LITHOLOGY DETERMINATION IN A COAL EXPLORATION DRILLHOLE USING STEINER WEIGHTED CLUSTER ANALYSIS

BENCE Á. BRAUN1–ARMAND ABORDÁN1*, 2–NORBERT P. SZABÓ1, 2

1Department of Geophysics, University of Miskolc
*armand.abordan@yahoo.com
2MTA-ME, Geoengineering Research Group, 3515, Miskolc-Egyetemváros, Hungary

Abstract: For a successful exploration project, the determination of lithology is crucial. For determining the lithology of a coal exploration drillhole, a non-hierarchical cluster analysis is presented in which the most commonly used measure of distance, the Euclidean distance, is altered by Steiner weights. Using the Steiner weighted Euclidean distance makes the method more robust, thus affected less by outliers. We compare the results of the improved method of cluster analysis (MFV-CA) with the interpreted borehole geophysical logs, mud logs and the traditional cluster analysis based on the Euclidean distance (TCA). The comparative study shows that the new method gives better resolution and is less affected by outliers than the Euclidean distance based cluster analysis, and may give additional information to the interpretation of geophysical well logs and mud logs.

Keywords: cluster analysis, most frequent value, Steiner weight, well-logging methods, coal exploration

1. INTRODUCTION

The exploration and production of hydrocarbons, fresh water, ores, and coals are becoming more important as humankind is advancing. The main goal of applied earth sciences nowadays is to find, accurately locate and estimate the volume and quality of these raw materials in the subsurface. For this purpose, many geophysical and geological methods can be used, two of which are utilized in this paper, i.e. well logging and geostatistics. The combination of these tools for data processing and interpretation is already well known in earth sciences. In this paper, a novel approach for the modification of a multivariate geostatistical method is presented.

Cluster analysis, a frequently used multivariate statistical method, is applied to classify objects based on a set of measured variables into a number of different groups by placing similar subjects in the same group. In this paper, the traditional cluster analysis (TCA), which is based on the Euclidean distance, is modified by incorporating the most frequent value (MFV) method [1]. Modifying the Euclidean distance with Steiner weights makes the procedure robust and therefore less affected by outlying data. The improved statistical method is tested and compared with the interpreted well logs, mud log and the TCA on two well logs measured from a coal
explore drillhole. In both cases, the cluster analysis modified by Steiner weights (MFV-CA) proves to be superior to the conventional cluster analysis.

2. TRADITIONAL METHODS OF CLUSTER ANALYSIS

Cluster analysis is a multivariate statistical method that aims to group data objects into groups based only on the information found in the data that describe the objects and their relationships. The goal is to collect the objects into groups in a way that the objects within a group are alike and are very different from objects outside the group. The greater the homogeneity within a group and the bigger the difference between groups, the better the clustering is. The objects are grouped based on some defined distance metric.

Based on the above definition, it can be seen that if the distance between two objects is small then they are similar and should be placed in the same group; if the distance between two objects is great then they are different and cannot be put into the same group. If the distance between two objects is too great (outliers), then the non-resistant nature of the method should be taken into consideration, meaning that the method is greatly affected by outliers, because in this case data with different orders of magnitude may greatly influence the result of the estimation [2]. A datum can be considered an outlier if it is orders of magnitude greater or smaller than the average value range of the dataset and it shows a sudden change in the dataset such as a Dirac delta function. The goal of clustering is to have minimal distances within groups and maximal distances between groups [3].

