Geologia Croatica
Journal of the Croatian Geological Survey and the Croatian Geological Society

2017 | 70/2 | 73–78 | 7 Figs. | 3 Tabs. | www.geologia-croatica.hr

Cluster defined sedimentary elements of deep-water clastic depositional systems and their 3D spatial visualization using parametrization: a case study from the Pannonian-basin

Janina Horváth, Szabolcs Borka and János Geiger

University of Szeged, Department of Geology and Paleontology, Hungary; (th.janina@geo.u-szeged.hu)

doi: 10.4154/gc.2017.06

Crossref Similarity Check

Article history:

Manuscript received January 31, 2017
Revised manuscript accepted April 24, 2017
Available online June 28, 2017

Abstract

Many multivariate statistical techniques have the ability to handle large data sets or a great number of parameters. Therefore, these multivariate statistical approaches are widely used in clastic sedimentology for facies analysis. Furthermore, most of the techniques which try to separate more or less homogeneous subsets can be subjective. This subjectivity raises several questions about the significance and confidence of clustering. The goal of this study is to optimize clustering and to evaluate the proper number of clusters needed in order to describe sedimentary and lithological facies through common characteristics. Also, with the interpretation of the clusters, the parametrized geometry adds further but quasi-subjective information to a 3D geological model. Two assumptions must be met: (1) well-definable geometries must correspond to the architectural elements (2) it is assumed that exactly one sedimentary or lithological facies belongs to each structural element and the flow properties are determined by these structural elements. This approach was applied to the clastic depositional data from a Miocene hydrocarbon reservoir (Algyő field, Hungary) to demonstrate the fidelity of the clustering method yielding an optimum of five cluster facies. The revealed clusters represent lithological characteristics within a (delta fed) submarine fan system. The paper deals with two stressed clusters in particular, showing sinusoid channels which were recognizable and measureable using parametrisation.

Keywords: cluster analysis, deep-water depositional system, geo-object, optimised clustering

1. INTRODUCTION

The goal of the study is to identify genetically similar depositional units by separating them with a clustering technique. Besides, the study focused on the optimization of the separated sedimentary elements by analysing the optimal number of clusters.

These separated units reflected in particular the lithological and petrophysical properties. Moreover, the analysed rock body does reflect that in the lithification stage of sandstone diagenesis, the applied petrophysical properties were still determined by the depositional genetics. One of the most important consequences of this finding is that the separated units are able to represent depositional facies with some additional parametrized geometry information about spatially extended clusters.

The cluster units may be the „cornerstones" as structural elements of a 3D-facies model. During the spatial visualization, the goal was to use methods which could handle and honour the geometries of depositional structural elements. The parametrized geometry adds an extra but quasi-subjective information to this 3D geological model. During the clustering two assumptions must be met: (1) well-definable geometries must correspond to the architectural elements (2) it is assumed that exactly one sedimentary or lithological facies belongs to each structural element and the flow properties are determined by these structural elements.

This paper demonstrates the method through a case study. The study area is located in the Algyő sub-basin of the Pannonian-basin geographically belonging to the Great Hungarian Plain. According to the paper by (GRUND & GEIGER; 2011; BORKA, 2016) this study area was characterized as sequences representing a prodeltaic submarine fan (Fig. 1).

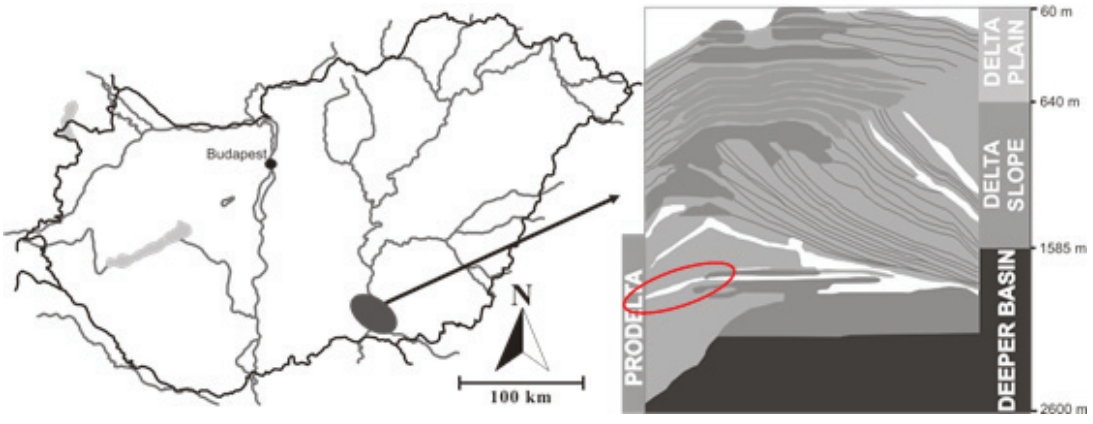

Figure 1. The Algyő delta sequences macrosedimentary model (based on BÉRCZI, 1988).

