AUTOMATIC ESTIMATION OF THE NUMBER OF SOIL PROFILE LAYERS USING BAYESIAN INFORMATION CRITERION

Jarosław Kurek, Michał Kruk, Piotr Bilski and Simon Rabarijoely and Bartosz Świderski

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
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
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

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

Abstract

In this study, Bayesian Information Criterion algorithm is utilized for the estimation of number of soil profile layers. In order to collect data, several probes are performed by geotechnical specialists in Warsaw University of Life Sciences (WUoLS) campus. Then soil profiles have been manually generated by geotechnical experts. It lets us to compare the results of novel automated method presented in this paper to real soil profile manually generated by geotechnical engineers. The database has been generated based on values derive from a probe CPT applied by geotechnical experts. Examination and accuracy calculation of the proposed method is presented and compared to reference real soil profile obtained by experts group.

Keywords: *geotechnical probes, soil categorization, BIC, Bayesian Information Criterion, clustering*

1. Introduction

Soil is a useful building material because it has the shear strength that can itself and other loadings. Otherwise, the same material may become very weak that it can no longer support itself and it can fail. Geotechnical specialists should predict the loading on a soil, its strength and determine whether it will be safe building construction. It is mean that to have all required documents, permits (e.g. building permit) to start building process of special construction, examination of soil by geotechnical engineers is mandatory. The regular way to assess soil profile in order to obtain all mandatory documents required to start building process is to perform traditional method, requiring drilling boreholes to collect soil samples which is well know method and good practice currently but long and expensive. This regular method is widely used and is based on analyzing charts of

collected data by the geotechnical experts manually (Marchetti 1980). They can analyze samples in laboratory to obtain satisfied accuracy but it takes much time (weeks) and of course is quite expensive. To reduce time and cost of operation, additional devices are used such us Cone Penetration Test (CPT) probe and The Flat Dilatometer Test (DMT) probe. These probes let us to reduce the number of requiring drilling boreholes but provide additional variables obtained from mentioned probes. The soil analysis becomes faster and cheaper than the traditional method, requiring drilling boreholes. But the main issue of the approach using probes CPT and DMT is human expert's knowledge and database (charts) to assign values to appropriate type of soil. The Database should be relevant to the place when drilling is located, because a lot of countries have different type of soil layers and values derived from probes CMT and DMT should be locally calibrated which is enormous hard process (expensive and long). To build database (charts) covered e.g. Poland multiple probing are required. But it would take too much time and is too expensive. The new approaches methods of classification, based on the expert's knowledge (Zhang and Tumay (1996)) are proposed. But when the latter is unavailable, additional approaches of extracting information from the measurement data sets are required.

Computer algorithm is able to create soil profile layers and classifies the layers to appropriate soil type. If we create that algorithm the output can be analyzed by geotechnical specialists to verify accuracy of generated profile. Created in this way database can be used as the soil identification module for further geotechnical system. Similar works were done before (Hashash et al. (2004), Shahin et al. (2005)), but new approaches must be proposed.

The measurement data have been acquired from Warsaw University of Life Sciences campus (WUoLS) during expansion of the university. Before the university obtain building permit to start build the new campus building objects, the multiples tests have been performed. Geotechnical specialists collect measurement data from CPT and DMT probes and also using traditional approaches such as drilling boreholes. The latter let us to treat as the reference method to compare with new the novel method presented in this paper and calculate the accuracy ratio.

2. Generate database

In this paper only data derive from CPT probe have been concerned and take into consideration in numerical tests. 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 CPT probe is depicted in the Figure 1.

Figure 1 View of CPTU tip (Mayne et al., 1995) (a) and DMT blade (b) used in the situ tests

The cone penetration test (CPT) is a method used to obtain the geotechnical engineering properties of soils and delineating soil stratigraphy. Mayne (2007) said that the cone penetration test soundings can be used either as a replacement or complement to conventional rotary drilling and sampling methods. CPT probe has undergone tremendous development in recent years as an in-situ site investigation tool. CPT test is the common used method of in-situ soil testing (Brouwer, 2007). The main advantage of used CPT in comparison to regular drilled boreholes method is fact that CPT test is performed without disturbing the ground, it provides information about soil type, geotechnical parameters like shear strength, density, elastic modulus, rates of consolidation and environmental properties.

