AUTOMATED SOIL PROFILE GENERATION METHODS ON THE BASIS OF DMT AND CPT DATA

Michał Kruk, Jarosław Kurek, Piotr Bilski and Simon Rabarijoely

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul. Nowoursynowska 159, 02-767, Warsaw, Poland
Michał Kruk michal_kruk@sggw.pl

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul. Nowoursynowska 159, 02-767, Warsaw, Poland
Jarosław Kurek jaroslaw_kurek@sggw.pl

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences – SGGW, ul. Nowoursynowska 159, 02-767, Warsaw, Poland
Piotr Bilski piotr_bilski@sggw.pl

Faculty of Engineering and Environmental Sciences, Warsaw University of Life Sciences – SGGW, ul. Nowoursynowska 159, 02-767, Warsaw, Poland
Simon Rabarijoely simon_rabarijoely@sggw.pl

Abstract

The paper presents automated methods of the soil profile generation on the basis of DMT and CPT data gathered from the Warsaw University of Life Sciences campus. Knowledge about the structure of the soil is important not only for the researchers, but also for engineers planning foundation of the new buildings. To properly design the building structure, the detailed information about the type and parameters of the soil must be determined. The traditional approach to this task was drilling boreholes in the ground to obtain soil samples. They could be then analyzed in the laboratory, so the information about the soil types in the test site was known with high precision. Unfortunately, the cost of drilling boreholes is high, therefore numerous attempts to obtain faster and cheaper methods are proposed. Application of geotechnical probes is more convenient, faster and cheaper than boreholes, therefore they supplement the boreholes during the in situ investigations. Based on the measured physical parameters and the diagram (called nomogram) connecting their values with the soil type, the profile can be generated by the human expert. Unfortunately, knowledge stored in nomograms is applicable only to specific geographical location. Using the diagram for soils with different geological history is prone to inaccuracies and identification errors. Therefore data analysis methods are implemented.

The main task of this work is to find the automated method to determine the structure of the soil. This method must be equal or even better than human expert to be accepted. In the work we proposed two solutions of this task and compared them with the human experts and with the soil samples obtained from the boreholes. The first one is based on the gradient analysis. It is composed of the Gaussian, average and median filtering and gradient or Laplacian zero-crossing search. The main problem in this method is smoothing the data and removing outliers (measurements significantly different than the neighbouring ones) – if we use bad parameters to filtering, we obtain too little or too many soil layers. The second method is based on the cluster analysis. The main problem in such methods is to find automatically the number of clusters. To do this we compared existing methods with our solution which is based on the gradient analysis.

The experiments had two aims. The first one was confrontation the soil profiles generated by the algorithms with the ones generated in the laboratory from the boreholes. In the optimal situation both

profiles should be identical. The second one was to confirm the accuracy of the profile by the geotechnical expert.

Keywords: *soil profiles, soil categorization, soft computing methods, clustering, dilatometer test.*

1. Introduction

The modern geotechnical exploration methods are widely supported by the computer technologies. The hardware used in these methods (invasive probes such as CPT and DMT) is helpful to obtain and store measurement data. In this approach it is easier to use the computer algorithms in analyzing the obtained data. They increase speed of geotechnical parameters' calculation. On the other hand, the widely used method is the analysis of gathered data using charts by the human expert (Marchetti 1980). The traditional and the most reliable approach to this task is drilling boreholes in the ground to obtain soil samples. Then they can be then analyzed in the laboratory, so the information about the soil types in the test site may be known with high precision. Unfortunately, the cost of drilling boreholes is high and it is very intrusive method, therefore attempts to minimize its usage are made.
The computer algorithm can analyze the collected measurement data automatically and as the output produces the soil profile. They can be analyzed by the geotechnical engineer, who verifies the accuracy of the generated profile or uses the information to calculate geotechnical,indexes. Such knowledge can be further used to develop the automated soil profile generation system, classifying the geotechnical layers based on the measured quantities at particular depths. Similar works were done before (Hashash et al. 2004, Shahin et al. 2005), but new approaches must be proposed.
The paper presents an automatic approach for soil profile generation using selected clustering methods. The measurement data are obtained from Warsaw University of Life Sciences (WUoLS) campus, Warsaw, Poland. The data gathered for the computer algorithm (implemented in Matlab enviroment) are obtained using DMT and CPT probes. The research presented here is a continuation of the experiments published before (Rabarijoely et al. 2007) and Rabarijoely and Bilski 2009). The expected result of the research is the method for soil profile generation which accuracy can be similar or better to the human expert. In this paper we combine the differential methods with the clustering algorithms and compare them to the automated clustering methods.

