EFFICIENCY OF SELECTED STATISTICAL CRITERIA IN DETERMINATION OF GEOTECHNICAL PARAMETERS FROM CPTU

ZBIGNIEW MŁYNAREK, JĘDRZEJ WIERZBICKI

Department of Geotechnics, A. Cieszkowski Agricultural University, Poznań, Poland

WALDEMAR WOŁYŃSKI

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Abstract: Determination of homogeneous geotechnical layers is one of the most important tasks in geotechnical designing. To achieve this a number of different statistical methods can be used. In the paper, the results of using a few of these methods have been presented and discussed. The differences between final geotechnical models of subsoil have been highlighted. As the result of analysis carried out, a two-step procedure was proposed for obtaining geotechnical layers, using statistical criteria.

1. INTRODUCTION

A comprehensive assessment of strength and deformation parameters of soils found in construction subsoil is obtained from in-situ and laboratory tests. The comprehensiveness of the assessment is connected with the fact that laboratory testing yields a point score evaluation of geotechnical parameters, while in-situ testing, e.g., using static testing such as CPTU, makes it possible to present the continuous picture of changes in parameters describing shear strength and changes in constrained moduli. On the other hand, in order to ensure rational dimensioning of the foundation of the object to be constructed on the subsoil analyzed, it is necessary to cluster results obtained from in-situ and laboratory tests. The objective of clustering is to isolate geotechnically homogeneous layers in the subsoil. Two crucial issues arise. The first concerns the selection of a clustering method and the recognition of this method as the most effective. In this type of assessment, it ought to be verified whether the clustering obtained differs from the other methods in terms of statistical criteria. The other problem pertains to the consistency of clustering in terms of strength and deformation criteria with the third criteria, concerning the genesis and stratigraphy of soils. It is an interesting fact that frequently it is the third criterion which is the starting point in clustering, and criteria crucial for foundation dimensioning, such as the above mentioned strength and deformability of subsoil, are secondary. In such an approach, strength and deformation parameters are ascribed to the previously isolated soil layers, diversified in terms of their lithology and genesis. The subject of this paper is the analysis of the role of the above mentioned criteria and efficiency of clustering.

Z. MŁYNAREK et al.

2. CPTU AND CLUSTERING OF DATA

The static probing method is especially suitable for the identification of subsoil stratigraphy and the assessment of strength and deformation parameters of soils found in the subsoil [3], [7]. Many classification systems have been developed to assess subsoil stratigraphy [3], which nevertheless need to be adapted to the local geological situation. When adapting a given system, different processes of sedimentation in a given area, e.g., Poland, and related overconsolidation and macrostructure effects need to be considered [4], [9]. An essential element in the appropriate application of each classification system is the establishment of the so-called representative CPTU parameters, obtained on the basis of penetration characteristics [7]. First studies on the subject (e.g., MŁYNAREK, LUNNE [8]) were published in the 1980's and highlighted the possibility of using statistical methods for that purpose. In these methods, clustering is based on standard CPTU parameters – corrected cone resistance q_t , or normalized cone resistance q_n , the coefficient of friction R_f (or normalized frictional resistance FR), the coefficient of excess pore pressure B_q and normalized cone resistance q_t .

Out of the above mentioned parameters those used to isolate subsoil layers in the lithological sense include q_t or Q_t as well as R_f or FR and B_q , while primarily the parameters q_n and Q_t are used to assess parameters of shear strength and constrained moduli. In this paper, the efficiency of the clustering methods applied and the analysis of significance of differences between representative parameters obtained from individual methods are presented on the basis of subsoil composed of soils from several geological formations.

3. STATISTICAL METHODS USED IN THE ANALYSIS

For the purpose of grouping data, many of more or less sophisticated cluster algorithms can be used. In our analysis, we choose three of them: two hierarchical and one combinatorical algorithms.

The basic component of each cluster algorithm is a similarity or dissimilarity measure between observations. In fact, a proper choice of this, to a large degree, determines the final result. Usually the dissimilarity measures are some type of distances, for example, classical Euclidean distance, or rarely used the cosine distance. His type of distance uses as a dissimilarity measure the angle between two vectors.

