RGNZ/MGPB

Comparative analysis of the K-nearest-neighbour method and K-means cluster analysis for lithological interpretation of well logs of the Shushufindi Oilfield, Ecuador

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
UDC: 550.8
DOI: 10.17794/rgn.2022.4.13

Original scientific paper

Franklin Gómez [1, 2]; Yetzabbel Flores[1]; Marianna Vadászi [1]

[1] *Faculty of Earth Science and Engineering, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary*

[2] *Petroleum Department, Earth Science Faculty, Escuela Politécnica Nacional, 170517, Ecuador, Ladrón de Guevara E11·253*

Orcid: Gómez - https://orcid.org/0000-0003-4367-4972; Flores - https://orcid.org/0000-0003-0365-8951; Vadászi - https://orcid.org/0000-0003-0649-311X

Abstract

The lithological interpretation of well logs is a fundamental task in Earth science that can be accomplished with the application of various machine learning algorithms. The current investigation attempts to evaluate the performance of the K-nearest-neighbour Density Estimate (KNN) and K-means cluster analysis methods for predicting lithology in a dataset of logs measured in the siliciclastic reservoir of the Shushufindi Oilfield of Ecuador. The comparison of lithological interpretation is assembled using classical methods, such as qualitative interpretation and density-neutron cross plot. The lithological interpretation results showed that the supervised method KNN has a higher fitting level with the comparison interpretation data (87.3%, 1145 m predicted of 1311.1 m interpreted) than the results of the K-means method (71.6%, 939.7 m predicted of 1311.1 m interpreted). The geological nature of the reservoir creates a level of a discrepancy because of the near geophysical responses between limestone and intermedia grain size rocks. The possibility of controlling this in the KNN algorithm makes it preferable for usage in these types of reservoir lithological interpretation.

Keywords:

machine learning; reservoir; lithology; Shushufindi

1. Introduction

For Earth science, the use and interpretation of well logs are fundamental for purposes such as stratigraphic interpretation, petrophysics characterization, rock mechanics behaviour and more. In the petroleum industry, their use is essential to estimate the pay zone and reservoir type. Frequently, well logs provide information for exploration, drilling, production, and reservoir management activities (**Bassiouni, 1994**). The working flow of this analysis lists stratigraphy interpretation as an exhaustive and time-consuming step, which is the reason for the constant development of diverse machine learning methods to be applied under the principle of geophysics response of lithology to various logging methods. Traditionally, the research has focused on increasing the accuracy of lithology interpretation using debris logging, core logging, and cross plot techniques (**Zhao et al., 2017**). For example, **Cheng et al. (2016)** established the cross plot method and logging curve calculation method to identify siliciclastic rocks, and **Khamees et al. (2021)** used the neutron-density cross plot, acoustic-density cross plot, and M–N cross plot with the same aims. Previous investigation results show the satisfactory use of these traditional methods in lithology identification, but with the emergence of deep learning, the application of machine learning methods is ensuring more dynamism and time-efficient workflows for the task (**Ali et al., 2021**).

Two popular methods are the K-nearest-neighbour density estimate (KNN), and K-means clustering. They allow discretising of known information, such as petrophysical properties, in order to predict them in unknown zones, as **Ali et al. (2021)** do in shear Sonic logs. However, The KNN and K-means methods have different approaches. KNN uses learning information to predict new events, while K-means performs clustering to categorise the information and define the properties of a future event (**Amonkar et al., 2022; Troccoli et al., 2022**). In this study, KNN and K-means machine learning methods are applied for the lithological interpretation of seven boreholes in Shushufindi Oilfield. The target dataset includes natural gamma (GR), bulk density (RHOB), deep resistivity (RD), and photoelectric factor (PEF) geophysical registers. The aim is to evaluate each method's performance in comparison with the results of a traditional cross-plot method and to analyse their applicability for a fast lithological description of a large set of data in a reservoir dominated by siliciclastic sedimentary sequences, but with the presence of a few well-mapped limestone layers.

Corresponding author: Franklin Gómez
e-mail address: oljfrank@uni-miskolc.hu

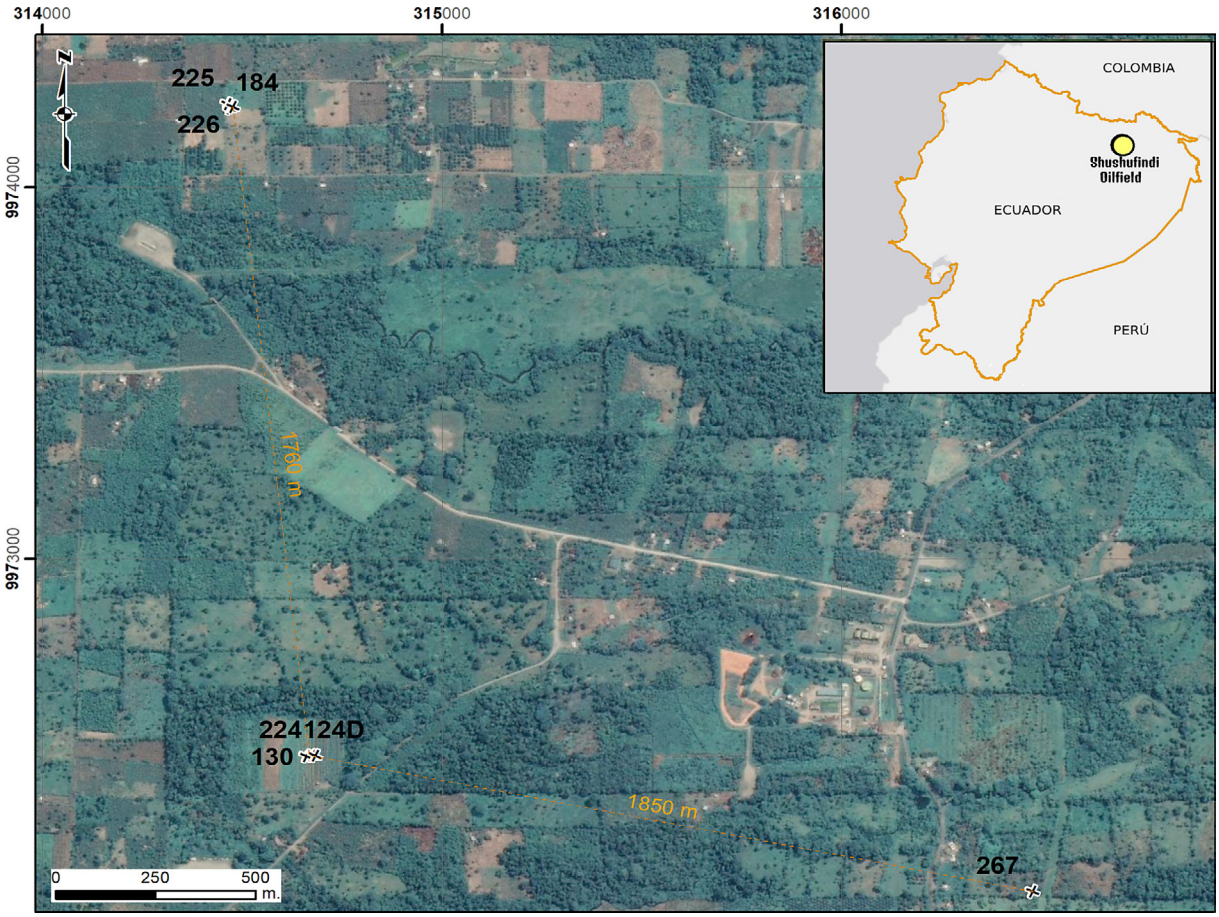

Figure 1: Location of the investigated wells within the Shushufindi Oilfield, NE of Ecuador. Datum WGS84-18S