The CPT test is based on pushing a cylindrical steel probe into the ground at a constant rate of e.g. 20 mm/s and measuring the resistance to penetration. The standard penetrometer has a conical tip with 60° angle apex, 35.7-mm diameter body. During CPT tests 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 variables in this paper.

The data input have been acquired from Warsaw of University of Life Sciences university campus. Based on collected data soil profiles have been made by geotechnical specialists. There were both drilling boreholes and CPT and DMT tests. Created soil profiles in form of cross-section charts are very helpful as reference to check the accuracy of presented novel approach. Measurement data have been gathered from two main places where the two new building currently exists (number 34 and 37). Hence the input data file derives from these mentioned buildings has codename CPTn_34 and CPTn37 where n is number of test in place when appropriate buildings have been set up. Every input data file contains data set in form of nxm matrix, where n is a number of rows (depth resolution) and m is a number of columns. The first column represents depth variable and rest of columns represent values of the measured parameters. Structure of a typical file is depicted in Table 1.

Table 1: Example of data input file No. CPT1_34

Depth [m]	**q_c [MPa]**	**f_s [MPa]**	**R_f [%]**
1	0.8	0.046667	5.833
1.2	1.3	0.033333	2.564
1.4	10	0.2	2.0
1.6	1.9	0.2466667	12.982

1.8	0.6	0.16	26.667
2.0	0.4	0.046667	11.667
…	…	**…**	**…**

In the Table 2 all data file codenames have been presented which have been taken into consideration during numerical experiment presented in this study.

Table 2: Codename of input data.

No.	Chart name	Filename	Number of soil layer
1	Model 34_pi-i	CPT2_34	5
2	Model 34_pii-ii	CPT4_34	4
3	Model 34_piii-iii	CPT1_34	3
4	Model 34_pniv-iv	CPT3_34	4
5	Model 34_pvi-vi	CPT5_34	4
6	Model 37_pi-i	CPT4_37	5
7	Model 37_pii-ii	CPT1_37	5
8	Model 37_piii-iii	CPT2_37	5
9	Model 37_piv-iv	CPT5_37	5
10	Model 37_pv-v	CPT4_30	5
11	Model 37_pvi-vi	CPT3_30	3

Example of created soil profile is depicted in Figure 2. Based on below chart a few soil layers can be observed. According the geotechnical experts' knowledge, the tiny layers should be removed because it probably is caused by obstacles (probe meets the rock).

