0	1.0 - 10.4	27.5 - 34.8	0.06 - 58.2
Niederfinow	2 - 25	17	3.0 - 27.0	2.1 – 5.3	18.5 - 27.5	0.04 - 0.33
Witzeeze	10 - 30	5	11.0 - 36.0	4.2 - 6.2	19.6 - 24.7	0.09 - 0.22
Ems	3 - 15	15	5.0 - 21.0	2.4 - 17.5	21.4 - 32.0	0.06 - 0.31
Niederfinow	10 - 15	2	9.0 - 40.0	--	21.6 - 24.7	0.66

and which were either formed by glacier erosion or deposition or located at the margin of the ice sheet. In the area of Northern Germany three ice age periods are reliably identified: Elster ice age, Saale ice age, Weichsel ice age; the oldest glaciation, the Elsterian, reached furthest south. Only in the west, the ice of the second glaciation, the Saalian, advanced beyond the Elsterian limits. During the last glaciation, the Weichselian, the ice sheet did not cross the Elbe River (Ehlers et al. 2011). Accordingly, different stratigraphic units should be considered during soil testing and for the definition of characteristic values.

In general, Glaciolacustrine sediments are cohesive soils. In terms of grain size analysis, they are classified as weakly sandy to sandy clays or silts according to DIN EN ISO 14688-1:2020-11. According to DIN 18196:2011-05, they are soils of the classes CL, OL or ML. With a liquid limit w_L ranging from 20 % to 90 % (on average 46 %), the plasticity of either clay and silt also ranges from low to high. Due to local lignite streaks and lenses, Glaciolacustrine sediments are weak to moderate organic; annealing losses between 1 % and 18 % (on average 4 %) are determined.

2.2 *Shear strength and compressibility properties*

This section briefly introduces the most important tests and parameters used for subsequent analyses. For a detailed description of the employed laboratory tests, it is referred to the respective standards (DIN EN ISO 17892-5:2017-08, DIN EN ISO 17892-9:2018-07). A summary of the employed test data is provided in Table 1. For few locations limited data are available emphasising the importance of prior knowledge for future engineering projects.

The machine learning algorithm is first applied to selected descriptive soil parameters. The liquid limit w_L is determined via the fall cone test. The plastic limit w_P is determined by repeated rolling of an ellipsoidal-sized soil mass. Both procedures and the calculation of the plasticity index $I_P = w_L - w_P$ are defined in DIN EN ISO 17892-12:2018-10. The clay content is obtained from sieve and sedimentation tests (DIN EN ISO 17892-4:2016-04) and the organic content results from tests with the loss-on-ignition method (DIN 18128:2002-12).

By means of oedometer tests, the axial compressibility and deformation of the soil is investigated. A cylindrical sample is deformed uniaxially. A metal ring prevents the specimen from deviating sideways. In the case of the presented investigations, a specimen was commonly tested against eight load levels which doubled after each step and ranged from 17.1 kN/m² to 1021.1 kN/m². Subsequently, the load was relieved and, then, the specimen reloaded up to a maximum of 2040.4 kN/m. In total, results of 90 oedometer tests are available for analyses. From the first load cycle, the stress-dependent oedometer module E_{oed} is obtained as the ratio of change in stress and change in vertical deformation. In the same way, the stress-dependent oedometer module $E_{s,r}$ is obtained from the reloading cycle.

Triaxial tests allow to investigate shear strength, stress-strain relationship and effective stress paths of a soil or rock specimen under compression. The presented data encompasses isotropic consolidated drained triaxial compression (CIDC) and isotropic consolidated undrained triaxial compression (CIUC) tests. In total 88 triaxial tests which three sub-specimens each were conducted; 33 tests of CIUC and 55 tests of CIDC. Both tests procedures allow to determine the effective shear parameters, effective cohesion c' and effective friction angle φ', which are of particular relevance for engineering practice.