2. Overview of the study area and dataset

Shushufindi Oilfield is a huge anticline reservoir of about 35 km in length and 6 to 7 km wide (**Biedma et al., 2014**). It is located in the Ecuadorian Amazon Basin, east of the South American country (see **Figure 1**). The Late-Cretaceous transpression event configured the basin morphology. It occurred before the subduction-induced uplift of the Andean Range and the formation of the fore-arc basin. The mature giant Shushufindi Oilfield, in the Sacha-Shushufindi Corridor, has a geological structure similar to a flower that developed due to the compressional events of the Early Cretaceous and Turonian ages (**Estupiñan et al., 2010**).

The dominant lithological classes in the stratigraphic column of the reservoir are siliciclastic sediments (see **Figure 2**), with a secondary volume of limestones. The productive layers in Shushufindi Oilfield are the Napo "U" sandstone, Napo "T" sandstone, and Hollín sandstone (**Ramirez, 2020**). Meanwhile, Tena Basal and the superior Hollín are secondary reservoirs, scarce accumulations of oil can also be found in "A" and "B" limestones of the Napo Formation (**Baby et al., 2004; Salazar, 2014; Renss, 2016; Tomalá, 2020**).

The dataset comprises four well-logging curves measured in seven wells located relatively close to each other (see **Figure 1**). The labels for the lithological classes are gained from bibliographic geological descriptions of the reservoir (**Biedma et al., 2014; Zhang and Li, 2016; Ramirez, 2020**). The geophysical logs used in the analyses include gamma ray (GR), deep resistivity (RESD), bulk density (RHOB), and photoelectric (PEF) curves (see **Figure 3**). Each curve is a continuous variable mapped at 0.15 m vertical definition, and its particular measure magnitude unit, constituting a dataset of 8,603 individual data points. **Table 1** presents the preliminary statistics of the dataset.

Table 1: Summary statistics of the Shushufindi Oilfield target dataset

Statistics	Depth (m)	GR (°API)	RESD (Ω.m)	RHOB (g/cm³)	PEF
MEAN	2960.9	84.93	20.88	2.49	3.23
STD	96.4	47.46	42.23	0.15	0.92
MIN	2767.0	8.32	0.83	1.34	1.58
25%	2892.4	49.01	4.83	2.40	2.39
50%	2951.2	76.90	9.35	2.51	3.25
75%	3018.6	116.09	22.28	2.58	3.88
MAX	3220.4	266.11	920.02	2.93	6.75

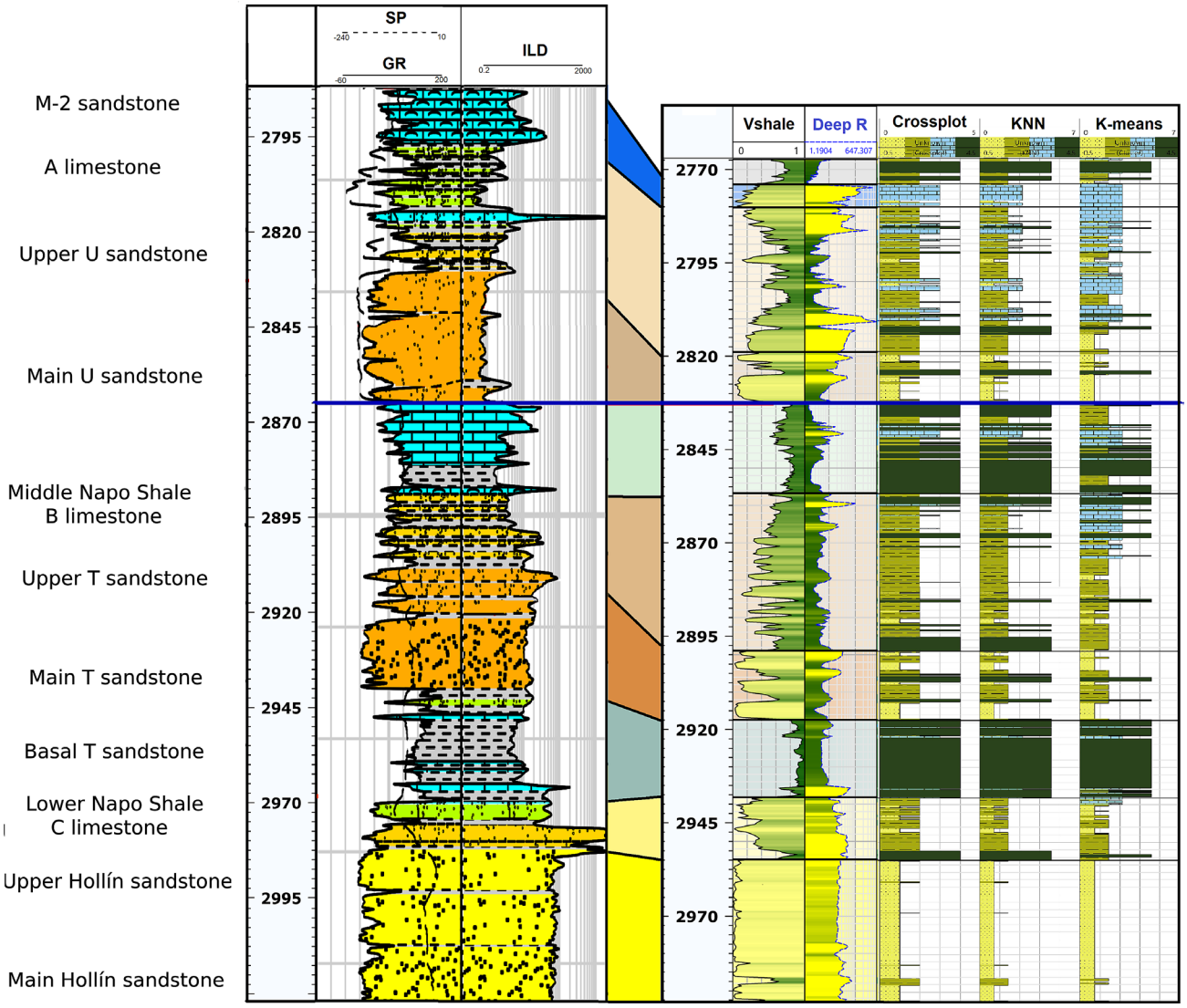

Figure 2: General stratigraphic column of Shushufindi Oilfield (left) adapted from **Biedma et al. (2014), Estupiñan et al. (2010)** and **Ramirez (2020)**, and correlated with the lithological interpretation of Well SF_130 (right)

3. Methods

In this study, two machine learning methods are applied over a dataset of continuous variables to use a quantitative approach with categorical results. The correlation between specific lithology and geophysical response is the fundamental basis for the application of the following methods.