This paper deals in particular with the details of the optimization of the clustering results and their spatial interpretation. Hence, the method chapter comprises the following: the parameters used, distribution of the data, the problems during the clustering such as the correlation between different parameters, how to define the proper number of separated clusters and the interpretation of spatial clusters.

Why does this paper focus on these problems? Usually it is complicated to determine the adequate number of clusters since the most essential parameter of clustering algorithms is to determine the number of clusters and the validity of the clustering. Clustering is an unsupervised technique so the researcher has little or no information about the number of clusters. At the same time, the number of clusters is a required parameter so this is a general problem as old as cluster analysis itself. Of course, geological knowledge about the field and information about the core samples can give a rough number of clusters. In addition, the following questions may arise: does the method have the ability to segregate all groups in the property space or not, are the created subsets adequately „homogeneous" or not? In this case the „homogeny groups" means that the cluster analysis divides data into groups when the main information in the groups is not the description of the linked objects, but rather their relationship GAN et al., 2007. The most common problem is when we separate too many – however homogeneous – groups, and we are not able to label all of them geologically. In contrast, if we have a small number of clusters, they can be relatively too heterogeneous and in this case, it is also hard to define them geologically.

2. METHODS

2.1. Data pre-processing

The clustering technique focused on the determination of lithology and facies based on four variables coming from interpreted well-logs: porosity, permeability, sand and shale contents, and also based on some core samples, which provided additional information. These core samples were also available from one well, which included continuous data from a thickness interval of about 35 metres. These samples acted as sign-posts in the interpretation of cluster results to define lithofacies.

The core analysis was presented by BORKA (2016). According to the core analysis, part of a typical mixed sand-mud submarine fan complex, with quasi-inactive parts (zones of thin sand sheets and overbank), channelized lobes (persistent sandstones in them may denote distributary channels), and a main depositional channel were revealed. However, due to the low number of core samples it was difficult to extend the lithological information to the whole area which contains 141 wells. Hence, the interpreted logs were used to define the lithology types and facies in the case of clustering.

Usually clustering does not require normal transformation but most clustering algorithms are sensitive to the input parameters and to the structure of the data set. The clustering may be more efficient if a good structure exists for the transformed variable, which can approximate the symmetric distribution. It should be close to symmetry prior to entering cluster analysis (TEMPL et al., 2006). Significant skewness could be measured in the distribution of the variables, especially in the shale content and permeability (Fig. 2 base on Eq.1.). A principal component analysis (PCA) was applied as pre-processing for the clustering which also required a normal distribution.

$$y = x^{\alpha} = \begin{cases} \dfrac{(x+\alpha_2)^{\alpha_1} - 1}{\alpha} & \alpha \neq \mathbf{0} \\ log(x+\alpha_2)\ \alpha = \mathbf{0} \end{cases} \qquad \text{Eq.1}$$

Box-Cox transformations (BOX & COX, 1964) of all single variables do not guarantee symmetry of the distribution, but more closeness to them (ASANTE & KREAMER, 2015; TEMPL et al., 2006). The applied transformation is a modification of the power transformation by BOX & COX (1964). This modified power transformation is defined for those cases when variables are negative or equal to zero (Eq.1) (SAKIA, 1992).

Between the porosity (FIAP) and permeability(PERM) variables and also the sand (VSND) and shale volume (VSHA) the correlations were significant (coefficient was 0.82 and -0.71). Hence, the PCA was used to reduce redundancy and create new components (the first component is based on permeability and porosity and the second component is based on sand content and shale content).