2.3 *K-means clustering*

During field classification and laboratory tests, different sediment or soil types may not always be clearly distinguishable or it may not be clear which type dominates the specimen. However, this distinction may be important in order to provide reliable information on soil characteristics and variability. Machine learning tools can assist in classifying sediments based on a number of objective criteria.

K-means clustering is one of the simplest and well-known unsupervised machine learning algorithms. It belongs to a family of algorithms which were developed independently by researchers from different disciplines (MacQueen 1967, Steinhaus 1956, Lloyd 1982). Main advantages are its simplicity as well as its scalability to different sample sizes. K-means clustering partitions *n* data points in *k* clusters based on their distance to the nearest mean, the cluster centroid. The algorithm minimizes the within-cluster variances which is commonly expressed via squared Euclidean distances.

In simplified terms, the algorithm has three main steps: Firstly, the number of cluster centroids must be provided by the user. The learning process then starts with a group of randomly selected centroids. Subsequently, the algorithm changes the positions of the centroids iteratively until either the difference between old and new centroids reaches a threshold or the defined number of iterations has been reached. The new centroids are computed as the mean value of all of the samples assigned to each previous centroid (Pedregosa et al. 2011).

The presented k-means clustering analyses use Python with the machine learning tools provided by the package scikit-learn (Pedregosa et al. 2011). To account for the different scales of the soil characteristics, the data was normalised before running the analyses. For normalisation the L^2 vector normalisation scheme was employed, which is based on the distance of the vector coordinate from the origin of the vector space.

3 PRIOR KNOWLEDGE ON COMPRESSIBILITY AND SHEAR STRENGTH

3.1 *Results of k-means clustering analyses*

Figure 2 visualises the results of the k-means clustering analyses. Based on normalised w_L and normalised I_P two clusters are identified. When using normalised clay content and normalised I_P three clusters are found. In the case of the two-cluster solution one cluster features material of moderate w_L and moderate I_P, whereas the other group features material of high w_L and high I_P. The three-clusters result features clusters of low, moderate and high clay content and I_P. Besides the above discussed clusters, the images show strong correlations between the w_L, I_P and clay content and I_P, which corresponds to results in literature.

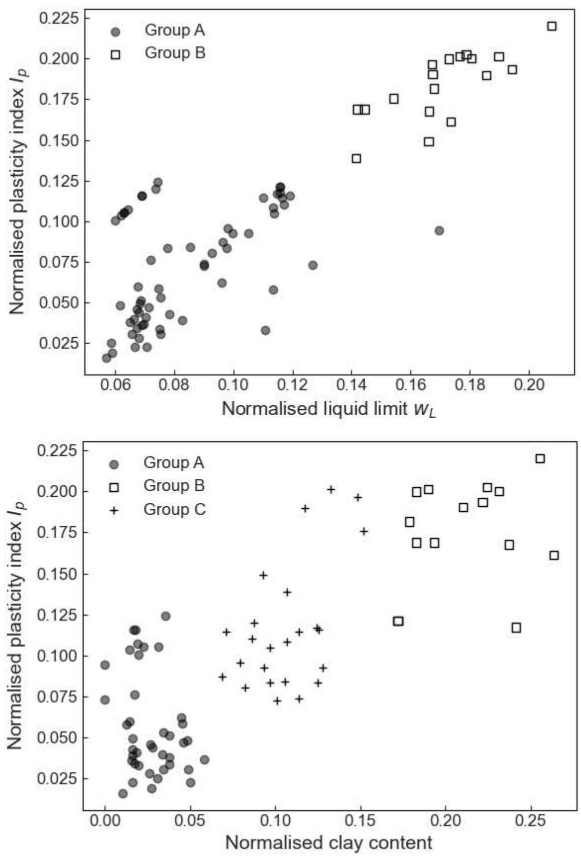

Figure 2. Results of K-means clustering analyses. Descriptive properties of the Glaciolacustrine sediments are normalised. Subsequently, clusters are identified by their distance to two or three circular centroids.