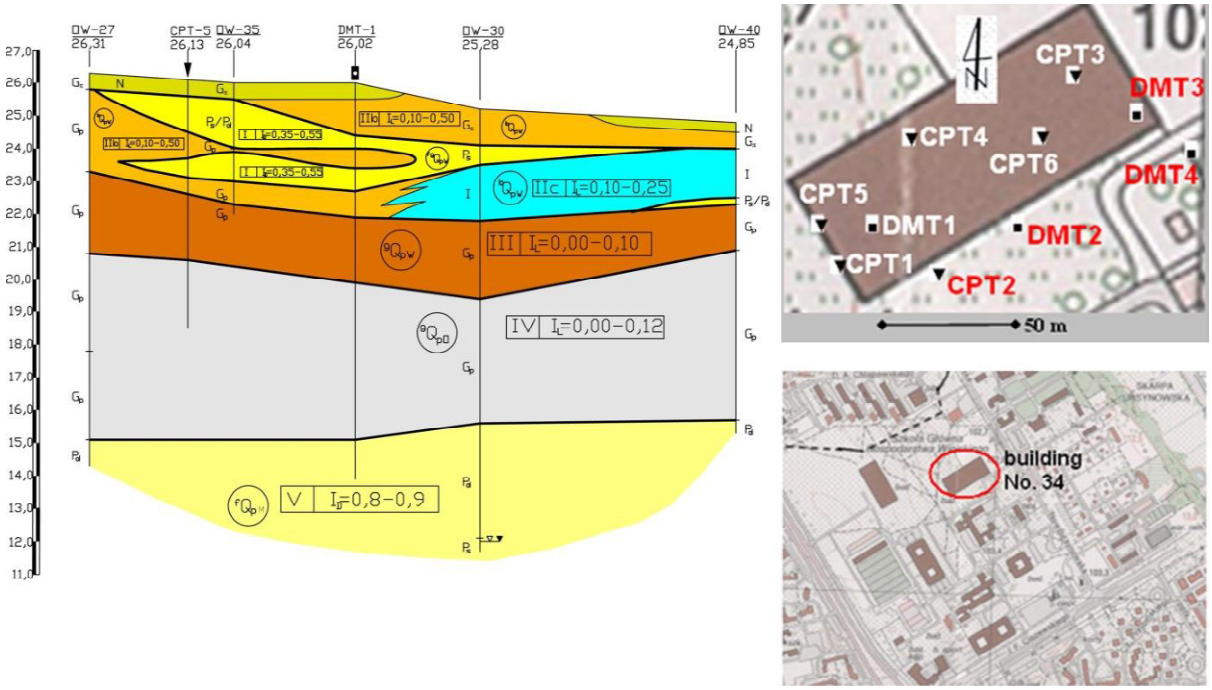

Figure 2 A typical geotechnical cross-section chart: OW – borehole, CPT – cone penetration test, DMT – Dilatometer test; (N – fill, Gp – sandy clay, Pd –fine sand, w_n – moisture content, I_D – relative density, I_L –liquidity index)

2. Algorithm of soil layers number estimation

To build special unsupervised module which would automatically generate the soil profile based on data measurement derives from CPT and DMT probes, two stages are required. The first stage is based on estimation how many soil layers exist in place when CPT and DMT have been applied. The second stage is regarding to determine to which geotechnical soil type belongs appropriate layer. In this paper the first stage has been analysed. In presented algorithm number of soil layers reflects the number of clusters found in database.

If we do not know a priori the number of clusters (soil layers), we can ask how many clusters are needed. One method to choose this number of clusters is to use the minimum value of the Bayesian Information Criterion (BIC) (Kaufman, 1990) which is employed in this study. The BIC is derived from Bayes' theorem (Stigler 1982) and is used to determine which probability-based mixture model is the most appropriate.

The calculation BIC score is based on generation of the probability density function of the d-dimensional gaussian distribution which is presented below:

$$y = f(x,\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} e^{-\frac{1}{2}(x-\mu)\,\Sigma^{-1}(x-\mu)'}$$

where x and μ are 1-by-d vectors and Σ is a d-by-d symmetric positive definite matrix. While it is possible to define the multivariate normal for singular Σ, the density cannot be written as above. Only random vector generation is supported for the singular case. Note that while most textbooks define the multivariate normal with x and μ oriented as column vectors, for the purposes of data analysis software, it is more convenient to orient them as row vectors. Example of 2-dimensional gaussian distribution for 100 random points is depicted in the Figure 3. There were two entries parameters applied to generate Figure 3:

$$mean = [1 \quad 3]$$

$$\text{variance} - \text{covariance matrix} = \begin{bmatrix} 5 & 1 \\ 1 & 3 \end{bmatrix}$$