3.1. Lithology interpretation

Some of the main traditional lithological interpretation methods are the neutron-density cross plot and the M-N lithology cross plot. Consequently, the traditional lithological interpretation of logs is applied to construct the performance control parameter curve, which is used to monitor the machine learning methods' performance and to supply the unavailable hard lithological data (**Khamees et al., 2021**). Based on these two traditional lithological methods, four lithological classes are distinguished: clean sandstone, intermedia-grained rock, limestone, and shale rich rock. The four defined lithological classes are interpreted for each log combining qualitative analysis of the curve shape (**Zhao et al., 2015; Erlström and Sopher, 2019; URL 1; URL 2; URL 3**) and the neutron-density cross plot (**Basal, 1998; Mamaseni et al., 2018; Imamverdiyev and Sukhostat, 2019**).

3.2. K-nearest neighbour (KNN) classifier

The K-nearest neighbour is a non-parametric method of estimating a probability density function. The algorithm estimates a function that predicts the rock type (z) according to the log-well registered values. Every interpreted rock category x is a p-dimensional random variable X. This means the interpretation of the rock type will depend on the pre-established variables z. The d(x,z)

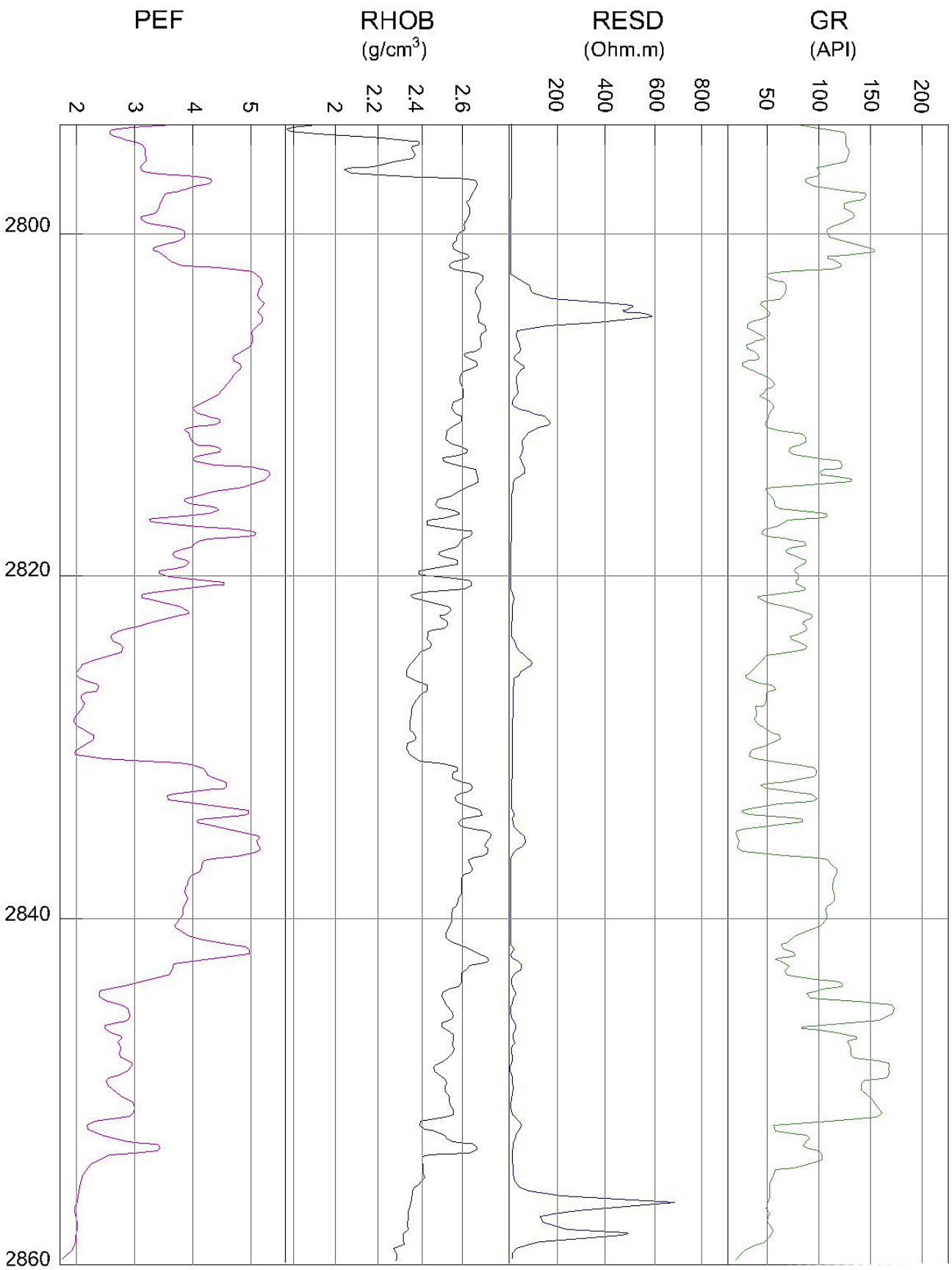

Figure 3: Log sections of well SF_267. Conventional log curves are used to predict the lithology in the well location.

represents the Euclidean distance between x and z. X is an example of z, consequently x is the rock type interpreted before the prediction (**Mitra et al., 2002; Delavar, 2022**). The hypersphere of radius r about z is designated by **Equation 1**.

$$A_{r,z} = \{X \setminus d(x,z) \le r\} \quad (1)$$

where:

A_r – volume of the hypersphere,
R – radius of the hypersphere,
x – categorical class,
X – variable.

Then **Equation 2** defines the density function:

$$f_N(z) = \frac{k(N)}{N} * \frac{1}{A_{rk(N),z}} \quad (2)$$

where:

$f_N(z)$ – function f to estimate z with N,
$k(N)$ – sequence of positive integers from $x1$ to xN,
x – rock type interpreted in a set of data.

The method is sensitive to the scale difference between variables in multidimensional space, so standardisation is required to eliminate the effect of scale differences in both training and test sets (**Zhang et al., 2022**). A specific explanation of neighbourhood classifiers can be found in **Hu et al. (2008)**. To select the effective value (n) of k, the error curves are plotted for different values of k (see **Figure 4**) for the training and test data until the test error curve stabilises at the optimum value (**URL 4**).

