Studia Geotechnica et Mechanica, Vol. XXVII, No. 3–4, 2005

USE OF CLUSTER METHOD FOR IN SITU TESTS

Z. MŁYNAREK, J. WIERZBICKI

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

W. WOŁYŃSKI

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

1. INTRODUCTION

Cluster analysis is a statistical technique used for dividing data into groups of similar traits [2]. This method can be used in these fields where the tests provide a set of data with a priori unknown division into groups and undefined distribution of variables, which make the method different from discrimination analysis. This makes it very useful for all kinds of subsoil research.

Grouping of data with cluster analysis can be carried out in a vertical profile [3] [1], but there are also examples of using it in separating areas of soil with similar parameters on a given plane [5]. In both cases, the method of calculations is the same, and only the researcher's approach changes: one can treat this task as an axial or plane problem. The quality of the picture obtained depends on the quality of the data collected and frequency of measurements. All kinds of in situ tests, particularly static penetration test (CPT, CPTU), open up wide possibilities. As indicated by [3], the CPTU results supported by cluster analysis can constitute a basis for separating homogeneous layers in subsoil. Among the most important advantages of this technique of profile analysis are: no need for initial knowledge on the number of layers obtained in the final stage, possibility of taking many variables into account at the same time and relatively simple calculating techniques.

2. THEORETICAL BASIS OF CLUSTER ANALYSIS

The parameters registered during cone penetration test, i.e. q_c – cone resistance (or corrected resistance q_t), f_s – sleeve friction, and u_2 – pore pressure, are not directly used in cluster analysis. Their values depend on the depth of measurement, i.e. on stress state in the subsoil. Therefore, the values measured in the same subsoil at two depths will not be identical, and both measurements will not be included in the same group. Hence, direct parameters from CPTU must be normalized by a vertical compo-

nent of geostatic stress. For creating homogeneous groups it is the best to use the following CPTU parameters:

- normalized cone resistance Q_t

$$Q_t = \frac{q_t - \sigma_{v0}}{\sigma'_{v0}},$$

- pore pressure parameter B_q

$$B_q = \frac{u_2 - u_0}{q_t - \sigma_{v0}},$$

- friction ratio R_f

$$R_f = \frac{f_s}{q_t} \cdot 100\%,$$

where:

q_t – corrected cone resistance,
f_s – sleeve friction,
u_2 – pore pressure behind the cone tip,
u_0 – hydrostatic pressure,
σ_{v0} – vertical stress,
σ'_{v0} – effective vertical stress.

The values obtained in such a way, from now on called test parameters, can be used to carry out cluster analysis along a given profile.

Let $X_1, \ldots, X_p$ be variables characterizing the parameters obtained from a cone penetration test, and:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n \tag{1}$$

be a vector consisting of observations of the parameters studied at the depth i. In order to eliminate the effect of different measurement units of each variable, usually data standardization is made, i.e. one moves from the initial variables $X_1, \ldots, X_p$ to new standardized variables $Z_1, \ldots, Z_p$, while $\mathbf{z}_i = (z_n, \ldots, z_{ip})$ has the form of:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \quad i = 1, \ldots, n; \quad k = 1, \ldots, p \tag{2}$$

and

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}, \quad k = 1, \ldots, p, \tag{3}$$

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2, \quad k = 1, ..., p\,. \tag{4}$$

The algorithm of hierarchic agglomeration used requires an appropriate measure of the distance or the similarity between pairs of the original items represented, as a rule, by *p*-variable vectors and two groups, each consisting of one or more original observations. The algorithm assumes creation of subsequent distance or similarity matrices, each of the dimension smaller by one. Therefore, the first matrix has the dimension *n*, and the last one 1. In cluster analysis, various distance or similarity measures are used. Among the distances most often applied are Euclidean and Mahalanobis distances. As a similarity measure between observation vectors at given depths there was assumed, just as in the article by HEGAZY and MAYNEY [3], cosine of an angle determined as:

$$\mathbf{d}_{ij} = \cos(\mathbf{z}_i, \mathbf{z}_j) = \frac{\sum_{k=1}^{p}\mathbf{z}_{ik}\mathbf{z}_{ij}}{\sqrt{\sum_{k=1}^{p}\mathbf{z}_{ik}^2 \sum_{k=1}^{p}\mathbf{z}_{jk}^2}}, \quad i, j = 1, ..., n. \tag{5}$$