The basic idea of hierarchical cluster methods is to collect observations into clusters by combining the closest observations or the closest clusters to a larger ones. The hierarchical algorithms usually are simple and fast, but give medium-quality results. Undoubtedly, the most popular of them is agglomerative algorithm. In the first step of this algorithm, each observation forms its own cluster. In the next step, we find the closest pair of clusters and merge them into a single cluster, so that the number of clusters reduces by one. So, apart from dissimilarity measure between observations, we must have some kind of distance between clusters. In our computation, we used a typical average method. Note that in the first few steps, clusters may consist of one observation only, but after a few steps the clusters become bigger and bigger and consist of many observations. We merge clusters until a predefined number of clusters is achieved.

The basic idea of the second hierarchical algorithm used by us is to join observations to form the shortest dendrite. The shortest dendrite is a graph in which the sum of lengths of edges linking all the nodes is minimal. In this graph, the nodes represent the observations, and the lengths of edges linking them are equal to the distance between them. There is a very fast algorithm to construct the shortest dendrite. Note that in this algorithm we use only the distances between observations, not between clusters.

The third method is not hierarchical one. Its main idea is to minimize the observations within clusters for the fixed number of clusters. Unfortunately such a kind of optimization algorithm does not exist. All known algorithms reach the local minima only. So typically we run an algorithm several times and choose the best results, that is division with minimal observations within clusters. In our computation, we repeat it 100 times. If we use the Euclidean distance between observations as a dissimilarity measure, the most popular of such an algorithm is the so-called *k*-means algorithm. In the first step, we assign each observation randomly to one of the fixed number of clusters. In the next step, we compute the means of each of the clusters and reassign each observation to the cluster with the closest mean. In such a way, we reduce the observations within clusters. After a few steps the means of the clusters do not change anymore to finish the computations. This algorithm is not so fast, but it usually gives very good results.

4. OBJECT OF STUDIES

The efficiency of individual cluster analysis methods was assessed on the basis of data collected during geotechnical studies of subsoil conducted as the part of the analysis of foundation for grain elevators in Ząbkowice Śląskie in southern Poland.

Surface subsoil layers in the area under study are composed of nonlithified Pleistocene and Holocene formations (figure 1). In these layers, the deluvial and glacilacustrine deposits are found. Silty clays and silts with sandy interbeddings predominate here. Their consistency is plastic or soft plastic. These formations lie on overconsolidated glacial deposits, developed in the form of glacial tills and fluvioglacial sands. Cohesive formations of the lower part of the profile are stiff.



Fig. 1. Stratigraphic profile of the area under study

5. ANALYSIS OF RESULTS

The analysis was based on two CPTU tests carried out at the distance of approx. 30 m from each other (figure 2a). Test points were selected in such a way that the effects of cluster analysis applied to isolate geotechnical layers on the basis of CPTU studies could be compared with the results of laboratory analysis.



Fig. 2. An example of CPTU results at the Ząbkowice Śląskie test site (a); the effect of averaging the parameters (*FR* and q_n) used in cluster analysis (CPTU No. 11) (b)

Preparation for clustering of data from in situ tests was conducted following the procedure presented by MŁYNAREK et al. [5]. Recorded values of CPTU results (q_c , f_s and u_2) were corrected and normalized in order to obtain the form of net cone resistance q_n and normalized values of cone resistance Q_t and friction ratio FR. In com-

parison to earlier attempts of data clustering [2], [6] it was decided to exclude the parameter B_q , whose variation in the profile of the soils analyzed was negligible. However, additionally the net cone resistance q_n was used in cluster analysis. This resulted in the inconsistency observed in practice, appearing at times between geotechnical layers, isolated on the basis of traditional parameters Q_t and FR, and in the conclusions from the analysis of strength and deformation parameters of subsoil, which are determined primarily from the value of net cone resistance. Increasing the number of parameters included in clustering was to yield more homogeneous geotechnical layers, also in terms of analysis of bearing capacity of subsoil. The values of the above mentioned parameters were next averaged at 0.2 m intervals, eliminating in the analysis single extreme values and limiting the set of data to a size facilitating efficient mathematical processing of data (figure 2b).

