Characterization of well logs using K-mean cluster analysis

Amjad Ali · Chen Sheng-Chang

Received: 13 October 2019 / Accepted: 25 April 2020 / Published online: 6 May 2020
© The Author(s) 2020

Abstract
The identification process of different lithologies, hydrocarbons, and water-saturated zones in oil and gas industries involves petrophysical studies that are carried out by geoscientists using different software packages. This study aims to propose a method by integrating mean cluster analysis and well logs to identify dominant lithologies, pore fluids, and fluids contact. For this purpose, initially, K-mean cluster analysis is applied to density log and P-wave velocity data of three wells in order to group them into different clusters. Based on centroids of each cluster, different lithologies have been identified. The density log equation has been utilized to compute the porosity of each cluster, and the mean of each density log cluster is used as matrix density. Next, sonic log equation has been inverted to compute the fluid velocity and the mean of each P-wave velocity cluster is used as matrix velocity. For the fluid density, sonic and density log equations are jointly inverted to compute the fluid velocity of each cluster. The fluid bulk modulus and acoustic impedance are computed using fluid density and velocity. Based on the results of K-mean cluster analysis, different lithologies (shale, sandstone, and limestone) have been recognized successfully. In well-1, hydrocarbon and water-saturated zones are successfully identified and fluids contact has been established in the zone of interest. However, well-2 and well-3 did not show any indications of the presence of hydrocarbon in the respective zones.

Keywords K-mean cluster analysis · Density log · P-wave velocity · Fluid velocity · Fluid bulk modulus · Acoustic impedance

Introduction
In the oil and gas exploration industry, generally, petrophysicists interpret well logs on the basis of their previous professional experience to identify hydrocarbon, water saturation zones, and fluids contact. The interpretation of well logs changes from person to person, and this mainly depends on the experience and knowledge of a geoscientist while the reserves estimation is a fundamental part of exploration activity. The characterization of pore fluids and fluids contact in a reservoir is very important for volumetric computation of reserves estimation in a hydrocarbon reservoir (Chombart 1960).

Over the period of time, numerous techniques have been developed for reservoir characterization. For example, Gutierrez et al. (2000) used a rock physics model to identify pore fluids from the sonic log and Mukerji et al. (2001) integrated rock physics and seismic data through geostatistics in order to reduce the uncertainty of seismic reservoir characterization while Mark Sams (2001) used the geostatistical inversion for lithological and impedance modeling for reservoir characterization. Afterward, Galikeev and Davis (2003) merged 4D seismic attributes and geostatistics for thin carbonate reservoir characterization while González et al. (2006) integrated rock physics and seismic inversion using multipoint geostatistics for reservoir characterization. Later on, Jiang and Yang (2012) incorporated stochastic inversion and petrophysical properties for a better understanding of subsurface rock properties. Furthermore, MacAllister et al. (2014) used rock physics with AVO analysis to identify stratigraphic reservoir and Nicholls et al. (2014) utilized 3D and 4D rock physics parameters to get a better picture of the reservoir and enhance the production. Moreover, Jing et al. (2015)
did the quantitative studies and integrated the rock physics with seismic to identify hydrocarbon-saturated zone. During the same period of time, Yang et al. (2015) combined the elastic properties with AVO for quantitative studies to identify potential hydrocarbon reservoir and Bredeesen et al. (2015) incorporated rock physics with AVO to identify the source and reservoir rocks while Wang et al. (2016) used three new rock physics parameters approximation from AVO for deep reservoir studies and Xiwei and Zhu (2016) used stochastic impedance inversion using the well and seismic data with the control sedimentary facies. On the other side, Huidong et al. (2016) used the multi-point geostatistical model for reservoir characterization of a sand body and Jie et al. (2016) utilized AVO and seismic tuning for the identification of thin layer hydrocarbon reservoir. Recently, Gomes de Mello e Silva and Beneduzi (2017) applied an empirical method to sonic log to identify pore fluid in a siliciclastic reservoir and Ali and Al-Shuhail (2018) did the joint inversion of P-wave velocity and impedance to identify fluids contact. Zhou et al. (2017) used the asymptotic equation for fluid identification in a carbonate reservoir.