Hence, a similarity matrix is obtained:

$$\mathbf{D} = (d_{ij}). \tag{6}$$

The hierarchic agglomeration is a step-by-step method. In the previous step, it is assumed that each observation creates a separate cluster. In the next step, the two nearest (closest) observations are connected (in a sense of selected distance or similarity measure) in order to obtain $n - 1$ clusters (groups). In each subsequent step, the number of clusters decreases by one to create finally one cluster. There are many methods of determining (in each step of the procedure discussed) the distance (similarity) measure between two groups. Among the most often used are: the method of single linkage, average linkage, and Ward's method. In further part of this paper, two of them are applied: single and average linkage methods. In the former, the distance (similarity) measure between two groups is determined as the greatest value of similarity between observations belonging to different groups. In the latter method, this distance (similarity) is calculated between centroids of clusters (mean values of all observations in a given group). This method does not cause any excessive ("artificial") inclusion of data by already existing groups which can happen in the method of single linkage.

The following four different methods of clustering were discussed:

the method based on angle cosine and average linkage (avcos),

the method based on angle cosine and single linkage (singcos),

the method based on Euclidean distance and average linkage (aveuk),

the method based on Euclidean distance and single linkage (singeuk).

3. COMPARISON OF THE SELECTED METHODS OF CLUSTER ANALYSIS

Application of the presented methods of cluster analysis was tested on data from geotechnical subsoil tests made under the corn silos in Borek Strzeliński [7]. In the subsoil tested, cohesive sediments predominate. They were formed as moraine loams of the Odra glaciation and are considered to be overconsolidated soils. Among moraine sediments there are present sandy interbeddings and layers of fluvioglacial coarse and medium sands typical of this facies.

3.1. SEPARATION OF GROUPS IN TEST PROFILE

A section of a selected CPTU profile between the depths of 1.2 and 14.2 m was analyzed. Read outs of test parameters were averaged to obtain the values from every 20 cm of the profile (figure 1).

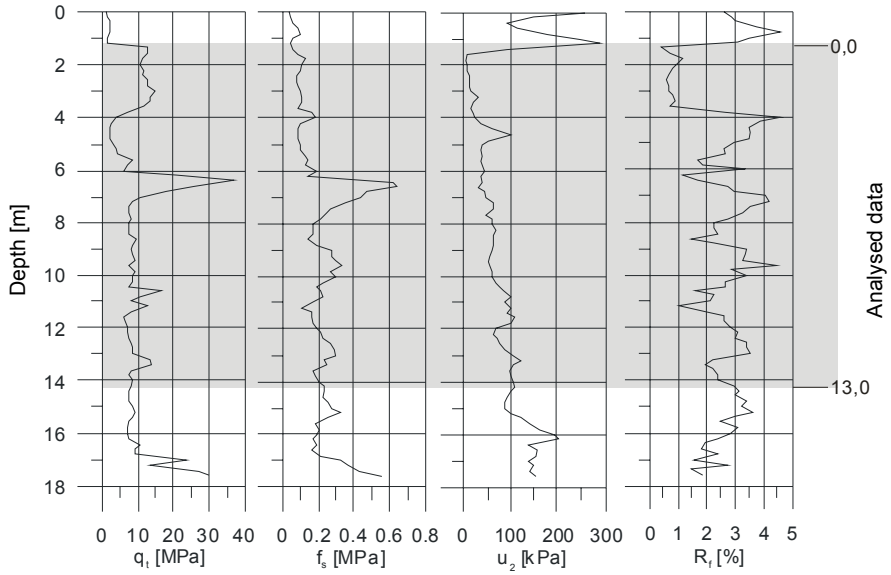

Fig. 1. Averaged test results of cone penetration

Cluster analysis for the selected profile was carried out taking into account first of all two parameters: Q_t and B_q, and then, additionally, R_f. The grouping yielded a number of possible solutions, from one cluster common for all data to maximum number of clusters which was different for each method considered (the table). In each solu-

tion, the number of the so-called free data, i.e. such that they did not group the others, was registered.

Following the assumptions concerning the expected geological structure, the number of 6 clusters in a profile of 13 m thickness was accepted as an initial (minimum) one. As results from the analysis, both procedures being based on the method of a single linkage gave clearly more free data (at the same number of groups) than the methods being based on average linkage (figure 2).