The results of CPTU sampling prepared in such a way were clustered based on three methods discussed in chapter 3. As was already mentioned, the application of hierarchic methods (cosine and Euclidean) resulted in several solutions (from 1 to *n* clusters), out of which the most suitable are to be selected. For this purpose different criteria may be applied, two of which were discussed by MŁYNAREK et al. [6]. The *k*-mean method and the dendrite method require the researcher to assume a priori the number of isolated clusters. In the case of analyzed data clustering, the results obtained by hierarchic methods were used to determine an estimated number of expected clusters. Thus, taking account of the criterion established by CALIŃSKI and HARABASZ [1] and MŁYNAREK and WIERZBICKI [5], the expected number of geotechnical layers in the subsoil was selected first on the basis of the results of hierarchic methods, and next this number was included in the calculations of the dendrite and *k*-mean method. To provide a better accuracy of the assumption adopted, calculations using both latter methods were also made for the number of layers larger and smaller than the original values.

The number of isolated clusters were analyzed separately for both test points. Evaluation of the VRC values obtained (CALIŃSKI and HARABASZ [1]) showed that in the case of point 11, the first maximum appears at 3–4 isolated clusters. In turn, including the set of criteria given by MŁYNAREK and WIERZBICKI [5] showed that a better solution would be to isolate 4 rather than 3 clusters. Analogously, in the case of the point 3 it was decided that the data accumulated are grouped to form 3 clusters. It should be stressed here that the number of isolated clusters constitutes a certain minimum, meeting the criteria adopted. According to the methodology applied, subsoil may be divided into statistically more homogeneous groups. However, in practice this means a multiple increase in the number of layers in the profile, which in turn unnecessarily complicates the model of geological and engineering structure of subsoil [6]. Taking into consideration the assumptions of further analysis, aiming at a comparison of efficiency of individual methods, it was assumed that all the methods indicate the same number of layers as optimal. In the analyzed case of a simple geological structure this assumption is obvious. However, it should be mentioned that in

the case of a more complex arrangement of geological formations, the optimum number of clusters may be significantly different for individual methods.

6. ASSESSMENT OF CLUSTERING EFFICIENCY USING SELECTED METHODS

The analysis of clustering efficiency used both CPTU parameters applied during cluster analysis and basic geotechnical parameters, i.e., a constrained modulus M_0 and non-drainage shear strength s_u determined on their basis.

The analysis of results was performed in two stages. The first stage included the determination of such statistics of CPTU parameters as the mean and standard deviation as well as the determination of 95% confidence intervals within established clusters. Mean values were next ascribed to specific measurement depths, according to the division obtained using the method applied. As a result probing profile was obtained, in which measured values of parameters were replaced by mean values, characteristic of a given layer. A visual analysis alone, especially of the characteristic values of a net cone resistance (q_n) , shows significant differences between results given by individual methods (figure 3).



Fig. 3. Characteristic values of q_n obtained on the basis of different clustering methods against the background of origin values of q_n (CPTU No. 11)

Among the hierarchic methods, it is the one using Euclidean distance which averages to the highest degree the whole profile, distinguishing only the layers differing most in terms of the parameters applied in the analysis. For the same part of the profile, in which the Euclidean method indicates one layer, the method using the cosine distance distinguishes three layers, while the method of *k*-means and the dendrite method – two layers each. The differences between individual solutions are well visible in the diagram presenting the changes in standard deviations of the parameter q_n with depth (figure 4).

The adoption of the solution supplied by the cosine method makes it possible to reduce definitely standard deviations in the upper and central parts of the profile in comparison to the Euclidean method. However, this is done at the expense of the lower part of the profile, in which the parameter q_n is characterized by a very high standard deviation. In turn, the application of the *k*-mean method gives an opposite effect – the upper and central parts of the profile are included in one layer with a relatively high standard deviation, while the lower part of the subsoil analyzed exhibits a definitely lower variability. The dendrite method, as the method of *k*-means, divides the subsoil into two layers, but with a very similar standard deviation, although markedly lower than that in the case of the Euclidean method. It should be mentioned here that when following the criterion of MŁYNAREK and WIERZBICKI [5], at this stage the analysis was conducted only for variation observed within the so-called primary clusters, i.e., the most numerous ones.