MATLAB script implements the following steps for machine learning applying the KNN method (**Pratama, 2019**): First, the target dataset is examined to create the training classifier set using the established parameters of the variables. Then, the best number of K is selected by the nearest-neighbour number and metric distance. Finally, cross-validation assesses the performance of the method before the algorithm is applied to the additional blind wells.

3.3. K-means cluster analysis

K-means cluster analysis is a simple not-supervised statistical method that orders the objects of a multivariate dataset into groups using the information of similarities given by metric distance. The method is sensitive to the scale difference between variables, so normalisation of the data set is required. Then, each object is designated into a non-overlapping group of great homogeneity and large differences from other groups. For further explanation, see **Ali and Sheng-Chang (2020)**, **Isfan et al. (2021)** and **Szabó et al. (2021)**. Mathematically, the method is expressed by **Equation 3**:

$$J = \sum_{j=1}^{k} \sum_{i=i}^{n} x_i^{(j)} - c_j^{\,2} \quad (3)$$

Where:

J – objective criterion,
x_i – analysed objects, $i= 1,..., n$,
c_j – cluster centroid, $j= 1,.., k$,
k – optimal number of clusters.

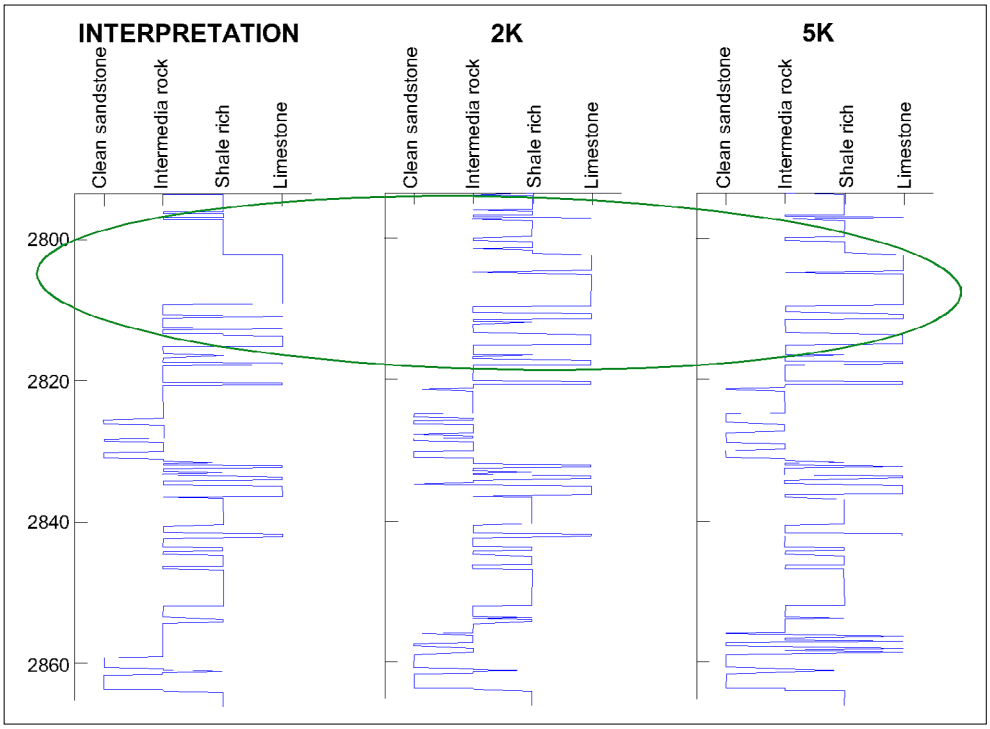

Figure 4: Comparison of the lithological interpretation curve (left) and lithological prediction results of the Nearest Neighbour Number method. Lithological results of well SF_267 plotted for values of K=2 (centre) and K=5 (right).

The objective criterion converges at the minimum sums of square deviation of objects *xi*, from the cluster centroid *cj*. The "City Block" distance metric is incorporated for this study. Here, each cluster centroid is the component-wise median of the points in the cluster, as **Equation 4** expresses:

$$D_1 = \sum_{k=1}^{N} \left| x_k^{(i)} - x_k^{(j)} \right| \quad (4)$$

Where:

D – sum of lengths between points,

$x_k^{(i)}$ – distance of x_k from the centroid in component *i*,

$x_k^{(j)}$ – distance of x_k from the centroid in component *j*.

In K-means clustering, the optimum number of clusters (*k*) is selected using the Elbow curve method (see **Figure 5**). The algorithm computes the sum of distance for different values of k, and each pair of parameters is plotted in a curve where the inflexion point may be the optimum value of *k* (**URL 5; Troccoli et al., 2022**). The K-means clustering algorithm starts heuristically choosing the first centroids. This means it starts with a randomly selected set of centroid locations. Next, it runs a series of iterations to test different solutions until the lowest sum of distances among the iterations is obtained and the local minimum solution is found (**URL 6**).

The K-means clustering algorithm is applied using MATLAB, following several steps (**Szabó et al., 2019; Ali and Sheng-Chang, 2020**): first, the optimum number of clusters (k) is selected by minimisation of the SSE (Sum of Squared Error) considering the prior lithological information. Then, the K-means technique groups the values into each cluster and reiterates until the data points have been distributed to their nearest centroid (see **Figure 6**). Finally, to validate the similarity between values in each cluster, a silhouette distribution is plotted (**Isfan et al., 2021**).

4. Results

The literature describes the Hollín Formation as clean siliciclastic sandstone in the stratigraphic column of the Ecuadorian Amazon Basin. This is easily identifiable in

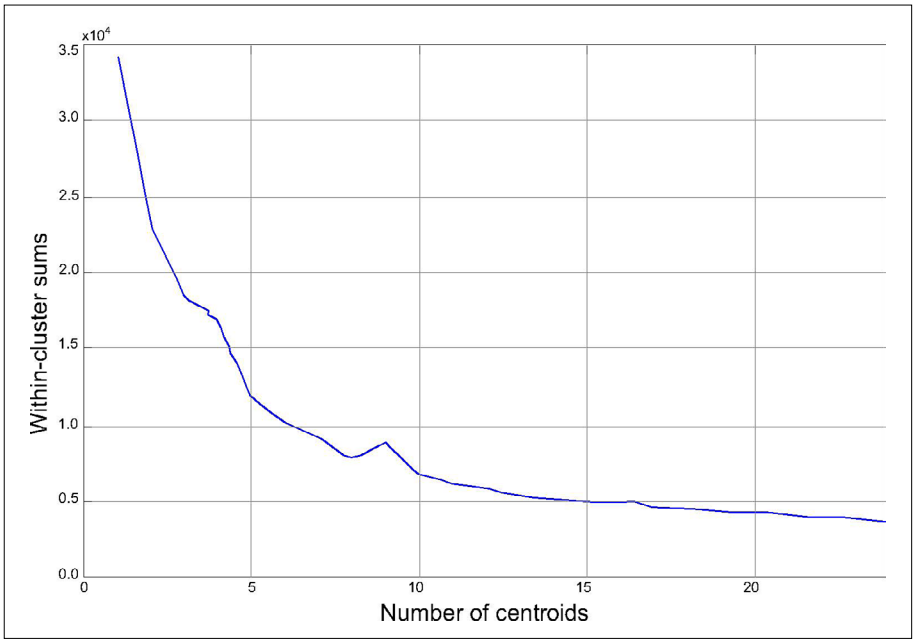

Figure 5: Elbow curve of the target dataset used to determine the optimum number of clusters

