Evaluation of mineralogy per geological layers by Approximate Bayesian Computation

Vianney Bruned, Alice Cleynen, André Mas, Sylvain Wlodarczyck

Evaluation of mineralogy per geological layers by Approximate Bayesian Computation

Vianney Bruned[1], Alice Cleynen[2], André Mas[2], and Sylvain Wlodarczyck[1]

[1]Schlumberger Petroleum Services, Montpellier, 34000, France.
[2]IMAG, Univ Montpellier, CNRS, Montpellier, France.

Abstract

We propose a new methodology to perform mineralogic inversion from wellbore logs based on a Bayesian linear regression model. Our method essentially relies on three steps. The first step makes use of Approximate Bayesian Computation (ABC) and selects from the Bayesian generator a set of candidates-volumes corresponding closely to the wellbore data responses. The second step gathers these candidates through a density-based clustering algorithm. A mineral scenario is assigned to each cluster through direct mineralogical inversion, and we provide a confidence estimate for each lithological hypothesis. The advantage of this approach is to explore all possible mineralogy hypotheses that match the wellbore data. This pipeline is tested on both synthetic and real datasets.

Keywords : Mineralogical Inversion; Inverse problem; Wellbore log; Approximate Bayesian Computation; Clustering

1 Introduction

One of the main goals of reservoir evaluation is the determination of petrophysical parameters like porosity, permeability or water saturation. In order to get an accurate estimation of these parameters, a complete characterization of the lithology or the nature of the rocks is necessary. The petrophysicist proceeds to the analysis of wellbore logs which often requires the input from an expert. Indeed, petrophysical inversion of wellbore logs yields a selection of minerals or fluids belonging to the formation usually with more unknowns (the mineralogy) than measurements (the logs). In a bulk density-neutron porosity cross-plot, an expert may identify the presence of gas, limestone or an exotic mineral. But these choices may not always be obvious from a direct lecture of the logs.

We can summarize roughly the characterization of the lithology based on conventional logs using a classical inversion approach like Elan or Multi-min (described in Mayer et al. [1980], Quirein et al. [1986], Cannon et al. [1990] or Peeters et al. [1991]) by the following steps:

- Definition of the zones of the well.
- Selection of the mineralogical components and fluids for each zone, and computation of the lithology using an inversion approach.
- Tuning the physical parameters (endpoints) of the components (when needed).

The petrophysicist will iterate the last two points until they reach a model fitting the data or having a good match with core data. Tuning the endpoints is often necessary for the shaly component. For instance, the variability of the endpoints of the kerogen is very high through the geographical areas: the gamma ray ranges between 500 API to 4000 API.

We propose a Bayesian approach to select minerals in a stratum. Forward modeling liberates us from the underdetermination of the classical inversion problem using conventional logs. Bayesian inversion methods have been mainly used to solve inverse problems in the domain of rock-physics or geophysics (Tarantola [2005]). Many methods use a Bayesian framework to solve the amplitude versus offset (AVO) inversion and to obtain petrophysical attributes like porosity, water saturation or volume of clay (Grana [2016], Xu et al. [2016], Rimstad et al. [2012]).

Other petrophysical approaches to retrieve interesting features have been tested in Qin et al. [2017], da Costa et al. [2008], Sanchez-Ramirez et al. [2010]. These methods mainly focus on the uncertainties of the logs and the analysis of the posterior distribution of the petrophysical outputs. Except in Yang et al. [2013], the minerals considered in the Bayesian inversion are very limited (Shale, Sand, Water, Hydrocarbron and sometimes Carbonate) and the model linking the minerals to the logs are quite simple. Moreover, these methods often have a number of unknowns equal to the number of logs available.

In this paper, we assume that the geological layers are known (we may refer to historical references Wolf and Pelissier-Combescure [1982], Moline et al. [1991] and Ye and Rabiller [2000] or to the more recent Rebelle and Lalanne [2014] for a description of some segmentation algorithms for layer detection). Hence we assume that these geological layers have a constant mineral composition, and we describe the model in Section 2.1. Unlike previous works, our model is flexible enough to allow complex relationships between the logs and the volumes, and we do not restrict ourselves to a low number of minerals. We introduce a three-step method to get the different lithological hypotheses per layers in order to solve the underdetermined mineralogical inversion problem. First, we use an Approximate Bayesian Computation (ABC) technique to get the posterior probability of the lithology (see Section 2.2.1). Then we use a density-based clustering algorithm on this posterior probability to distinguish the different lithological hypotheses, as described in Section 2.2.2. Finally, for each of these hypotheses, we propose a method to tune the endpoints based on the resolution of a global optimization problem. Our method is therefore entirely automated, and proposes different plausible lithological hypotheses as well as some confidence estimate for each of them. We illustrate its performance on both synthetic (see Section 3) and real datasets (see Section 4.)

2 Methodology

2.1 Model and Context

In mineralogical inversion, a crucial step is the choice of the components of the lithology. Indeed, starting from d logs and using a classical inversion imply a maximum of $d+1$ minerals or fluids in the inversion model. But usually, M the number of mineral components is such that $M \gg d$, so a model selection is needed. Here, we denote the elements of the lithology by $V \in \mathbb{R}^M$ or volume (volumetric fraction) and the logs by $L \in \mathbb{R}^d$. V represents the volumetric fraction of the minerals and the fluids. We assume that for each depth n, we have:

$$L_n = G\left(V_n\right) + X_n, \tag{1}$$

where $X_n \sim N\left(0, \Sigma\right)$ ($\mathcal{N}$ is a multivariate normal distribution and $\Sigma \in \mathcal{M}\left(\mathbb{R}^d\right)$ is the covariance matrix) and G an operator from $\mathbb{R}^M$ to $\mathbb{R}^d$. Besides, the volumes $V_{i,n}$ are constrained by:

$$\sum_{i=1}^{M} V_{i,n} = 1,$$
$$V_{i,n} \geq 0 \text{ for all i.}$$

Our aim is to select the minerals that may appear within the geological stratum. Computing the exact amount of the volumetric fraction of the minerals is performed in the final step. Notice that whenever the number of selected components is M with $M < d + 1$, a classical inversion program can be run to obtain the exact volumetric fraction. The physical parameters/endpoints of the different minerals or fluids are fixed.

We consider that d is around 4-5, we often have a triple combo (gamma ray: GR, neutron porosity: ϕ_N, bulk density: ρ_b) plus the resistivity R_t and the photoelectric factor pef or the sonic Δ_t. Petrophysical models involve usually between 10 and 15 components. We order them in different classes of minerals: Shale, Sand, Carbonate. The porosity ϕ, containing the different fluids (water, oil or gas), is added to the model. This classification is important for the prior of the lithological model. Table 1 illustrates an example of a lithological model with the main families of minerals.

Family	Components
Sand-Mica	Quartz, Plagioclase, Mica, Feldspar
Carbonate	Calcite, Dolomite, Ankerite
Clay/Shale	Illite, Chlorite, Smectite, Kaolinite...
Porosity	Water, Oil or Gas
Others	Halite, Anhydrite, Pyrite..

Table 1: Example of lithological model

2.2 Method

We propose in this section a two-step method based on Approximate Bayesian Computation or ABC (see Marin et al. [2012] for a review of this Bayesian method) and density-based clustering to get the lithological hypothesis on a given stratum. Figure 1 summarizes the proposed methodology.