Fig. 4. Changes in standard deviation of parameter q_n in primary clusters, depending on the clustering method applied (CPTU No. 11)

As we explained in the point 2, the clustering of data in order to isolate soil layers in the subsoil being homogeneous in terms of strength and rigidity is crucial. The analysis included undrained shear strength, which identifies the strength of cohesive soils, and constrained modulus, describing the rigidity of cohesive and non-cohesive subsoils. The values of these parameters were calculated as follows:

$$s_u = \frac{q_n}{N_{kt}},\tag{1}$$

$$M = \alpha \, q_n, \tag{2}$$

where: q_n – net cone resistance, N_{kt} – cone factor, α – factor of finding M.

The results of analysis pertaining to the division into layers performed in terms of the parameters q_n , Q_t and FR are presented in table 1.

Table 1

Method	Cluster number	Constrained modulus M (MPa)			Non-drainage shear strength s_u (kPa)		
		\overline{X}	σ	$\pm 95\%$	\overline{X}	σ	±95%
Euclidean	1	25.2	13.0	25.4	-	_	_
	2	29.9	16.2	8.1	181.2	97.9	49.3
	3	68.9	0.8	14.4	_	_	_
	4	56.0	48.4	869.5	_	_	_
Cosine	1	20.7	15.0	23.1	_	_	_
	2	55.3	23.3	19.2	334.9	141.2	116.6
	3	10.2	5.1	5.3	62.1	31.1	31.9
	4	30.1	7.3	6.4	182.4	44.0	39.0
k-means	1	45.6	7.9	6.5	276.3	48.0	39.6
	2	13.0	10.2	21.3	_	_	_
	3	20.1	11.1	7.5	121.6	67.3	45.6
	4	44.2	20.5	16.8	_	_	_
Dendrite	1	68.9	0.8	14.4	_	-	_
	2	26.4	15.3	23.5	_	_	_
	3	21.2	11.7	7.5	128.6	71.2	45.5
	4	45.0	10.8	9.3	272.6	65.3	56.5

Mean values, standard deviations and 95% confidence intervals of selected geotechnical parameters, determined on the basis of the results of individual clustering methods (CPTU No. 11)

The results obtained confirm conclusions obtained from the comparison of mean CPTU parameters. On the other hand, these results strongly emphasize how large the range of error may be, resulting only from the averaging of soil parameters within isolated layers. This error measured by standard deviation considerably exceeds 10%, and the incorporation of 95% confidence intervals shows that in practice it is not possible to apply this measure of variation to the analysis of bearing capacity of subsoil at 95% confidence, but at a definitely lower one. Thus, this confirms the observation by MŁYNAREK et al. [6], who report that homogeneity of isolated layers, satisfying from the viewpoint of statistical analysis, requires a very detailed division of the profile. This in turn leads to the extension of the scale of complexity to the geological model and at the same time to the appearance of considerable discrepancies between the lithostratigraphic model and the strengthdeformation model. A certain confirmation of the adequacy of the assumed division into geotechnical layers is the fact that the analysis of the significance of differences conducted using the t-Student test (for clusters with normal distributions) and the Wald-Wolfovitz test (in the case of samples with a distribution inconsistent with the normal one) showed that apart from one case, i.e., the coefficient FR of layers 2 and 3 in the point 11, the differences between distinguished clusters are significant. This statement pertains obviously to the assumption of isolated clusters as the minimum number, i.e., it limits unilaterally the model of geologic structure.

Significant information on the efficiency of the clustering methods applied is supplied by the analysis of mean values of geotechnical parameters established for the whole profile, taking into consideration the obtained divisions into layers. For this purpose in both testing points, weighted means were calculated for two geotechnical parameters (M_0 , s_u), standard deviations and 95% confidence intervals. The adopted weight factor was the depth of an isolated geotechnical layer. In this way, as a result of the analysis, some information was obtained on the total variation of a given parameter in the whole profile. The results are presented in table 2.