Table

The maximum number of clusters in each method

Method	Avcos	Singcos	Aveuk	Singeuk
Number of clusters	16	15	15	12

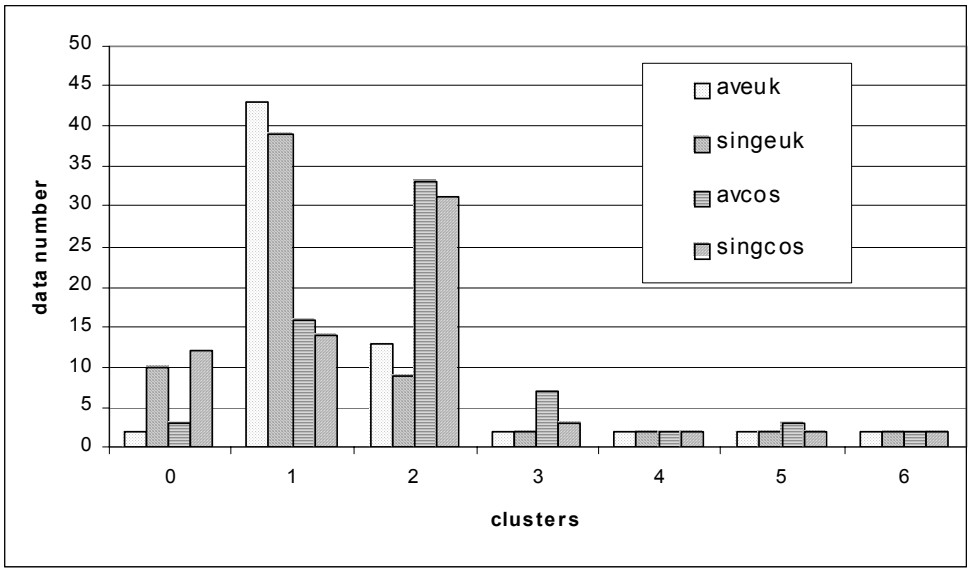

Fig. 2. Data number in clusters obtained in different methods (cluster "0" gathers the free data)

For each separated cluster mean (characteristic) values of test parameters, standard deviations and coefficients of variation (cv) of these samples were determined (figure 3). Independently of the method of calculation, two main (the greatest) clusters are distinguished: 1 and 2. Analyzing the values of cv of the parameter Q_t it is clear that both methods based on Euclidean distance as a measure of similarity yield a zero cluster with the variability of almost twice as great as that in the methods based on an an-

gle cosine. Comparison of size of the clusters and their variability shows a clear dependence of the size reduction on a decrease in the coefficient of variation (e.g. groups 1 and 2).