In the present world, artificial intelligence and computer machines are playing a very important role. Due to their wide range of applications, researchers are turning towards the combination of artificial intelligence and computer machines. They are seeking their applications in every field of science. For example, Lin and Salisch (1994) presented a mathematical model by combining principle component and cluster analysis along with Bayes discriminant analysis to determine different lithologies, porosity, and permeability. Euzen et al. (2010) integrated cluster analysis and electrofacies in order to identify an unconventional hydrocarbon prospect in Upper Mannville incised valley. Dudley et al. (2016) combined the borehole data and cluster analysis with real-time microseismic data to characterize an unconventional reservoir. Cluster analysis is an unsupervised machine learning algorithm of clustering/grouping a large dataset into significant subgroups so that the data points in same class have same characteristics and different from another subgroup. It was first introduced by Queen (1966). It has a number of applications in various fields, for example data mining (Fayyad et al. 1996), compression of data, and quantization of vector (Gersho and Gray 2012). Generally, K-mean cluster analysis is used for continuous datasets (Fukunaga 2013; Duda et al. 1973). The purpose of clustering is to discover important patterns in large datasets (Wagstaff 2012). K-mean cluster analysis is easy and simple to understand, and it is fast and robust to cluster large dataset. In this study, in order to minimize the human error during the interpretation of well logs and to get the better reliable results, we are going to integrate K-mean cluster analysis and well logs to identify dominant lithologies and zone of interest in the well logs. It can be a helpful tool to interpret large well logs datasets and it can reduce human error in order to get better and reliable results.

Methodology

In this study, we have incorporated three wells logs for the computation purpose. The physical structure of the methodology can be seen in Fig. 1. K-mean cluster analysis is a well-known clustering algorithm because of its easy implementation and efficiency (Nazeer and Sabastian 2009). K-mean clustering is an unsupervised learning algorithm, and the main aim of K-mean clustering is to partition n number of observations into K number of clusters. For the numerical dataset, the center of each cluster is represented by the mean/centroid. In each cluster, every observation belongs to the nearest mean. Mathematically, K-mean cluster analysis can be written as (Wang et al. 2012);

\[
J = \sum_{j=1}^{K} \sum_{i=1}^{n} (x_i - C_j)^2
\]

here in Eq. 1, J is an objective function, K is number of clusters, n is the number of observations, \(x_i\) is the observation, and \(C_j\) is centroid for cluster j. K-mean cluster analysis works in the following way;

1. First, we need to select the number of clusters (K) randomly.
2. Based on the number of clusters, the K-mean cluster will divide the data points into subsets and it will allocate the centroids to each subset or cluster.
3. Then we will redefine the number of clusters (K) and K-mean will compute the clusters again and assign the data points to their nearest centroids.
4. The second and third step will be iterating until the arithmetic means/centroids do not change any more.

At first stage, we need to select an arbitrarily number of clusters for the given dataset and then compute the distance between each data points and cluster and allocate it to nearest cluster. Now update the number of clusters and its average. Then redefine the number of cluster and compute the new averages and assign the data points to the new cluster based on their closest cluster. Repeat this process, until there is no change in centroids.

K-mean clustering is a very efficient and robust method, but we need to predefine the number of clusters (K). There is no optimal number of clusters, and the best approach is to compute multiple clusters and compare the results of each cluster analysis and choose the best one.
In this study, initially, K-mean clustering is applied to the density log and P-wave velocity data of three wells to group them into different clusters and to get the mean of each cluster. The mean value of each cluster is given in Table 1. After K-mean clustering of each well, the density log equation is used to compute the porosity of each density log cluster, as (Alger et al. 1963):

\[
\phi = \frac{\rho_m - \rho_b}{\rho_m - \rho_f}
\]  

(2)

where \(\phi\) is porosity in fraction, \(\rho_m\) is matrix density in g/cm\(^3\) and is the mean of each density log cluster, as given in Table 1, and \(\rho_b\) is bulk density value from density log in g/cm\(^3\), \(\rho_f\) is the fluid density value, and here it is considered as 1.1 g/cm\(^3\) value of brine water.

After this, the moving average filter is used to remove any outliers from the porosity values and to make the porosity of each cluster smooth and acceptable.