Let the vectors $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ denote two multivariate observations from a population with $p$ random variables $X_1, \ldots, X_p$. In well log analysis, $X_i$ denotes a physical variable measured in the borehole by the $i$-th logging tool. In a more detailed form, the $i$-th and $j$-th observations are $\mathbf{x}^{(i)} = \{x_1^{(i)}, \ldots, x_p^{(i)}\}^T$ and $\mathbf{x}^{(j)} = \{x_1^{(j)}, \ldots, x_p^{(j)}\}^T$, which represent two so-called objects in the data space, respectively. In order to group the objects into clusters a measure for the similarity of elements needs to be defined. To determine the similarity between two objects, distance measures can be used. The TCA uses the Euclidean distance

$$D_e(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt{((\mathbf{x}^{(i)} - \mathbf{x}^{(j)}))'(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}))}, \tag{1}$$

By weighting it with the covariance matrix, we get the Mahalanobis distance

$$D_m(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt{((\mathbf{x}^{(i)} - \mathbf{x}^{(j)}))'\mathbf{S}^{-1}(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}))}, \tag{2}$$

where $\mathbf{S} = \mathbf{C}^\top \mathbf{C}/(n - 1)$ is the covariance matrix derived from the standardized data matrix $\mathbf{C}$. 

There are several different methods that can be used for cluster analysis. The main difference among the different types is whether the set of clusters is hierarchical or non-hierarchical. The main property of hierarchical clustering is that it permits clusters to have sub-clusters. It also has the advantage over non-hierarchical clustering that prior to the analysis it does not require the specific number of clusters to be created during the process. The main drawback of hierarchical clustering is that it needs a lot of computing power; therefore, it is suitable only for smaller data sets. It can further be divided into agglomerative and divisive methods. In case of the agglomerative method each object starts as an individual cluster, then the two closest are combined repeatedly until all objects are in the same cluster. After calculating all cluster solutions, the optimum number of clusters can be chosen. On the contrary, the divisive method starts with one cluster and with the same strategy as the above method in a reverse order separates objects until all are separated.

The non-hierarchical or partitioning clustering is fundamentally different from hierarchical clustering. This kind of clustering requires the specific number of clusters to be created prior to the process. It is done by a partitioning algorithm and the number of objects in each cluster is computed during the process. It has the feature of handling big data sets well; however, the result is affected by the initial selection of centroids.

The SSE (Sum of the squared error) can be used for estimating the optimal number of clusters. It measures the distance of each data point to the closest centroid and then calculates the total sum of squared errors

\[
SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2 \left( \mathbf{c}_k, \mathbf{x}_k^{(i)} \right),
\]

where \(d\) gives the distance between the \(k\)-th vector belonging to the \(k\)-th cluster (with \(n_k\) elements) and \(k\)-th centroid \(\mathbf{c}_k\) \((k = 1, 2, ..., K)\). The centroid is calculated by the mean of objects forming a cluster

\[
\mathbf{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_k^{(i)}.
\]

The value of SSE also describes scatter, which reaches a limit with the increasing number of clusters [2]. For the optimal solution, the smallest SSE value and the smallest number of clusters should be chosen because choosing a greater value hardly adds any more information but it makes the interpretation more difficult.

From the non-hierarchical clustering methods, the most commonly used is K-means clustering (Figure 1). A prototype-based technique attempts to find a user-specified number of clusters \(K\), which are represented by their centroids. The first step is to assign \(K\) initial centroids, where \(K\) is a user-specified parameter then each point is assigned to the closest centroid and each collection of points assigned
to a centroid is a cluster. The centroids are recalculated based on the points that are assigned to the clusters. We repeat this until all centroids remain the same, as seen in Figure 1 [3].

![Figure 1](image)

3. MOST FREQUENT VALUE BASED CLUSTER ANALYSIS

The most frequent value method is based on the weighting of data, calculating the weighted average. The weighted average is in most cases more advantageous than the arithmetic means because of its robustness (being less affected by outliers). The goal of cluster analysis is to have minimal distance within the clusters and maximal among the clusters, therefore it is necessary to weight data more that are closer to each other and weight less those which are farther [4].