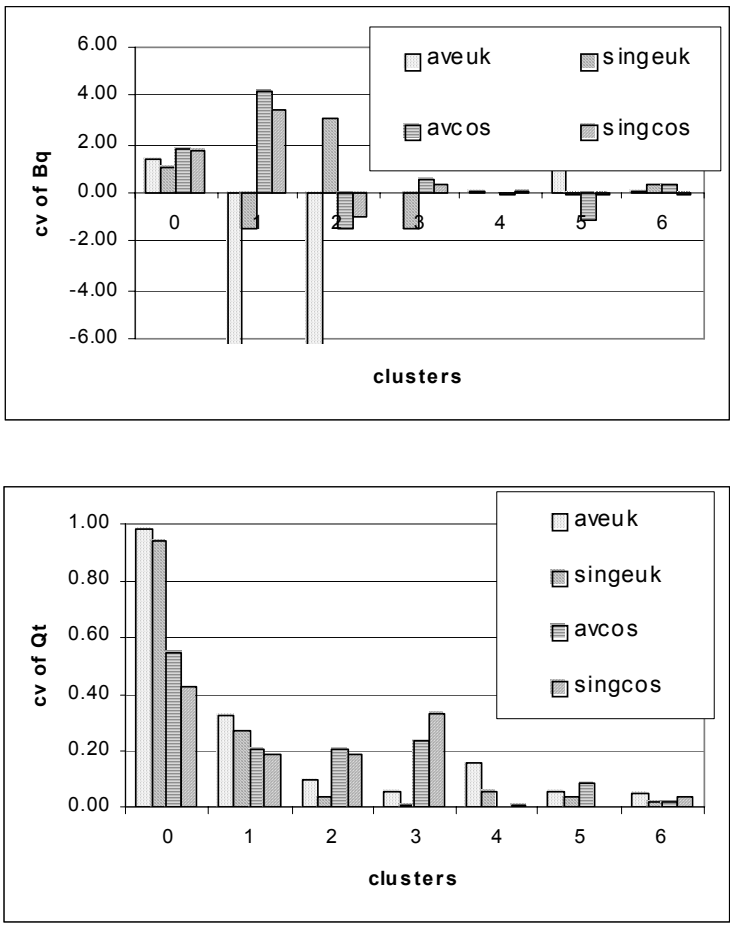

Fig. 3. Coefficients of variation (cv) of B_q and Q_t in the clusters obtained

After increasing the number of separated clusters from 6 to 8 a marked decline in cv within the clusters was observed (figure 4).

In this case, the coefficient of variation of Q_t only for one cluster (the methods based on an angle cosine) clearly exceeded 0.2. A characteristic feature of using the methods that are based on measurement of Euclidean distance is an almost twofold increase in the number of free data compared to the case of separating 6 clusters.

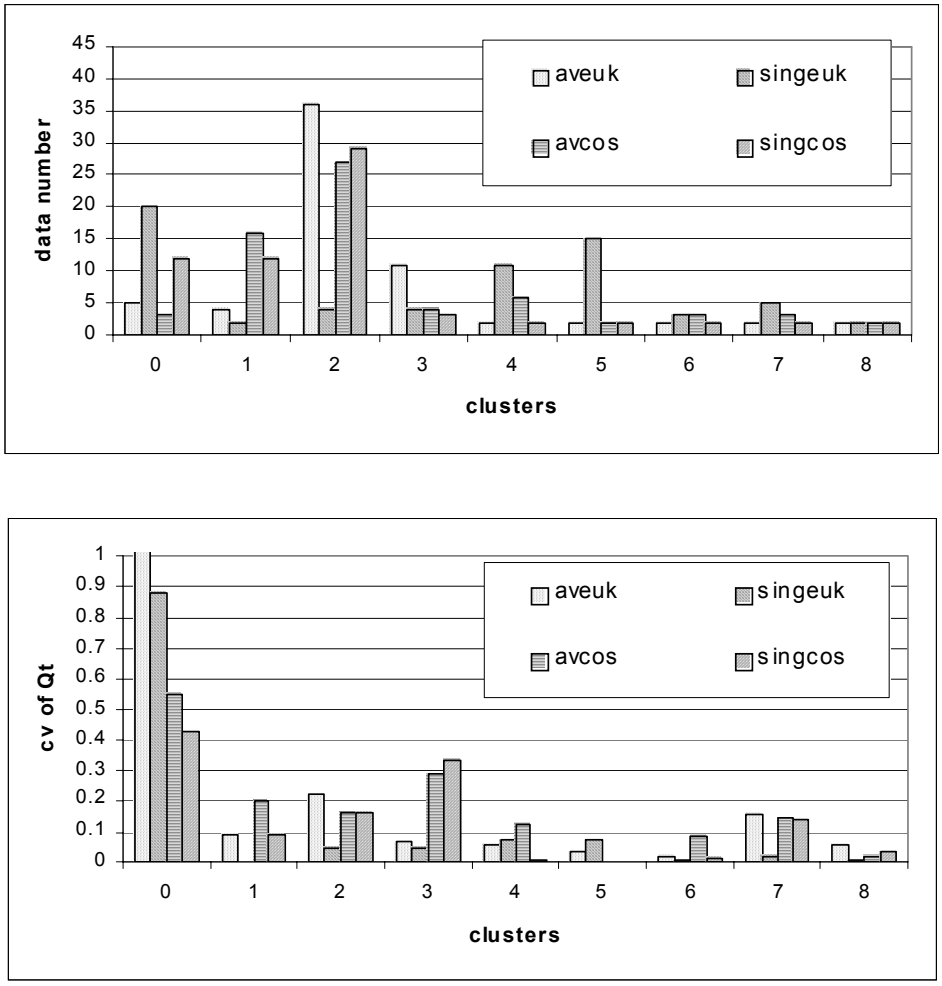

Fig. 4. Data number and coefficient of variation of Q_t in the case of 8 clusters

During analysis of the profile of interest by means of an avcos method, 6, 9, 10 and 11 clusters were subsequently separated (figure 5). It is worth noticing that after exceeding the number of 10 clusters, a very rapid increase in the number of free data takes place. Hence, it should be expected that from a practical point of view the maximum number of clusters which should be separated in a profile is 10. In each case of separating 6, 9 or 10 clusters, domination of three clusters was observed, which is very clear in the case of separating 6 clusters. This indicates regrouping of data while maintaining a certain predominating division of the profile studied.