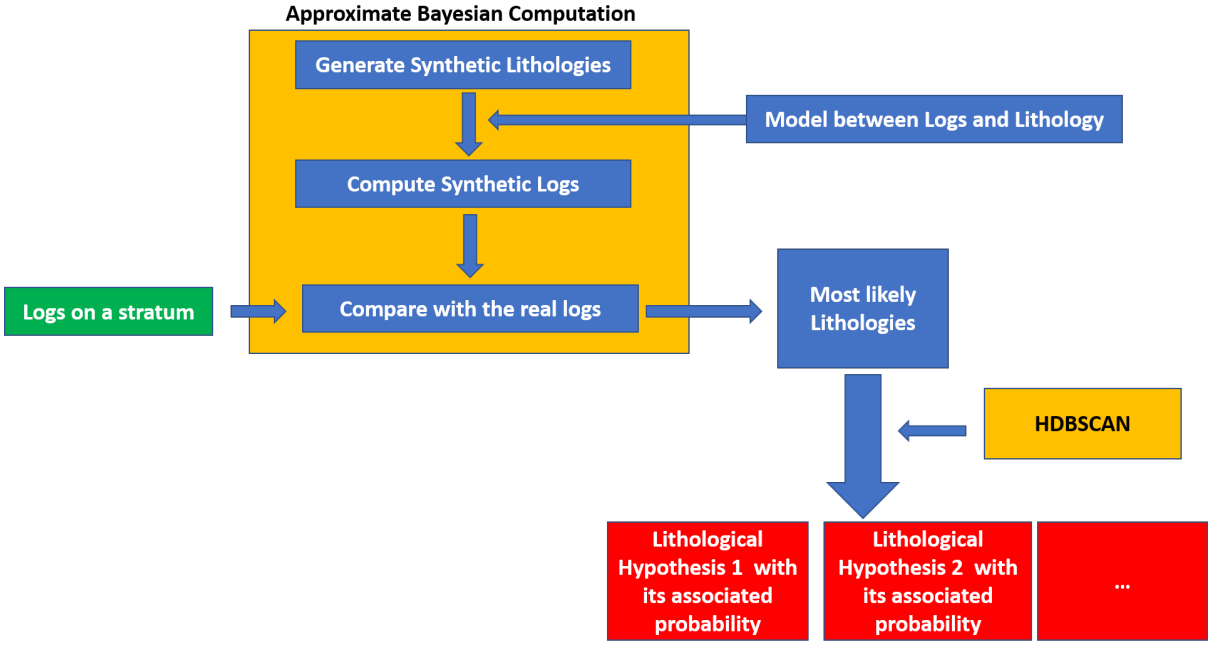

Figure 1: Global methodology: from the logs of a layer to the petrophysical hypothesis. The inputs, the methods used and the outputs are respectively in green, yellow and red.

2.2.1 Bayesian inference with ABC

About the Dirichlet distribution The Dirichlet distribution of order k, $\mathrm{Dir}(\alpha)$ where $\alpha \in \mathbb{R}^k$, allows us to generate a vector $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$ in the $k-1$ simplex, or in other words: $\sum_{i=1}^{k} x_i = 1$ and $0 \leq x_i$. Its probability density function is given by:

$$f\left(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k\right) \frac{1}{B(\alpha)} \prod_{i=1}^{M} x_i^{\alpha_i - 1},$$

where B is the beta function. The expectation of each element of $\mathbf{x}$ is $\mathrm{E}[x_i] = \frac{\alpha_i}{\sum_{i=j}^{k} \alpha_j}$. The variance and the covariances are proportional to $\frac{1}{\left(\sum_{j=1}^{k} \alpha_j\right)^2}$.

When $\alpha_1 = \ldots = \alpha_k = 1$, $\mathrm{Dir}(\alpha)$ is the uniform distribution on the $k-1$ simplex. When $\alpha_1 = \ldots = \alpha_k = 0.1$, sparsity appears over the simplex, the corners and the edges of the simplex have more density mass. If the $\alpha_i > 1$ ($\alpha_1 = \ldots = \alpha_k = 5$), a mode will appear clearly on the location of the average. Figures 2 and 3 illustrate these cases in $\mathbb{R}^3$.

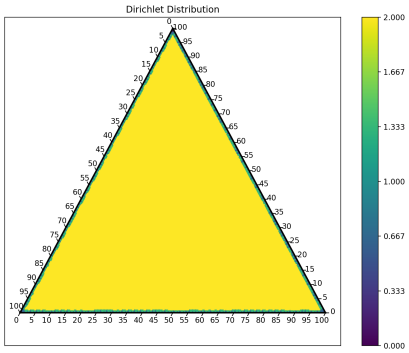

Figure 2: Ternary Density plot of a $\mathrm{Dir}(1, 1, 1)$. It is a uniform distribution over the simplex of $\mathbb{R}^3$.

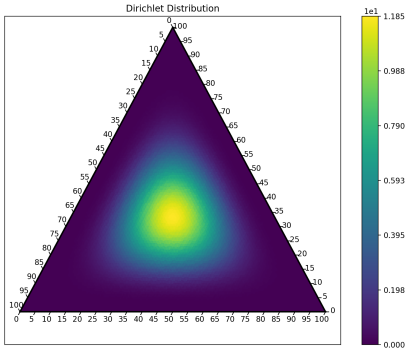

Figure 3: Ternary Density plot of a $\mathrm{Dir}(5, 5, 5)$. We observe a mode at $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

The inference model is based on a Dirichlet distribution as prior for the mineralogy (V) and introduces a Gaussian noise for the logs (L). The Dirichlet distribution fulfills the closure constraints of the volumes. The model can be specified as :

$$\begin{cases} V \sim \mathcal{D}\left(\alpha_1, \ldots \alpha_M\right), \\ L \sim \mathcal{N}\left(G\left(V\right), \Sigma\right). \end{cases} \tag{2}$$

Once for all, we set the known operator G as linear and bounded, so $G \in \mathcal{M}_{d,M}\left(\mathbb{R}\right)$ the set of rectangle real-valued matrices with d rows and M columns. The elements of G are the cells $\mathcal{M}_{d,M}$ and stand for the endpoints. We can rewrite this model in terms of likelihood function:

$$\begin{cases} p\left(V\right) = \dfrac{1}{B(\alpha)} \displaystyle\prod_{i=1}^{M} V_i^{\alpha_i - 1}, \\ p\left(L|V\right) = \exp\left[-\dfrac{1}{2}\left(GV - L\right)^T \Sigma^{-1} \left(GV - L\right)\right]. \end{cases} \tag{3}$$

Using Bayes' rule, the likelihood function of the combined model can be written as:

$$\begin{aligned} p(V|\,L) &= \frac{p(L|V)\,p(V)}{p(L)}, \\ p(V|\,L) &\propto \exp\left[-\frac{1}{2}(GV-L)^T\,\Sigma^{-1}\,(GV-L)\right]\frac{1}{B(\alpha)}\prod_{i=1}^{M} V_i^{\alpha_i-1}. \end{aligned} \tag{4}$$

Because $d \ll M$, the matrix G is not invertible so we cannot compute the maximum a posteriori (MAP) estimation of the lithology V. To compute the posterior distribution of the lithology we need a numerical method. Bayesian frameworks have been used to solve ill-posed inverse problem because they handle the non-uniqueness of the solution (see Stuart [2010]). In this study, we adopt the Approximate Bayesian Computation: it is a simple rejection method. Let θ be the parameter of interest, $\mathbf{y}$ the observations, ϵ a tolerance level, π the prior distribution of the model, and ρ a distance function. The original algorithm of ABC is given below:

Algorithm 1 General ABC

for $i = 1$ to J **do**
  repeat
    Generate θ' from the prior distribution $\pi(.)$
    Generate $\mathbf{z}$ from the likelihood $f(.|\theta')$
  until $\rho(\mathbf{z}, \mathbf{y}) \leq \epsilon$
  set $\theta_i = \theta'$
end for

For our problem, we adapt Algorithm 1. At each depth n we state :

Algorithm 2 Adapted ABC

for $i = 1$ to J **do**
  Generate V' from the prior distribution
  Generate the corresponding $L^\star$ from the model
  if $\forall 1 \leq i \leq d$, $\|L_i^\star - L_{n,i}\| < \delta_i$ **then**
    accept V'
  end if
end for

Where the δ_i are defined for each log and should be calibrated and J is around 10^6. We apply the ABC procedure for all the depth of the layer and get a posterior distribution on the layer of the lithology. Due to the non-uniqueness of the inverse problem, the posterior probability on the layer of lithology is often multi-modal. For instance, if two shales similar in terms of physical parameters appear in the lithological model, it will be hard to differentiate them.

2.2.2 Clustering on the results of ABC

In order to retrieve the most probable lithology hypothesis on the given stratum, we perform a density-based clustering algorithm on the ABC results. The aim of the clustering is to distinguish the modes that may appear in the results of ABC. Ideally we seek a density-based clustering algorithm detecting the outlying lithologies and with a data-driven tuning of the number of clusters. In the literature, several generic density-based algorithms may be found such as density-based spatial clustering of applications with noise (DBSCAN) Ester et al. [1996], ordering points to identify the clustering structure (OPTICS) Ankerst et al. [1999] and Hierarchical DBSCAN or HDBSCAN Campello et al. [2013]. We choose HDBSCAN for many reasons:

- Conversely to DBSCAN, the distance value (main parameter of DBSCAN) disappears. The main parameter is the minimum size of the cluster.
- Identification of the outliers or noise is available in DBSCAN but not in OPTICS.
- The *hdbscan* package in Python (McInnes et al. [2017]) is quite efficient, could be run in parallel and offers low computational time. The size of a stratum could reach sometimes hundreds of feet so we may have 100 000 points to cluster.