Then P-wave velocity is computed from the sonic log as:

\[
V_p = \left( \frac{10^6 \cdot 3.28}{V_t + \frac{1 - \phi}{V_m}} \right)
\]  

(3)

In Eq. 3, \(V_p\) is P-wave velocity in m/s and DT are sonic log values.

Here the sonic log equation has been used to compute the fluid velocity (Tixier et al. 1959):

\[
\frac{1}{V_p} = \phi \frac{1 - \phi}{V_f} + \frac{1}{V_m}
\]  

(4)

where \(V_p\) is P-wave velocity computed from the sonic log in m/s, \(\phi\) is porosity in fraction, \(V_f\) is fluid velocity in m/s, and \(V_m\) is matrix velocity in ma/s.

Sonic log equation (Eq. 4) is inverted to compute the fluid velocity of each cluster of P-wave velocity data, as:

\[
V_f = \frac{V_p * (\phi - \phi^2) - V_m * \phi}{(1 - \phi)}
\]  

(5)

here in Eq. 5, \(V_f\) and \(V_p\) are fluid velocity and P-wave velocity in m/s, respectively. \(V_m\) is matrix velocity, and it is the mean of each cluster of P-wave velocity data, as given in Table 1. \(\phi\)
is porosity in fraction and fluid velocity is computed for each cluster, and then, moving average filter is used to remove any outliers.

For the computation of fluid density, the density log Eq. 2 and sonic log Eq. 4 are inverted as:

$$\rho_f = \rho_m - \frac{(\rho_m - \rho_b) \cdot V_p}{V_f} \left(\frac{V_m - V_f}{V_m - V_p}\right)$$  \hspace{1cm} \text{(6)}

The fluid density is computed for each cluster, and then, moving average filter is used to remove any outliers. The fluid bulk modulus is computed as:

$$K_f = (V_f^2 \cdot \rho_f \cdot 1000)$$ \hspace{1cm} \text{(7)}

here in Eq. 6, $K_f$ is fluid bulk modulus in Pascal.

Generally, acoustic impedance is the product of the density and velocity of rocks. In this the study to confirm the fluids contact, we used the fluid density and fluid velocity in order to establish and confirm the fluids contact, as:

$$\text{AI} = V_f \cdot \rho_f \cdot 1000$$ \hspace{1cm} \text{(8)}

In the study, the mean values of porosity and density of each cluster are provided in Table 1.

### Results

In this study, three different well logs data have been used for the computation purpose. Initially, K-mean clustering is adopted for the cluster analysis of density log and P-wave velocity data in order to group these two datasets into different clusters and to get the mean of each cluster. The mean of each cluster of density log and P-wave velocity data is used as matrix density and matrix velocity for further computation, respectively. The porosity, fluid velocity, fluid bulk modulus, and AI are computed in order to identify main lithologies, a potential hydrocarbon reservoir, pore fluids, and fluids contact.

### Cluster analysis and porosity computation

K-mean cluster analysis clustered the density log and P-wave velocity data of well-1 and well-2 into 9 and well-3 into 6 clusters, respectively, as shown in Figs. 2, 3, 4, 5, 6, and 7. For the porosity computation, the density log equation (Eq. 2) has been used to compute porosity for each cluster of density logs. Figures 8, 9, and 10 are correlation matrix plots between porosity and density of each well, respectively. For each cluster, K-mean cluster analysis gave us the centroid. Based on the centroid of each cluster, lithological discrimination has been done; centroids along with their respective lithologies are tabulated in Table 1. The standard published