The next equation is assumed to have high values of \( n \) and we have a symmetric density function. The weighted average is calculated with the \( \varphi(x) \) symmetric weight function, where \( T \) denotes the point where the \( \varphi(x) \) weight function has its maximum value

\[
M_n = \frac{\sum_{i=1}^{n} x_i \varphi_i}{\sum_{i=1}^{n} \varphi_i},
\]

\[
\varphi_i = \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M_n)^2},
\]

where \( (x_i - M_n) \) denotes the differences within the cluster, \( \varphi \) is a characteristic weight-function and \( \varepsilon \) is the dihesion, the shape parameter of the weight function.

The \( M_n \) is called the most frequent value, which can be calculated iteratively from the combination of equations (5) and (6), because the weights are independent of \( M_n \).
Lithology Determination in a Coal Exploration Drillhole Using Steiner Weighted... 55

[1]. Thus during the calculation of \( M_n \) we have to define \( M_n \) iteratively in such a way that it satisfies the equation

\[
M_n = \frac{\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M_n)^2} x_i}{\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M_n)^2}}.
\]  

(7)

For the first iteration (\( k=1 \)) \( M_n \) is substituted with the mean of the data and the dihension \( \varepsilon \) is calculated from Equation (8), the sequential number \( k \) of the iterations and the clusters beyond possibility

\[
\varepsilon_1 \leq \frac{\sqrt{3}}{2} (\max(x_i) - \min(x_i)).
\]  

(8)

In the beginning, we choose a greater value for \( \varepsilon \) and then even the outlying data are assigned a relatively large weight. In the following iteration steps, the values of \( \varepsilon \) and \( M_n \) are derived from each other based on Equations (9) and (10) [1]

\[
\varepsilon^2_{k+1} = \frac{3 \sum_{k=1}^{n} \frac{(x_i - M_k)^2}{\varepsilon_k^2 + (x_i - M_k)^2}}{\sum_{i=1}^{n} \frac{1}{\varepsilon_k^2 + (x_i - M_k)^2}},
\]  

(9)

\[
M_{n,k+1} = \frac{\sum_{k=1}^{n} \frac{\varepsilon_{k+1}^2}{\varepsilon_{k+1}^2 + (x_i - M_{n,k})^2} x_i}{\sum_{k=1}^{n} \frac{1}{\varepsilon_{k+1}^2 + (x_i - M_{n,k})^2}}.
\]  

(10)

At the end of the procedure the small value of \( \varepsilon \) ensures that data close to \( M_n \) contribute to the result with larger weight and outliers with smaller weight or not at all. The least-squares procedure works with 100% efficiency if the error distribution is Gaussian. This is not surprising because the least-squares estimate is the best when the distribution is Gaussian. Unfortunately, its efficiency diminishes sharply to zero for error distributions having longer tails. Therefore, the least-squares principle should not be applied for any type of distribution other than the Gaussian type. In contrast, the MFV procedure is very highly efficient (>90%) regardless of the distribution type. Therefore, the general high robustness of the MFV procedure is proven [5].

Based on tests the MFV weighting improves the result of the TCA, and being a robust statistical procedure it can be applied very effectively for lithology determination. As a starting point for the Steiner weights, first we take the Euclidean distance as in Equation (1) and then we modify it, which gives the robust algorithm of MFV-CA as seen in Equation (12). The values of the Steiner weights that are computed in the most frequent value method are given by the following equation
where \( W_{ii} \) is the \( i \)-th diagonal element of the weight matrix, which gives the difference between the \( i \)-th data and \( k \)-th centroid, and \( \varepsilon \) is calculated in an inner iteration cycle.

Applying the Steiner weights in the TCA will improve its performance. Combining Equations (1) and (11) we get Equation (12), the Euclidean distance altered by the most frequent value based weight function

\[
D_{MFV}(x_k^{(i)}, c_k) = (x_k^{(i)} - c_k)^T W^{(MFV)} (x_k^{(i)} - c_k),
\]

where \( k = 1,2,3...K \). As a result, data farther from the centroid will affect the formation of clusters with smaller weight, while the contribution of the elements closest to \( M_n \) is the greatest.