The clusters are validated using further characteristics of the data. It can be shown that the two-cluster solution corresponds well with the stratigraphic units of the Glaciolacustrine sediments (Figure 3). When plotting the data into Casagrande's plasticity chart it can be observed that sediments that were deposited during Weichsel and Saale glaciation are classified as Group A; sediments deposited at the end of the Elster glaciation are classified as Group B. Weichselian and Saalian sediments cannot be clearly distinguished from each other in Casagrande's plasticity chart. Additional parameter investigations did not provide a satisfactory differentiation either. Reasons for this may be, among other things, the material composition or perhaps uncertain data.

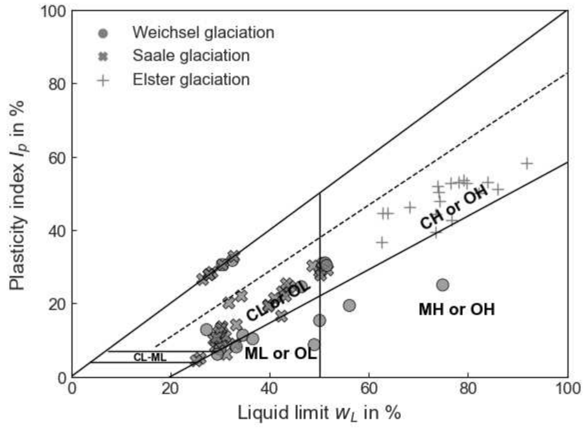

Figure 3. Two-cluster solution plotted against Casagrande's plasticity chart. The clusters seem to correspond with the stratigraphic layers.

The presented k-means clustering analyses are showing promising results. Yet, it must be noted that at present available data of Elsterian sediments primarily contains specimens of the so-called Lauenburg clay, which is a rather distinct, often in literature mentioned clay with varying amounts of silt. In the case of the Saale and Weichsel sediments, the stratigraphic units cannot always be clearly determined from the available boring and geotechnical reports. For single specimens, no information on stratigraphy is available. Bedload analyses were only carried out in few cases, so that the existing classification is usually based on the judgement of senior engineers which are familiar with the regional geology. Still, this may result in erroneous categorisations. Additionally, it should be noted that k-means clustering is a rather simple algorithm. It assumes that data points belonging to a cluster are evenly distributed and located circular around the centroid. In fact, clusters may also have different shapes like an ellipse. The presented methodology may therefore benefit from the application of more advanced algorithms such as Mean-shift clustering or Gaussian Mixture Models.

In summary, further investigations are certainly required to confirm the determined clusters. The result may be improved by adding further data. The differentiation into three clusters cannot be justified on the basis of the existing soil characteristics and metadata. This does not necessarily mean that this classification is not accurate, but, at present, the defining properties may not be traceable within the existing data. Additionally, measurement uncertainties may alter the results. Despite these drawbacks, subsequently, the variability of Glaciolacustrine sediments is analysed within the two identified clusters. Table 2 summarises the point statistics of few classification properties for each group.

3.2 *Typical ranges of mean and variability (COV) of compressibility properties*

Due to the geological history of origin, it can be assumed that the material was covered by a thick Pleistocene ice cover. With altitudes of 500 m to 1000 m the material is geologically preloaded and, thus, over-consolidated. Typical ranges of the reloading module $E_{s,r}$ are explored in Figure 4 and Figure 5 as a

Table 2. Summary of count and point statistics, mean, standard deviation (std) and coefficient of variance (COV), of classification properties for the two detected clusters.