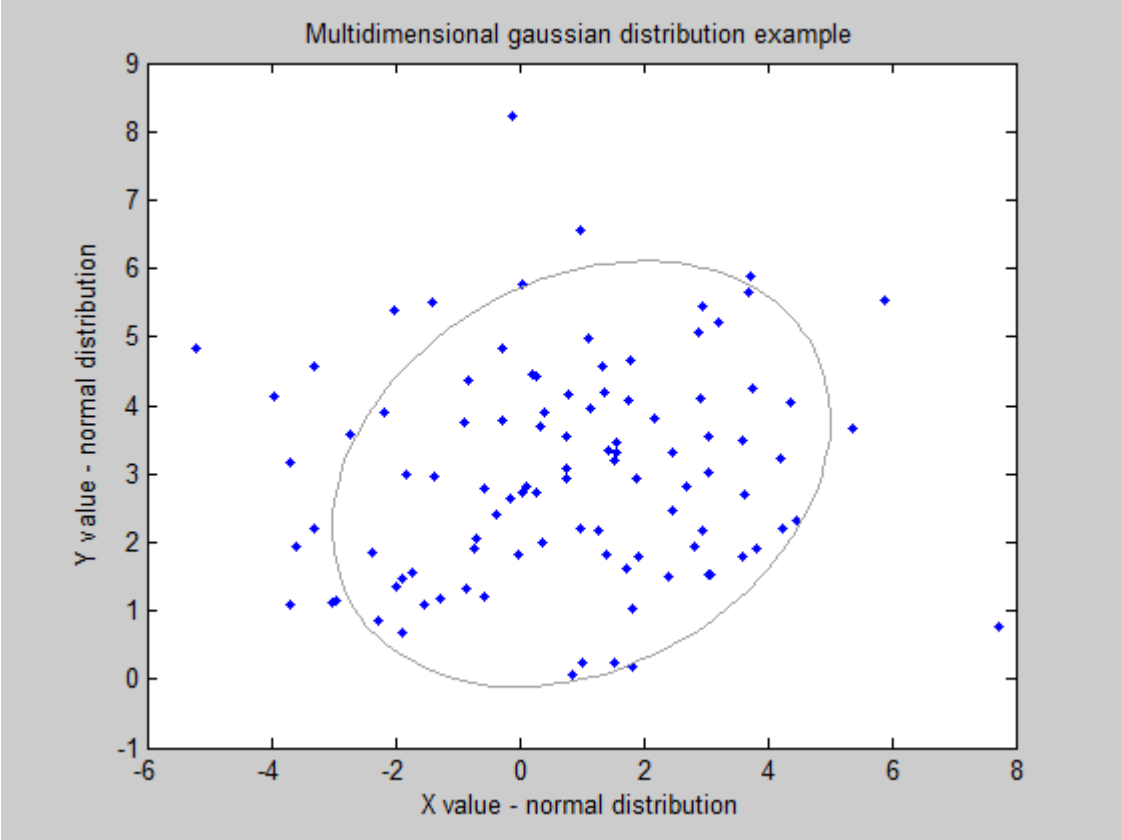

Figure 3 : Multidimensional gaussian distribution example.

We can apply reverse approach, based on input data we have tried to find the Maximum Likelihood (ML) estimator of the mean and variance parameters. The ML estimator of the covariance is biased, but this bias is small (of order 1/n). Example of that approach is depicted in the Figure 4. Based on previous generated data the estimator of the mean and variance parameters are the following:

$$mean = [1.28 \quad 2.93]$$

$$\text{variance} - \text{covariance matrix} = \begin{bmatrix} 6.69 & 1.66 \\ 1.66 & 3.2 \end{bmatrix}$$