4. APPLIED WELL LOGGING TOOLS AND THE LITHOLOGY

To test the MFV-CA procedure and compare it to the TCA, a coal exploration drill-hole in Hungary is used. Based on the mud log and the interpreted well logs the dominant rock types of the area are sandstone, siltstone, mudstone, marl, coal, carbonaceous clay, marly sandstone, silty sandstone, and diabase.

In the investigated drillhole the layer thicknesses vary from a few tens of cm to 25 m. The well logs were measured at 0.1 m intervals. The tools and the measurement interval were selected based on the highly varying lithology and layer thicknesses. The following well logging tools were selected as an input for the cluster analysis: spontaneous potential (SP), natural gamma-ray intensity (GR), shallow apparent resistivity (RES10), medium apparent resistivity (RES40), gamma-gamma (density) (DEN) and neutron porosity (NPOR). With these tools the detection of coal is relatively unambiguous, since coals show high density-porosity, high neutron-porosity, high resistivity and low natural gamma radiation, which increases with shale content [6]. Lithological boundaries are usually picked manually or with software, but cluster analysis automatically provides the layer boundaries based on the created clusters.

5. INTERPRETATION OF WELL LOGS WITH STEINER-WEIGHTED CLUSTER ANALYSIS

5.1. Test on well log-1

Based on the SSE diagram (Figure 2), the optimal number of clusters is 4; for more than 4 clusters the algorithm created non-existing and transitional rock types. Figure 3 contains the interpreted well logs, mud log, TCA and the MFV-CA.
Lithology Determination in a Coal Exploration Drillhole Using Steiner Weighted...

The brightness index (BI) of coals indicate their rank; the brighter the coal the higher its carbon content and the lower its ash content is [7]. It can be seen that in the interval of 576–577.5 m the interpreted well logs and the mud log defined sandstone; however, both types of cluster analysis defined siltstone. In the depth range of 577.5–587.5 m all methods defined the same rock type, sandstone. In the interval of 587.5–588.5 m most probably coaly mudstone is present, based partly on the interpreted well logs and the cluster analysis with the modified Euclidean distance; however, the mud log suggests siltstone with interbedded coal layers and the cluster analysis based on the Euclidean distance suggests coaly siltstone. The area between 588.5–592.5 m is defined as coal by the mud log and the cluster analysis, while the well logs suggest carbonaceous clay at some intervals. In the 592.5–595.5 m interval the well logs defined coal, although both the cluster analysis and the mud log suggest interbedded carbonaceous clay layers. This reinterpretation changes the predicted lithology of this coal seam, thus the extent of the economically extractable coal seam is reduced. This might be acceptable for a coal seam of this magnitude, but in the case of a bigger coal seam or an oil field these reinterpretations can greatly influence the volume estimations and therefore the economic decisions. Therefore, it is not advisable to rely only on the interpretation of well logs and the mud log; the results of the cluster analysis should also be considered, because it numerically interprets the lithology and thus can either confirm or confute the original estimates. In the depth range of 595.5–597 m all methods suggest the presence of siltstone. The diabase starting from 597 m is interpreted as siltstone by the cluster analysis. Defining the diabase by cluster analysis is somewhat uncertain because the properties of this layer are not described well.

Figure 2
Optimal number of clusters for well log 1
Figure 3
Results of cluster analysis for well log 1
5.2. Test on well log-2