The goal of the PCA method was to create new components which are able to preserve as much of the variance of the original variables as possible. Besides this, it was important that the new latent variables are able to combine optimally the weighted observed variables. The first component retained 90.65% of total variance of porosity and permeability and the second component

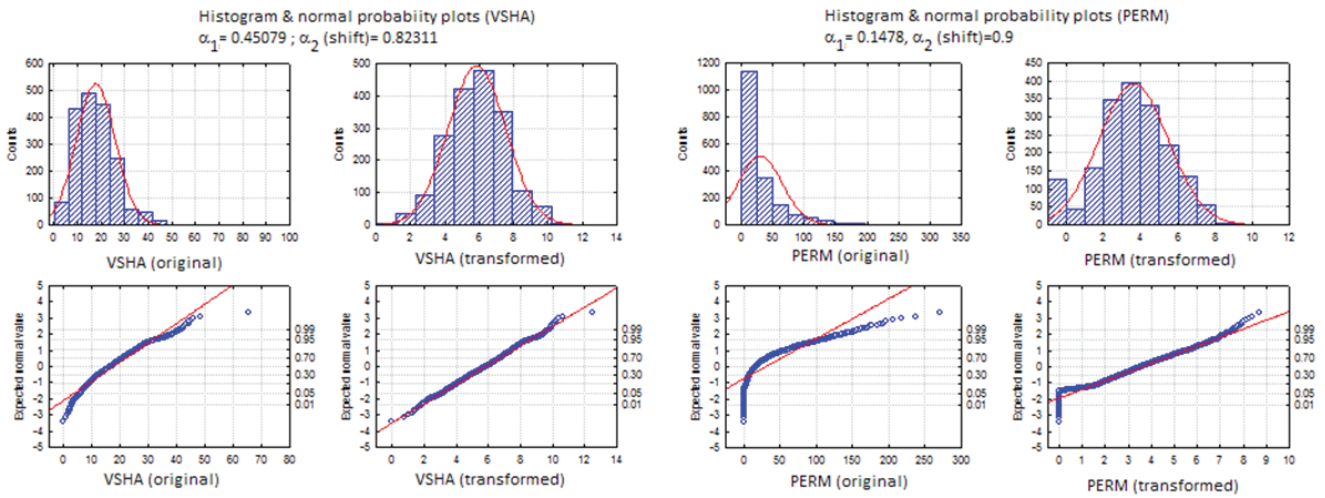

Figure 2. Results of Box-Cox transformation.

also retained a similar large percentage (85.55%) of the total variance of sand and shale volume. The PCA required a normal distributions as well.

The clustering was done on the new PCA components. One of the neural network clustering (NNC) techniques was applied in the separation of the data set. This clustering method was applied because NNC was used in similar problems to characterization of clastic sedimentary environments (e.g. HORVÁTH, 2015; HORVÁTH & MALVIĆ, 2013).

In the initial settings the size of the training set was fixed at 70% for all data points. For the validation and testing, 15-15% of the whole set was used, evenly divided. These three subsets were collected by the network in a random way to avoid bias. The learning rate of NNC clustering converged monotonically in the [0,1] interval from the first to the last training cycle. The start value was specified as 0.05 and 0.002 for the end value.

The initial number of clusters was determined to be a low value, which resulted in a robust lithofacies. Then the number of clusters was increased from value 3 to 8 one by one. (Three subset was set as minimum number of clusters in the separation according to the core samples. That suggested at least lithofacies – sand-, silt- and marlstone are separable as clusters in the data set.) But the selection of the proper number of clusters from the possible solutions is not trivial.

2.2. Selecting the proper number of cluster solutions

A number of authors have suggested various indexes to solve these problems but this means that usually the researcher is confronted with crucial decisions such as choosing the appropriate clustering method and selecting the number of clusters in the final solution. Numerous strategies have been proposed to find the right number of clusters and such measures (indexes) have a long history in the literature. The study focused on trying to determine the right number of clusters and to analyse some suggested sum of squares indexes (called WB indexes). The „leave-one out" (LOO) classification method was used in the discriminant function analysis (DFA) as cross validation (ASANTE & KREAMER, 2015).

To determine the stable number of clusters the DFA with LOO cross validation technique was used. A cluster structure was declared stable if DFA predicted at least 80% of the members in each cluster grouping. This threshold was set based on practical observations. Overall cross-validated results for each clustering of stable clusters range from 88.0-91.9%.

To select the optimal number of clusters in the final solution, a statistics test based on the sum of squares was applied. Since a single statistics test method cannot be depended upon, additional methods were used (ASANTE & KREAMER, 2015). There are several suggested indexes depending on the sum of squares (Eq.2–5):

Hartigan (1975): $$Ht = log\frac{SS_b(K)}{SS_w(K)} \quad \text{Eq.2}$$

Explained variance: $$ETA_K^2 = \frac{SS_b(K)}{SS_t} \quad \text{Eq.3}$$

Proportional reduction of error: $$PRE_K^2 = \frac{SS_w(K)}{SS_w(K-1)} \quad \text{Eq.4}$$

F-Max statistics: $$F - Max = \frac{\frac{SS_b(K)}{K-1}}{\frac{SS_w(K)}{n-K}} \quad \text{Eq.5}$$

Eq.5 is equal to the CALINSKY & HARABASZ index (1974) which is called the variance ratio criterion (VRC). Well-defined clusters have a large SSb (Sum of Squares between groups) and a small SSw (Sum of Squares within groups). The larger the VRC ratio, the better the data partition is. So the optimal number of clusters is determined by maximum VRC. Eq.2 is the Hartigan index, the so-called crude rule of thumb which is able to estimate the optimal number of clusters with the minimum value of second differences.