The two main parameters of HDBSCAN are: the number of points m_{pts} used for the core distance (DBSCAN, OPTICS) and the minimum size to form a cluster m_{clSize}. Usually, $m_{clSize} = m_{pts}$ and having a smaller m_{pts} implies fewer points detected as noise. Here is a description of the algorithm:

Algorithm 3 HDBSCAN

`Compute` $\forall x_p$ `the core distance` $d_{core,m_{pts}}(x_p)$
`Build a minimum spanning tree (MST) using the mutual reachability distance` $d_{mreach,m_{pts}}(x_p, x_q)$
`Derive the cluster hierarchy from the MST. Transform the MST into a dendrogram`
`Condense the dendrogram using` m_{clSize}

The core distance of a point x_p, $d_{core}(x_p)$, is the distance from x_p to its m_{pts}-nearest neighbor (including x_p). The mutual reachability distance between two points x_p and x_q in $\mathbf{X}$, with regards to m_{pts}, is defined as $d_{mreach}(x_p, x_q) = \max\{d_{core}(x_p), d_{core}(x_q), d(x_p, x_q)\}$. An output of HDBSCAN is the condensed tree which gives a hierarchical visualization of the data in form of a tree. The Excess of Mass (EOM) criterion described in Campello et al. [2015] allows a flat clustering by selecting the stable leaves of the condensed tree as clusters. Figure 4 is an example of condensed tree. The height of the leaves represents the local density.

With HDBSCAN, defining the minimum size of a cluster is mandatory, we choose $m_{clSize} = m_{pts}$ between 1 and 5% of the number of points. This parameter has to be calibrated. We can order the different clusters of lithology by their size and/or the error of reconstruction of a classical inversion method when the number of lithological component obeys $M \leq d + 1$.

We can define an empirical probability $\hat{p}_i$ for each lithological hypothesis:

$$\hat{p}_i = \frac{m_i}{N_{ABC}}, \tag{5}$$

where N_{ABC} is the number of lithology selected by the ABC part and m_i is the size of the cluster i found by HDBSCAN. Note that $\sum \hat{p}_i < 1$ because HDBSCAN considers a part of the lithologies as noise.

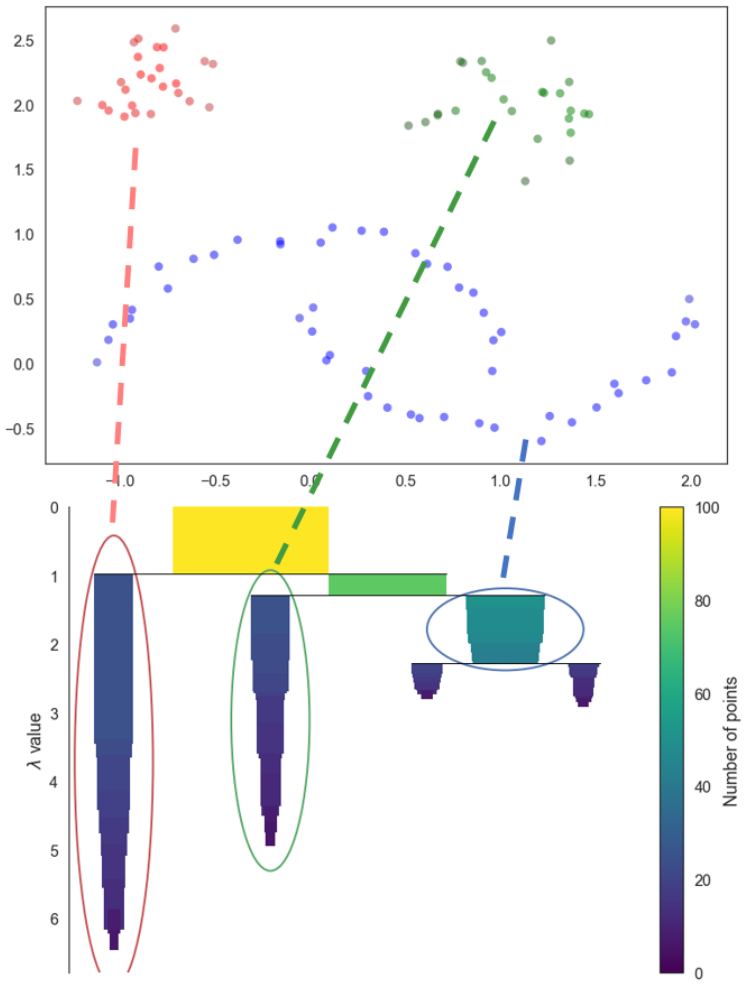

Figure 4: Example of a condensed tree on the two moons dataset (toy example of the *hdbscan* library in python McInnes et al. [2017]). We have the three coloured clusters found by HDBSCAN. The width of different nodes and leaves represent the number of elements. The encircled nodes and leaves are the clusters selected by the EOM criterion. The λ axis represents the density of the cluster.

3 Results on synthetic data

3.1 Synthetic example and Monte Carlo sampler

We use below a synthetic generator of logs that may be roughly summarized in three steps. First, we generate an average lithology using a Dirichlet distribution, and then we add a Brownian bridge to get a synthetic lithology over the synthetic stratum. Finally, we transform the lithology into logs using equation 1 and we add noise.

In our test with ABC, we select $M = 10$ lithological components: 3 carbonates (Calcite, Ankerite, Dolomite), 2 sands (Quartz, N-Feldspar), 4 shales (Illite, Kaolinite, Chlorite, Smectite) and water for the porosity. Here we do not use a simple Dirichlet distribution where all the α_i are equal. Indeed, the porosity is generally below 35% and we assume that a mix of minerals of the same family is not probable. For this reason, we draw first $V_{water} \sim \mathcal{U}[0, 0.35]$, then the proportion of the families (carbonate, shale and sand) $(V_{sand}, V_{shale}, V_{carbonate}) \sim (1 - V_{water})\,\mathcal{D}(1, 1, 1)$. Finally, we draw the minerals according to a Dirichlet distribution with a small alpha parameter equal to 0.1. For instance, we generate the volumes of the carbonate family: $(V_{calcite}, V_{dolomite}, V_{ankerite}) \sim V_{carbonate}\mathcal{D}(0.1, 0.1, 0.1)$. Indeed, we prefer a sparse

model fitting the reality: sparsity is a well-known feature of most lithological databases. We defined a lithological sampling model. An example of prior distribution on the lithology using the sampling model is shown in Figure 5. Besides, we generate 1 million lithologies for ABC.

Volume	Chlorite	Illite	Kaolinite	Smectite	Quartz	Calcite	Dolomite	XWater	XOil
Sandy	-	-	-	-	80	-	-	20	-
Sandy Oil	-	-	-	-	80	-	-	5	15
Shaly-Sand 1	-	40	-	-	40	-	-	20	-
Shaly-Sand 2	-	80	-	-	10	-	-	10	-
Shaly-Sand 3	-	-	-	80	10	-	-	10	-
Shaly-Sand 4	40	40	-	-	10	-	-	10	-
Shaly-Carbonate 1	-	-	-	60	-	20	-	20	-
Shaly-Carbonate 2	-	-	30	-	-	50	-	20	-
Shaly-Carbonate 3	-	20	20	-	-	-	40	20	-
Carbonate-Shaly	-	20	-	-	-	30	30	20	-
Carbonate	-	-	-	-	-	40	40	20	-

Table 2: Different lithologies tested. The lines correspond to the alpha parameter. They are similar to the proportion.

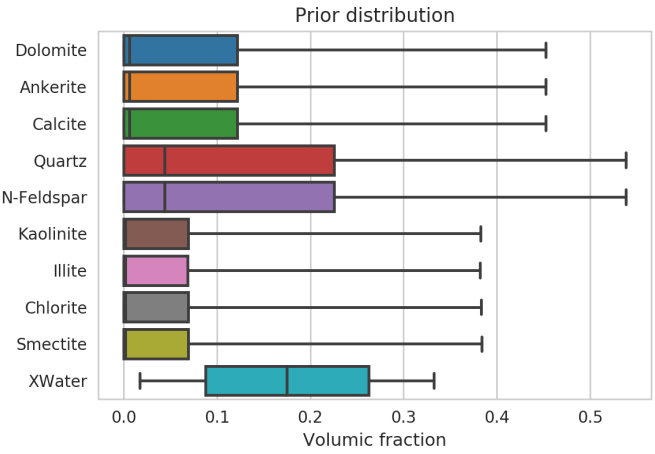

Figure 5: Example of prior distribution of the lithology. On the boxplots, we represent the percentile 5 and percentile 95 as the whiskers (applied to all the box plots).