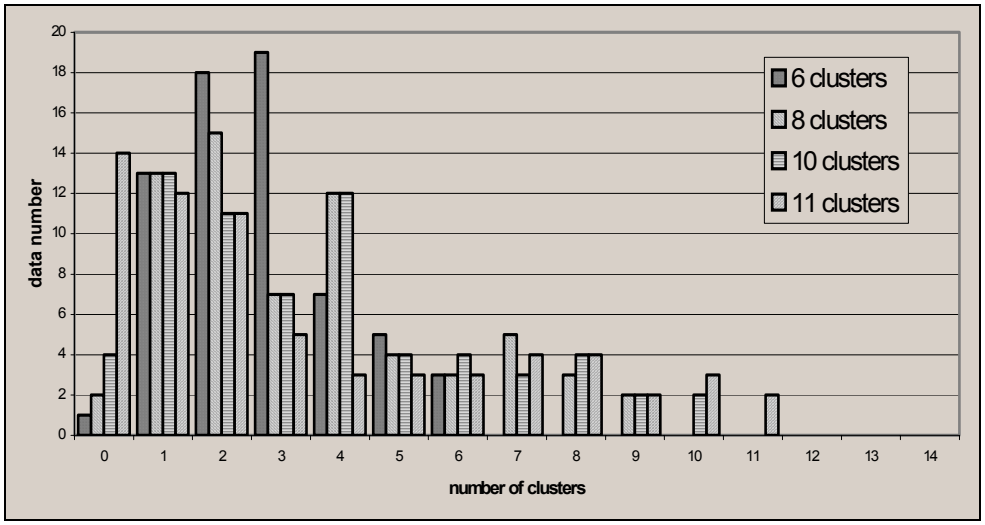

Fig. 5. Changes in data number within clusters, depending on an increase in the cluster number

Introduction of the third variable, i.e. R_f , into the analysis brought about a marked decrease in the number of free data and only a slight increase in cv of the parameter Q_t of the most numerous clusters. The cv of the index R_f is on the level similar to that of Q_t and reaches its maximum value of 0.3. A characteristic feature is the observation that in none of the main clusters the highest variability indices of both parameters occur simultaneously. As earlier, also this time the parameter B_q has clearly the highest variability. Separating additional clusters (to the total of 9 and 10) caused a decrease in cv of each parameter; however, this prevented us from obtaining main clusters with low coefficient for the parameters R_f and Q_t (below 0.2). It was also observed that R_f reveals much greater homogeneity in the separated clusters than Q_t. Hence, it could be supposed that the profile studied is more differentiated with respect to the kind of soil than to strength parameters.

3.2. SELECTION OF THE METHOD

The analysis does not indicate definitely an advantage of one of the methods over the others. However, observations allow some conclusions which could be helpful in eventual choice of the clustering method.

The number of clusters, hence the suggested number of geotechnical layers separated in a profile, must be considered individually for each case. The criteria facilitating the choice can be as follows: the number of free data, coefficient of variation of a cluster, and size of clusters (e.g. separation of the main clusters if indicated as a result of geological assumptions).

At a small number of clusters, the methods using a single linkage yield more homogeneous groups, but this results in much greater number of free data. Hence, the picture of subsoil only seems to be more clear since a great number of free data distorts its effectively.

The choice of an angle cosine as a measure of similarity results in even more size of the main clusters than in the case of Euclidean distance. In the case of a great number of clusters (which warrants lower variability coefficients within the clusters), the former method yields less favourable results (higher variability coefficients) than the latter. At the same time the methods being based on Euclidean distance are more sensitive to an increase in the number of groups which results in rapid rise in free data.