Table 2

	Method	Constrained modulus			Non-drainage shear strength		
Point		M (MPa)			s_u (kPa)		
		\overline{X}	σ	$\pm 95\%$	$\overline{\mathbf{X}}$	σ	$\pm 95\%$
3	Euclidean	35.0	18.8	14.1	202.2	107.4	61.4
	cos	35.9	15.6	14.6	193.6	85.8	68.8
	k-means	37.1	16.1	9.5	202.6	97.9	76.4
11	Euclidean	31.2	16.3	33.2	181.2	97.9	49.3
	cos	32.9	13.2	12.6	184.1	68.8	59.1
	k-means	30.4	10.8	9.2	185.0	59.4	43.1
	dendrite	30.5	11.6	10.2	181.2	69.1	49.5

A listing of weighted mean values of constrained modulus and non-drainage shear strength in the profiles analyzed, depending on clustering method (CPTU Nos. 3 and 11)

A significant conclusion resulting from the analysis of table 2 is the statement that in the case of two tests, the Euclidean method determines the mean with the highest standard deviations and the widest confidence interval. This pertains both to the variation in the constrained modulus M_0 and the undrained shear strength s_u . In the case of the other methods, conclusions are no longer so obvious. In testing point No. 11, the lowest standard deviation and the widest 95% confidence interval are found for the method of kmeans. The results of the dendrite method differ slightly, while those of the cosine method differ significantly. In contrast, in the testing point No. 3 it was found that the cosine method generated layers with the lowest mean standard deviation. The difference between this method and the method of k-means is not big, which is confirmed by the analysis of 95% confidence intervals. In this case, the method of k-means yields narrower confidence intervals of the mean weighted constrained modulus and undrained shear strength. When analyzing the measures of variation of parameters determined in the profile it needs to be emphasized that mean weighted values of constrained modulus and undrained shear strength in both testing points differ insignificantly between individual methods (by no more than approx. 5%).

Z. MŁYNAREK et al.

In order to illustrate the changes in the structure of the geotechnical profile, which takes into consideration only the variation in rigidity and shear strength, based on the already performed lithostratigraphic division (on the basis of the results from the *k*-means method), the values of constrained modulus and shear strength were determined along the whole profile. Subsequently these values constituted the basis for the identification of clusters based on *k*-means. Then, within such isolated layers, mean values of geotechnical parameters were calculated, together with their standard deviations and 95% confidence intervals. Mean values of constrained modulus together with standard deviation are presented in figure 5.



Fig. 5. Comparison between characteristic values of constrained modulus M, obtained on the basis of the 1st step clusters (a) and the 2nd step clusters (b) (CPTU No. 11)

A comparison with fiducial values of constrained modulus, determined initially (clustering based on three CPTU parameters – "the 1st step" clustering), indicates distinct differences in the division of the subsoil. These differences are least significant in the lower part of subsoil, in which only an additional isolation of several layers with a higher modulus is obtained. A fundamental difference is visible in the central and upper parts of the subsoil analyzed. The original division of subsoil assumed the existence of one layer in this place, whereas analysis based on the criterion of homogeneity of the constrained modulus divides the subsoil into two layers. It should be stressed that this division is similar to the one based on three CPTU parameters on the grounds of the constitutes the mean of the newly distin-

guished layers. The analysis of standard deviation and of 95% confidence intervals also leads to similar conclusions (table 3).

Table 3

Mathad	Charten Ma	Constrained modulus M (MPa)			
Method	Cluster No.	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$		± 95%	
	1	44.24	4.19	3.82	
k maana	2	60.35	7.29	13.48	
<i>k</i> -means	3	29.33	4.30	3.72	
	4	9.69	3.96	3.43	
Weighted mean		30.39	4.45	4.58	

Mean values, standard deviations and 95% confidence intervals of constrained modulus M in individual clusters – the 2nd step clustering (CPTU 11)

A general conclusion drawn from the analysis is the statement that the division of subsoil into layers, based on the three CPTU parameters treated jointly (which is the basic, 1st step division), may significantly differ from the division based solely on one geotechnical parameter (a detailed, 2nd step division). Although a detailed division is characterized by a markedly smaller margin of error (when we analyze a selected parameter), for obvious reasons it may not reflect the characteristics of other geotechnical parameters of subsoil. The basic division is thus the most universal solution, which does not mean that in a given situation it is the best possible one.

The above conclusion is very well documented by the prepared geotechnical profiles. The construction of these profiles was based on the results of clustering performed using all the four methods, conducted on the basis of CPTU parameters, as well as the results of clustering of the method of *k*-means based on determined geotechnical parameters (figure 6). The profiles obtained prove that the application of different clustering methods and different initial parameters in consequence leads to the adoption of different models of the subsoil geological and engineering structure.