2. The input data

The soil measurements were taken at Warsaw University of Life Sciences (WUoLS) during the expansion of the university. Before new buildings could be established, throughout soil investigation had to be performed. Therefore multiple tests were conducted, using the presented probes and traditional methods, i.e. boreholes. The latter as the most accurate was treated as the reference method. The data for the experiment were gathered from geotechnical investigations by CPT and DMT probes.
The cone penetration test (CPT) is a standard and well established method widely used to recognize and analyze geotechnical conditions (Lunne et al. 1997, Młynarek 2007, Huang A & Mayne 2008). The probe is presented in Fig. 1. It is inserted into the ground with the constant speed of 2cm/s. During that process the measurement data of four parameters are obtained: depth (d), the resistance of the cone (q_c), sleeve friction resistance (f_s) and friction coefficient (R_f). We used the first three values in the presented experiment.

Figure 1 : The cone penetration test probe

The Flat Dilatometer Test (DMT), developed in Italy in 1980, is currently used in dozens countries both for research and practical applications. Wide diffusion of the DMT lies on the following reasons:

- Simple equipment and operation.
- High reproducibility.
- Cost effectiveness.
- Variety of penetration equipment (Totani et al. 2001)

The dilatometer consists of a steel blade having a thin, expandable, circular steel membrane mounted on one face. When at rest, the membrane is flush with the surrounding flat surface of the blade. The latter is connected, by an electro-pneumatic tube running through the insertion rods, to a control unit on the surface (Fig. 2a). The control unit is equipped with pressure gauges, an audio-visual signal, a valve for regulating gas flow (provided by a tank) and vent valves. The blade (Fig. 2b) is advanced into the ground using common field equipment, i.e. push rigs normally used for the cone penetration test (CPT) or drill rigs. Pushing the blade with a 20 ton penetrometer truck is the most effective (up to 100 m of profile per day). The test starts by inserting the dilatometer into the ground. Soon after the penetration, the operator inflates the membrane and takes, in about 1 min, two readings: the A pressure, required to just begin to, move the membrane ("lift-off"), and the B pressure, required to move the center of the membrane 1.1 mm against the soil. A third reading C ("closing pressure") can also optionally be taken by slowly deflating the membrane soon after B is reached. The blade is then advanced into the ground of one depth increment (typically 20 cm) (Totani et al. 2001).

Figure 2 : The DMT control unit (a) and the probe (b)

The tests were made in the areas where new buildings were to be established. As the final result the geotechnical cross-section charts were made by the human experts. The example of such a chart is presented in Fig.3. These charts are useful for the presented algorithm examination and the knowledge from it will be used to develop artificial intelligence classification system. To develop and test the algorithm, CPT and DMT data sets were used. They are called “CPT1”, “CPT2”, “DMT1”, “DMT2”, and so on. Each data set is a $n \times m$ matrix, where n is a number of rows (number of depths at which the measurements are taken) and m is a number of columns. The first column contains information about the depths, while the other columns contain values of the measured parameters.

The data gathered by CPT and DMT probes are presented in Fig.4. It is easy to observe a lot of extremes. In the most of them should be the change of the sort of soil. The smallest extremes should be removed because they contains noise – for example, they are generated when the probe hits the

rock. One of the task of described experiment is to remove this noise and at the same time to save useful information. It was done by using low pass filters such as Gaussian filter.