2.3. Interpretation of the separated clusters

The goal of the clustering method is to define „cluster facies" endowed with lithological and petrophysical parameters and the extension of these separated clusters based on multiple-point (cell-

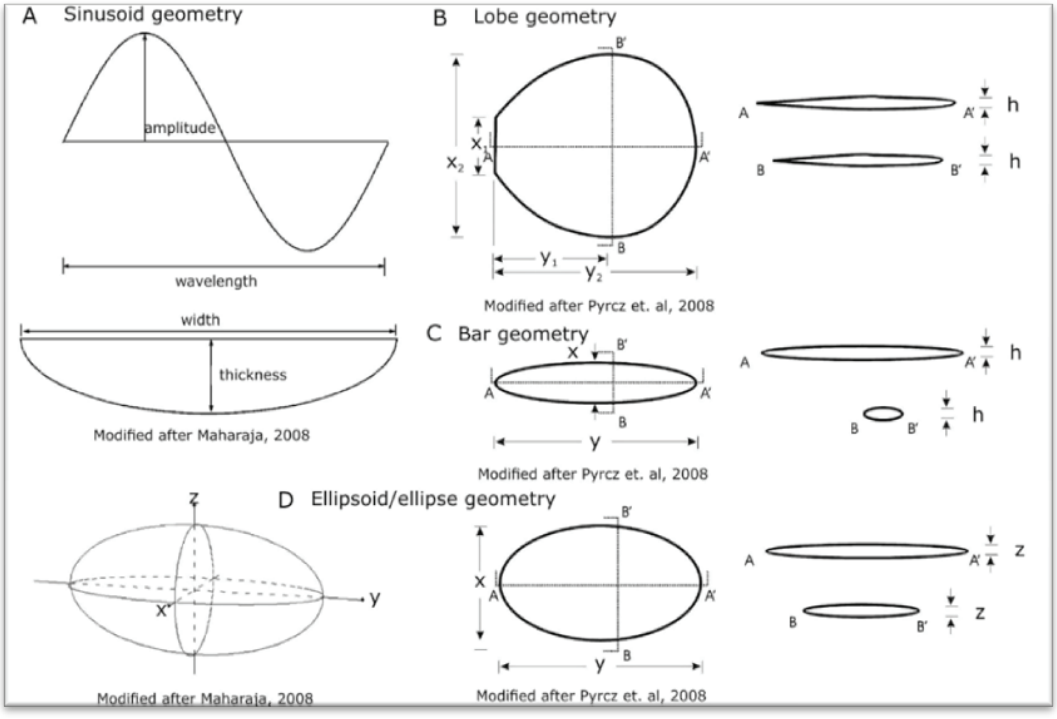

Figure 3. Parameters of geo-objects (modified after PYRCZ et. al, 2008; MAHARAJA, 2008).

based), object-based and process mimicking (non cell-based) algorithms. They are able to handle the additional geological information e.g. geometry.

Including these additional parameters as inputs is based on the consideration that the flow properties in a clastic reservoir are mostly determined by the geometries and the lithofacies of ancient sub-environments. The latter means that these methods can use only categorical variables, and exactly one lithofacies belongs to an ancient sub-environment.

The parameter of geometry can be regarded as a quasi-subjective geological data. Although the method of measurement has widespread literature, the final result moderately depends on the practitioner. Moreover the defined geometry possesses a distribution (mean, minimum and maximum values etc.), but it isn't as verifiable as the parameters and results of variogram-based algorithms.

Currently several parametric shapes (i.e. geo-bodies, geo-objects) are available. These geo-bodies are generalized shapes mimicking the true architectural elements.

In case of deep-water submarine complexes, the following geo-bodies correspond to the sub-environments: (1) Sinusoid objects: braided or meandering channels (NORMARK, 1970; READING & RICHARDS, 1994); (2) Lobe objects: channelized or unchannelized lobes (MUTTI, 1985; READING & RICHARDS, 1994); (3) Bar objects: mouth-bar at terminus of main depositional valley on the lower part of upper-fan (FGS) (NORMARK, 1970); (4) Ellipsoid objects: crevasse splays attached to channels (PYRCZ et. al, 2008; MAHARAJA, 2008).