	Clay content	Organic content	w_L	w_P	I_P
	in %	in %	in %	in %	in %
Group A (Weichselian / Saalian sediments)					
count	61	43	61	53	53
mean	16.44	3.84	37.38	21.18	18.98
std	14.21	1.82	10.23	7.02	9.01
COV	0.86	47.24	27.36	33.16	47.49
Group B (Elsterian sediments)					
count	18	8	18	18	18
mean	53.47	6.52	75.56	27.02	48.54
std	14.72	2.65	7.90	5.25	5.61
COV	0.28	40.64	10.45	19.43	11.55

function of the mean soil stress σ_m. The quality of the tests was ensured by a review of laboratory reports. The water content of the test specimens at the start of the test ranged between 15 % and 40 %; the porosity between 30 % and 75 %.

A linear least-squares regression was conducted to determine the stress-dependent $E_{s,r}$. Equations are provided in the top right corner of the figures. The R^2 value is used as a measure of how well observed outcomes are predicted by the model. In the presented studies, a moderate fit of equations to data is observed. Particularly Group B is characterised by uncertainty in slope and constant. Yet, due to physical considerations a negative slope, as it could be derived from the confidence intervals, must be rejected.

On average, a large variability of $E_{s,r}$ is observed. Among other things, this may be due to inhomogeneous soil specimens, slightly modified test procedures or different test operators. However, the figures illustrate that the differentiation of stratigraphic units by means of cluster analyses results in more precise soil characteristics. In particular for the Elsterian sediments, the compressibility is much lower than for Weichselian / Saalian sediments. Without differentiation the characteristic compressibility of Elsterian sediments and its variability would likely be overestimated. In the case of Weichselian and Saalian sediments, on the other hand, the increase of $E_{s,r}$ is likely to be underestimated.

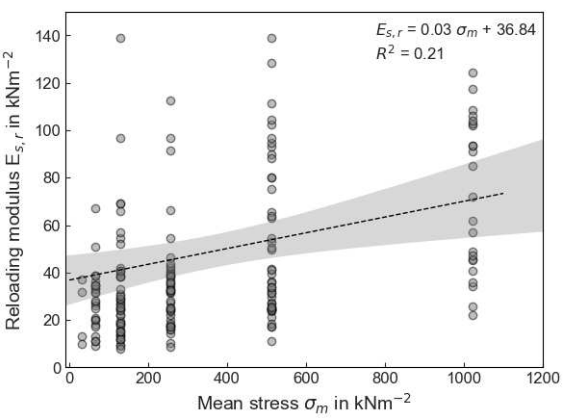

Figure 4. Stress-dependent modulus $E_{s,r}$ (reloading) over mean soil stress σ_m for the complete data set. The grey shaded area visualises the 95 % confidence interval for the linear regression.

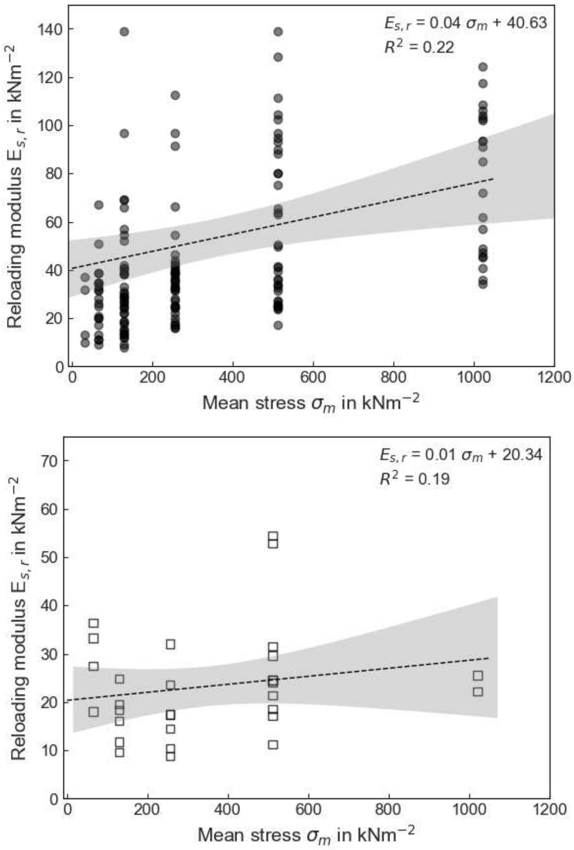

Figure 5. Stress-dependent reloading modulus $E_{s,r}$ over mean soil stress σ_m for three data combinations. The top image shows the analysis with data points that were assigned to Group A; the bottom image shows Group B data points. The grey shaded area visualises the 95 % confidence interval of the linear regression.