Here are some examples of distributions that the defined sampling model generates over 1 million samples. The histogram of the total volume of shale is presented in Figure 6, the probability near one is very low because of the water distribution which is not correlated to the mineral volumes. Figure 7 displays the histogram of the volume of illite generated by the sampling model. An explanation of the sampler parameter is that an important variance for the volume fits well the reality. Indeed, according to the Dirichlet distribution with $\alpha_1 = \cdots = \alpha_{10} = 1$, $\mathrm{Var}[V_i] = 0.008185$ and if $\alpha_1 = \cdots = \alpha_{10} = 0.1$ then $\mathrm{Var}[V_i] = 0.045$. With the defined sampling model $\mathrm{Var}[V_i] \simeq 0.02$ (depending on the family of the mineral) except for the water $\mathrm{Var}[V_{XWater}] = 0.01$).

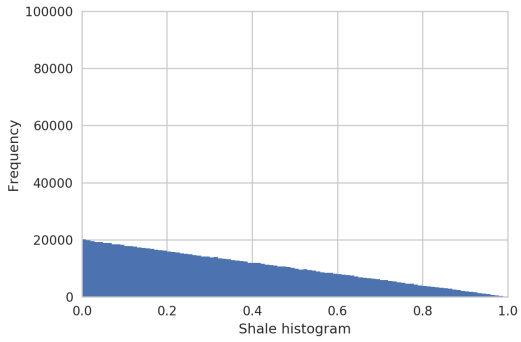

Figure 6: Histogram of the shale volume generated by the sampling model.

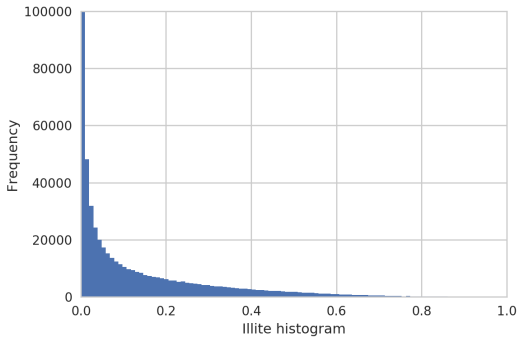

Figure 7: Histogram of the Illite volume generated by the sampling model. The frequency around 0 is near 600 000 samples.

3.2 Results

We show here the results on the Shaly-Sand 1 (see Table 2) example with a stratum of 250 samples and with the following logs: gamma ray GR, bulk density ρ_b, neutron porosity ϕ_N. On average we have $V_{Illite} = 0.34$, $V_{Quartz} = 0.49$ and $V_{Water} = 0.17$. With a fixed rejection factor δ on the logs ($\delta_{GR} = 12$ API, $\delta_{\rho_b} = 0.05$ G/C^3, $\delta_{\phi_N} = 0.03$ V/V), we obtain an average of 3000 lithologies per depth. In Figure 8, the boxplot of all the volumes selected by ABC on the stratum is provided. We see that the volumes of Illite, Smectite, Quartz, and N-Feldspar have a large variance. We remark too that the volumes of carbonate are low. For more details on these distributions, we look at the histograms of the first and second components of the principal component analysis (PCA) of the results. The percentage of variance explained by the first and second components are respectively 40% and 33%. Figure 9 displays a cross-plot of the two components and the associated histograms. We notice the multi-modality of the first projection with a main mode and the bimodality of the second projection. An explanation is that the modes of the first projection accounts for the competition between Kaolinite and Smectite. The second projection shows the competition between the Quartz and the N-Feldspar. After the

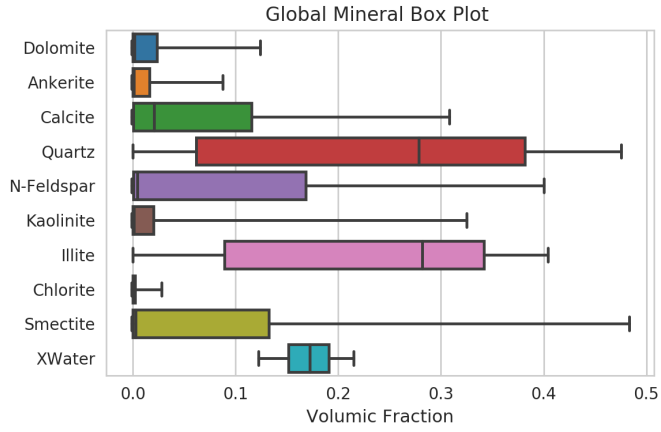

Figure 8: Boxplot of the volumes selected on the stratum by ABC.

ABC step, we apply HDBSCAN in order to find the most plausible lithological hypothesis. We fix the minimum number of samples to form a cluster to 5% of the number of the lithology selected by ABC. In this case, HDBSCAN finds 3 clusters and classifies around 35% lithologies as noise. Figures 10 illustrates the results of HDBSCAN on the first and second component of

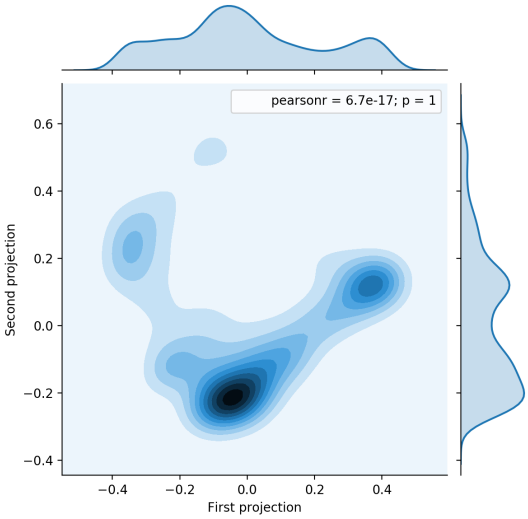

Figure 9: Density cross-plot of the first and second component of the PCA.

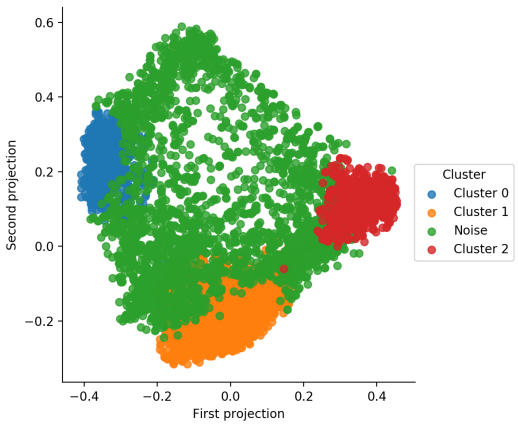

Figure 10: Cluster found by HDBSCAN (PCA projection).

the PCA of the volumes selected previously. We display the clustering on the cross-plot between the first and the second projections of the PCA. We distinguish clearly the clusters.

The boxplots of the three clusters are shown in Figures 11, 12 and 13. Figure 12 is the boxplot of the cluster with some Illite and Quartz corresponding to the synthetic lithology. In this cluster, the average of the three main components are: $\bar{V}_{Illite} = 0.30$, $\bar{V}_{Quartz} = 0.37$ and $\bar{V}_{XWater} = 0.17$. The presence of Calcite, with an average volume $\bar{V}_{Calcite} = 0.05$ explains why we underestimate the real volume of Quartz and Illite. A possible reason for this fact relies on the flat clustering of HDBSCAN. It tends to advantage large clusters and as consequence, we still have some competitions between Carbonate and Quartz which may have a similar response to the logs. Cluster 0 (Figure 11) and Cluster 2 (Figure 13) are other possible lithological hypotheses where respectively Illite is replaced by another shale, Smectite, and Quartz is replaced by another sand, N-Feldspar.