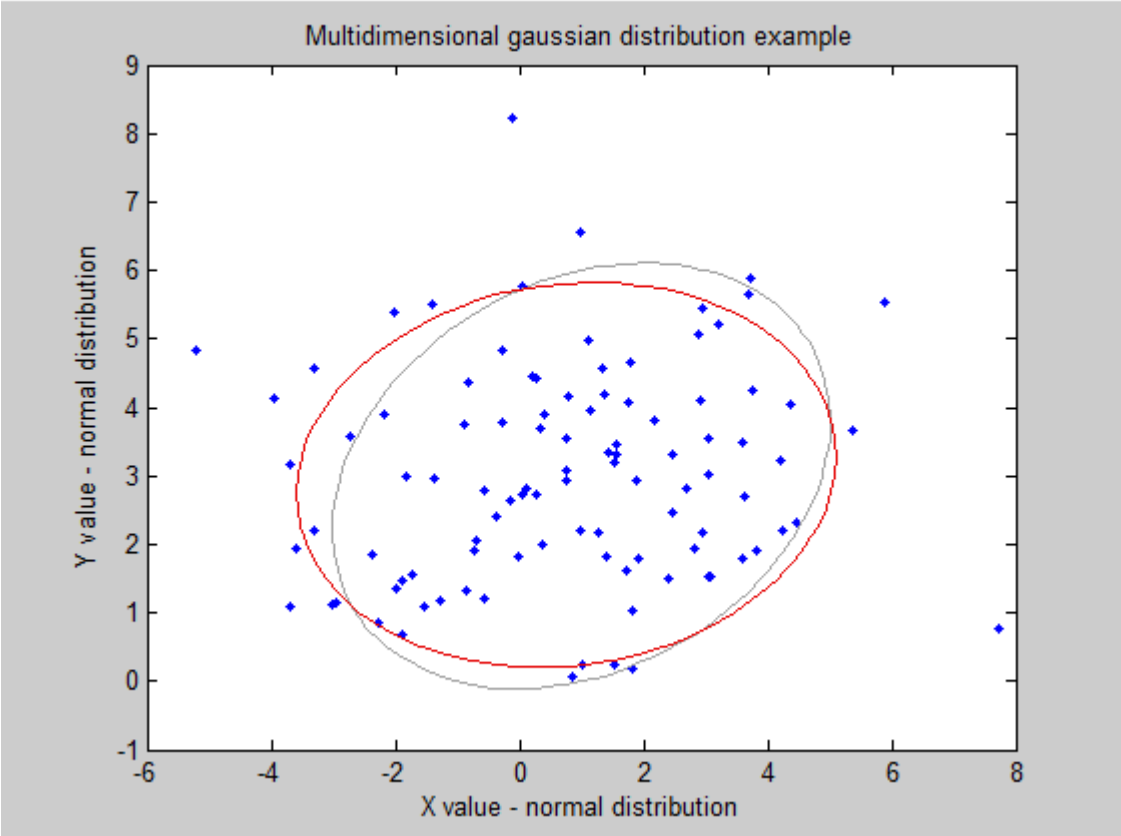

Figure 4 :Model estimation example. (Plot the estimated distribution in red)

In the automatic estimation of number of layers in soil profile approach by means of BIC method. generation of mixture distribution is mandatory. It is build based on a several gaussian distributions which are mixtured in some proportions e.g. will be set respectively to 40% and 60%. Example of mixture distribution for 3 distributions are depicted in the Figure 5. Initially default proportion was 33% for each distribution but after esitmated distribtution will not significantly different.

The general formula to estimate number of soil layers based on CPT test i the following (Raftery 1982):

$$\mathrm{BIC} = -2\ln(\mathtt{L}) + v\ln(n)$$

where n is the number of data points, $\mathtt{L}$ is the likelihood of the parameters to generate the data in the model and v is the number of free parameters (means and standard deviations) in the gaussian mixture model. The BIC takes into account both the fit of the model to the data and the complexity of the model. A model that has a smaller BIC is preferred. It means that the smallest BIC value is the preffered number soil profile layers for the input data.