| S/no | Cluster | Mean value ($\rho_m$) (g/cm$^3$) | Mean value ($V_m$) (m/s) | Lithology   |
|------|---------|---------------------------------|--------------------------|-------------|
| Well-1 | Cluster-1 | 2.013 | 1671 | Shale  |
|       | Cluster-2 | 2.066 | 1768 |        |
|       | Cluster-3 | 2.067 | 2062 |        |
|       | Cluster-4 | 2.15  | 2369 |        |
|       | Cluster-5 | 2.153 | 2582 |        |
|       | Cluster-6 | 2.46  | 2781 | Sandstone |
|       | Cluster-7 | 2.699 | 3604 | Limestone |
|       | Cluster-8 | 2.487 | 3029 | Sandstone |
|       | Cluster-9 | 2.513 | 3397 |        |
| Well-2 | Cluster-1 | 2.129 | 2190 | Shale  |
|       | Cluster-2 | 2.286 | 2541 |        |
|       | Cluster-3 | 2.295 | 2920 |        |
|       | Cluster-4 | 2.389 | 3175 |        |
|       | Cluster-5 | 2.53  | 3456 | Sandstone |
|       | Cluster-6 | 2.556 | 3641 |        |
|       | Cluster-7 | 2.614 | 3391 |        |
|       | Cluster-8 | 2.621 | 3302 |        |
|       | Cluster-9 | 2.644 | 3710 |        |
| Well-3 | Cluster-1 | 2.12  | 1910 | Shale  |
|       | Cluster-2 | 2.169 | 2480 |        |
|       | Cluster-3 | 2.511 | 3297 | Sandstone |
|       | Cluster-4 | 2.568 | 4863 |        |
|       | Cluster-5 | 2.452 | 3519 |        |
|       | Cluster-6 | 2.482 | 3402 |        |
value of matrix density of quartz is 2.65 g/cm$^3$ (for sandstone) and calcite is 2.71 g/cm$^3$, and it uses for limestone. Matrix density actually represents the mineral density of rock, and mostly rocks are heterogeneous. So instead of using the published values for matrix density and matrix velocity, the centroid of each cluster is used as matrix density ($\rho_m$) and matrix velocity ($V_m$) for further computation, respectively. For all the three wells, the values of $\rho_m$ and $V_m$ of each cluster are given in Table 1.

For well-1, based on the mean of each cluster, four main lithologies are distinguished. From surface to 1500 m, the mean of the first 5 clusters are $\leq 2.2$ g/cm$^3$ so it is categorized as shale; then from 1500 to 1750 m, the mean of sixth cluster is 2.5 g/cm$^3$ and it is categorized as sandstone;
next, from 1750 to 2000 m, the mean of seventh cluster is 2.7 g/cm³ and it is categorized as limestone; and last, from 2000 to 2500 m, the mean of eighth and ninth clusters is ≤ 2.5 g/cm³ and they are categorized as sandstone, as shown in Fig. 8. So, there are three main zones of interest based on lithology distribution which are starting from 1500 to 2500 m. Porosity in the zones of interest is ranging from < 5 to 30%, as shown in Fig. 8.

For well-2, based on the mean of each cluster, two main lithologies are identified. From surface to 1500 m, the mean of first 4 clusters is ≤ 2.2 g/cm³ so it is categorized as shale, while the mean of remaining clusters has ≥ 2.5 g/cm³; therefore, they are categorized as sandstone dominant lithology. In the sandstone dominant part, the porosity is ranging from < 1 to 15%, as shown in Fig. 9.

For well-3, based on the mean of each cluster, two main lithologies are identified. From surface to 450 m, the mean of the first two clusters is < 2.2 g/cm³ so it is categorized as shale, while the mean of remaining clusters has 2.5 g/cm³; therefore, they are categorized as sandstone dominant lithology. In the sandstone dominant part, the porosity is ranging from < 1 to 25%, as shown in Fig. 10. In Table 2, for all three wells, porosities with their corresponding depths and lithologies are tabulated.