After removing the 35 % of lithology considered as noise, we can order the different hypotheses by the number of points:

- Cluster 1 ($\hat{p}_1 = 40\%$). Main minerals: Quartz/Illite/Water.
- Cluster 2 ($\hat{p}_2 = 15\%$). Main minerals: N-Feldspar/Illite/Water.
- Cluster 0 ($\hat{p}_0 = 10\%$). Main minerals: Quartz/Smectite/Water.

Here the classification by the size of the hypothesis gives a good result: Cluster 1 is the lithological hypothesis (Quartz/Illite/Water) matching the real lithology. Our methodology can provide either most probable lithological hypotheses or empirical distributions for the components common to all hypotheses. For instance, we can define the empirical distribution of the water $\hat{f}_{Water}$:

$$\hat{f}_{Water} = \frac{\sum m_i \hat{f}_{i,Water}}{\sum m_i}, \tag{6}$$

where $\hat{f}_{i,Water}$ is the empirical distribution of the water in cluster i and m_i is the size for cluster i. Figure 14 shows the empirical distribution for our synthetic case. The mode of this distribution is closed to the real average value of the water of the synthetic data (red vertical line on the figure).

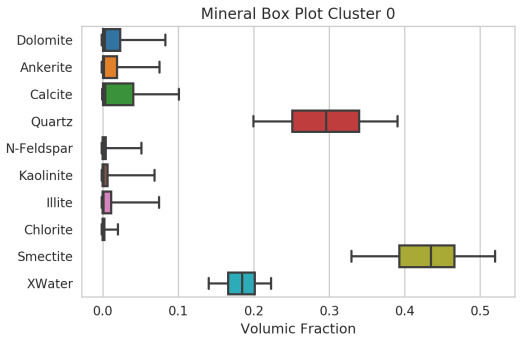

Figure 11: Boxplot of the volumes of cluster 0 (HDBSCAN).

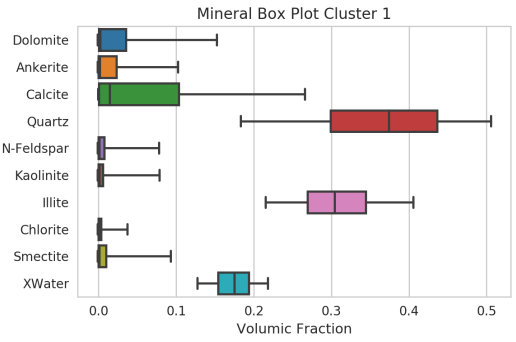

Figure 12: Boxplot of the volumes of cluster 1 (HDBSCAN), the main cluster.

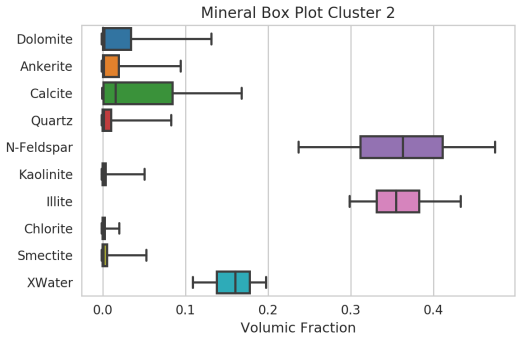

Figure 13: Boxplot of the volumes of cluster 2 (HDBSCAN).

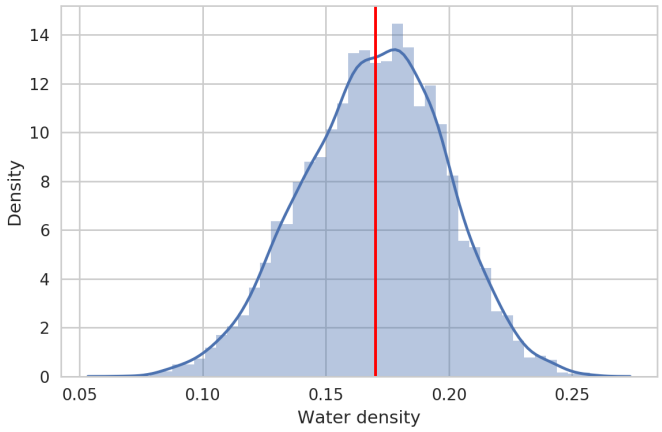

Figure 14: Empirical distribution of the water volume. We can derive from this plot an optimistic, average and pessimistic scenarii (percentile 10, median and percentile 90). The true average volume of water is 0.17 (vertical line in red).

In Figure 15, we present the synthetic lithology and the corresponding logs used to perform a Bayesian inversion per layer. In the last three tracks, the different hypotheses found by HDBSCAN (average of the minerals per hypothesis) are given.

For the other synthetic cases, corresponding to 2 minerals of different families and water, we are able to determine the right lithology. The volumes found are close to the reality. For the mix of carbonate, sand or shale, the true hypothesis is not necessary the main one (it depends on

the choice of the δ and the α parameter). Changing the parameter alpha has some consequences on the results. Moreover, when 3-4 components (shaly-sand-carbonate) are detected and one of them is around 10% it seems difficult to identify the smallest one.

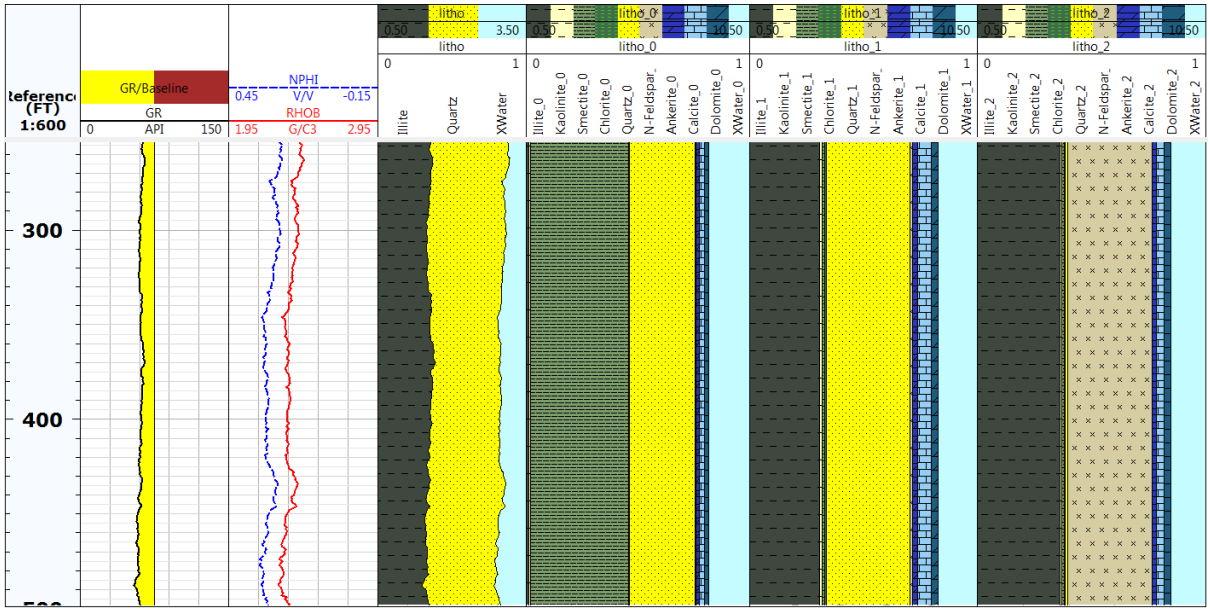

Figure 15: Synthetic lithology with its logs and results of our methodology. Track 1: GR. Track 2: ρ_b and ϕ_N. Track 3: Corresponding lithology. Track 4: Lithological hypothesis of cluster 0. Track 5: Lithological hypothesis of cluster 1. Track 6: Lithological hypothesis of cluster 2.

4 Case study

The dataset presented in this section comes from the Kansas Geological Survey, Oil and Gas Well Database (`http://www.kgs.ku.edu/Magellan/Qualified/index.html`). We present here a well from the Wellington Field: well 1-28 (API number: 15-191-22590). From this well, we get 9 logs: the gamma ray GR, the resistivity R_t, the compressional slowness Δ_t, the bulk density ρ_b, the neutron porosity ϕ_N, the photoelectric factor pef, the uranium concentration U, the thorium concentration TH and the potassium concentration K. We have also a computed lithology and the mud log to check our results. These data are part of the South-central Kansas $C0_2$ project.