Figure 3 shows the measureable parameters of these geo-bodies.

Sinusoid geometry should be characterized by: amplitude, wavelength, width, thickness and sinuosity (ratio of true streamline length (on the interval of wavelength) and wavelength) of the geo-body. Lobe geometry: mouth (x1), width (x2), length to largest width (y1), total length (y2), thickness (h) of the geo-body. Bar geometry: width (x), length (y) and thickness of the geo-body. Ellipsoid/ellipse geometry: semi-principal-axes (x, y, z) of a triaxial ellipsoid.

3. RESULTS

3.1. Results of clustering optimization

The analysis of cluster stability by DFA has eventuated in several stable cluster results (thresholds in excess of 80%); however, according to cross validation the 5 cluster solution was determined to be optimal. Based on LOO, 91.9% of the cross-validated grouped cases are correctly classified. The analyses of differences reduction between training error and validation error also showed the same optimum. The difference-plot (Fig. 4-a) reached the elbow point at the case of the five cluster solution. In practice the error rate was acceptable if it was appropriately low and good fit (the training-test-validation error rate approximated each other but the validation error was slightly higher than the training error).

Table 1. Test statistics results for estimating the number of clusters.

No.clust.	3	4	5	6	7	8
PRE_K^2	0.681758	0.782905	0.848698	0.867727	0.878515	0.904526
ETA_K^2	not defined	0.317831	0.304513	0.123945	0.081557	0.214392

In addition, the plot of Hartigan values (Fig. 4-b) and F-max(F) values (Fig. 4-c) determined a similar 'best fit' in the case of the five cluster solution.

From the explained variance values ETA_K^2, the three cluster solution explained 68% of the variance in the dataset; the four cluster solution explained ~78% and so (Tab. 1). The table shows that the increment in the proportional reduction of error ETA_K^2 significantly stopped from cluster five. Also the PRE_K^2 values sharply decreased from cluster five.

3.2. Interpretation of clusters

According to the optimality analysis the five cluster solution was approved. During the interpretation of these five clusters they were also matched to the lithological description of core samples (Fig. 5). In figure 5 the 0-cluster facies (black colour in the lithofacies from NNC) shows the impermeable units which were omitted from the analyses. According to this comparison – between the lithofacies coming from clusters and genetic lithofacies coming from core samples – together with the statistical characters (Tab. 2) the clusters were labelled: (1) siltstones and marls, interbedded sandstones; (2) spatially dispersed, low permeability sandstones; (3) alternation of siltstones and sandstones; (4) silty sand; (5) massive sandstones.

Of course, the goal was not to define „cluster facies" as simple lithological types. The spatial extension of clusters can also show well-defined depositional geometries.

2 out of the 5 clusters were chosen with the highest porosity, sand-content and permeability values (clusters 4 and 5). Table 2 summarized the group average of two clusters chosen from the five (clusters 4 and 5). The purpose of the visualization was to examine what geometries are shown by clusters 4 and 5.

A quasi-3D model (flatted to the impermeable argillaceous marlstone seal) was constructed by Voxler 3's FaceRender module. In this case cluster 4 and 5 show two sinusoid geo-bodies at 13 metres beneath the seal (Fig. 6). Direct measurement isn't available in Voxler 3, so from the same depth, sand and porosity contour maps using kriging estimation were used for the parametrization.

The two results show good similarity (Fig. 6), although one is based on discrete values, and the other is based on continuous

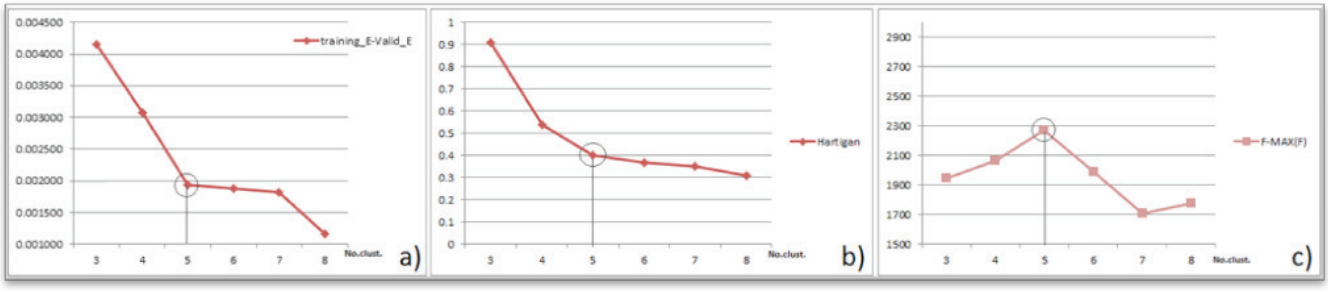

Figure 4. a) Difference plot based on NNC; b) plot of Hartingan indexes, c) F-max(F) plot