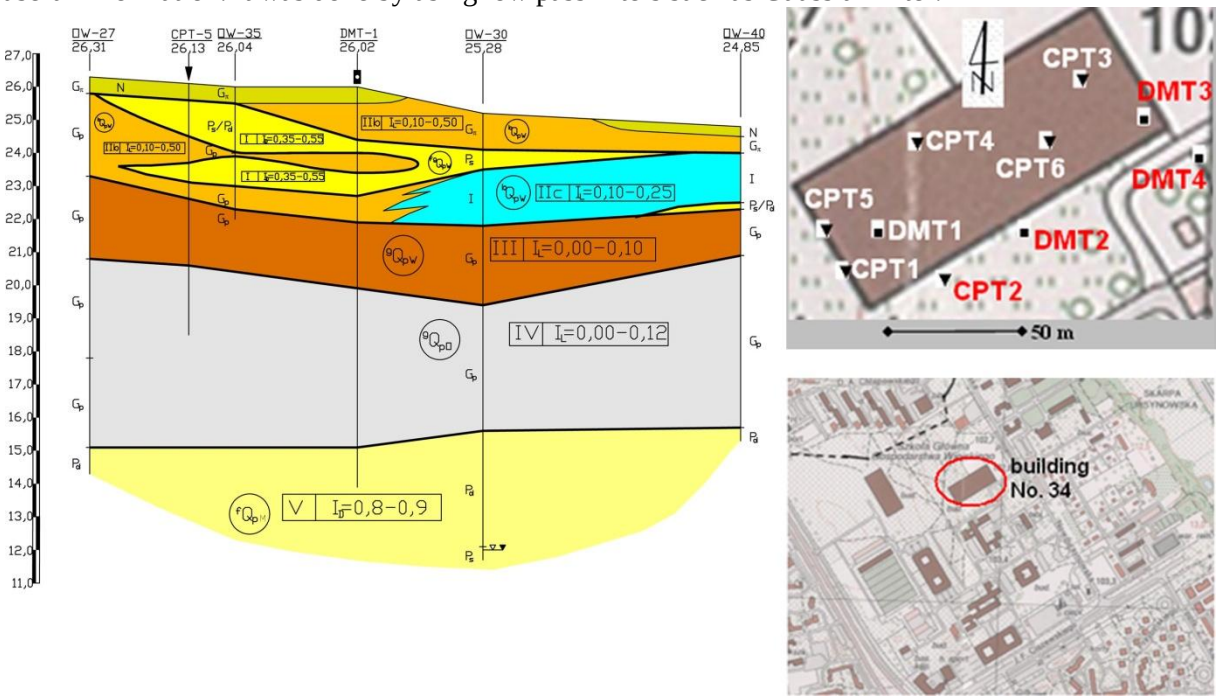


Figure 3. A typical geotechnical cross-section: OW – borehole, CPT – cone penetration test, DMT – Dilatometer test; (N – fill, Gp – sandy clay, Pd –fine sand, w_n – moisture content, I_D – relative density, I_L –liquidity index)

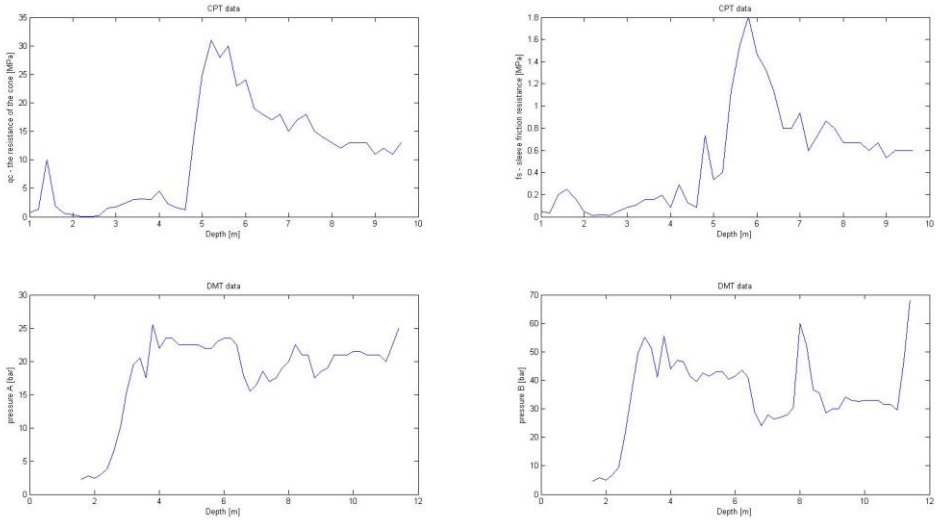


Figure 4. Results of the CPT and DMT measurements

3. Applied methods

The first step was improving the resolution of data set. Because of small data set (about 50 per one measurement) the resolution is insufficient. To make better resolution the cubic spline interpolation was performed on the all data sets. It is a piecewise continuous curve, passing through each of the values in the table. We start from a table of points $[x_i, y_i]$ for $i=0,1,\ldots,n$ for the function $y=f(x)$, which makes $n+1$ points and n intervals between them. There is a separate cubic polynomial for each interval, containing its own coefficients:

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

Together, these polynomial segments are denoted *S(x)*, the spline. Since there are *n* intervals and four coefficients for each we require a total of *4n* parameters to define the spline *S(x)*. We need to find *4n* independent conditions to fix them. Two conditions for each interval are taken from the requirement that the cubic polynomial match the values of the table at both ends of the interval:

$$S_i(x_i) = y_{i,}\, S_{i+1}(x_{i+1}) = y_{i+1}$$

These conditions result in a piecewise continuous function.
We still need *2n* more conditions. Since we would like to make the interpolation as smooth as possible, we require that the first and second derivatives also be continuous:

$$S'_{i-1}(x_i) = S'_i(x_i),\ S''_{i-1}(x_i) = S''_i(x_i)$$

These conditions apply for *i=1,2,…,n-1*, resulting in *2(n-1)* constraints. So we need two more conditions to completely fix the spline:

$$S''_0(x_0) = 0, S''_{n-1}(x_n) = 0$$

With *4n* coefficients and *4n* linear conditions it is straightforward to work out the equations that determine them. The conditions can be reduced easily to a tridiagonal system with the coefficients c_i as unknown variables. Once solved, the remaining coefficients are easily determined. Cubic splines are popular because they are easy to implement and produce a curve that appears to be seamless. As we have seen, a straight polynomial interpolation of evenly spaced data tends to build in distortions near the edges of the table. Cubic splines avoid this problem, but they are only piecewise continuous, meaning that a sufficiently high derivative (third) is discontinous. If the application is sensitive to the smoothness of derivatives higher than second, cubic splines may not be the best choice.
The second step was the data set smoothing. It was necessary, because there is a lot of noise (represented by small extremes on the chart). The noise appeared when the probe hit the rock. In our experiment the best results were obtained using the Gaussian filter. It can be described by the following equation:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

The Gaussian function is non-zero for $x \in (-\infty, \infty)$ and would theoretically require an infinite window length. However, since it decays rapidly, it is often reasonable to truncate the filter window and implement the filter directly for narrow windows, in effect by using a simple rectangular window function. In our algorithm we calculated the optimal windows size in experimental way. It should be equal to 150. The figure 5 presents the Gaussian window of this size.