In order to determine the parameter range and COV of the regression parameters, subsequently, a bootstrapping approach was applied. Bootstrapping is a common non-parametric method for the assessment of errors in a statistical estimation problem. In simple terms, bootstrapping assigns measures of accuracy (bias, variance, confidence intervals, prediction error, etc.) to sample estimates in order to obtain a sample that complies with predefined statistics (Efron, 1982, 1993). In the case of the presented study, the $E_{s,r}$ values were re-sampled 50 times. In each sampling loop, ten samples were drawn with replacement for each mean stress level and, subsequently, a linear regression was conducted. The number of samples is supposed to ensure that the slope is maintained. However, this may not always be true, particularly in the case of Group B. Additionally, since the presented equations describe the stress-dependent mean $E_{s,r}$, the equations may be shifted along the y-axis till they meet the desired safety level required for the definition of characteristic values. This could be, for instance, the 95 % prediction band. Prediction bands contain, with a pre-specified probability, a new observation from the same population from which the sample is drawn. The respective 95 % lower limit (LL) and upper limit (UL) band and a summary of regression statistics are given in Table 3.

It can be observed that the variability expressed via the coefficient of variance (COV) of the slope *a* is higher in Group B (Elsterian sediments) than in Group A (Weichselian / Saalian sediments), whereas the variability of the constant *b* is higher in Group A. Larger R^2 values indicate that, on average, the linear regression in Group A reaches a better fit than in Group B. Special attention should be paid to the large standard deviation (std) of Group A (std = 8.52 kN/m²) which must be considered when trying to define a conservative $E_{s,r}$.. The variability of *a* and *b* for the complete data set is comparable to that of Group A. Naturally, the provided 95 % limits of the complete data are wider than in Group A or Group B. In addition, due to the large variability of test results, the 95 % prediction bands represent a rather conservative estimate of $E_{s,r}$.

Table 3. Summary of point statistics, mean, standard deviation (std), coefficient of variance (COV), 95 % lower limit (LL) and upper limit (UL), derived from bootstrapping analyses for linear regression of the form $E_{s,r}\ [kN] = a \cdot \sigma_m + b$.

Parameter	*a*	*b*	R^2
Complete data set			
mean	0.030	36.633	0.493
std	0.006	8.829	0.188
COV	0.186	0.241	0.382
95 % band (LL/UL)	0.030/0.036	-59.380 / 133.066	--
Group A (Weichselian / Saalian sediments)			
mean	0.029	40.185	0.446
std	0.005	8.522	0.171
COV	0.180	0.212	0.383
95 % band (LL/UL)	0.032 / 0.039	-61.004 / 142.260	--
Group B (Elsterian sediments)			
mean	0.003	21.773	0.105
std	0.002	1.100	0.058
COV	0.574	0.051	0.550
995 % band (LL/UL)	0.006 / 0.010	-2.101 / 43.790	--

3.3 *Typical ranges of mean and variability (COV) of shear strength*

Based on failure points obtained from triaxial tests, effective friction angle ϕ' and effective cohesion c' are determined with Eq. (1) and Eq. (2) via the p'-q' - diagram where $p' = (\sigma_1{}' + \sigma_3{}')/2$ and $q' = (\sigma_1{}' - \sigma_3{}')/2$. The quality of triaxial test results was ensured by a review of laboratory reports. The water content of the test specimens at the start of the test ranged between 18 % and 50 %; the porosity between 33 % and 50 %.