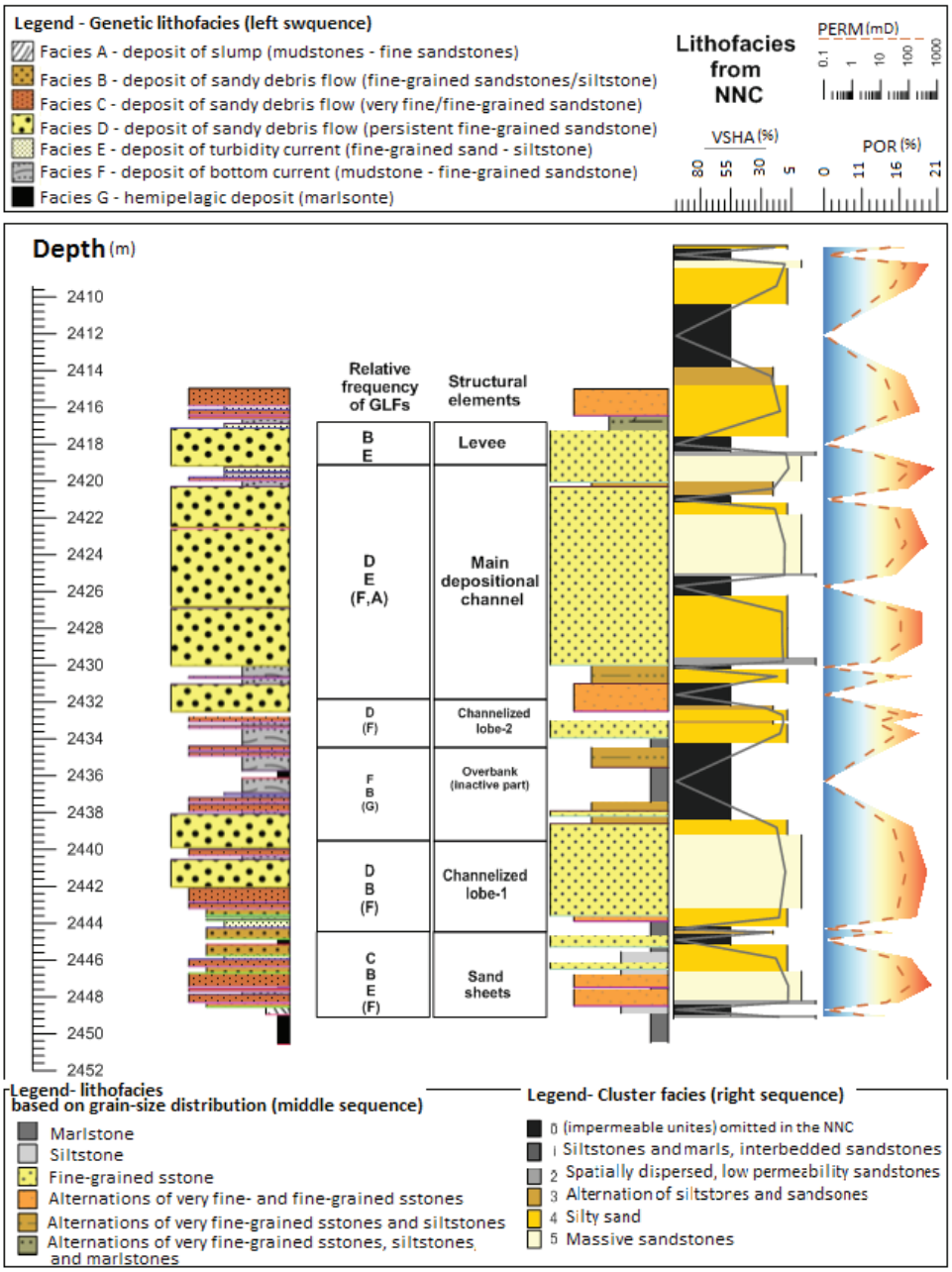

Figure 5. Comparison of NNC lithofacies (right sequence) with genetic lithofacies (left sequence) and lithofacies based on grain-size distribution (middle sequence) (based on BORKA, 2016).