**Pore fluids and fluids contact identification**

For the fluid velocity computation, Eq. 5 has been used to compute the fluid velocity for each cluster of all three wells. Initially, K-mean cluster analysis is applied to P-wave velocity data to cluster it into 9 and 6 clusters, respectively, and to get the mean of each cluster, as shown in Figs. 5, 6, and 7. Instead of using the published values of matrix velocity from the literature for the computation, the mean of each cluster of P-wave velocity data is used as matrix velocity. Figure 11 is the cross-plots of fluid velocity, bulk modulus, and acoustic impedance of well-1. Based on data distribution, the green color (fluid velocity) is ranging from 1250 to 1800 m/s and it is identified as brine water-saturated zones while the data distribution of red color (fluid velocity) is ranging from 0 to 800 m/s and it is identified as hydrocarbon-saturated zones. The cross-plots of fluid bulk modulus and acoustic impedance also endorse the results of fluid velocity, as shown in Fig. 11. Based on fluid velocity, bulk modulus, acoustic impedance, the hydrocarbon-saturated zones suggest that it is live oil. Using fluid bulk modulus and acoustic impedance, two fluid contacts are marked at two different depths, first at 1500 m and second at 1950 m.
The results of fluid velocity, bulk modulus, and the acoustic impedance of well-2 and well-3 did not give us any prominent indication about the presence of hydrocarbon reservoir in the sandstone dominant part, as shown in Figs. 12 and 13. This shows that the well-2 and well-3 were not successful exploratory wells. The fluid velocity range in these two wells exhibits that there are some traces of hydrocarbon present.

**Conclusion**

Well logs interpretation is an important part of exploration activity in order to identify different lithologies, zone of interest, and volumetric computation of reserves estimation. In exploration industry, these objectives are mainly achieved through the experience of a geoscientist, which varies from person to person. Human error always exits during the interpretation process. The main objective of this study is to overcome this human error through integration of K-mean cluster analysis with well logs to identify main lithologies, the zone of interest, pore fluids, and fluids contact. In the absence of prior knowledge of the subsurface lithologies and formations, all the objectives have been successfully achieved through this study. At the first stage, K-mean clustering is applied to density log and P-wave velocity data to make the clusters. Based on the centroids of each cluster, the dominant lithologies are identified. The porosity and fluids velocity of each cluster are computed. Finally, fluid bulk modulus and AI are computed for each cluster to confirm the existence of pore fluids and fluids contact. This study shows that:

- K-mean cluster analysis is an easy and robust algorithm to implement.
- K-mean cluster analysis is a good tool to identify the main subsurface lithologies.
• Based on the centroids of cluster analysis, we can identify the zone of interest in well logs.
• K-mean cluster analysis is a good tool to minimize human error during the interpretation of well logs.
• The standard published value of matrix density actually represents the mineral density of rock, and mostly rocks are heterogeneous. So instead of using published data for matrix density and velocity, we can use the mean of each cluster as matrix density and velocity.
• Inversion of fluid velocity, fluid bulk, and acoustic impedance is good tools, and if they are integrated with K-mean cluster analysis, these parameters can be helpful to identify the pore fluids and fluids contact.

![Well correlation matrix of well-3 among depth, density, and porosity with diagonal histogram and correlation coefficient](image)

**Table 2** Porosity with their corresponding lithologies and depths in each well

| S/no | Depth range (m) | Lithology | Porosity range |
|------|----------------|-----------|----------------|
| **Well-1** | | | |
| 1 | 0–1500 | Shale | 0–30% |
| 2 | 1500–1750 | Sandstone | 1–30% |
| 3 | 1750–2000 | Limestone | 1–27% |
| 4 | 2000–2500 | Sandstone | 1–13% |
| **Well-2** | | | |
| 1 | 0–1500 | Shale | 0–35% |
| 2 | 1500–3130 | Sandstone | 1–16% |
| **Well-3** | | | |
| 1 | 0–425 | Shale | 0–21% |
| 2 | 1200–1940 | Sandstone | 1–25% |
Fig. 11 Fluid velocity, bulk modulus, and acoustic impedance well-1
**Fig. 12** Fluid velocity, bulk modulus, and acoustic impedance of well-2
Fig. 13 Fluid velocity, bulk modulus, and acoustic impedance of well-3

References

Alger RP, Raymer LL Jr (1963) Formation density log applications in liquid-filled holes. J Pet Technol 15(03):321–332. https://doi.org/10.2118/435-PA

Ali A, Al-Shuhail AA (2018) Characterizing fluid contacts by joint inversion of seismic P-wave impedance and velocity. J Pet Explor Prod Technol 8(1):117–130. https://doi.org/10.1007/s13202-017-0394-3

Bredesen K, Jensen EH, Johansen TA, Avseth P (2015) Seismic reservoir and source-rock analysis using inverse rock-physics modeling: a Norwegian Sea demonstration. Lead Edge 34(11):1350–1355. https://doi.org/10.1190/tle34111350.1