For Well 1-28, we use the segmentation algorithm PELT (Killick et al. [2012]) to identify the different beds where we assume a fixed lithology. As inputs for PELT, we took GR, ρ_b, ϕ_N and pef and a Gaussian model is used. Figure 16 gives the results of the segmentation

In Figure 17, we display the lithology provided by the Kansas Geological Survey, Oil and Gas Well Database, and also the mud log (description of the mineralogy) with the geological age. We notice that the segmentation algorithm provides layers coherent with the change of lithology given by the experts.

We apply our methodology on the different segments of PELT using $d = 4$ logs (GR, ρ_b, ϕ_N, pef). We use the same lithological model as in the synthetic case with $M = 10$ components: Water, 3 carbonates (Ankerite, Dolomite, Calcite), 2 sands (Quartz, N-Feldspar) and 4 shales (Illite, Chlorite, Smectite and Kaolinite). The hyper-parameters of ABC are tuned the following

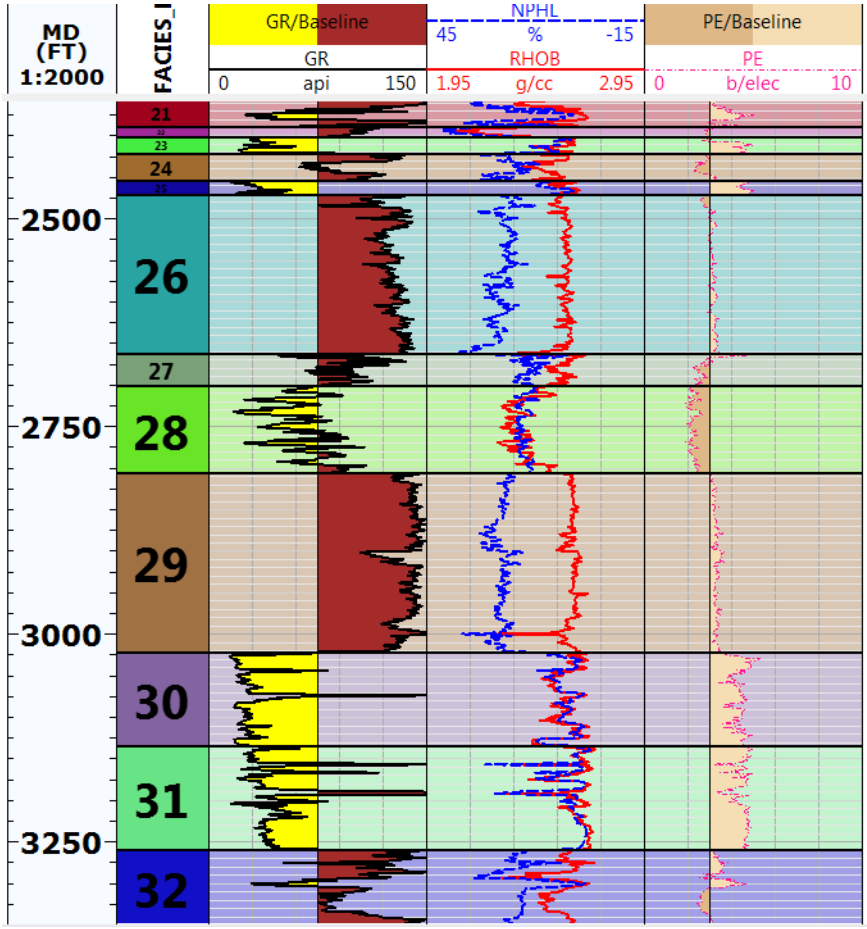

Figure 16: Well 1-28 with the zonation provided by PELT. Track 1: Zonation from PELT. Track 2: GR. Track 3: ρ_b and ϕ_N. Track 4: pef.

way. We choose $J = 10^6$ and we set the rejection thresholds for the different logs:

$$\begin{cases} \delta_{GR} = 50 \text{ API} \\ \delta_{\rho_b} = 0.05 \text{ G}/C^3 \\ \delta_{\phi_N} = 0.03 \text{ V/V} \\ \delta_{pef} = 0.2 \text{ b/e} \end{cases} \tag{7}$$

The parameter δ_{GR} is quite important since the values of the Gamma Ray are not necessarily corrected and in a classical mineralogical inversion program, the weight of this log is very light. Because of the rejection method used, it is possible that we do not select in a segment several lithologies if our lithological model does not fit the data. In this case, we set a threshold of an average minimum number of lithology per depth selected by the ABC step to launch the clustering algorithm otherwise we do not give any hypothesis. We fix this quality parameter of the lithological model to 50. In Figure 18, we display the main lithological hypothesis found by our method for each zone. In parallel, we performed a classical petrophysical inversion using more logs (8 logs available) and with minerals selected accordingly to the information given by the Kansas study. In general, the first hypothesis matches quite well the main minerals present in the layers. Sometimes, it may be difficult to distinguish minerals from the same family: Quartz and N-feldspar (layers 30 and 31) or the different clays (layers 27 and 33). Another point is that we are not precise in terms of percentage: in layer 42, we should have around 70% of dolomite, in our main hypothesis the volume of Dolomite is around 50%.

Focusing on the layers 37 to 41 (Figure 19), we do not provide any lithological hypothesis for

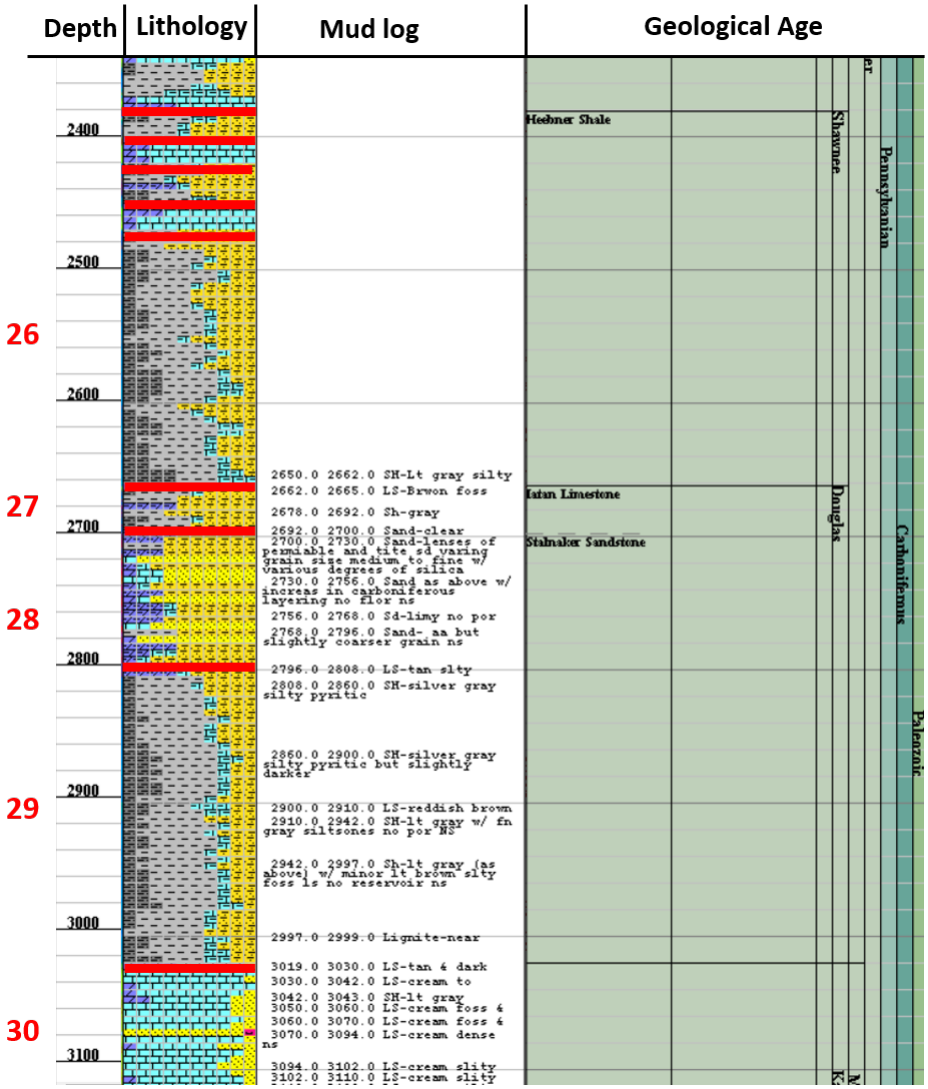

Figure 17: Well 1-28. Well analysis provided by the CO_2 project. From the left to the right, we have the depth, the lithology computed, the mud log (brief description of the mineralogy) and the geological age. The horizontal red lines correspond to the change points found by PELT and the numbers are corresponding to the segmentation of Figure 16. The change points are linked to an important change in the lithology.

the layers 37, 39 and 40. It is due to the lack of lithology selected by ABC because of a Gamma Ray too high (around 250 API) or a photoelectric factor too low (around 1.6 b/e) which do not fit out lithological model with 10 elements. Another issue is visible on layer 41: the main hypothesis is a large volume of sand (around 80%) and a small volume of clay (around 10%) whereas the classical inversion program is more balanced (mixed of Quartz and Illite). This kind of error is due to the fact that we select lithologies only in the range of depth where there is a majority of quartz; and we do not pick any lithology in the shaly area because of a high GR. Layer 41 has large variations of its lithology. We must check that the main lithological hypothesis is well distributed on the layer.