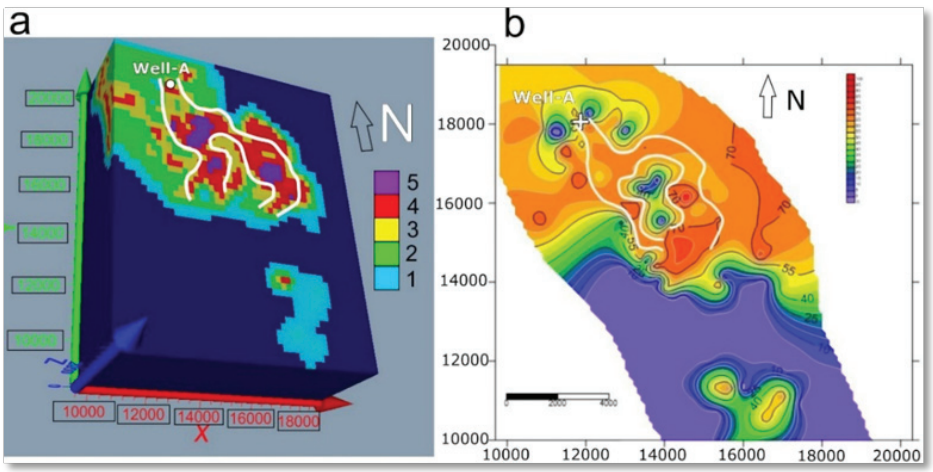

Figure 6. Picture 'a' shows two sinusoid geo-objects related to clusters 4 and 5; picture 'b' shows the same shapes in a sand-content contour map. The two slices are from the same depth, at 13 metres beneath the seal.

Table 2. Mean of clusters 4 & 5 based on the original data.

	Cluster-4	Cluster-5
POR (%)	18.39	20.25
PERM(mD)	32.24	87.16
VSHA (%)	15.30	8.79
VSND (%)	65.93	71.23

variables. Therefore, measurement on the contour maps was valid. The measured parameters are shown in Figure 7.

The geometric values are summarized in Table 3. The sinusoid geo-objects could be well tracked through approximately 45 slices i.e. contour maps (0.4 metres/1 slice). This means that thicknesses of both of the bodies are 18 metres (0.4 m x 45).

Core samples of Well-A were available from this depth. These can be characterized by massive, structureless fine sandstones with ripped intraclasts. They are deposits of sandy debris flows (SHANMUGAM, 2006) related to distributary channels or the proximal part of lobes. The GR and SP logs show cylindrical shapes which usually denotes channels (READING & RICHARDS, 1994).

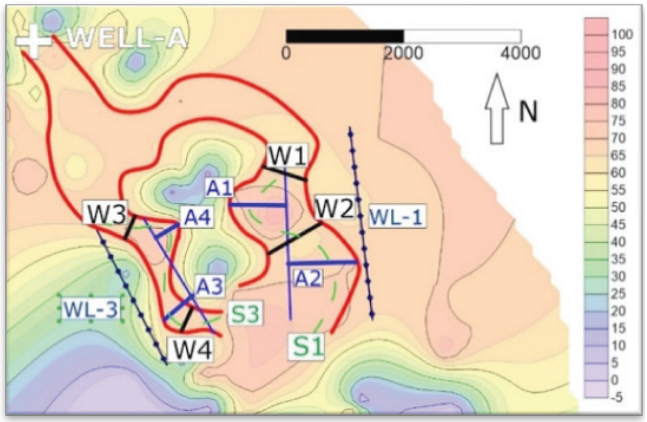

Figure 7. Notations with number 1 and 2 belong to the right sinusoid geo-object, while 3 and 4 belong to the left sinusoid geo-object; A – amplitude, W – width, WL – wavelength, S – length of streamline

Table 3. Measured values of the sinusoid geo-objects dimension: meter, except the SIN (ratio).

Right geobody								abbreviations
A1	A2	WL1	W1	W2	S1	SIN	TH1	
637	775	2156	496	685	2935	1.36	18	
Left geobody								
A3	A4	WL3	W3	W4	S3	SIN	TH3	
310	309	1658	277	286	2358	1.42	18	

A – amplitude
W – width
WL – wavelength
S – length of streamline
TH – thickness
SIN – sinuosity

4. SUMMARY

The study demonstrated that the transformed variables by the Box-Cox and PCA process reduced the impact of skewness and the redundancy in variables to avoid misclassification. The NN clustering with the final settings is validated using the DFA LOO method. Members in each cluster grouping were validated by over 80% prediction. Evaluation of optimal cluster solution relied on more WB indexes. All of them determined the „best fit clustering" with a five number of clusters solution.