Simultaneous application of three test parameters caused separation of an additional main group. At the same time the effect of increase in the number of groups on determined variability coefficients was clearly weaker. This indicates difficulties which can be encountered during attempts at characterizing subsoil on the basis of different parameters. Doubtlessly more clear picture can be obtained while considering parameters previously grouped with respect to an actual design task.

4. SEPARATION OF GEOTECHNICAL LAYERS

The clusters obtained can constitute the basis for separating geotechnical layers in the subsoil studied. A comparison of the results obtained using two grouping methods which are based on average linkage for the test point selected, is given as an example. In such a case, geotechnical layers were separated simultaneously based on the three CPTU parameters discussed. For each of the 6 clusters mean values of the parameters were accepted as characteristic ones. The subsoil picture obtained by means of the two methods on the one hand reveals significant similarities, but on the other one it shows some differences (figure 6).

In the upper part of the profile, a layer of a high Q_t (1 in the avcos method and 2 in the aveuk one) is clearly marked. The middle and lower parts of the profile are also characterized by a similar Q_t obtained in both methods. However, in the case of the parameter R_f, the difference is clear. In the aveuk method, the layer 1 strongly predominates, while in the avcos, two almost identical layers (2 and 3) are separated. These layers have different values of the friction coefficient R_f, therefore, indirectly, it can be assumed that they belong to different soil types. In turn, in subsurface part of the profile, the avcos method allows us to identified one layer (No. 4), while the aveuk – several layers, differing in all parameters. In the case of both methods, between 6 and 7 m of the profile there is a layer of a high Q_t (No. 6). However, it is not as homogeneous as the layers 1 (avcos) and 2 (aveuk). Characteristic interbeddings in the lower part of the profile are also detected by both methods. The main difference be-