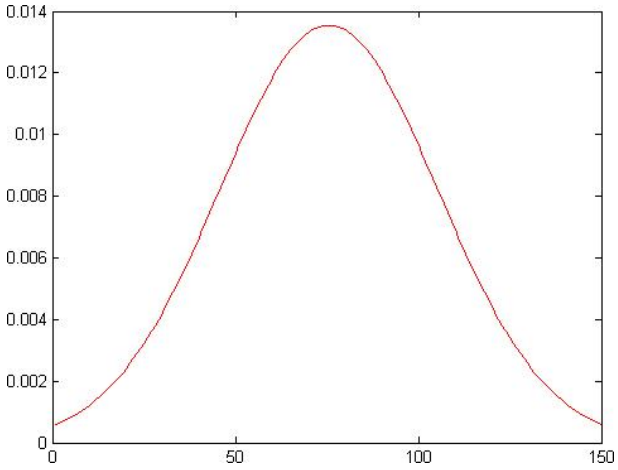
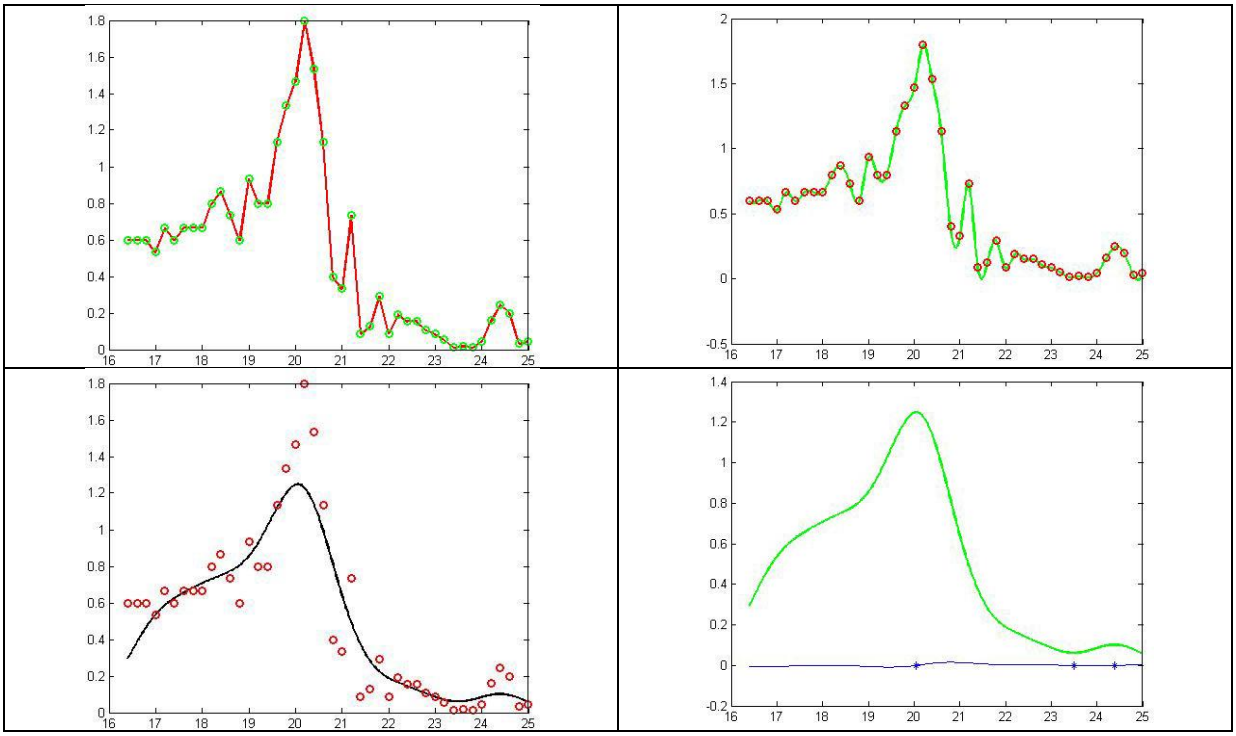


Figure 5. The Gaussian window

The results of smoothing by Gaussian filter and cubic spline interpolation are presented in Fig. 6. To find the borders of soil profiles we checked the zero crossing of the first derivative:

$$S'(x) = 0$$



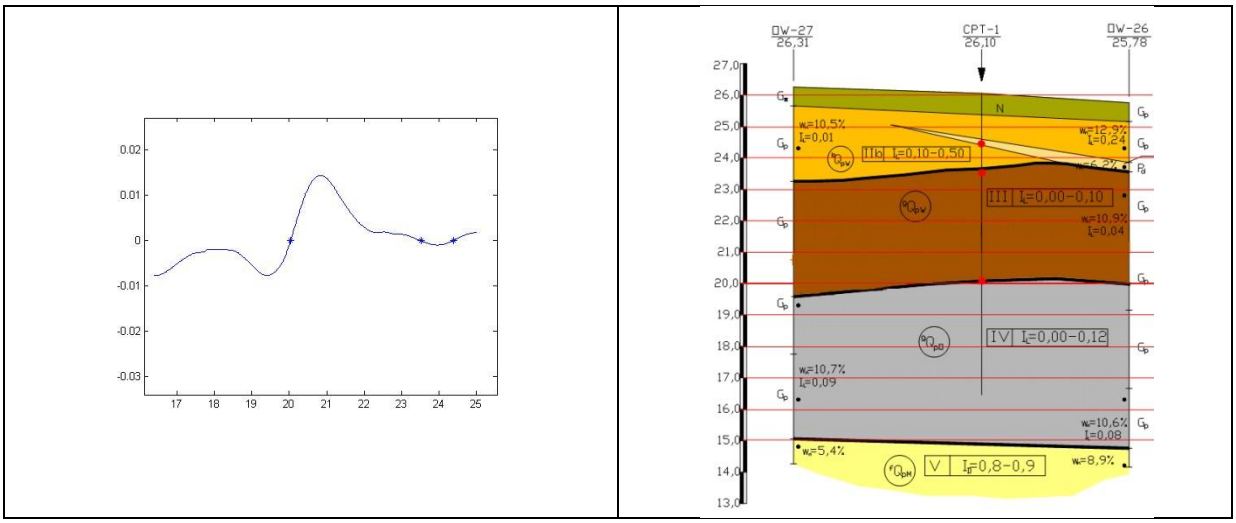


Figure 6. Results of the CPT measurements a) result of cubic interpolation b) Gaussian filtering c) Gaussian filtering and its derivative (blue line) d) first derivative and its zero crossing e) geotechnical cross-section

It is easy to observe in Fig. 6 that the first border of two types of soils was avoided. It was caused by the measurement which was started at depth equals 1 meter.
To compare described method we used a few clustering method with automatic estimation of the number of clusters. In these methods each cluster should represent the different soil category. To perform cluster analysis we used *k*-means algorithm and Silhouette, Calinski-Harabasz and Hartigan algorithms (M. Yan 2005) were used to obtain the number of clusters. Each cluster should represent one soil profile. The maximum and minimum value of the depth of the points in the cluster should describe the range of the soil profile.

3.1 K-means clustering

The well known *k*-means clustering is a typical example of partitioning techniques. It has become one of the most popular clustering methods because it is computationally easy to implement and is generally accessible in the statistical software and clustering packages. Various algorithms have been developed to search for the optimal partition of the *k*-means clustering, which are frequently referred to as *k*-means algorithms since it involves the calculation of the mean (centroid) of each cluster. We only introduce the *k*-means algorithm which we used. In (Forgy 1965) suggested a *k*-means algorithm consisting of the following steps (M. Yan 2005):

- Start with *k* randomly-selected initial centers (seed points). Obtain the initial partition by assigning each object to its closest center.
- Recompute the centroids with the current arrangement of objects.
- Assign each object to the cluster with the nearest centroid. The centroids remain unchanged for an entire pass through the set of objects.
- If no movement of an object has occurred during a pass, stop. Otherwise, repeat step 2 and step 3.

3.2 Silhouette statistic

Kaufman and Rousseeuw (Kaufman and Rousseeuw 1990) proposed the silhouette index as to estimate the optimum number of clusters in the data. The defnition of the silhouette index is based on the silhouettes introduced by (Rousseeuw), which are constructed to show graphically, how well each object is classified in a given clustering output. To plot the silhouette of the *m*-th cluster, for each object in C_m, calculate *s(i)* as:

$$b(i) = \min_{C \neq C_m} d(i, C)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$is the average dissimilarity of object *i* to all other objects in the *m*-th cluster and $d(i, C)$ = is the average dissimilarity of object *i* to all other objects in cluster $C, C \neq Cm$

The silhouette index, denoted by $\dot{s}(g)$, is defined as the average of the $s(i)$ for all objects in the data. $\dot{s}(g)$ is called the average silhouette width for the entire data set, reflecting the within-cluster compactness and between-cluster separation of the clustering. Compute $\dot{s}(g)$ for $g = 1, 2, ...$ The optimum value of g is chosen such that $\dot{s}(g)$ is maximized over all g:

$$G = \arg\ max_g\, \dot{s}(g)$$

3.3 Calinski and Harabasz's method

This approach determines ^G by maximizing the index $CH(g)$ over g, where $CH(g)$ is given by

$$CH(g) = \frac{B(g)/(g-1)}{W(g)/(n-g)}$$

and $B(g)$ and $W(g)$ are the between- and within-cluster sum of squared errors, calculated as the trace of matrix *B* and *W*, respectively. $CH(g)$ is only defined for g greater than 1, since $B(g)$ is not defined when g = 1. Here W and B are defined as follows:

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x}_m)((x_{ml} - \bar{x}_m)',$$

$$B = \sum_{m=1}^{g} n_m (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})', \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where $\bar{x}_m$ is the cluster mean, g is the number of the clusters.