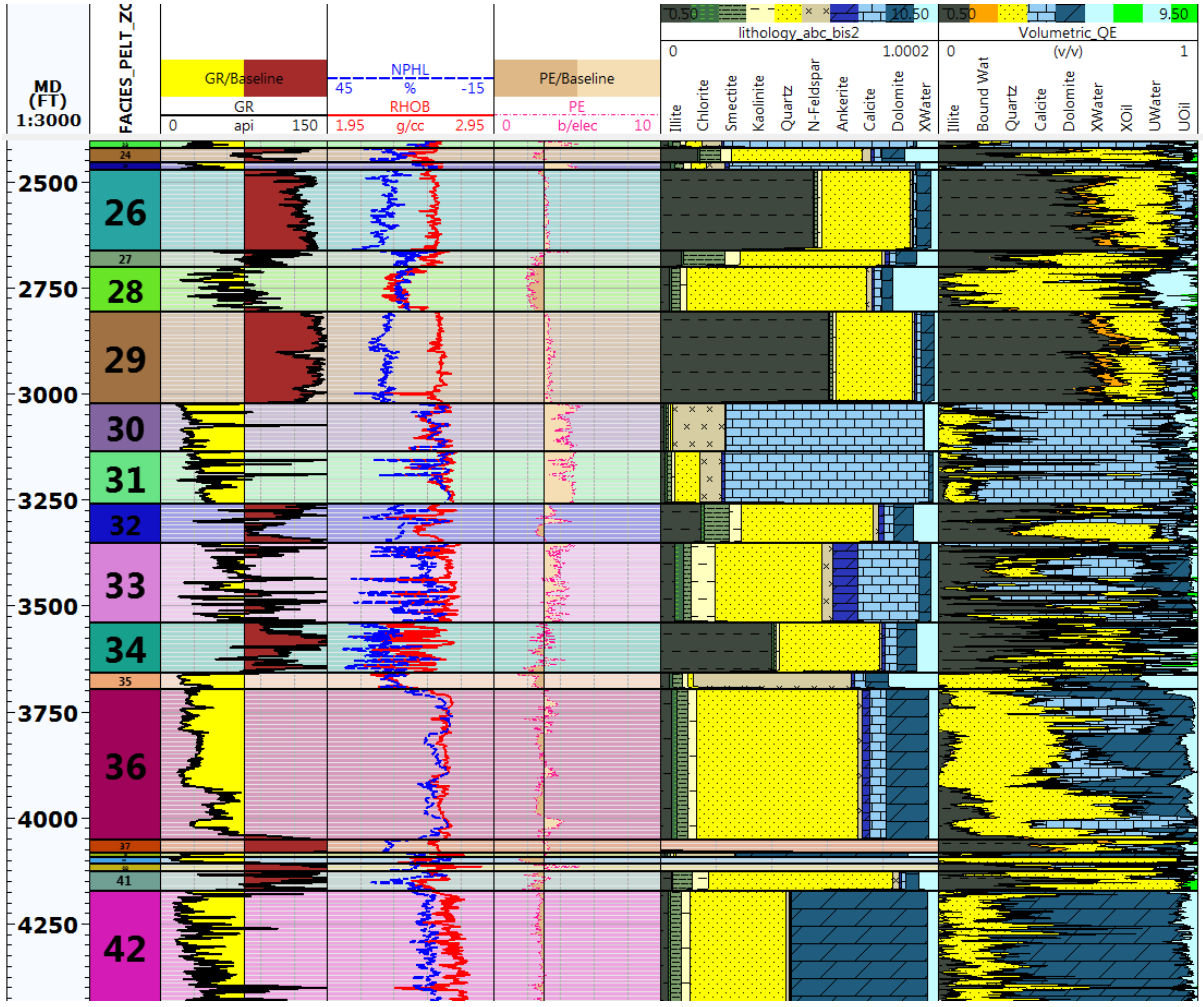

Figure 18: Well 1-28. Results of our methodology. Track 1: zonation obtained by PELT. Track 2: GR. Track 3: ρ_b and ϕ_N. Track 4: pef. Track 5: Results of our methodology (displaying the mean lithology of the first hypothesis). Track 6: Lithology obtained by a classical solver using more logs. We notice that the first hypothesis matches quite well the results of the classical inversion even if we are not precise in terms of percentage.

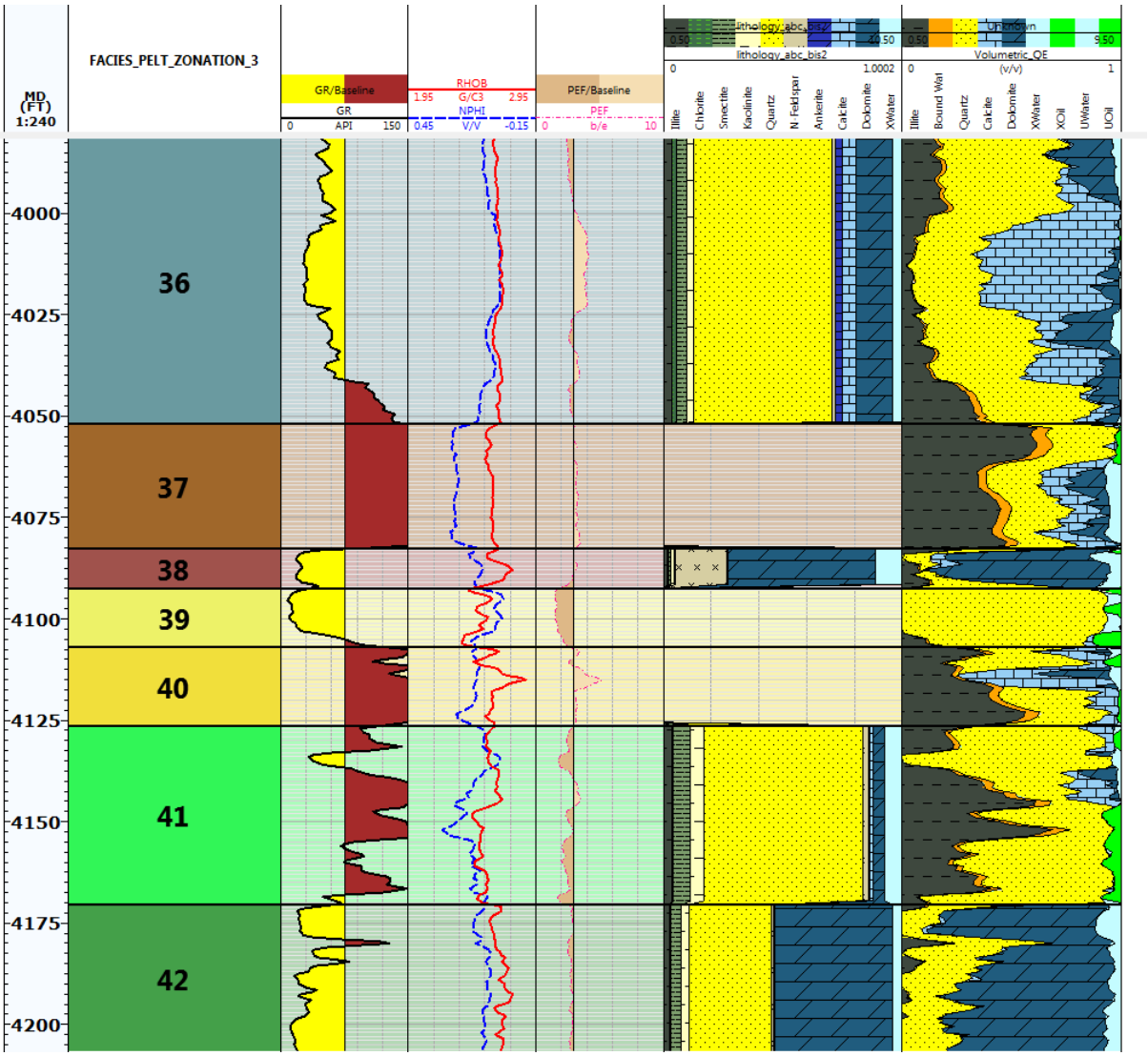

Figure 19: Well 1-28. Results of our methodology focusing on few layers. Track 1: zonation obtained by PELT. Track 2: GR. Track 3: ρ_b and ϕ_N. Track 4: pef. Track 5: Results of our methodology (displaying the mean lithology of the first hypothesis). Track 6: Lithology obtained by a classical solver using more logs. In layers 37, 39 and 40, we do not provide any hypothesis due to the few lithologies selected by ABC step.