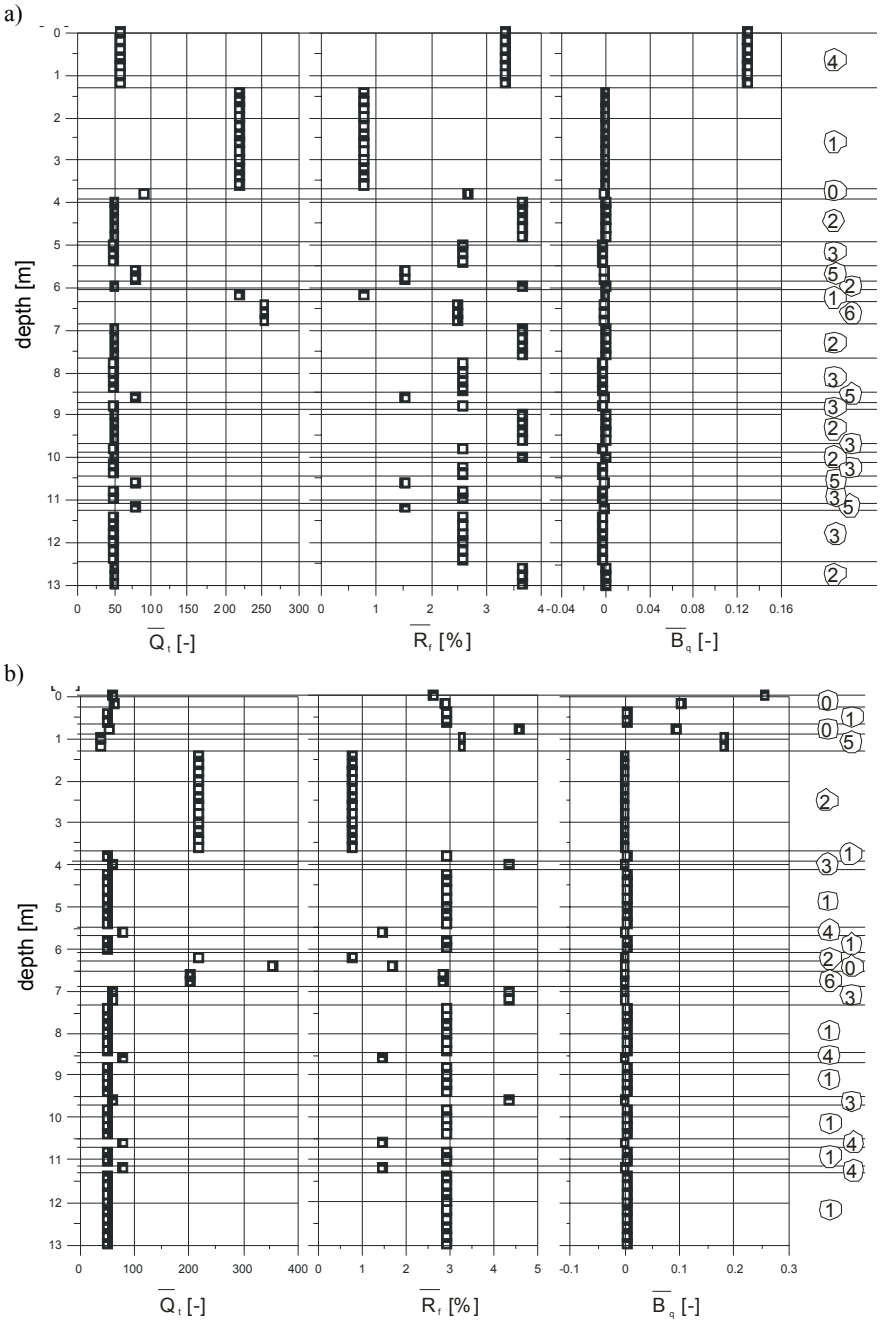

Fig. 6. Characteristic values of CPTU parameters for 6 clusters derived by avcos (a) and aveuk (b) methods

tween application of both methods is related to presentation of the middle and lower parts of the subsoil as a layered one (the avcos method) or relatively homogeneous (the aveuk method). To assess the importance of this difference, its effect on the result interpretation, i.e. identification of soil type and its characteristic geotechnical parameters, was presented. This evaluation was made based on an example of placing each layer on the Robertson classification diagram [8] (figure 7).

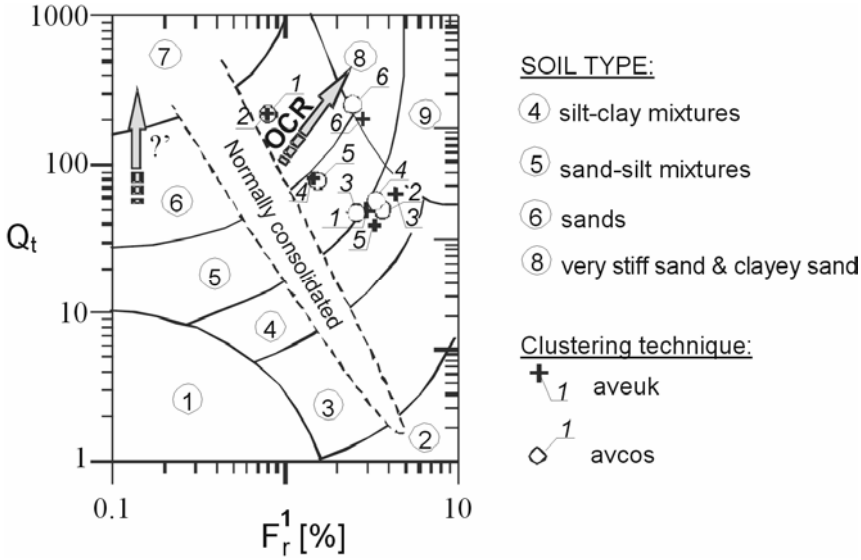

Fig. 7. Location of geotechnical layers obtained by avcos and aveuk methods on the Robertson [8] classification chart

Both methods allowed us to classify the layers 1 (avcos) and 2 (aveuk) as slightly overconsolidated sands and the layer 6 as strongly overconsolidated sands. Such an interpretation is in accordance with stratigraphic position of the layers and confirms that the assumptions accepted for the genesis of the area tested are well-founded. Also the layers 5 (avcos) and 4 (aveuk) identified as interbeddings in the middle and lower parts of the profile were classified as overconsolidated sand–silt mixtures (loamy sands). According to the earlier observations, some differences are observed between the layers 2 and 3 (avcos) and in the same part of the profile, layer 1 (aveuk). Treating subsoil as one layer we would define it as consisting of overconsolidated loams or loamy sands. In the case of separating an additional layer No. 3, the subsoil structure should be described as a layered one. In such a case, however, the type of soil does not change completely, the layer 2 is made of loam and sandy loam, and the layer 3 of sandy loam and loamy sand. Therefore, it can be assumed that in the example discussed, application of both methods allows effective and realistic separation of geo-

technical layers. The method based on angle cosine enables detecting an additional group of layers, but this is of rather more geological-stratigraphic importance compared to geotechnical one.