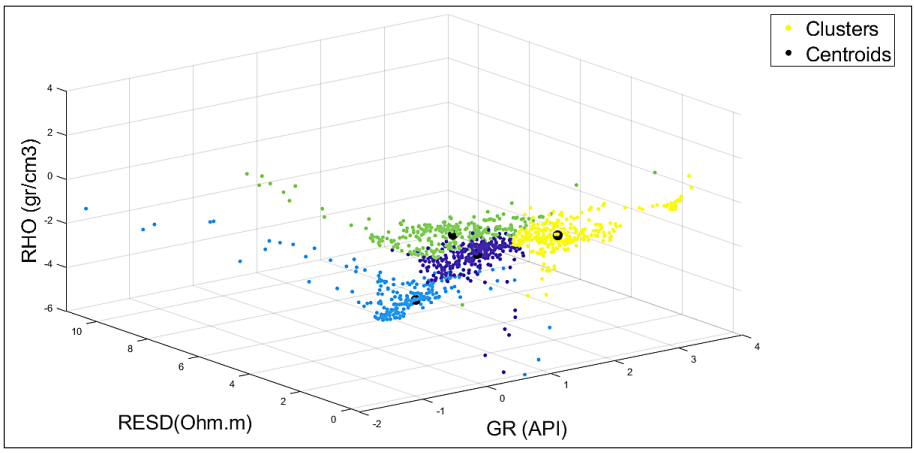

Figure 6: K-means clustering of the target data set: 3D diagram of point distribution into four clusters

Table 2: Parameters of the variables of each lithological class for the Shushufindi Oilfield with defined limits to calculate the training data set of KNN method

Rock type	GR (API)	Density (g/cm^3)	Deep resistivity (Ohm.m)	Photoelectric factor ()
Clean Sandstone	0 >, =< 40	1.52 >, < 2.69	0.9 >, =< 921	< 4.5
Intermedia Rock	40 >, < = 100	1.11 >, < 2.77	0.9 >, =< 700	< 4.5
Shale Rich	100 >	1.17 >, < 2.88	0.83 >, =< 294	< 4.5
Limestone		< 2.71		>= 4.5

the well logs by the monotonous shape with occasional spikes reflecting the small granulometric variation within the formation, and the lower shale volume of the registers. The Napo Formation has the richest shale rocks in its lower and middle members (see **Figure 2**). The main limestone layers separate facies, and they are clearly defined from the characteristic geophysical response values in neutron porosity, gamma ray, deep resistivity, and density log curves. Finally, the shaly sandstone and sandy shale are considered the intermedia lithology group, complementing the sedimentary sequence of the basin. Intermedia rock layers present strong identifiable characteristics in the shale volume curve.

Considering the qualitative analysis and the neutron-density cross plot results, the defined rock classes are clean sandstone, intermedia rocks, limestone, and shale rich rock. The parameter values of each class defined for each variable (log type) in this reservoir are summarised in **Table 2**. More detailed information about lithological interpretation can be found in **Asquith (1982)**.

The confusion charts display the interpretation results of each method, considering all the values in the target dataset. They summarise the distribution of concordance and discrepancy between the prediction and the lithological interpreted control parameter for each class. For example, the KNN method agrees with 91.4% of the data assigned as clean sandstone in the control parameter. This percentage represents 85.1% of the total predicted data by the method for this class. The remaining 14.9% reassigns data that is interpreted as intermedia rock to this class (see **Figure 7**).

The KNN method's global results agree with the control parameter in 7,513 data points (1145 m) from the total 8,603 (1311.1 m) in the target dataset. This represents agreement of 87.3%. An important fact to note is that the limestone class has 71.9% agreement, even though it represents 82.5% of the total predicted values in this class (see **Figure 7**).

The K-means method global results agree with the control parameter in 6,166 data points (939.7 m) from the total 8,603 in the target dataset. This represents 71.6% agreement between them. Here, the limestone class has 90.3% agreement with the control parameter, even though it represents only 36.6% (114.8 m) of the

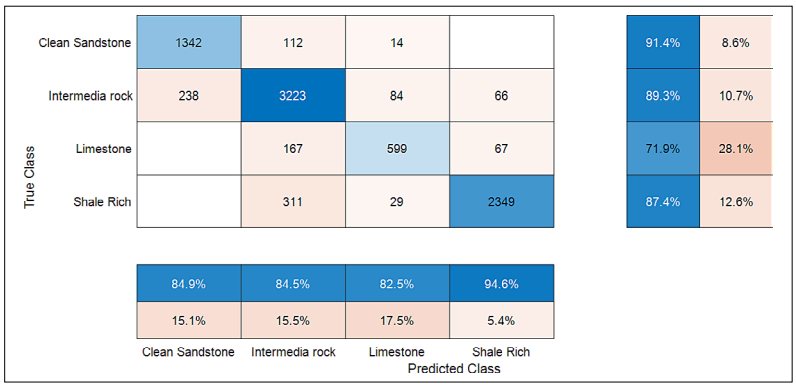

Figure 7: Confusion chart of the KNN method results showing correlation between prediction vs. interpretation of the target data set. Blue represents the percentage of predicted data assigned to the same class as the interpretation, red is the percentage of data that differed.

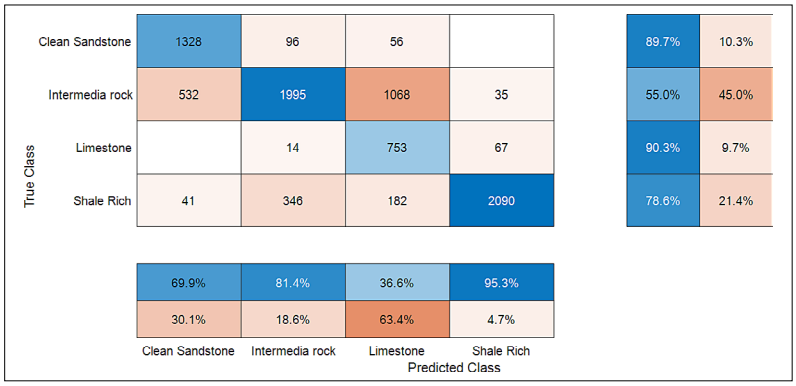

Figure 8: Confusion chart of the K-means method results showing correlation between prediction vs. interpretation of the target data set. Blue represents the percentage of predicted data assigned to the same class as the interpretation, red is the percentage of data that differed.