3.4 Hartigan's method

Hartigan (Hartigan 1975) proposed the following index

$$Har(g) = \left[\frac{W(g)}{W(g+1)} - 1\right]/(n - g - 1)$$

Intuitively, the smaller the value of $W(g)$, the higher similarity between objects which have the same cluster memberships. For the fixed values of g and $W(g)$, $Har(g)$ will be sufficiently large if and only if $W(g + 1)$ is sufficiently small. Thus, the idea is to start with $g = 1$ and to add a cluster if $Har(g + 1)$ is signifficantly large. The distribution of $Har(g)$ can be approximated by the F-distribution, which provides an approximated cut-off point. A simpler decision rule suggested by Hartigan is to add a cluster if $Har(g) > 10$. Hence, the cluster number is best estimated as the smallest g, $g = 1, 2, ...$, such that $H(g) \leq 10$.

4. Results of experiments

The described methods were compared with the soil profiles generated in the laboratory from the boreholes. In the optimal situation both profiles should be identical. Table 1 presents the result of numerical experiments and compares it with boreholes.

Table 1: Experimental results

n_{sp} / loc	BHT	Alg	Silhouette	CaH	Hartigan
CPT 1	6	5	3	8	8
CPT 2	8	7	2	7	4
CPT 3	6	**6**	5	10	8
CPT 4	7	6	4	10	6
CPT 5	7	**7**	2	9	5
CPT 6	5	**5**	2	10	5
DMT 1	6	**6**	2	10	8
DMT 2	7	**7**	2	5	6
DMT 3	7	**7**	2	8	7
DMT 4	7	6	3	10	7

In Table 1 n_{sp} is the number of soil profiles, "BHT" is the bore hole test, "Alg" is our described algorithm based on interpolation and first derivative, "CaH" – Calinski and Harabsz's method, "loc" – localization of the probe test. It is easy to observe that Silhouette and Calinski and Harabasz's method are useless in this experiment. The best results were obtained by our algorithm. Only Hartigan gave similar (but still worse) results. Generally, dissimilarities between our algorithm and reference method (bore hole test) are caused by two factors – the first one is that probe test started always from one meter depth. All layers above were avoided. The second one is the resolution of the probe test. The measurements were saved each 0.2 m. In the case of thin layer it may be avoided or treated as noise. Interesting fact is that all differences are equal one which was caused by described two factors.