$$\sin\phi' = \tan\alpha' \qquad (1)$$

$$c' = k/\cos\phi' \qquad (2)$$

As shown in Figure 6 (top) two different strength envelops are determined for Group A and Group B. Figure 6 (bottom) shows the linear regression with the complete data set. In contrast to the analysis of $E_{s,r}$, the regression for the shear parameters can only be obtained by the analysis of a dependent parameter pair. Failure during triaxial tests always results from a specific constellation of $\sigma_1{}'$ and $\sigma_3{}'$ on a continuous scale. Therefore, the above used bootstrapping approach is not applicable to the evaluation of shear parameters. Table 4, thus, only includes the regression parameters, the corresponding shear parameters and the 95 % limits for the two groups and the complete data set.

Table 4 immediately reveals a significant shortcoming: the linear regression results, on the one hand side, in negative and, on the other hand side, rather high values of c'. Negative c' values are physically wrong and should not be considered in a design. The discussed shortcomings may result from minor deficiencies during the test procedure, which are difficult to retrace retrospectively. Another reason for inaccurate estimates of ϕ' and c' is the regression model itself. Experimental data have shown that the strength envelops for soils are nonlinear. Nevertheless, the linear Mohr–Coulomb strength parameters are widely applied in engineering practice. Yet, at the same time it is pointed out that both regressions achieve high R^2-values. A finer clustering may improve the estimates of ϕ' and c'.

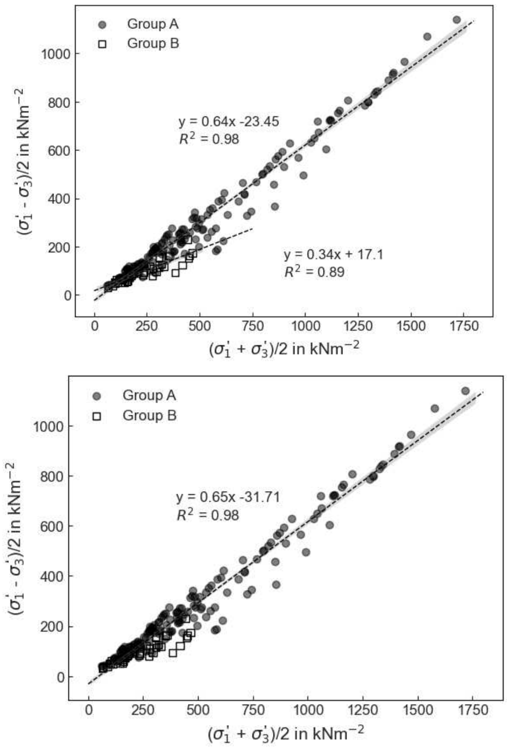

Figure 6. Extended shear diagram of the triaxial tests. The top image shows the analysis with two groups, the bottom image uses the complete data set. The grey shaded area visualises the 95 % confidence interval for the linear regression.

Table 4. Summary of point statistics, 95 % lower limit (LL) and upper limit (UL) derived from linear regression of the form $y = a \cdot x + b$

Parameter	*a*	ϕ'	*b*	*c'*	R^2
Complete data set					
Regression	0.65	40.54	-31.71	(-41.73)	0.98
95 % band (LL/UL)	0.65 / 0.65	40.17 / 40.47	-120.16 / 56.74	(-157.23) / 74.58	--
Group A (Weichselian / Saalian sediments)					
Regression	0.64	39.80	-23.45	(-30.51)	0.98
95 % band (LL/UL)	0.64 / 0.65	39.87 / 40.17	(-111.45) / 64.55	(-145.20) / 84.47	--
Group B (Elsterian sediments)					
Regression	0.34	19.90	17.1	18.18	0.89
95 % band (LL/UL)	0.34 / 0.35	19.82 / 20.18	(-23.78) / 57.98	(-25.27) / 61.77	--

A closer review of the results in Table 4 indicates that the variability of ϕ' and c' expressed via the 95 % limits does not differ significantly between Group A (Weichselian / Saalian sediments) and Group B (Elsterian sediments). A larger variability can be observed for c' due to the drawbacks discussed in the previous paragraph. The highest variability of c' can be observed for the complete data set, whereas ϕ' is affected moderately. Compared to $E_{s,r}$, the variability of the test results within the groups is small. Thus, the confidence intervals of both regressions are narrow. Moreover, it can be observed that Group A dominates the regression. This may be due to the fact that for group B only investigations at lower stress levels are available.