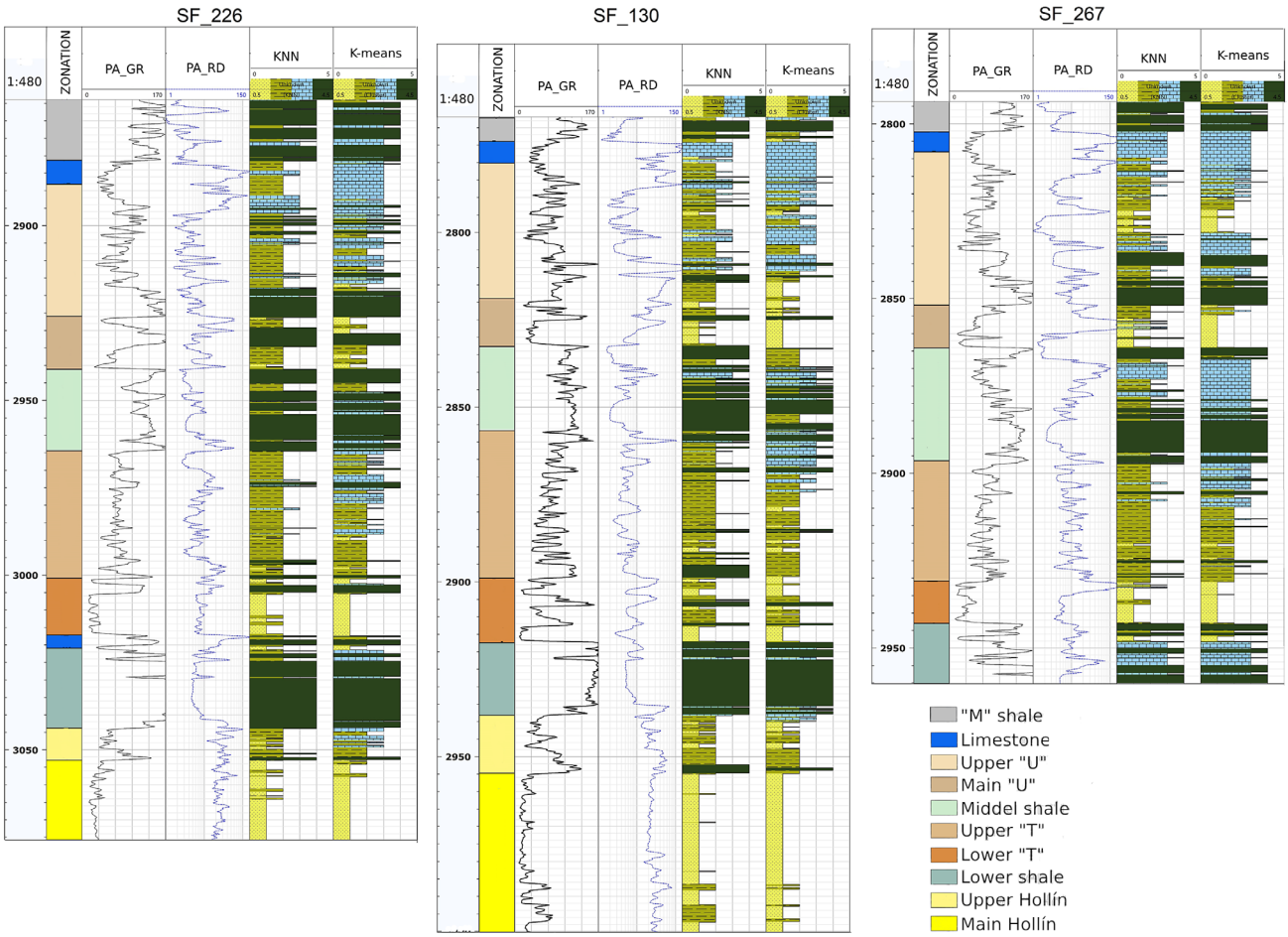

Figure 9: Stratigraphic cross-section correlating wells SF_226, SF_130, SF_267. Gamma ray and deep resistivity curves are plotted next to the lithological interpretation results of KNN and K-means methods.

total predicted values for this class. The K-means shares the cluster of limestone class with values interpreted as intermedia rock at 51.8% (1068 points or 163 m), shale rich at 8.8% (182 points or 27.7 m), and clean sandstone at 2.7% (56 points or 8.53 m) (see **Figure 8**).

Figure 9 is an example of a graphical representation of the predicted stratigraphy. Here, the discrepancy of the limestone class (light blue blocks) is visible in a correlated cross-section of the boreholes SF_226, SF_130, and SF_267.

5. Discussion

According to **Figure 7** and **Figure 8**, the KNN method used for lithological prediction with 5K has higher effectiveness in lithological determination than the K-means method. The key problem with the K-means method is the confusion between limestone and intermedia rock. In interpretation, the separation between limestone and another rock type is based only on the photoelectric log, whereas K-means uses all borehole logs for grouping. We corroborate this in **Figure 8**, where the K-means method confuses 1,068 points of intermedia rock with limestone. On the other hand, the KNN method estimates a function that predicts the litho-types using all the interpreted data, as it is a supervised statistical method. For the same litho-type, the KNN method confuses only 84 points of intermedia rock with limestone.

Human inadvertence limits well log interpretation, especially when a huge quantity of data is analysed. Consequently, the KNN method makes well log analysis more effective. Furthermore, the KNN method can be considered a tool for well log interpretation, just as classical cross plot techniques are nowadays, with the advantage that any exceptional case could be added to the first interpretation in order to use it as training data.

The intermediate rock litho-type represents a gradual change from sandstone to shale, within which sandy shale and shaly sandstone litho-types may exist. The methods are applied so that these types of subcategories are grouped, since neither the KNN prediction nor the K-means method can be applied to differentiate these categories. The difference between these categories is subjective to the percentage of clay.

For the KNN method, we analyse the number of neighbours that best corresponds with reduced discrepancies in prediction and lithological interpretation. This is a requirement when the KNN method is applied, since

it is based on the adequate search for neighbours towards the point to be predicted. With the Shushufindi Oilfield data, the number of neighbours that gives the best results is 5K (see **Figure 4**). The KNN method depends a great deal on the learning data. If training data is correct in the results, we guarantee a prediction with an agreement of over 80%. However, the method cannot be applied alone, since greater certainty must be had in the interpretation. For this reason, it is recommended as a complementary method in stratigraphic interpretation practices.