5. CONCLUSIONS

In the context of the analysis, a general conclusion can be drawn that cluster theory methods can be a statistically objective and effective way in searching for geotechnically homogeneous layers in subsoil. At present, however, the procedure of gradual construction of the layers starting with one-axis grouping (in a single test profile) and passing to plane grouping at a constant level of the stresses σ_{v0} in subsoil with resulting spatial picture of homogeneous subsoil layers is inevitable. A starting point of constructing geotechnical layers should be appropriately selected method of in situ soil testing. It is doubtless that the most favourable way is the cone penetration test (CPTU) since the parameters measured in this strength test provide information about history of subsoil load (OCR) [9], differentiation due to grain-size distribution, strength and deformation parameters and conditions of subsoil drainage. Analysis indicated that to obtain a full picture of the changes in these geotechnical parameters it is worthwhile considering three parameters: q_c (Q_t), R_f, B_q in clustering. A particular position in separating the layers in subsoil takes the parameter B_q. In separating homogeneous layers, the variability coefficient of B_q was always the highest which means that B_q is the most variable parameter. This is a very beneficial result of the analysis because B_q is the coefficient sensitive to the changes in pore pressure in subsoil, hence to the changes in drainage conditions. This fact is very significant in making decisions about choosing the method of interpretation of the shear strength parameters. This problem is also encountered when soils of differentiated or anisotropic structure, e.g. varved clays, occur in a subsoil. Therefore, applying the parameters Q_t and R_f, it is possible to separate genetically homogeneous layers of these clays differing from others in their genesis or grain-size distribution, while with the parameter B_q it is possible to separate additional zones of differentiated directions of filtration. The homogeneous layers identified according to the criteria assumed by a geotechnician from CPTU parameters will also be described in terms of their strength and deformation, since knowing CPTU parameters one can use interpretation methods which allow determination of strength parameters and deformation moduli [4], [6].

The cluster analysis methods discussed facilitate objective qualitative and, which is important, quantitative analyses of test profiles. Hence, they constitute very useful tool in geotechnical design. The analysis must include a conclusion which underlines geotechnican's role in the process of layer identification. The goetechnician should decide which of the parameters should play the main role in formation of homogeneous lay-

ers.

REFERENCES

[1] BALBI D.J., SABOYA F., *CPTU-Soil Profile Interpretation Based on Similarity Concept*, Proc. 2nd International Site Characterization Conference, Porto, Portugal, 2004.
[2] EVERITT B., *Cluster Analysis*, Halsted–Wiley, N.Y., 1974.
[3] HEGAZY Y.A., MAYNE P.W., *Objective site characterization using clustering of piezocone data*, Journal of Geotechnical and Geoenvironmental Engineering, Vol. 12, 2002.
[4] LUNNE T., ROBERTSON P.K., POWELL J.J.M., *Cone Penetration Testing in geotechnical practice*, Reprint by E & FN Spon, London, 1997.
[5] MŁYNAREK Z., LUNNE T., *Statistical estimation of homogeneity of North Sea overconsolidated clay*, Proc. of International Conference on Statistical and Application Probability, Vancouver, Canada, 1987.
[6] MŁYNAREK Z., TSCHUSCHKE W., GOGOLIK S., *W sprawie wyznaczania modułów odkształcenia podłoża budowlanego metodą statycznego sondowania i dylatometrem Marchettiego*, Inżynieria Morska i Geotechnika, Gdańsk, 2003, No. 3–4.
[7] MŁYNAREK Z., TSCHUSCHKE W., GRAF R., GOGOLIK S., *Dokumentacja geotechniczna w sprawie warunków gruntowych i wodnych terenu przeznaczonego pod budowę elewatorów zbożowych w miejscowości Borek Strzeliński*, Opracowanie Hebo-Poznań, 2002, No. 201.
[8] ROBERTSON P.K., *Soil classification system using the cone penetration test*, Canadian Geotechnical Journal, 1990, 27(1).
[9] WIERZBICKI J., *Wykorzystanie techniki sondowania statycznego do oceny wskaźnika przekonsolidowania niektórych osadów plejstoceńskich*, Acta Scientiarum Polonorum; formatio Circumectus, 2002, 1–2, Kraków.