5 Discussion

We propose a new methodology giving promising results on both synthetic and real examples. The method is fully automated and delivers lithological hypotheses weighted by posterior probabilities. The method needs few logs to produce results, and augmenting the number of logs reduces the uncertainties associated with the hypotheses. We could use other Bayesian inference methods such as Markov chain Monte Carlo (MCMC), but ABC may convince the reader by its simplicity of parametrization and implementation. However, during the tests, we identified several issues :

- A problem of identifiability with many minerals in small proportion. We are not able to detect the presence of many minerals of the same family when no log differentiates them. For instance, differentiating carbonate without pef is not easy.
- Detecting specific minerals such as Kerogen, Pyrite, or Anhydrite may be tricky because adding them into a lithological model always improves the fit and consequently leads to a bias. Our method tends to select them in small proportion. This kind of problem could be solved with better prior on the lithological model.
- The final selection of minerals is not very clear when we select more minerals than available logs. For instance, if we have a hypothesis with M components in small proportion (less than 5% in average) and $M > d+1$, d being the number of logs, we cannot run a classical inversion program to get the lithology depth by depth.
- We do not integrate hydrocarbons in the lithological sampling model. The problem is that when we introduce oil or gas in a lithological model using the triple combo (GR, ρ_b, and ϕ_N), they are usually selected even if they are not present. The resistivity equation is needed to determine the difference between water and hydrocarbons, but its parametrization is difficult.
- Some clusters selected by HDBSCAN are located on a subsection of the stratum. This problem is due to the segmentation part or the definition of the stratum. Strata with changing or unstable lithologies may yield surprising results: a multiplication of lithological hypotheses covering different subsections of the stratum.
- The method may lead to the absence of results because of a lithological sampling model not adapted or noisy data that do not fit in the δ_i of ABC.
- The calibration of the HDBSCAN parameter (essentially m_{pts}) can be tricky. The default parameter of 5% of the size of the data seems to work in our tests, but a more complex study is needed.

6 Conclusion

Some of the issues mentioned above in the Discussion section may be overcome. We give below, as a conclusion, some tracks for potential improvement.

In order to deal with the lack of identifiability and the presence of minerals in small proportion in our hypotheses, a solution relies in grouping the minerals by family before the clustering phase and apply the clustering on the family. It should reduce the uncertainty on the presence of some mineralogical families. The size of the stratum could play a role in the determination of the lithology if the logs on the stratum are not homogeneous. Indeed, noisy logs will produce less dense regions in the solution space (lithology) so the clustering algorithm will have difficulties to find lithological hypotheses. Thus, instead of using only a stratum of a single well, applying

our method on a global stratum (the same stratum from different wells) should eliminate or differentiate more lithological hypotheses. For the determination of the water saturation S_w (ratio water/porosity), a possible solution to handle the problem is to first have a rough idea of the porosity using a model without hydrocarbons; and then use the resistivity equation on the first estimation of the porosity to determine the presence of gas or oil. We can iterate the process to improve the precision of the method by adding the selected hydrocarbon in the inference model. It is an imitation of the petrophysicist's workflow. We can also improve the lithological sampling model with a more complex Bayesian hierarchy by adding prior on the α or the δ_i. For the absence of lithologies selected by ABC, a variant of the method consists in selecting the k first lithologies generated by the lithological sampling model which minimizes the error between the real logs and the generated one. Choosing $k = 500$ gives satisfactory results on synthetic data. Even if the Bayesian inference will always deliver results, the quality of these results can be dubious: the lithologies can be selected because of only one log (for instance the bulk density). The reliability of the results is an issue. This concern could eventually be attacked by raising alarms when the data are poor or warning the users when their lithological hypotheses are not appropriate.

Bibliography

Ankerst, M., M. M. Breunig, and J. Kriegel, H. P. et Sander (1999). Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 49–60.

Campello, R., D. Moulavi, and J. Sander (2013). Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, 160–172.

Campello, R. J., D. Moulavi, A. Zimek, and J. Sander (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 5.

Cannon, D. E., G. R. Coates, et al. (1990). Applying mineral knowledge to standard log interpretation. In *SPWLA 31st Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts.

da Costa, E. F., F. S. Moraes, C. Loures, L. Geraldo, et al. (2008). An automatic porosity and saturation evaluation based on the inversion of multiple well logs. *Petrophysics 49*(03).

Ester, M., H. P. Kriegel, and X. Sander, J. et Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd 96*(34), 226–231.

Grana, D. (2016). Bayesian linearized rock-physics inversion. *Geophysics 81*(6), D625–D641.

Killick, R., P. Fearnhead, and I. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*(500), 1590–1598.

Marin, J. M., P. Pudlo, and R. J. Robert, C. P. et Ryder (2012). Approximate bayesian computational methods. *Statistics and Computing 22*(6), 1167–1180.

Mayer, C. t. et al. (1980). Global, a new approach to computer-processed log interpretation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

McInnes, L., J. Healy, and S. Astels (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software 2*(11), 205.

Moline, G., P. Drzewiecki, and J. Bahr (1991). Identification and characterization of pressure seals through the use of wireline logs: A multivariate statistical approach. *AAPG Bulletin (American Association of Petroleum Geologists);(United States) 75*(CONF-910403–).

Peeters, M., R. Visser, et al. (1991). A comparison of petrophysical evaluation packages: Logic, flame, elan, optima, and ultra. *The Log Analyst 32*(04).

Qin, R., H. Pan, P. Zhao, Y. Liu, and C. Deng (2017). Bayesian inversion of well logs for petrophysical properties estimation. In *International Geophysical Conference, Qingdao, China, 17-20 April 2017*, pp. 1067–1070. Society of Exploration Geophysicists and Chinese Petroleum Society.

Quirein, J., S. Kimminau, J. La Vigne, J. Singer, F. Wendel, et al. (1986). A coherent framework for developing and applying multiple formation evaluation models. In *SPWLA 27th Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts.

Rebelle, M. and B. Lalanne (2014). Rock-typing in carbonates: A critical review of clustering methods. *Society of Petroleum Engineers*.

Rimstad, K., P. Avseth, and H. Omre (2012). Hierarchical bayesian lithology/fluid prediction: A north sea case study. *Geophysics*.

Sanchez-Ramirez, J. A., D. Wolf, C. Torres-Verdín, A. Mendoza, G. L. Wang, et al. (2010). Synthetic and field examples of the bayesian stochastic inversion of gamma-ray, density, and resistivity logs. In *SPWLA 51st Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta Numerica 19*, 451–559.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Wolf, M. and J. Pelissier-Combescure (1982). Faciolog - automatic electrofacies determination. *Society of Petrophysicists and Well-Log Analysts*.

Xu, C., Q. Yang, and C. Torres-Verdín (2016). Bayesian rock classification and petrophysical uncertainty characterization with fast well-log forward modeling in thin-bed reservoirs. *Interpretation*.

Yang, Q., C. Torres-Verdín, et al. (2013). Joint stochastic interpretation of conventional well logs acquired in hydrocarbon-bearing shale. In *SPWLA 54th Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts.

Ye, S.-J. and P. Rabiller (2000). A new tool for electro-facies analysis: multi-resolution graph-based clustering. Paper PP, *Transactions of the SPWLA 41st Annual Logging Symposium*, Dallas, Texas, 4-7 June.