**Figure 4**
Optimal number of clusters for well log 2

**Figure 5**
Results of cluster analysis for well log 2
As shown in Figure 4, five clusters were created during the clustering for well log 2. Figure 5 contains the interpreted well logs, mud log, the TCA and the MFV-CA. In the interval of 628–632.5 m the rock type is either siltstone or sandstone (it is uncertain which cluster denotes which rock type; there might be grain size transition between the two). Down to 649 m, there is a thick coal seam, which is interbedded with up to 1 m thick carbonaceous clay layers, according to the cluster analysis. The interpreted well logs show fewer carbonaceous clay layers, but suggests interbedded mudstone layers up to 1.5 m thick. Based on the mud log and the cluster analysis, more carbonaceous clay layers intersect this thick coal seam, therefore the whole seam cannot be economically extracted. The tuff layer at the bottom of the log is defined by all methods but at slightly different depths. The description of the mud log only correlates at some intervals with the cluster analysis, which is probably caused by the lower rank coals in the interval that are grouped to carbonaceous clay layers.

6. COMPARISON OF TRADITIONAL CLUSTER ANALYSIS AND STEINER-WEIGHTED CLUSTER ANALYSIS

The cluster analysis modified by Steiner weights works well in case of shorter intervals where diverse layers and thin layering is present. The SSE value can usually be used effectively to estimate the number of clusters to be created. If it proves to be insufficient, an additional cluster can be added. The shallow apparent resistivity tool at most intervals read a monotonic trend, therefore it did not affect the clustering algorithm much. The lower rank of coals is closely related to the grouping of clusters into carbonaceous clay. Where similar rock types follow each other (clay–carbonaceous clay–clay) the cluster analysis based on the Euclidean distance modified by Steiner weights exerts an adequate smoothening effect. Both clustering algorithms have difficulties differentiating sandstone and siltstone (this can be due to grain size boundary or a badly graded rock). When processed with the Euclidean distance, thicker layers are smoothed out completely and the separation of thin interbedded layers is not possible, only the depth range of the layer is acceptable. CPU time for both methods, even for big data sets, is negligible.

The main goal of these methods is to determine the lithology precisely with the minimal number of clusters. The presented cluster analysis with the modified Euclidean distance gives a more robust estimation, which later can be used as a starting model for an inversion procedure. Weighting cluster analysis based on the Euclidean distance with the MFV procedure allows us to determine the lithology with a minimal number of clusters and with good accuracy and resolution. This new procedure makes it easier to accurately map the layers or reservoirs in the subsurface, and therefore better volume estimates can be made, which has significant economic importance.
7. TEST OF ROBUSTNESS

In Figure 6, four different log intervals can be seen. In these logs two outliers each were added to the data set (Table 1), for both the classical cluster analysis with the Euclidean distance and for the one weighted by the Steiner weights. During robustness testing we define how much a data point needs to be altered for the clustering algorithm to handle the object as a different cluster.

Comparing the results, it can be seen that the cluster analysis based on the Euclidean distance modified by the Steiner weights is more robust and has better noise rejection capability than the classical cluster analysis using the Euclidean distance, because it groups an object to a new cluster only with greater outliers. These tests verify the more robust nature of the non-hierarchical cluster analysis using the Euclidean distance modified by the Steiner weights presented in this paper. This exerts a smoothening effect on the logs where thin and similar rocks are next to each other (e.g. clay – carbonaceous clay – clay) and where an outlier is present. This smoothening is not applicable for extreme values.

<table>
<thead>
<tr>
<th></th>
<th>GR (1)</th>
<th>N. POR. (2)</th>
<th>RES40 (3)</th>
<th>SP (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural gamma ray</td>
<td>36.000</td>
<td>54.500</td>
<td>9.590</td>
<td>43.630</td>
</tr>
<tr>
<td>Neutron porosity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium apparent resistivity</td>
<td>36.383</td>
<td>52.098</td>
<td>9.681</td>
<td>43.558</td>
</tr>
<tr>
<td>Spontaneous potential</td>
<td>83</td>
<td>162</td>
<td>28</td>
<td>132</td>
</tr>
<tr>
<td>Limit of outlier</td>
<td>55</td>
<td>127</td>
<td>21</td>
<td>102</td>
</tr>
</tbody>
</table>