The model based on the Euclidean method differs considerably from the other models. Cohesive soils in this model are incorporated almost solely into one layer – sandy clay with the mean constrained modulus of approx. 30 MPa. In the other models, in this layer two or three additional layers were isolated. Soils with both lower and higher constrained moduli were isolated. Apart from two main layers of cohesive formations, depending on the method adopted, additionally loamy sands or firm sandy clays are isolated in profile 11. It should be stressed that in the upper part of the profile 3, when using the *k*-means method, layers defined by other methods as non-cohesive soils, with low bearing capacity parameters, were included in the cohesive layer with increased bearing capacity. Such an effect results from relatively high (for non-cohesive soils) FR values being used during clustering and responsible – to a high degree – for the interpretation of lithology of the soils analyzed.



Fig. 6. Geotechnical cross-sections obtained on the basis of different clustering techniques

7. CONCLUSIONS

The analyses conducted allow several interesting generalizations to be made:

• There are classification systems, which are constructed on the basis of CPTU testing parameters. They may be good starting points for the identification of soils found in the subsoil and for the determination of drainage conditions in the subsoil.

• However, in order to separate from subsoil the zones homogeneous in terms of strength and deformations, a separate clustering procedure needs to performed, based on statistical criteria.

• Among the clustering methods analyzed, i.e., hierarchical agglomeration (Euclidean and cosine), hierarchical dendrite and *k*-means, the latter may be considered most

effective. This method does not lead to excessively detailed division of subsoil into layers and the resulting division clearly differentiates layers in terms of shear strength and constrained modulus. However, geotechnics engineer should take the decision on the adoption of a specific method. It is in terms of geotechnics that the decision may be taken a priori what level of differentiation needs to be adopted for constrained moduli or undrained shear strength to solve the geotechnical problem under consideration.

• The division of subsoil into homogeneous zones in terms of the parameters, which represent shear strength, e.g., s_u and constrained modulus, is very valuable in the case of pile foundation or direct foundation. This type of division makes it possible to assess promptly the location of zones inappropriate for the location of pile head or for the determination of the level for direct foundation.

ACKNOWLEDGEMENT

This paper is supported by the founds intended for science in the years 2005–2007 as research project No. 0562/P04/2005/29.

REFERENCES

- CALIŃSKI T., HARABASZ J., A dendrite method for cluster analysis, Communication in Statistics, 1974, Vol. 3, 1–27.
- [2] HEGAZY Y.A., MAYNE P.W., Objective Site Characterization Using Clustering of Piezocone Data, Journal of Geotechnical and Geoenvironmental Engineering, 2002, Vol. 12.
- [3] LUNNE T., ROBERTSON P.K., POWELL J.J.M., Cone Penetration Testing in geotechnical practice, reprint by E & FN Spon, London, 1997.
- [4] MLYNAREK Z., TSCHUSCHKE W., WOŁYŃSKI W., The CPTU classification chart for the post-flotation sediments based on statistical criteria, Proc. of the 15th International Conference on Soil Mechanics and Geotechnical Engineering, Istambul, 2001, Vol. 1. Balkema Publishers, 459–462.
- [5] MLYNAREK Z., TSCHUSCHKE W., WIERZBICKI J., WOLYŃSKI W., Statistical criteria of deformation of homogenous geotechnical layers, Proc. of the 16th International Conference on Soil Mechanics and Geotechnical Engineering, Osaka, 2005, Japan, 725–728.
- [6] MŁYNAREK Z., WIERZBICKI J., WOŁYŃSKI W., Use of cluster method for in situ tests, Studia Geotechnica et Mechanica, 2005, Vol. XXVII, No. 3–4, 16–27.
- [7] MLYNAREK Z., Limitation to application of empirical relationships from in-situ tests, oral presentation on XIII Danube-European Conference on Geotechnical Engineering, Ljubljana, 2006.
- [8] MLYNAREK Z., LUNNE T., Statistical estimation of homogeneity of North Sea overconsolidated clay, Proc. of International Conference on Statistical and Application Probability, Vancouver, 1987, Canada.
- [9] POWELL J.J.M., Technical session 1c: In-situ testing; General Report, Proc. of 16th International Conference on Soil Mechanics and Geotechnical Engineering, Osaka, 2005, Millpress-Rotterdam, 2971–2981.