Also, it is represented that in a case of mature field (dense hard data) 'optimized' clusters (i.e. lithofacies) can show geometric features. Clusters with the highest porosity, permeability and sand content – which may denote the most active part of the submarine fan – correspond to a sinusoid structural element (i.e. channel). The parameters of this geo-object can be used as input data of multiple-point or object based simulations.

ACKNOWLEDGEMENT

This paper was supported by the New National Excellence Program of the Ministry of Human Capacities.

REFERENCES

ASANTE, J. & KREAMER, D. (2015): A New Approach to Identify Recharge Areas in the Lower Virgin River Basin and Surrounding Basins by Multivariate Statistics.– Mathematical Geosciences, 47/7, 819–842. doi: 10.1007/s11004-015-9583-0

BÉRCZI, I. (1988): Preliminary sedimentological investigation of a Neogene Depression in the Great Hungarian Plain.– In: ROYDEN, L.H., & HORVÁTH, F. (eds.): The Pannonian Basin: A study in basin evolution, AAPG Memoir, 45, 107–116.

BORKA, SZ. (2016): Markov chains and entropy tests in genetic-based lithofacies analysis of deep-water clastic depositional systems.– Open Geosci., 8, 45–51. doi: 10.1515/geo-2016-0006

BOX, G.E.P. & COX, D.R. (1964): An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211–252.

CALINSKI, T. & HARABASZ, J. (1974): A dendrite method for cluster analysis, Communications in Statistics, 3, No. 1, 1–27. doi: 10.1080/03610927408827101

GAN, G., MA, C. & WU, J. (2007): Data Clustering: Theory, Algorithms, and Applications, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania, 466 p. doi: 10.1137/1.9780898718348

GRUND, SZ., & GEIGER, J. (2011): Sedimentologic modelling of the Ap-13 hydrocarbon reservoir, Central European Geology, 54/4, 327–344. doi: 10.1556/CEuGeol.54.2011.4.2

HARTIGAN, J.A. (1975): Clustering Algorithms– John Wiley and Sons, Inc., NY, USA 351 p.

HORVÁTH, J. (2015): Depositional facies analysis in clastic sedimentary environments based on neural network clustering and probabilistic extension.– Unpubl. PhD Thesis, University of Szeged, 118 p.

HORVÁTH, J. & MALVIĆ, T. (2013): Characterization of clastic sedimentary environments by clustering algorithm and several statistical approaches – case study, Sava Depression in Northern Croatia.– Central European Geology, 56/4, 281–296.

MAHARAJA, A. (2008): *TiGenerator*: Object-based training image generator, Computers and Geosciences.– Elsevier, 34, 1753–1761. doi: 10.1016/j.cageo.2007.08.012

MUTTI, E. (1985): Turbidite systems and their relations to depositional sequences.– In: ZUFFA, G.G. (ed.): Provenance of Arenites. D. Reidel Publishing Company, 65–93. doi: 10.1007/978-94-017-2809-6_4

NORMARK, W.R. (1970): Growth patterns of deep sea fans. AAPG Bulletin, 54, 2170–2195.

PYRCZ, M.J. & DEUTSCH, C.V. (2014): Geostatistical reservoir modelling.– Oxford University Print, 2nd edition, University of Oxford, 448 p.

PYRCZ, M.J., BOISVERT, J.B. & DEUTSCH, C.V. (2008): A library of training images for fluvial and deepwater reservoirs and associated code. Computers and Geosciences, Elsevier, 34, 542–560. doi: 10.1016/j.cageo.2007.05.015

READING, H.G. & RICHARDS, M. (1994): Turbidite systems in deep-water basin margins classified by grain size and feeder system.– AAPG Bulletin, 78, 792–822.

SAKIA, R.M. (1992): The Box-Cox transformation technique: a review.– The Statistician, 41, 169–178. doi: 10.2307/2348250

SHANMUGAM, G. (2006): Deep-Water Processes and Facies Models: Implications for Sandstone Petroleum Reservoirs.– Elsevier, 1st ed., Amsterdam, The Netherlands, 498 p.

TEMPL, M., FILZMOSER, P. & REIMANN, C. (2006): Cluster analysis applied to regional geochemical Data: problems and possibilities.– Research report-CS-2006-5, Vienna University of Technology, 39 p.