Additional information on the variability of ϕ' and c' are derived from the analysis of each test series comprising of three specimens (see Table 5). Again, different material properties are observed within the two groups. The differentiation into clusters reduces the group-inherent variability. However, in contrast to the analysis in Table 4, c' only takes positive values.

In summary, based on the currently available data it can be assumed that if not differentiated between the stratigraphic layers, the shear strength of Group B (Elsterian sediments) is overestimated; c', on the other hand, is overestimated for Group A (Weichselian / Saalian sediments).

Table 5. Summary of point statistics from analyses of each test series

	Complete data set		*Group A (Weichselian / Saalian sediments)*		*Group B (Elsterian sediments)*	
	ϕ'	*c'*	*ϕ'*	*c'*	*ϕ'*	*c'*
	in °	in kN/m²	in °	in kN/m²	in °	in kN/m²
mean	31.01	16.83	33.22	18.29	23.18	11.67
std	7.85	20.5	6.77	21.79	6.36	14.54
COV	0.25	1.22	0.20	1.19	0.27	1.25

4 CONCLUSIONS AND OUTLOOK

This paper presents a data-driven methodology for a differentiation of soil types that serves the multivariate character of soil data. With the aid of the machine learning algorithm k-means clustering two soil types with different material properties are identified. Subsequently, point statistics of material properties such as compressibility and shear strength are determined for each soil type separately.

The results of the statistical analyses show that a differentiation into soil types reduces the variability within a cluster and, thus, allows for a more precise estimate of characteristic values. Moreover, it can be shown that mean and COV are clearly affected by the dominant soil type.

Further investigations should validate the clusters that have been identified. This can be done, for example, by applying more advanced cluster algorithms. Besides that, additional data or the consideration of different soil characteristics may assist in specifying clusters more accurately. Based on these supplementary studies, the determined mean and variability of the investigated soil types should be reviewed.

On a broader basis, it is recommended to support generic databases which store data in a structured, machine readable manner. Only in this way, geotechnical engineering will benefit from recent and future developments in the field of data science.

5 REFERENCES

Bond, A. 2011. A procedure for determining the characteristic value of a geotechnical parameter. Vogt, N., Schuppener, B., Straub, D. and Bräu, G. (Ed.): *Geotechnical Safety and Risk. ISGSR 2011*. Karlsruhe, Bundesanstalt für Wasserbau, 419–426.

DIN 18128:2002-12. Baugrund - Untersuchung von Bodenproben - Bestimmung des Glühverlustes. Beuth, Berlin.

DIN 18196:2011-05. Erd- und Grundbau - Bodenklassifikation für bautechnische Zwecke. Beuth, Berlin.

DIN EN 1997-1:2014-03: Eurocode 7 - Entwurf, Berechnung und Bemessung in der Geotechnik - Teil 1: Allgemeine Regeln. Beuth, Berlin.

DIN EN ISO 14688-1:2020-11. Geotechnische Erkundung und Untersuchung — Benennung, Beschreibung und Klassifizierung von Boden — Teil 1: Benennung und Beschreibung. Beuth, Berlin.

DIN EN ISO 17892-12:2018. Geotechnische Erkundung und Untersuchung –Laborversuche an Bodenproben –Teil 12: Bestimmung der Fließ- und Ausrollgrenzen. Beuth, Berlin.