To improve the application of the K-means method, it is recommended to first separate the limestone from the other litho-types. This is because all logs are used in the method to place the centroids. However, limestone only requires a photoelectric log to be differentiated from the rest of the litho-types. Alternatively, we should only apply the method in siliciclastic reservoirs.

There is a margin of uncertainty for both methods, even if the confusion with limestone is not considered. With the KNN method, this margin of error represents 3.7% of the total points analysed, and for the K-means method this margin of error is 8.8% of the total points analysed. This allows us to recommend both methods for lithological interpretation, with the limitation that the K-means method is subject to inefficiency in differentiating limestone.

6. Conclusion

Lithology is interpreted from 8,603 points at a scale of 0.15 m of depth per point from 7 wells throughout the Shushufindi Oilfield. We identify mainly four litho-types: sandstone, intermedia rock, shale rich rock, and limestone. We corroborate the lithological interpretation with the stratigraphy that exists in the area (see **Figure 2**). The methods used for the lithological interpretation are qualitative and the classical methods with Neutron-Density cross plot. For the application of the KNN method, we interpret 1,532 points from well SF_130, representing 233.5 m of the stratum. The lithology of the total of the 8,603 points represents 1,311.1 m. Of the total points, the KNN method with 5K correctly predicts 87.3%, which means 7,513 points or 1,145 m. Of the total of 8,603 interpreted points, the K-means method correctly classifies 71.6%, which represents 6,166 points, or 939.7 m. We found a fundamental problem for using this method in the differentiation of limestone, since it confused 1,068 points that represent 163 m, which is 12.7% of the total interpreted information.

In conclusion, the KNN method was found to be the more effective method to interpret the lithology in the reservoir of the Shushufindi Oilfield. The KNN method requires more time for application than the K-means cluster analysis but allows us to avoid the ambiguity of the data by the parameterisation of the computation through the training data set. Meanwhile, the K-means method proved to be ineffective for this specific reservoir because of the high level of ambiguity due to confusion between the intermedia rock type and limestone class. Another feasible solution, out of the scope of this study, may be the application of a complex clustering technique, like hierarchical clustering, which sounds promising for lower error because of its subdivision of clusters.

Acknowledgement

Franklin Gómez would like to recognise and express his sincerest gratitude to the Escuela Politécnica Nacional Ecuador, for its sponsorship and assistance as he begins his doctoral studies at the University of Miskolc. Franklin Gómez and Yetzabbel Flores would like to recognise their sincerest gratitude to the Stipendium Hungaricum Foundation and Hungary as well for its sponsorship and assistance in the research's development.

7. References

7.1. Published works

Ali, A., and Sheng-Chang, C. (2020): Characterization of well logs using K-mean cluster analysis. Journal of Petroleum Exploration and Production Technology, 10, 6, 2245–2256. https://doi.org/10.1007/s13202-020-00895-4

Ali, M., Jiang, R., Maa, H., Pan H., Abbas, K., Ashraf, U., and Ullah, J. (2021): Machine learning - A novel approach of well logs similarity based on synchronization measures to predict shear sonic logs. Journal of Petroleum Science and Engineering, 203, 1-10. https://doi.org/10.1016/j.petrol.2021.108602.

Amonkar, Y., Farnham, D., and Lall, U. (2022): A k-nearest neighbor space-time simulator with applications to large-scale wind and solar power modelling. Patterns, 3, 3, 1-12. https://doi.org/10.1016/j.patter.2022.100454.

Asquith, G. B., and Gibson, C. R. (1982): Basic Well Log Analysis for Geologists. American Association of Petroleum Geologists, Tulsa, 117 p.

Baby, P., Rivadeneira, M., and Barragán, R. (2004): La Cuenca Oriente: Geología y Petróleo. Intitut de recherche pour le developpement, Petroecuador, Quito, 414p. (In Spanish – without English abstract)

Basal, A.M. (1998): Analytical treatment of neutron-density crossplot for shaly sand reservoirs. Journal of King Abdulaziz University: Earth Sciences, 10, 1, 115–142. doi: 10.4197/ear.10-1.8.

Bassiouni, Z., and Rhea, J. (1994): Theory, measurement, and interpretation of well logs. Henry L. Doherty Memorial Fund of AIME, Society of Petroleum Engineers, Dallas, TX, 372 p.

Biedma, D.F., Corbett, C., Giraldo, F., Lafournère, J. P., Marín, G. A., Navarre, P. R., Suter, A., Villanueva, G., and Vela, I. (2014): Shushufindi - Reawakening a giant. Oilfield Review, 26, 3, 42–58. https://www.slb.com/-/media/files/oilfield-review/4-reawake-english

Cheng, D., Yuan, X., Zhou, C., Tan, C., and Wang, M. (2016): Logging-lithology identification methods and their application; a case study on the Chang-7 Member in central-western Ordos Basin, NW China. China Petroleum Exploration, 21, 117–126. doi: 10.3969/j.issn.1672-7703.2016.05.0016

Delavar, M.R. (2022): Hybrid machine learning approaches for classification and detection of fractures in carbonate reservoir. Journal of Petroleum Science and Engineering, 208, Part A, 1-21. https://doi.org/10.1016/j.petrol.2021.109327

Erlström, M., and Sopher, D. (2019): Geophysical well log-motifs, lithology, stratigraphical aspects and correlation of the Ordovician succession in the Swedish part of the Baltic Basin. International Journal of Earth Sciences, 108, 4, 1387–1407. https://doi.org/10.1007/s00531-019-01712-y

Estupiñan, J., Marfil, R., Scherer, M., and Permanyer, A. (2010): Reservoir sandstones of the Cretaceous Napo Formation U And T members in the Oriente Basin, Ecuador: Links between diagenesis and sequence stratigraphy. Journal of Petroleum Geology. 33, 3, 221–245. https://doi.org/10.1111/j.1747-5457.2010.00475.x

Hu, Q., Yu, D., and Xie, Z. (2008): Neighborhood classifiers. Expert Systems with Applications: An International Journal, 34, 2, 866-876. https://doi.org/10.1016/j.eswa.2006.10.043

Imamverdiyev, Y., and Sukhostat, L. (2019): Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023

Isfan, I., Harsono, A., and Haris, A. (2021): Cluster Analysis of Lithology Grouping Trends using Principal Component Spectral Analysis and Complex Seismic Attributes. Makara Journal of Science, 25, 1, 21-27. https://doi.org/10.7454/mss.v25i1.1227

Khamees, L.A., Alhaleem, A. A., and Alrazzaq, A. (2021): Different methods for lithology and mineralogy recognition. Materials Today: Proceedings, 1-4. https://doi.org/10.1016/j.matpr.2021.04.531

Mamaseni, W.J., Naqshabandi, S. F., and Al-Jaboury, F. K. (2018): Petrophysical properties of the Early Cretaceous formations in the Shaikhan Oilfield/Northern Iraq. Earth Sciences Research Journal, 22, 1, 45–52. https://doi.org/10.15446/esrj.v22n1.66088