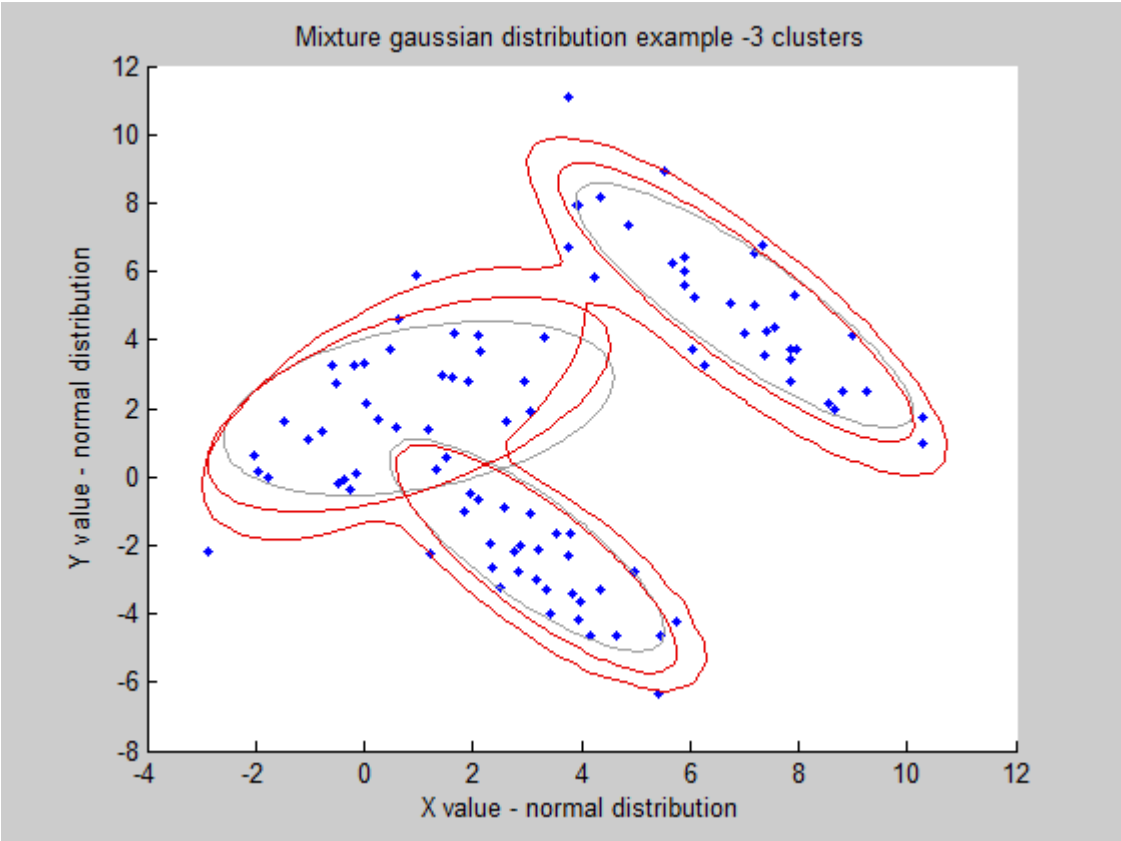

Figure 5 :Model estimation for 3-dimensional mixture distribution example. (Plot the estimated distribution in red)

To clarify the approach presented in this paper psudo code is depicted in the Figure 6 :

```
procedure EstimateLayers
input CPT_X
maxLayers=5
begin
        for k=1 to maxLayers
           mikstureDensityObject=GenerateGaussianMixtureDensity(k);
          AllModelsCollection[k] = learn(mikstureDensityObject, CPT_X);
          BICScoreCollection[k] = bic_score(AllModelsCollection[k], CPT_X);
        end
output=min(BICScoreCollection);
end
```

Figure 6 :Estimation of number of soil profile layers algorithm.

3. Numerical experiments

Based on presented algorithm 11 numerical experiments have been performed. The results have been presented in Table 3. The results are optimistic: for 11 numerical experiments algorithm estimates correctly 9 models. It means that input accuracy ratio is 81%. Geotechnical experts confirm that algorithm would be applied as a module in some geotechnical software to estimate number of soil layers in soil profile but they also are sceptic that input data was constrained to only 11 cases so very hard to say about future application in other places in Poland.

Table 3: Result of numerical experiments

No.	Chart name	Filename	Referenced No of soil layers	Calculated No of soil layers
1	Model 34_pi-i	CPT2_34	5	5
2	Model 34_pii-ii	CPT4_34	4	4
3	Model 34_piii-iii	CPT1_34	3	3
4	Model 34_pniv-iv	CPT3_34	4	4
5	Model 34_pnvi-vi	CPT5_34	4	4
6	Model 37_pi-i	CPT4_37	4	4
7	Model 37_pii-ii	CPT1_37	5	5
8	Model 37_piii-iii	CPT2_37	5	5
9	Model 37_piv-iv	CPT5_37	5	5
10	Model 37_pv-v	CPT4_37	5	4
11	Model 37_pvi-vi	CPT3_37	4	5

In the Figure 7 and 8 the results of applied algorithm have been depicted for filename CPT3_34, CPT5_34, CPT4_37, CPT1_37.