* = Euclidean distance modified by Steiner weights

Table 1
Statistics of well logs
8. CONCLUSION
According to our tests, the non-hierarchical cluster analysis based on the Euclidean distance modified by the Steiner weights (MFV-CA) presented in this paper provides highly acceptable results. It is efficient, robust and can determine the lithology with good resolution. It also has good noise rejection capability, as shown in Figure 6. The classical cluster analysis based on the Euclidean distance (TCA) can also pro-
vide acceptable results; however, its accuracy is worse and its lithology determination is less exact. The MFV-CA method as an additional tool for lithology determination, supplementing mud log and well log analysis, can be applied to verify and specify the results of the other two methods numerically. In the future we are planning to follow up our research by applying the MFV-CA method to unconventional reservoirs.

**ACKNOWLEDGEMENTS**

The authors are thankful for the support and data provided by Geo-Log Geophysical & Environmental Ltd.

**LIST OF SYMBOLS**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^{(i)}$</td>
<td>vector of $i$-th multivariate observation</td>
<td>–</td>
</tr>
<tr>
<td>$x^{(j)}$</td>
<td>vector of $j$-th multivariate observation</td>
<td>–</td>
</tr>
<tr>
<td>$X_i$</td>
<td>$i$-th observed well logging parameter</td>
<td>–</td>
</tr>
<tr>
<td>$D_e$</td>
<td>Euclidean distance</td>
<td>–</td>
</tr>
<tr>
<td>$D_m$</td>
<td>Mahalanobis distance</td>
<td>–</td>
</tr>
<tr>
<td>$S$</td>
<td>covariance matrix</td>
<td>–</td>
</tr>
<tr>
<td>$C$</td>
<td>standardized data matrix</td>
<td>–</td>
</tr>
<tr>
<td>SSE</td>
<td>sum of the squared error</td>
<td>–</td>
</tr>
<tr>
<td>$d$</td>
<td>distance between the cluster elements and its centroid</td>
<td>–</td>
</tr>
<tr>
<td>$c_k$</td>
<td>the centroid of the $k$-th cluster</td>
<td>–</td>
</tr>
<tr>
<td>$n_k$</td>
<td>number of objects in the $k$-th cluster</td>
<td>–</td>
</tr>
<tr>
<td>$\varphi(x)$</td>
<td>weight function in MFV procedure</td>
<td>–</td>
</tr>
<tr>
<td>$M_n$</td>
<td>empiric most frequent value</td>
<td>–</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>dihesion</td>
<td>–</td>
</tr>
<tr>
<td>$k$</td>
<td>number of iterations</td>
<td>–</td>
</tr>
<tr>
<td>$K$</td>
<td>number of clusters</td>
<td>–</td>
</tr>
<tr>
<td>$\varepsilon_1^{(MFV)}$</td>
<td>initial value of dihesion</td>
<td>–</td>
</tr>
<tr>
<td>$W_{ii}^{(MFV)}$</td>
<td>$i$-th Steiner weight coefficient</td>
<td>–</td>
</tr>
<tr>
<td>$D_{MFV}$</td>
<td>Steiner weighted Euclidean distance</td>
<td>–</td>
</tr>
<tr>
<td>$SP$</td>
<td>spontaneous potential log</td>
<td>mV</td>
</tr>
<tr>
<td>$GR$</td>
<td>natural gamma-ray intensity log</td>
<td>API</td>
</tr>
<tr>
<td>RES10</td>
<td>resistivity log with shallow penetration</td>
<td>ohmm</td>
</tr>
<tr>
<td>RES40</td>
<td>resistivity log with medium penetration</td>
<td>ohmm</td>
</tr>
<tr>
<td>DEN</td>
<td>gamma-gamma (density) log</td>
<td>g/cm$^3$</td>
</tr>
<tr>
<td>NPOR</td>
<td>neutron porosity log</td>
<td>%</td>
</tr>
</tbody>
</table>
REFERENCES