Mitra, P., Murthy, C.A., and Pal, S. K. (2002): Density-based multiscale data condensation. IEEE Transactions on Pattern analysis and Machine Intelligence, 24, 6, 734-747. https://doi.org/10.1109/TPAMI.2002.1008381

Pratama, H. (2019): Machine Learning: Using Optimized KNN (K-Nearest Neighbors) to Predict the Facies Classifications. Proceedings of the 13th SEGJ International Symposium, Tokio, Japan, 538–541. https://doi.org/10.1190/segj2018-139.1

Ramirez, F.A. (2020): Cretaceous Napo U and Napo T Sandstone Channels Accommodation Space Created by Erosion and Tectonism, Exploration and Development Implications on Western Ecuadorian Oriente Basin. 2020 AAPG Annual Convention and Exhibition Online Meeting. https://doi.org/10.1306/11351Ramirez2020

Szabó, N.P., Nehéz, K., Hornyák, O., Piller, I., Deák, C., Hanzelik, P. P., Kutasi, C., and Ott, K. (2019): Cluster analysis of core measurements using heterogeneous data sources: An application to complex Miocene reservoirs. Journal of Petroleum Science and Engineering. 178, 575–585. https://doi.org/10.1016/j.petrol.2019.03.067

Szabó, N.P., Braun, B. A., Abdelrahman, M. M. G., and Dobróka, M. (2021): Improved well logs clustering algorithm for shale gas identification and formation evaluation. Acta Geodaetica et Geophysica. 56, 711-729. https://doi.org/10.1007/s40328-021-00358-0

Troccoli, E.B., Cerqueira, G. A., Lemos, B. J., Holz, M. (2022): K-means clustering using principal component analysis to automate label organization in multi-attribute seismic facies analysis. Journal of Applied Geophysics. 198, 104555. doi: 10.1016/J.JAPPGEO.2022.104555.

Zhang, H., and Li, J. (2016): Lower Napo Shale Formation in the Oriente Basin: Analysis of Reservoir Formation Conditions and Prediction of Favorable Areas. Advances in Petroleum Exploration and Development, 11, 2, 6–15. https://doi.org/10.3968/8502

Zhang, J., He, Y., Zhang Y., Li W., and Zhang J. (2022): Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoir Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin. NW China. Energies. 15, 10, 3675. https://doi.org/10.3390/en15103675

Zhao, J., Wang, F., and Lu, Y. (2017): Application of Multivariate Membership Function Discrimination Method for Lithology Identification. Jsm 46, 2223–2229. doi:10.17576/jsm-2017-4611-24

Zhao, T., Jayaram, V., and Marfurt, K. (2015): TOC estimation in the Barnett Shale from triple combo logs using support vector machine. SEG Denver 2014 Annual Meeting. 1491–1495. https://doi.org/10.1190/segam2015-5922788.1

7.2. Unpublished reports

Renss, G. (2016): Alcance a la reevaluación del diagnóstico y plan de manejo ambiental del campo Shushufindi para la ampliación y perforación adicional de pozos de desarrollo en la plataforma Aguarico 05, ampliación de las plataformas Shushufindi 34, Shushufindi 51, Shushufindi Vol. 1. (In Spanish – without English abstract)

Salazar, A. B. (2014): Actualización de las reservas y ubicación de pozos de relleno para incrementar la producción del Campo Shushufindi. Escuela Politécnica Nacional, Quito, 273 p. (In Spanish – without English abstract)

Tomalá, S. M. (2020): Estudio del comportamiento de fluidos en procesos de inyección de agua como método de recuperación secundaria en la arenisca Napo "U" Inferior del Campo Shushufindi. Universidad Estatal Península de Santa Elena, Santa Elena, 36 p. (In Spanish – without English abstract)

7.3. Internet sources

URL 1: http://pages.geo.wvu.edu/~jtoro/Petroleum/15_logs.pdf (accessed 7t June 2022)

URL 2: https://www.geol.umd.edu/~jmerck/geol342/lectures/19.html (accessed 7t June 2022)

URL 3: https://www.kgs.ku.edu/PRS/ReadRocks/LDS.html (accessed 7t June 2022)

URL 4: https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/ (accessed 07th June 2022)

URL 5: https://www.analyticsvidhya.com/blog/2021/02/simple-explanation-to-understand-k-means-clustering/ (accessed 07th June 2022)

URL 6: https://www.mathworks.com/help/stats/kmeans.html#buefs04-3 (accessed 07th June 2022)

SAŽETAK

Usporedna analiza metode najbližega susjedstva vrste K te klasterske analize vrste K-sredine u svrhu litološke interpretacije bušotinske karotaže naftnoga polja Shushufindi u Ekvadoru

Litološka interpretacija karotaže predstavlja jednu od temeljnih interpretacija u geoznanostima, a moguće ju je ostvariti primjenom različitih algoritama strojnoga učenja. U ovome istraživanju testirane su metode najbližega susjedstva (engl. skr. NN) vrste K (za procjenu gustoće) i klasterske analize vrste K-sredine kod predviđanja litologije iz karotažnih podataka izmjerenih u siliciklastičnome ležištu naftnoga polja Shushufindi u Ekvadoru. Usporedba litološke interpretacije napravljena je korištenjem kvalitativne interpretacije te karotaže gustoće i neutrona. Rezultati su pokazali kako KNN bolje predviđa na temelju interpretiranih podataka (87,3 %, tj. 1145 m predviđeno od interpretiranih 1311,1 m) nego rezultati dobiveni klasterskom analizom K-sredina (71,6 %, tj. 939,7 m predviđeno od interpretiranih 1311,1 m). Geologija ležišta uvjetuje određenu razinu odstupanja zbog vrlo sličnih geofizičkih odgovora između vapnenaca i srednjozrnatih klastita. Mogućnost kontrole u KNN-u čini algoritam preporučljivim za litološku interpretaciju sličnih ležišta.

Ključne riječi:
strojno učenje, ležište, litologija, Shushufindi

Authors' contribution

Franklin Gómez (Petroleum Engineer and Master in Fluid Thermodynamics. Professor at Escuela Politécnica Nacional, Ecuador, student in the Mikoviny Samuel Doctoral School of the University of Miskolc, 3 years of experience in workover and hydraulic fracturing) elaborated the algorithm in Matlab of the KNN method and the lithological interpretation. He had the main idea of the article. **Yetzabbel Flores** (Hydrogeologist, student in the Mikoviny Samuel Doctoral School of the University of Miskolc) performed the lithological interpretation, produced many of the figures, and carried out the K-means cluster analysis algorithm in Matlab, as well as contributing to the literature review. **Marianna Vadászi** (Master of Business Administration in University of Miskolc, Faculty of Economics. PhD degree in Earth Sciences from the University of Miskolc, Mikoviny Sámuel Doctoral School of Earth Sciences) was responsible for the complete revision of the document and supervision of the doctoral research.