DIN EN ISO 17892-4:2016. Geotechnische Erkundung und Untersuchung - Laborversuche an Bodenproben - Teil 4: Bestimmung der Korngrößenverteilung. Beuth, Berlin.

DIN EN ISO 17892-5:2017. Geotechnische Erkundung und Untersuchung – Laborversuche an Bodenproben – Teil 5: Ödometerversuch mit stufenweiser Belastung. Beuth, Berlin.

DIN EN ISO 17892-9:2018. Geotechnische Erkundung und Untersuchung – Laborversuche an Bodenproben – Teil 9: Konsolidierte triaxiale Kompressionsversuche an wassergesättigten Böden. Beuth, Berlin.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R. 1993. *An introduction to the bootstrap.* Monographs on statistics and applied probability, 57, Chapman & Hall, New York.

Ehlers, J., Grube, A., Stephan, H.-J., Wansa, S. 2011. Pleistocene Glaciations of North Germany—New Results. *Quaternary Glaciations - Extent and Chronology - A Closer Look*, 15, 149–162.

Joint Committee on Structural Safety (JCSS). 2006. Probabilistic Model Code. Section 3.7: Soil Properties.

Kausch, F. 2020. Geotechnische Charakterisierung des Hamburger Baugrundes. https://www.hamburg.de/contentblob/1298120/152cdc0f47334bf187a6d836976b5fa1/data/d-geotechnik-text.pdf, 23.10.2020.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* 28 (2), 129–137.

Löfman, M., Korkiala-Tanttu, L. 2019. Variability and Typical Value Distributions of Compressibility Properties of Fine-Grained Sediments in Finland. Ching, J., Li, D.-Q. and Zhang, J. (Ed.): *Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR 2019)*, Taipei, Taiwan, 11-13 December 2019, Research Publishing (S) Pte. Ltd., 182–187.

Lumb, P. 1966. The Variability of Natural Soils. *Can. Geotech. J.,* 3 (2), 74–97.

Lumb, P. 1974. Application of Statistics in Soil Mechanics. Lee, I. K. (Ed.): *Soil Mechanics-New Horizons*, London, 44–111.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Le Cam, L. M. and Neyman, J. (Ed.): *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkley, University of California Press (Vol 1), 281-297.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. 2011. Scikit-learn. Machine Learning in Python.

Phoon, K.-K. 2017. Role of reliability calculations in geotechnical design. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards,* 11(1), 4-21.

Phoon, K.-K. 2019. Managing Risk in Geotechnical Engineering – from Data to Digitalization. Ching, J., Li, D.-Q. and Zhang, J. (Ed.): *Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR 2019)*, Taipei, Taiwan, 11-13 December 2019, Research Publishing (S) Pte. Ltd., 13–34.

Phoon, K.-K. and Kulhawy, F. H. 1999a. Characterization of geotechnical variability. *Can. Geotech. J.*, 36 (4), 612–624.

Phoon, K.-K. and Kulhawy, F. H. 1999b. Evaluation of geotechnical property variability. *Can. Geotech. J.,* 36 (4), 625–639.

Steinhaus, H. 1956. Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences*. Classe 3 (12), 801–804.

Uzielli, M., Lacasse, S., Phoon, K.-K., Nadim, F. 2006. Soil variability analysis for geotechnical practice. Phoon, K.-K., Hight, D. Leroueil, S. and Tan, T. (Ed.): *Characterisation and Engineering Properties of Natural Soils*, Taylor & Francis.

Wang, Y. 2017. *Selection of characteristic values for rock and soil properties using Bayesian statistics and prior knowledge*. Joint TC205/TC304 Working Group on "Discussion of statistical/reliability methods for Eurocodes" - Final Report (Sep 2017). http://140.112.12.21/issmge/TC205_304_reports/Characteristic%20values%202.pdf, 03.11.2020.

Wang, Y., Cao, Z. and Li, D. 2016. Bayesian perspective on geotechnical variability and site characterization. *Engineering Geology,* 203, 117-125.