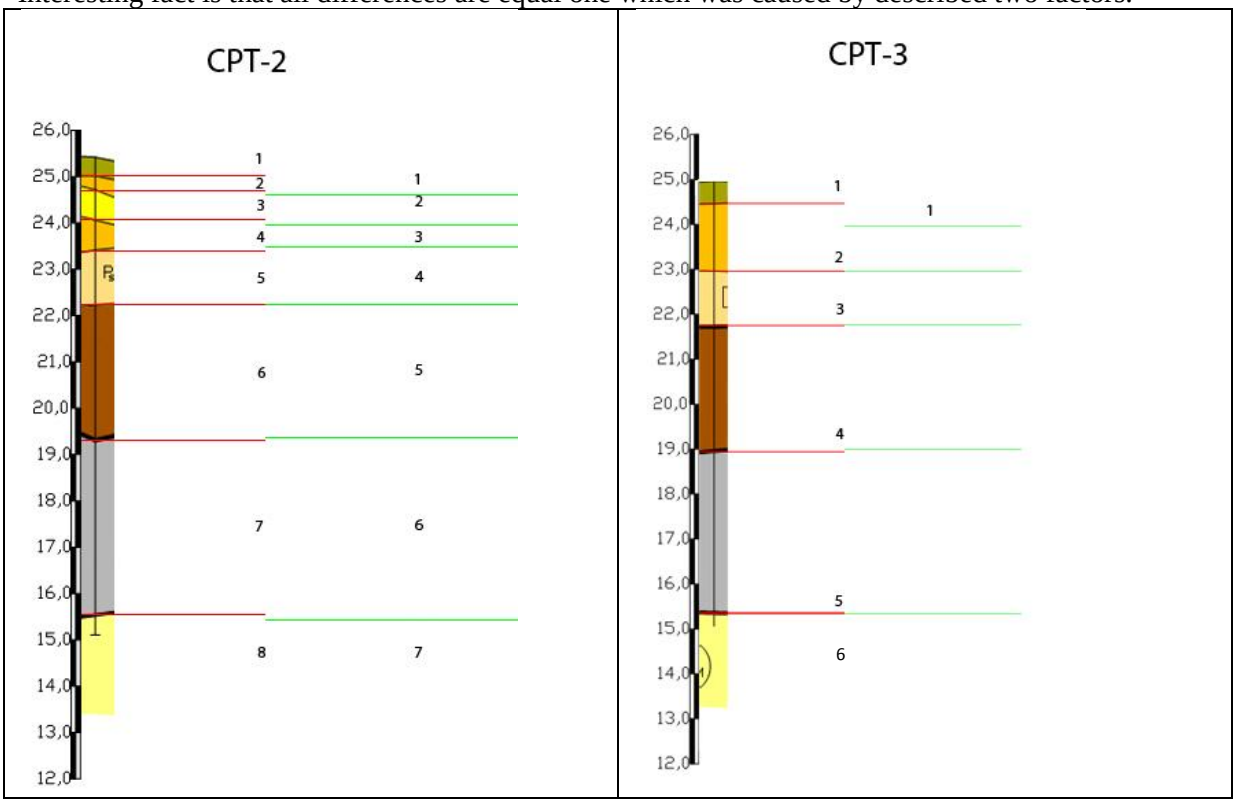


Figure 7. Results of experiments. Red lines refers to borehole test (our reference metod), green line refers to our algorithm

The Figure 7 shows the graphical results of experiments. It is easy to observe that measurements started from 1 meter depth. The charts shows that our algoirthm is more accurate when the profiles are wider.

5. Conclusions

The proposed method appears to be useful for the automatic generation of the soil profiles. The ones obtained after applying the clustering algorithm to the data from two different probes are comparable. The proposed methodology is an attractive alternative for the geotechnical engineers, replacing the nomograms and supplementing boreholes. The open question is the selection of the correct number of categories to catch the most important ones and the versatility of the approach. It must be checked if the knowledge extracted from one site is usable in other, i.e. it can be used to classify soils in other locations as well.

ACKNOWLEDGEMENTS

This work is supported by the Polish National Centre of Science, Grant No. 2011/03/D/ST8/04309.

References

Forgy E. W. (1965). Cluster analysis of multivariate data: Effciency vs. interpretability of classification. *Biometrics*, 21:768-769

Hartigan, J.. Clustering Algorithms. New York: Wiley, 1975.

Hashash, Y.M.A., Jung, S., and Ghaboussi, J. (2004). Numerical implementation of a neural network based material model in finite element analysis, *International Journal for Numerical Methods in Engineering*, *59*, 989-1005.

Huang A., Mayne P. W. (2008). Geotechnical and Geophysical Site Characterization. *Proc. of the 3 rd inter. Conf. on Site characterization*, Taipei, Taiwan. Published by: Taylor & Francis Group, London, UK

Lunne, T., Robertson, P.K., Powell, J.M. (1997). Cone penetration testing in geotechnical practice. *Blackie Academic and Professional*, London, England

Marchetti S. (1980). In Situ Tests by Flat Dilatometer. J. *Geotech. Eng*. Div., ASCE, 106, GT3, 299-321.

Młynarek Z. (2007). Site investigation and mapping in urban area. *Proc. of the 14th European Conference on Soil Mechanics and Geotechnical Engineering*. Madrid, Vol. 1, 175-202.

Rabarijoely S., Bilski P., Falkowski T. (2007). The usage of the graph clustering algorithm to the recognition of geotechnical layers. *Annals of Warsaw University of Life Sciences – SGGW. Ann. Warsaw Univ. of Life Sciences – SGGW, Land Reclam., No 38, 2007*, 57 - 68.

Rabarijoely S., Bilski P. (2009). Automated soil categorization using CPT and DMT investigations, *2nd International Conference on New Developments In Soil Mechanics and Geotechnical Engineering, 28-30 May 2009, Near East University, Nicosia, North Cyprus*

Shahin, M.A., Jaksa, M.B., and Mier, H.R. (2005). Neural network based stochastic design charts for settlement prediction, *Can. Geotech. Jour*. (42), 110-120.

Totani G., Marchetti S. , Monaco P. & Calabrese M. (2001). Use of the Flat Dilatometer Test (DMT) in geotechnical design, IN SITU 2001, *Intnl. Conf. On In situ Measurement of Soil Properties*, Bali, Indonesia

Yan M. (2005). Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion, *PhD Dissertation*, Virginia Polytechnic Institute