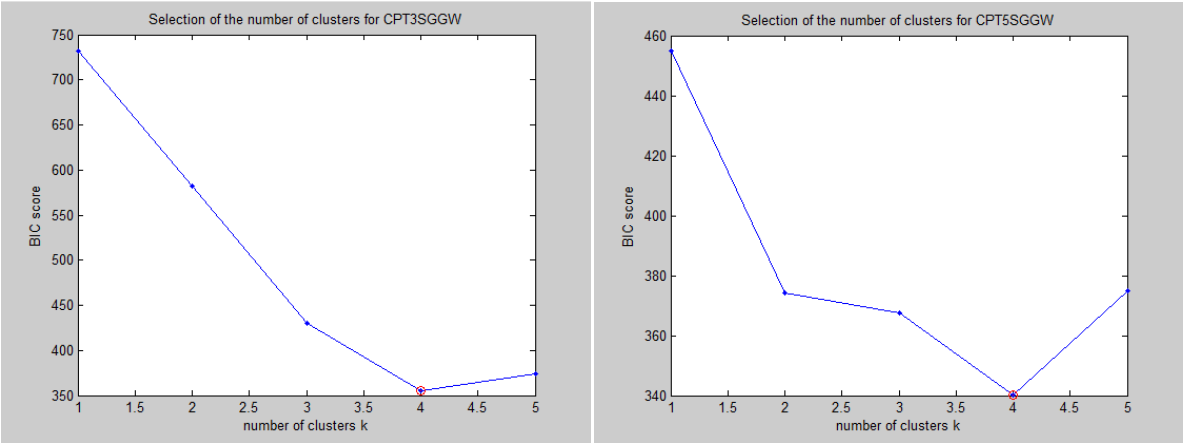

Figure 7 : The results of applied algorithm for filename CPT3_34 (CPT3SGGW), CPT5_34 (CPT5SGGW)

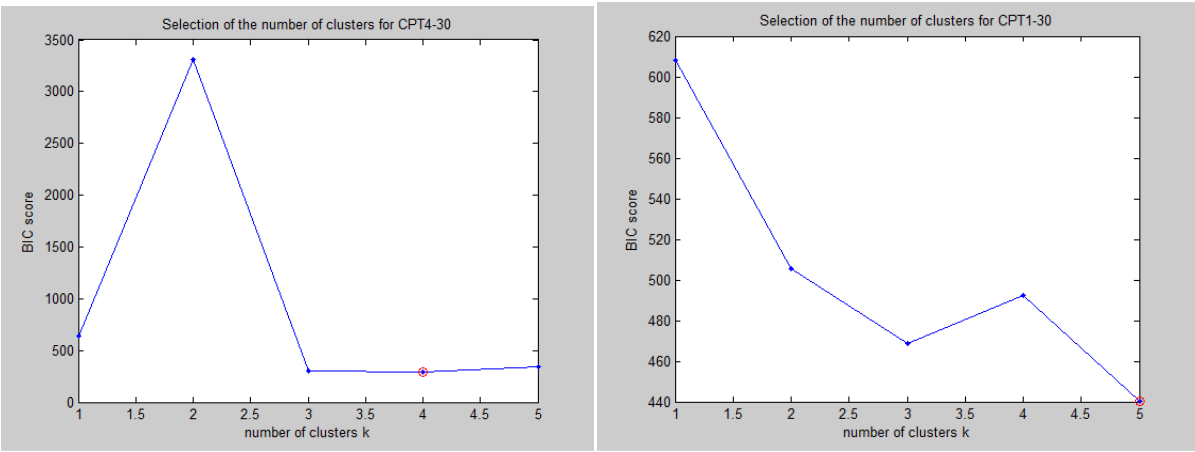

Figure 8 : The results of applied algorithm for filename CPT4_37 (CPT4-40), CPT1_37 (CPT1-40)

4. Conclusion

The proposed method appears to be useful for the automatic generation of the soil layers number. The proposed method has 81% accuracy ratio which is confirmed by geotechnical experts that could be applied as of one of the module in the system to generate soil profile in automatic way. But this approach should be verified by more than 11 cases to obtain more reliable accuracy ratio.

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

References

Brouwer, J. J. M. (2007). In-Situ Soil Testing. East Sussex: Lankelma
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
Kaufman L., Rousseeuw P.. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. 1990.
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*
Raftery A.. A note on Bayes factors for log-linear contingency table models with vague prior information. Journal of the Royal Statistical Society. 48(2): 249-250. 1986.
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.
Stigler. S. M. Thomas Bayes' Bayesian Inference. Journal of the Royal Statistical Society, Series A. 145: 250–258. 1982.
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*
ZHANG Z. and TUMAY M. 1996: Simplification of soil classification charts derived from the cons penetration test, Geotechnical testing Journal, Vol. 19, No. 2, pp. 203-216.
ZHANG Z. and TUMAY M. 1999: Statistical to fuzzy approach toward CPT soil classification, Journal of Geotechnical and Geoinveronmental Engineering, Vol. 125, No. 3, pp. 179-186.